Asian Journal of Pharmaceutical and Clinical Research

INNOVARE ACADEMIC SCIENCES
Knowledge to Innovation

Special Issue (April)

Online - 2455-3891
Print - 0974-2441

Full Proceeding Paper

PROSODY PREDICTION FOR TAMIL TEXT-TO-SPEECH SYNTHESIZER USING SENTIMENT ANALYSIS

VAIBHAVI RAJENDRAN*, BHARADWAJA KUMAR G

School of Computing Science and Engineering, Vellore Institute of Technology University, Chennai Campus, Tamil Nadu, India.
Email: vvaibavi@gmail.com

Received: 23 January 2017, Revised and Accepted: 03 March 2017

ABSTRACT

A speech synthesizer which sounds similar to a human voice is preferred over a robotic voice, and hence to increase the naturalness of a speech synthesizer an efficacious prosody model is imperative. Hence, this paper is focused on developing a prosody prediction model using sentiment analysis for a Tamil speech synthesizer. Two variations of prosody prediction models using SentiWordNet are experimented: one without a stemmer and the other with a stemmer. The prosody prediction model with a stemmer performs much more efficiently than the one without a stemmer as it tackles the highly agglutinative and inflectional words in Tamil language in a better way and is exemplified clearly, in this paper. The performance of the prosody prediction model with a stemmer has a higher classification accuracy of 77% on the test set in comparison to the 57% accuracy by the prosody model without a stemmer.

Keywords: Natural language processing, Prosody, Sentiment analysis, Tamil, Text-to-speech.

INTRODUCTION

Interaction being the biggest key to information gain or exchange plays a vital role in human life. In this Information and communication technology era, designing interactive computer systems that are effective, efficient, easy, and enjoyable to use is becoming increasingly important. Of the numerous ways explored by researchers to enhance human-computer interaction (HCI), text to speech or speech synthesis is one such modality which helps in developing better interfaces for HCI. Some of the applications of a text-to-speech (TTS) system include information retrieval system: Banking, telephony, marketing; readers: Vocal alerts in railway stations, airports, and other public places; screen readers: Manuscripts, e-mails; e-learning: Learning a new language or new topic; special equipment: Reading aid for visually challenged, speaking aid for vocally challenged.

A TTS system takes natural language sentences in textual form as input and provides the synthesized speech. The most important qualities of a speech synthesis system are naturalness and intelligibility. Naturalness describes how closely the output sounds such as human speech, while intelligibility is the ease with which the output is understood. Even though work has been extensively carried out for building a generic TTS system for any given language, the prevalence of some enthralling facts between human and machine intelligence always springs up an orbit of inefficiency by a certain margin. Such subjectivity of insolvable facts can be found right from the starting stage of text processing to the final stage of speech production.

A TTS system comprises a natural language processing (NLP) module and a digital signal processing (DSP) module. The NLP module is mainly responsible for performing the text processing while the DSP module is responsible for the generation of the final synthesized speech waveform. In most of the literature, lot of emphasis is always laid on the DSP module for improving the efficiency of the existing TTS systems. The NLP module has got very little attention, but it is necessary to be given equal emphasis to both NLP and DSP modules for the better performance of the TTS system. In a concatenative speech synthesis approach, the NLP module includes text normalization, grapheme to phoneme conversion, and prosody analysis. The DSP module includes the process of unit selection, prosody generation, and speech processing and speech waveform generation.

This paper is focused on developing a prosody prediction model for a Tamil TTS synthesizer. Tamil language is an official language for most of the people in Tamil Nadu, Puducherry, Andaman and Nicobar in India and also in countries such as Sri Lanka, Singapore, and Malaysia. Hence, developing a Tamil TTS will help the people in these region to interact in an easier and comfortable way with the computer. Prosody can be referred to the patterns of duration, intonation, and intensity in speech [1]. Prosody structures the flow of sequence of syllables, words, or phrases and helps in increasing the articulation and coarticulation required to make the synthesized speech more natural.

The rest of the paper is organized as follows: The next section briefs about the work related to the development of prosody models in the literature, Section 3 holds a detailed discussion about the two variants of prosody prediction models developed by us, and finally, the last section holds the discussion on our results and conclusion.

RELATED WORK

The prosody prediction and generation models are mostly centered on the three main parameters which define prosody. The three parameters are duration, intonation, and intensity patterns, and the models developed based on these parameters are named as duration models, intonation models, and intensity models, respectively [2]. Variance in duration patterns of speech unit utterances provides naturalness to speech. Duration model targets on modeling the time taken for uttering each individual speech unit and researchers have explored developing a duration model for Indian Languages using rule-based approach (Hindi), Classification and Regression Tree approach (Hindi and Telugu), and Neural Networks (Hindi, Telugu, and Tamil) [3,4]. The dynamics of fundamental frequency contour over time generally caused by vocal fold vibration is termed as intonation. A rule-based intonation model for Hindi and a statistical intonation model based on Neural Networks and SVM has been developed for Hindi, Telugu, and Tamil [5,6]. Intensity Models specific to Tamil and as well other Indian Languages is still unevolved. Various generic models such as ToBI, INSTINT, Tilt, Fujisaki, Klatt, and Tones are available but are yet to be adapted for Indian Languages. Most of these models in the literature revolve around acoustic, articulatory, and perceptual perspectives of prosody. Development of multimodal prosody models [6] and linguistic models are still unexplored for Indian languages and as well as many other languages.

PROSODY PREDICTION

Prosody can be referred to the patterns of duration, intonation, and intensity patterns in speech. Prosody structures the flow of sequence of syllables, words, and phrases. Thus, prosody helps in increasing the articulation and coarticulation required to make the synthesized speech more natural. In the context of a TTS system, prosody needs to be first predicted from the text and then generated. Prosody prediction generally refers to the process of understanding the emotion conveyed in a sentence or a phrase in the text. Prosody prediction from speech is much simpler in comparison to prosody prediction from the text. In the speech, we tend to acknowledge the variance in the speech utterances when a person is angry, sad, or neutral. A speaker may utter the sentence or phrase with a short duration but with a high intonation and intensity when he/she is angry. When the speaker is sad, he/she may utter the sentence or phrase with a longer duration, low intonation, and intensity. Hence, the way a person utters the sentence or phrase greatly depends on the state of mind of the person. Predicting this state of mind of the person is relatively easy in speech as we can analyze the three main parameters of speech: Duration, intonation, and intensity. Hence, the biggest challenge in prosody prediction from the text is to predict the state of the mind of the writer just from text.

Let us take an example phrase, "Oh! Really?" in the text, the speaker can actually utter it either in an excited way (positive) or in a depressed way (negative) depending on his/her state of mind. Sometimes, the context of the phrase can help us find out the state of mind of the speaker to a certain extent. Again, analyzing the context and demarcation of a context is in itself a crucial task when it comes to prosody analysis and is explained in detail in section 3.2. In general, there are four different perspectives from which we can view prosody and is given as follows:

- Linguistic.
- Articulatory.
- Acoustic and.
- Perceptual.

Researchers have already explored prosody prediction using articulatory, acoustic, and perceptual perspectives [3-6]. Hence, we look at predicting prosody from the lesser explored linguistic perspective. At linguistic level, the features will be words and its related information. The prosody prediction model is explained in detail in the following subsections. The input sentence is first subjected to a text normalizer and then presented to the prosody prediction model. The text normalizer model is discussed in subsection 3.1, and the prosody prediction model is discussed in subsection 3.2.

Text normalization

A text normalization model is required to improve the intelligibility of a TTS system by decreasing the number of unclear utterances by converting the non-standard words (NSW) into standard words. The process involved in performing the Tamil text normalization [7] is briefed below:

Step 1: Pre-processing: The input text document is subjected to identification of character encoding issues and possible multi-lingual issues. Incorrectly, encoded words/text are removed form the document.

Step 2: Sentence splitting: The pre-processed text document is then segmented into a list of sentences using sentence delimiters, which is mostly a '.' in Tamil language and sometimes exclamatory marks such as '?', '!'.

Step 3: Tokenization: Next, each segment is split into further smaller tokens (words) using the word delimiters (white space).

Step 4: Semiotic classification: Each token is now either identified into a standard or NSW.

1. The presence of numerals, special symbols, and notations within a word classifies the word to be an NSW.
2. Once we identify an NSW, it is subjected to a semiotic classifier.
3. A set of 8 classes have been taken into consideration for forming the semiotic classifier after analyzing the NSWs in Tamil. The eight classes dealt by the semiotic classifier are: Numbers, date, time, alphanumeric, abbreviations and acronyms, money, punctuations, and special characters.

Step 5: Verbalization: Each NSW token is classified into one of these 8 classes which helps us in processing these NSWs into a standard words.

Now, the text document processed by the text normalizer is given to the prosody prediction model for identifying the prosody prevalent in the text.

Prosody prediction

Text analysis for prosody can be performed at 3 levels:

- Document level: The whole text document can be taken into one single unit for analysis. However, in reality, articulation and coarticulation have a lot of variance within a paragraph itself, and hence, document level analysis may not be suitable for prosody analysis.
- Sentence level: Each sentence in the text document can be considered for analysis. This seems to be a better unit than an entire document as generating the prosody for each sentence according to the predicted prosody is feasible and will be effective.
- Phrase level: Each sentence can be seen as a set of phrases and can be considered for prosody analysis. Phrase level analysis is actually a good choice for prosody analysis but not for prosody generation. During prosody generation there might be inconsistencies in the structured flow of utterances within the same sentence, and hence, the synthesized speech may sound to be an unnatural sequence of word utterances. Possibilities of parts of the same sentence getting generated using different prosodic parameters seem more likely and it can eventually make the synthesized speech sound totally unnatural.

Thus, prosody prediction is performed at the sentence level using linguistic information of the words in the sentence.

Prosody prediction using sentiment analysis

Typically, to perform sentiment analysis on Tamil text using a computational approach, the first requirement is a prior polarity lexicon where the words are tagged with a prior polarity. [8]. A lexical resource which holds the words in a language along with the polarity tags is called as a SentiWordNet. A SentiWordNet for Tamil has been developed and is available for exploration [8]. This SentiWordNet contains four different lists, and each list corresponds to a different kind of polarity. The four different polarities based on which the Tamil words have been categorized are: Positive, negative, neutral and ambiguous. Using this SentiWordNet, the prosody analysis at sentence level is carried out with the aid of sentiment analysis and the Algorithm is given below:

For each word in the sentence:

1. Presence of the word is checked in the SentiWordNet list:
   a. If the word is present in the positive list, a high-positive score is retrieved.
   b. If the word is present in the negative list, a high-negative score is retrieved.
   c. If the word is present in the neutral list, a low-positive score is retrieved.
   d. If the word is present in the ambiguous list, a low-negative score is retrieved.
   e. If the word is not found in any of the lists, a zero score is given for the word. (Discrete values are placed for the high-positive, low-positive, high-negative, and low-negative scores).
2. The cumulative score for each sentence is calculated by adding the individual scores of each word in the sentence.
3. Now, for each sentence, if the sentence score is:
   a. Positive - positive (happy) prosodic parameters will be generated for the sentence.
   b. Negative - negative (sad) prosodic parameters will be

generatedfor the sentence.

c. Neutral - neutral (neither happy nor sad) prosodic parameters will be generated for the sentence.

Prosody prediction using sentiment analysis after stemming

The prosody prediction model using sentiment analysis works well on sentences with pure Tamil words scripted using proper Tamil written-grammar, but the model fails to retrieve a score for loan/foreign words and also for agglutinative and inflectional words. A morphological analyzer, lemmatizer, or stemmer can be used to find the root word, but a credible morphological analyzer or lemmatizer is unavailable for Tamil. (Efforts for building a robust morphological analyzer for Tamil is still ongoing [9]). Hence, we decided to make use of a stemmer to get the root word and then perform the sentiment analysis on those words. Stemming process refers to a crude heuristic process of chopping off the ends of words to obtain the root word by removing the affixes. We employed a stemmer [10] to obtain the root word and to overcome the agglutination and inflections on the words in the sentence. The algorithm for prosody prediction using sentiment analysis after applying the stemmer is given below:

For each word in the sentence:

1. Apply the stemmer and obtain the stemmed word.
2. Place the stemmed word in a new file against the original word.

For each stemmed word:

a. The presence of the word is checked in the SentiWordNet list:
   i. If the word is present in the positive list, a high-positive score is retrieved.
   ii. If the word is present in the negative list, a high-negative score is retrieved.
   iii. If the word is present in the neutral list, a low-positive score is retrieved.
   iv. If the word is present in the ambiguous list, a low-negative score is retrieved.
   v. If the word is not found in any of the lists, a zero score is given for the word.

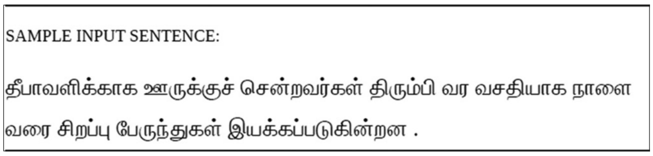
Discrete values are placed for the high-positive, low-positive, high-negative, and low-negative scores.

b. The cumulative score for each sentence is calculated by adding the individual scores of each word in the sentence.
c. Now, for each sentence, if the sentence score is:
   a. Positive - positive prosodic parameters will be generated for the sentence.
   b. Negative - negative prosodic parameters will be generated for the sentence.
   c. Neutral - neutral prosodic parameters will be generated for the sentence.

The performance of this refined model increases the probability of getting a word score for every word in the sentence and is discussed in the next section.

RESULTS AND DISCUSSIONS

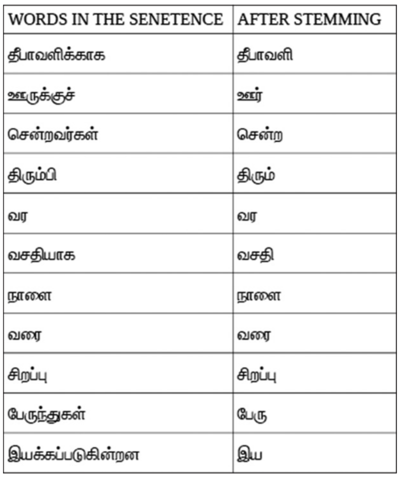
The results of the prosody prediction model with the usage of a stemmer are certainly better than the prosody prediction model without a stemmer. To understand the importance of the stemmer, let us take the sentence given in Fig. 1, it gets a sentence score of 5 when we apply the prosody prediction model with sentiment analysis alone and gets a sentence score of 10 when we apply the prosody prediction model using sentiment analysis with a stemmer. The sentence given in Fig. 1 apparently gets clearly classified as a positive sentence due to the usage of a stemmer which actually increases the probability of obtaining the word score for each word in the sentence. The stems of the words in the sentence given in Fig. 1 is tabulated in Fig. 2, the stems generated for some words are incorrect, and this is due to the complexities stirred by the morphological richness in Tamil language.



SAMPLE INPUT SENTENCE:

தீபாவளிக்காக ஊருக்குச் சென்றவர்கள் திரும்பி வர வசதியாக நாளை வரை சிறப்பு பேருந்துகள் இயக்கப்படுகின்றன .

Fig. 1: Sample input sentence



WORDS IN THE SENETENCE	AFTER STEMMING
தீபாவளிக்காக	தீபாவளி
ஊருக்குச்	ஊர்
சென்றவர்கள்	சென்ற
திரும்பி	திரும்
வர	வர
வசதியாக	வசதி
நாளை	நாளை
வரை	வரை
சிறப்பு	சிறப்பு
பேருந்துகள்	பேரு
இயக்கப்படுகின்றன	இய

Fig. 2: Stemmed words

Table 1: Performance evaluation of the prosody prediction models

Metric/model	Without stemmer	With stemmer
Number of word score retrieved	37	44
Prosody prediction accuracy (%)	57	77

To evaluate the performance of the prosody prediction models with and without a stemmer, a test set of 30 sentences with varied subject coverage is constructed. The prosody prediction model with a stemmer performs much better than the model without a stemmer and the results are tabulated in Table 1. The prosody prediction model without a stemmer provides an accuracy of 57% while the prosody prediction model with the stemmer provides a higher accuracy of 77% for the test set. The reason for the increase in the accuracy of the prosody prediction model with a stemmer is due to the higher number of word score retrieval and is achieved due to the affix removal by the stemmer. The number of words for which a word score was successfully retrieved is 37 in the prosody prediction model without a stemmer whereas the usage of a stemmer increased the number to 44 for the test set under consideration. The higher number of word score retrieval also provides a much optimal sentence score, and hence, the prosody prediction model with a stemmer performs definitely better than the prosody model without stemmer.

CONCLUSION AND FUTURE WORK

Since Tamil is a highly agglutinative and inflectional language despite the use of the stemmer a word score was not retrievable for some words due to the following reasons:

- Prevalence of loan/foreign word: Words which are in regular usage in Tamil language but has been borrowed from other languages. Such words are neither available in the SentiWordNet nor processable by the stemmer.
- Morphological richness of Tamil language: Although a stemmer has been employed to resolve the agglutination and inflection to yield the root word, the process is very complex. The intensity of Tamil language's inflectional nature can give rise to around 200 forms

of words from just one word. The inclusion of auxiliaries to the same word can give rise to more than 1800 forms of the word. In addition, if the word is combined with another word in the process of agglutination, the number of word formations cannot be ascertained at all [11,12].
- Confined the list of words in SentiWordNet: Listing all the occurrences of words in any language is actually an impossible task but the SentiWordNet used for obtaining the word score for prosody prediction is based on a confined set of words, and hence, a lot of words are not found in the SentiWordNet. A resource which allows us to perform sentiment analysis without any restraints on the input word is still unsubstantial for Tamil.

Usage of efficient morphological analyzers along with a SentiWordNet with more openness in terms of words can improve the performance of the model further. Identification of loan/foreign words and processing them separately will also help in enriching the performance of the prosody prediction models for Tamil using sentiment analysis.

REFERENCES

1. Bellur A, Narayan KB, Krishnan KR, Murthy HA. Prosody Modeling for Syllable-Based Concatenative Speech Synthesis of Hindi and Tamil. In: Communications (NCC), 2011 National Conference on 2011 January 28, IEEE. p. 1-5.
2. Reddy VR, Rao KS. Prosody modeling for syllable based text-to-speech synthesis using feedforward neural networks. Neurocomputing 2016;171:1323-34.
3. Rao KS, Yegnanarayana B. Modeling durations of syllables using neural networks. Comput Speech Lang 2007;21(2):282-95.
4. Rao KS, Koolagudi SG. Selection of suitable features for modeling the durations of syllables. J Softw Eng Appl 2010;3(12):1107.
5. Rao KS, Yegnanarayana B. Intonation modeling for Indian languages. Comput Speech Lang 2009;23(2):240-56.
6. Reddy VR, Rao KS. Two-stage intonation modeling using feedforward neural networks for syllable based text-to-speech synthesis. Comput Speech Lang 2013;27(5):1105-26.
7. Rajendran V, Kumar GB. Text processing for developing unrestricted Tamil text to speech synthesis system. Indian J Sci Technol 2015;8(29):1-10.
8. Das A, Bandyopadhyay S. Sentiwordnet for Indian Languages. China: Asian Federation for Natural Language Processing; 2010. p. 56-63.
9. Antony JB, Mahalakshmi GS. Challenges in morphological analysis of tamil biomedical texts. Indian J Sci Technol 2015;8(23):1.
10. Tamil Stemmer. Available from: http://www.github.com/rdamodharan/tamil-stemmer.
11. Kumar MA, Dhanalakshmi V, Soman KP, Rajendran S. A sequence labeling approach to morphological analyzer for Tamil language. Int J Comput Sci Eng 2010;2(6):1944-5.
12. Kumar MA, Dhanalakshmi V, Rekha RU, Soman KP, Rajendran S. A novel data driven algorithm for Tamil morphological generator. Int J Comput Appl 2010;6(12):52-6.