Classification and Regression Trees (CART) based Data Driven Approach for Prosody Duration Modeling in Sinhala Language

DDM Dolawatta[1], N.G.J. Dias[2], K.H.Kumara[2]
[1] External Examinations Branch, University of Kelaniya,
[2] Department of Statistics & Computer Science, University of Kelaniya.

ABSTRACT

A Text-to-Speech (TTS) Synthesizer or Text-to-Speech Engine is a computer based system that capable to read any text aloud with naturally. In TTS, the text might be inserted directly to the computer by an operator or an output file of an Optical Character Recognition (OCR) system of a scanned written text document. Prosody features play a major role when developing a TTS system. Getting the correct intonation, Stress and duration from written text is the most challenging problems for natural languages. The prosodic duration highly affect on machine generated synthetic speech's naturalness and intelligibility. Here we have used different features that are automatically derived from the text and affect on the duration pattern of the natural speech to be modeled the duration.

In this work, in order to develop generic models for prosodic synthesis in speech synthesis, we have selected a speech corpus of 150 possible sentences in Sinhala Language and recorded them according to the three intonation patterns angry, sadness and sarcastic with a female native speaker who is a well trained person in Drama and Theater. Both the waveform and the spectrogram were used to determine the segment (phoneme) boundaries, and the boundaries identified are confirmed by listening to the speech.

Each segment in the corpora was annotated with the following features together with the actual segment duration and finally generated the CART. Identity of the current phoneme, Identity of the preceding phoneme, the features considered are the Identity of the following phoneme, Position in the parent syllable, Parent syllable initial, Parent syllable final, Parent syllable position type, Number of syllables in the parent word, Position of parent syllable in the word, Parent syllables break information, Phrase length (number of words) and Position of phrase in the utterance. Above features were observed from similar worked carried out for other languages specially Asian languages [1].

Predictions of the segmental durations were done as follows. The decision tree (CART) was traversable starting from the root node, taking various paths satisfying the conditions at intermediate nodes, till the leaf node is reached. The leaf node contains the value of segmental duration prediction.

The duration model developed in this paper was implemented on the SINHALA Text-to-Speech synthesis system developed by K.H.Kumara of the University of Kelaniya[2]. This was built within the MBROLA Speech Synthesizer developed by the Facult´e Polytechnique de Mons (Belgium), based on the concatenation of diphones.

List of references

[1]. A New Prosodic Phrasing Model for Indian Language Telugu, N. Sridhar Krishna, Hema A. Murthy, Department of Computer Science and Engineering,
Indian Institute of Technology, Madras, Chennai - 600036

[2]. K.H.Kumara, Text-to-Speech Synthesis for Sinhala Language, MPhil Thesis 2009,
University of Kelaniya, Sri Lanka