ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

Volume 3, Special Issue 3, March 2014

2014 International Conference on Innovations in Engineering (ICIET'14)
On 21st & 22nd March Organized by
K.L.N. College of Engineering, Madurai, Tamil Nadu, India

Estimation Of Phrase Boundaries For Tamil Speech Synthesizer

Sivaram L P, Nagarajan T
Speech Lab, SSN College of Engineering, Chennai,Tamilnadu, India
Speech Lab, SSN College of Engineering, Chennai,Tamilnadu, India

ABSTRACT— Given any arbitrary text in a language, a text-to-speech (TTS) system is expected to produce high quality speech. One of the major language-specific information, in addition to list of phonemes is, the phrase boundaries in a given sentence, in the form of "comma" and other punctuations. Wherever a comma is present in the text, during parsing, the synthesizer will introduce a silence to represent it. This will improve the quality, and even in some cases, the proper meaning can be conveyed. However, for Tamil, in the written text, the phrase boundaries are not explicitly present, thus the quality of the HMM-based synthesizer is found to be poor, in the sense that, the individual words in the sentence sound very good, but as a sentence, it does not sound natural. For the language Tamil, estimating the phrase boundaries, from a given sentence, is still a research issue. A system without phrase boundary is built as a baseline system. Without any analysis carried out on given text, silence is introduced arbitrarily after each word, every two words, and every three words. Even though, there is an improvement in the naturalness in the synthesized speech, since phrase boundaries, in terms of pauses, are introduced arbitrarily, in many synthetic sentences the quality is annoyingly low. An analysis is carried out on word terminal syllables occuring at the phrase boundaries and the 50 most frequently occuring word terminal syllables are considered. Based on this analysis another system is built which gives phrase boundaries after the words that terminate in these syllables. Significant improvement is achieved when phrase boundaries are predicted using terminal syllables, however, certain phrase boundaries are not predicted due to absence of terminal syllables. So a final system is developed, where initially phrase boundaries are predicted based on the word terminal syllable and then if the number of words in each phrase exceeds a threshold, a new phrase boundary is introduced at the midpoint of each phrase. This system produces high quality speech with a mean opinion score (MOS) of 4.23.

KEYWORDS—phraseboundary, HMM-based speech synthesis.

I. INTRODUCTION

Text to speech synthesis system takes a given text as input and produces corresponding synthesized speech as output. From the given input text, as the first-step, the system is expected to convert the graphemes to phonemes (G2P). Then, the required phonemes/models are concatenated and speech is synthesized. In hidden Markov model (HMM)-based speech synthesis system, the required language-specific information has to be explicitly provided for better quality in synthesized speech. The quality of the HMM framework-based voice sounds highly intelligible. However, the naturalness is considerably less mainly due to the absence of explicit phrase boundaries available in the text to be synthesized. One of the major language-specific information, in addition to list of phonemes is, the phrase boundaries in a given sentence, in the form of ``comma" and other punctuations like exclamation, interrogation mark, etc. To perceive naturalness in the synthesized speech we need little longer silence in between the phrases than usual. The synthesizer will give silence wherever there is punctuation. It gives longer pause for period (.) and medium pause for comma (,) and very short pause for spaces. This will improve the quality, and in some cases, proper meaning is conveyed only if a silence is perceived after a phrase boundary. For example, many of us encounter a famous example of "Hang not leave him", which conveys extreme opposite meaning based on the punctuation "," as follows: "Hang not, leave him" and "Hang, not leave him". The former one conveys not to hang a person and to leave him,

whereas the later one conveys, to hang a person and not to leave him. One can say it is just a comma, but it makes the meaning completely different. This shows the importance of phrase boundaries and punctuations. Similarly in Tamil language also, phrase boundaries are important. But, for Tamil, in the written text, comma is not used in the written-text. For example, in the sentence, ைரக ைர யரதக அ ப சத ெகா ெசறன. No punctuations are here. One can read this sentence in different ways depending on his/her lung capacity. In general it will be tougher to read longer sentences without pauses in between. In the same sentence, if comma separates the phrases as follows, it can be read phrase by phrase, by influencing the reader to give pauses at the places of comma. ைரக , ைர ய ரத க , அ ப , சத ெகா ெசறன. This makes more sense and makes reading comfortable. In a similar way, the synthesized speech with phrase breaks sounds good. In this work, an attempt is made to automize the phrase boundary prediction in HMM-based speech synthesis system, based on word terminal syllables and number of words in the input text.

This paper is organized as follows. Section II discusses several existing approaches for phrase boundary prediction. Section III gives an overview of the proposed system. Section IV describes the speech corpus used for this work. Section V discusses the overview of HMM-based speech synthesis system. Section VI discusses the manual introduction of phrase boundary in the systems developed. Section VII discusses the working of automized phraseboundary prediction algorithm. Section VIII concludes the paper and Section IX discusses the future work.

II. EXISTING SYSTEM

The prediction of phrase boundaries in Indian language corpora, is a tough task, due to the absence of case markers. [1] describes the phrase boundary prediction in Indian language corpora. In this work, for Tamil language, during the training phase, phrase boundaries are manually annotated and during synthesis, the phrase boundaries are predicted using a decision tree. A feature termed as morpheme tag is defined. A set of morpheme units occuring at the end of a word, having a bearing on phrase boundaries are identified manually. Words containing these morpheme units are tagged accordingly. The morpheme tag feature is included in the feature list for predicting phrase boundaries. Phrase boundaries are introduced after morpheme tags. In [2], to identify the significance of phrase boundary, several experiments are carried out and terminal syllables are identified. The last syllable in the word is termed as terminal syllable. Using these terminal syllables phrase breaks are modeled and syntactic breaks are implemented. If the word ending of a word in the text has been marked as a syntactic break and the last syllable of the word (the terminal syllable) is among the list of the top 50 terminal syllables for that language (derived from the analysis), then that syntactic break is also a phrase break. In Hindi, the majority of the pauses are less than 80 ms in duration, for Telugu and Kannada the majority of the pauses range from a few milliseconds to 480 ms. The pause durations vary over a significant range within a language and also between languages. Experimentation is done with different thresholds (25 ms, 50 ms and 80 ms,) above which a pause is marked as a phrase break. Word boundaries that coincide with pauses greater than 80ms are marked as phrase breaks, while all other word boundaries are marked as syntactic break. In this current work an analysis is carried out on word terminal syllables occuring at the phrase boundaries and the 50 most frequently occuring word terminal syllables are considered.

III. PROPOSED SYSTEM

Several approaches to introduced phrase boundaries using morpheme tags, end syllables or terminal syllables, etc are analyzed from the literature. A HMM-based speech synthesis system is used here. Even though this HMM framework gives promising results in terms of intelligibility of the synthesized speech, naturalness is not good as expected. Here, an idea has been made to implement a generalized algorithm in this HMM-based TTS system, which introduces phrase boundaries in the synthesized speech automatically by introducing comma(,) in the input text. Word count in the input sentence is also taken into account along with the terminal syllables and the phrase boundaries will be introduced. Before implementing such an algorithm, as a preliminary work, different analyses are carried out on the word count of a text input. The experiments carried out are explained in Section V.

IV. SPEECH CORPORA

Forty minutes of speech data is collected from a native female Tamil speaker in laboratory environment using dynamic unidirectional microphone at a sampling rate of 16,000 Hz with a recording interface, in order to create the speech corpus. The text data for recording is taken from Tamil historical novel “Ponniyin Selvan”.

V. EXPERIMENTAL SETUP

HMM-based speech synthesis consists of a training and synthesis phase. In the training phase, spectral parameters and excitation parameters namely, Mel generalized cepstral coefficients (mgc) and log fundamental frequency (lf0) are extracted as feature vectors, from the speech data [3][5]. Context-independent monophone HMMs are then trained using these features and time-aligned phonetic transcriptions. For the HMM-based system, the context-dependent pentaphone is considered as the basic subword unit. The UTF-8 text is converted to a sequence of pentaphones. As in conventional HTS, the context-dependent models are initialized with a set of context-independent monophone HMMs. Tree based clustering of states is obtained as a result of state tying using common question set. In the synthesis phase, for the given text context-dependent label files are

generated and the required context-dependent HMMs are concatenated to obtain the sentence HMM. For the input text sentence spectral and excitation parameters are generated and a speech waveform is synthesized. This process is illustrated in figure 1.

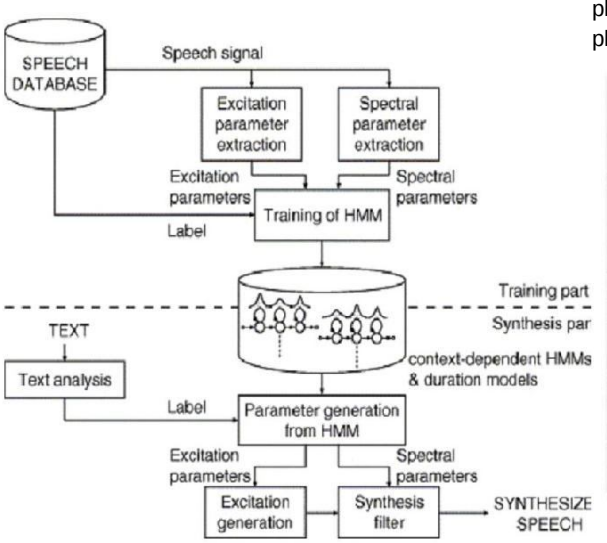

Fig. 1. HTS -Overview

VI. Manual Introduction of Phrase Boundaries

A detailed analysis on terminal syllables (end syllable of a word) has been carried out and the 50 most frequently occuring word terminal syllables list is prepared. In the proposed system, phrase boundaries are introduced based on the word count of each sentence. Before implementing such a system, analyses on word count based phrase boundaries are carried out by manually introducing them in sentences at different positions. Initially, a HMM-based speech synthesis system is developed without any phrase boundaries. Analysis on the synthesized speech, is carried out considering this system as a

baseline system. Apart from this, manual phrase boundaries are introduced arbitrarily as follows:

Without comma (as a baseline system)

With comma, after every word in the given sentence

With comma, after every two words, in the given sentence

With comma, after every three words, in the given sentence

After the annotation, four different HMM-based speech synthesis system is developed by means of training phase and synthesis phase as mentioned in Section V. Analysis is done on the synthesized speech and tabulated in Table 1.

A. Without Phrase Break

As a base line system a system without any phrase boundary i.e., the system having only spaces in between words is built, so that the synthesizer gives a minimum duration of silence eventually between all words without any phrase boundaries or special pauses. The absence of phrasebreak in utterance is shown in the figure 2.

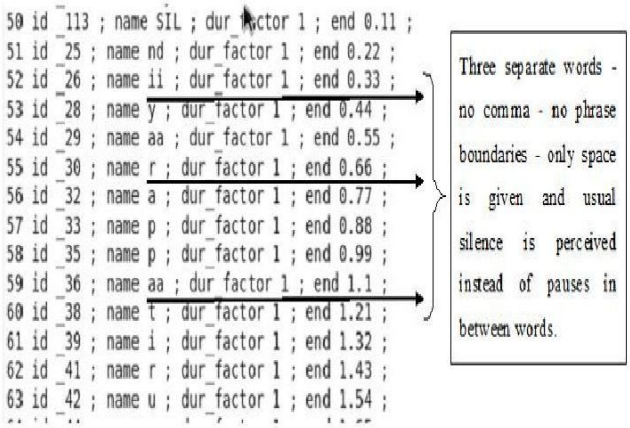

Fig. 2. Absence of phrasebreak in utterance

B. Phrase boundary after every word

A system with phrase boundary after each word is built, such that the synthesizer gives a pause after each words in the synthesized voice. The normalized text data is annotated with comma after each word as a phrase boundary and then transliterated. The occurrence of phrasebreak after every word in utterance is shown in the figure 3.

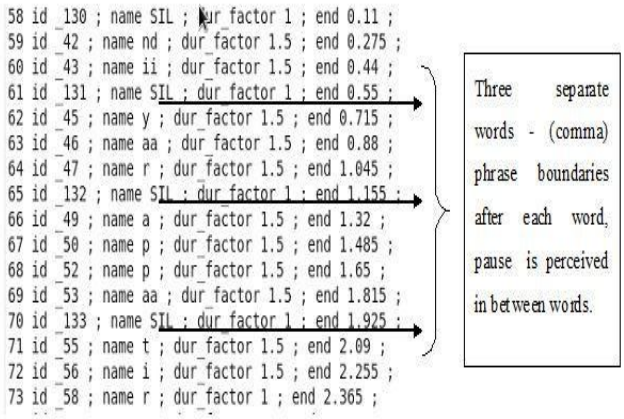

Fig.3 . Occurrence of phrasebreak in utterance

C. Phrase boundary after two word

A system with phrase boundary after every two words is built, such that the synthesizer gives a pause after every two words in the synthesized voice. The normalized text data is annotated with comma after every two word as a phrase boundary and then transliterated. The occurrence of

phrasebreak after every two word in utterance is shown in the figure 4.

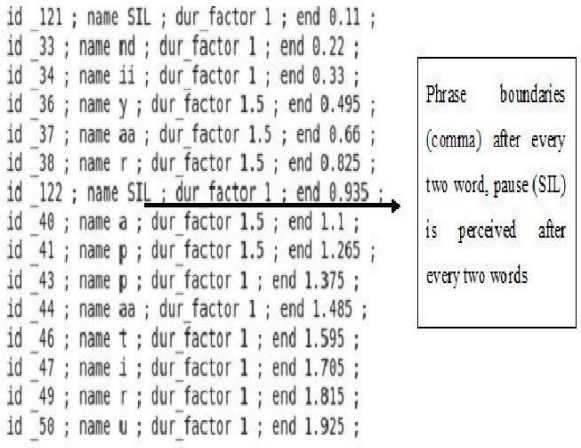

Fig. 4. Occurrence of phrasebreak in utterance

D. Phrase boundary after three word

A system with phrase boundary after every three words is built, such that the synthesizer gives a pause after every three words in the synthesized voice. The normalized text data is annotated with comma after every three word as a phrase boundary and then transliterated. The occurrence of phrasebreak after every three word in utterance is shown in the figure 5.

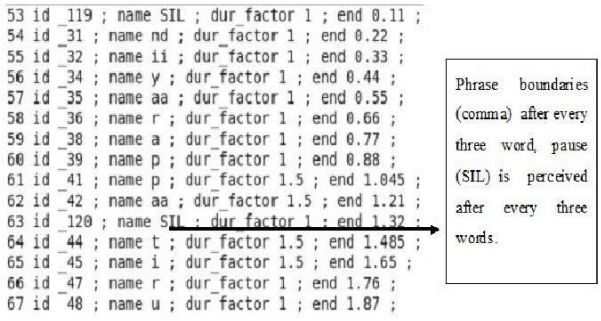

Fig. 5. Occurrence of phrasebreak in utterance

VII. AUTOMATIC PREDICTION OF PHRASE BOUNDARY

The output speech synthesized by the systems with phrase boundary is better than the system without phrase boundary. Even though we can perceive naturalness to certain extent from the above systems with default phrase boundary, it is not perceived in all the sentences. The naturalness of the synthesized speech varies for different sentences based on the number of words in a sentence. From the analysis carried out in above systems in Table 1, we can conclude that, the synthesized speech will convey proper sense with considerably good naturalness only by the combination of different length of phrase boundary. An analysis is carried out on word terminal syllables occuring at the phrase boundaries and the 50 most frequently occuring word terminal syllables are considered. Based on this analysis another system is built which gives phrase boundaries after the words that terminate in these syllables.

Table 1.
Analysis of synthesized voice

SYSTEM	No Phrase Boundary	Comma after 1 word	Comma after 2 word	Comma after 3 word
Place of Annotation	Only spaces are in between words.	Comma along with spaces in each sentence.	Space and comma used alternatively after each word.	Space is given after every word and comma in given after every three word
Naturalness	Poor. Longer and shorter pauses occur randomly and reduces the naturalness	Poor since longer pause is introduced after every word	Naturalness is considerably good for smaller sentences like 4 or 6 words. Fails mostly in sentences with odd word count.	Naturalness is comparatively good for most cases. Fails in few cases like smaller sentences and even length sentences.
Comments	1. Equal silence of smaller duration is perceived between words throughout the synthesized voice.	1. Longer silence after each word.	1. Short silence if space is used, longer silence if comma is used.	1. Short silence if space is used, longer silence if comma is used.
	2. No longer pauses or phrase boundaries in the synthesized voice.	2. Phrase boundary is introduced after every word manually.	2. Depending on word count the naturalness varies here	2. Naturalness is considerably good for longer sentences.

Significant improvement is achieved when phrase boundaries are predicted using terminal syllables, however, certain phrase boundaries are not predicted due to absence of terminal syllables. So a final system is developed, where initially phrase boundaries are predicted based on the word terminal syllable and then if the number of words in each phrase exceeds a threshold, a new phrase boundary is introduced at the midpoint of each phrase. The process of automized phrase prediction is depicted in the figure 6 below.

Fig. 6. Working of automated phrase prediction algorithm

VIII. PERFORMANCE ANALYSIS

The performance of the synthesized speech is evaluated by the traditional mean opinion score (MOS). MOS is a five-point grade scale, where score of 5 corresponds to excellent intelligibility and quality, and 1 corresponds to highly unintelligible and annoying. The evaluation is performed with 20 sentences, synthesized by each synthesizer played to 20 listeners, in a laboratory environment. The scores obtained for the proposed system with automated phrasebreak are compared with those obtained for system without phrasebreak. In addition to rating the overall quality of the synthesized speech, the listeners are also asked whether phrasebreaks are perceivable. The MOS of the proposed system with phrasebreak and the existing system without phrasebreak is 4.23 and 3.08 respectively. Thus, the naturalness and phrasebreak is effectively perceived in the proposed system.

IX. CONCLUSION

The synthesized speech has good naturalness than the existing systems with a MOS of 4.23. For small sentences, the phrase break is perceived at the word terminal syllable in the top 50 list. However, for longer sentences, the phrase break is perceived at the top word terminal syllable list as well as at different positions, depending upon the number of words in the sentence, as per the automatic phrase boundary prediction algorithm. Thus effective improvement in naturalness is perceived from the synthesized speech.

X. FUTURE WORK

In the future, an attempt will be made to automate the phrase boundary prediction by means of duration of silence in between phrases in the sentence. Another HMM-based speech synthesis system will be built based on this automated phrase boundary algorithm and the improvement in the naturalness will be evaluated subjectively.

ACKNOWLEDGMENT

The authors would like to thank Ms. Saranya M S, Ms. Ramani B, Ms. Anushiya Rachel G, Ms. Lilly Christina S for their contribution to the work.

REFERENCES

[1] A. Bellur, K. B. Narayan, R. K. K, and H. A. Murthy, "Prosody modeling for syllable-based concatenative speech synthesis of hindi and tamil," in National Conference on Communications (NCC), Banglore, January 2011,pp. 1–5.

[2] K. S. P. Anandaswarup Vadapalli, Peri Bhaskararao, "Significance of word-terminal syllables for prediction of phrase breaks in text-to-speech systems for indian languages," in Proceedings of 8th ISCA Speech Synthesis Workshop, Barcelona, Spain, September 2013, pp. 89–194.

[3] B. Ramani, S. L. Christina, G. A. Rachel, V. S. Solomi, M. K. Nandwana, A. Prakash, S. A. Shanmugam, R. Krishnan, S. K. Prahalad, K. Samudravijaya, P. Vijayalakshmi, T. Nagarajan, and H. Murthy, "A common attribute based unified hts framework for speech synthesis in Indian languages," in 8th ISCA Workshop on Speech Synthesis, Barcelona, Spain, August 2013, pp. 311–316.

[4] A. Black, H. Zen, and K. Tokuda, "Statistical parametric speech synthesis," in Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on, vol. 4, 2007, pp. IV–1229–IV–1232.

[5] Boothalingam, Ramani; Sherlin Solomi, V; Gladston, Anushiya Rachel; Christina, S Lilly; Vijayalakshmi, P; Thangavelu, Nagarajan; Murthy, Hema A. "Development and Evaluation of Unit Selection and HMM-Based Speech Synthesis Systems for Tamil." Communications (NCC), 2013 National Conference on. New Delhi: IEEE, feb-2013. 1-5.