3.23 An analysis of sound parameters for prosodic modeling in Sinhala text to speech synthesis

N.G.J. Dias, K.H. Kumara, [2]D.D.M. Dolawattha
Dept. of Statistics and Computer Science, University of Kelaniya,
[2] Centre for Open and Distance Learning University of Kelaniya

ABSTRACT

Speech synthesis is the artificial production of human speech. A computer system used for this purpose is called a speech synthesizer, and can be implemented in software and/or hardware. Text-to-Speech (TTS) is one of the speech synthesis technologies. Before a synthesizer can produce an utterance, several steps have to be completed. Among them, after computing the basic pronunciation from authographic text, prosody annotation should be performed.

Finding correct intonation, stress, and duration from written text is the most challenging problem for most of the natural languages. These features together are called prosodic or suprasegmental features and may be considered as the melody, rhythm, and emphasis of the speech at the perceptual level. Unfortunately, written text usually contains very little information of these features and some of them change dynamically during speech. However, with some specific control characters this information must be given (at least some extend) to the speech synthesizer to produce enough natural speech of the target language. On the other hand timing at sentence level or grouping of words into phrases correctly is difficult; in many languages, prosodic phrasing is not always marked in text by punctuation, and phrasal accentuation is almost never marked. If there is no breath pauses in speech or if they are in wrong places, the speech may sound very unnatural or even the meaning of the sentence may be misunderstood. As an example, in Sinhala, the input string " අම්ම ආවද? ” /αμμα α:ϖαδα/(Διδ μοτηερ χομε?) " can be spoken as three different ways changing the intonation patterns as angry, sadness and sarcastic; giving three different meanings to the listener. Here intonation means how the pitch pattern or fundamental frequency changes during speech. The prosody of continuous speech depends on many separate aspects, it may be twice as high as with male voice and with children it may be even three, such as the meaning of the sentence and the speaker characteristics and emotions. Therefore it is clear that prosody plays a major role in speech synthesis, and a deeper treatment of prosody is a must in any kind of speech synthesis.

In this work, in order to develop generic models for prosodic synthesis in speech synthesis, we selected 150 possible sentences in Sinhala Language and recorded them according to the above three intonation patterns (i.e. angry, sadness and sarcastic) with a female native speaker who is a well trained person in Drama and Theater. Then we computed various speech parameters for above 150X3 sentences using PRAAT speech processing tool developed by www.praat.org. Hence we found that for all above 150 sentences there is an incremental pattern in the duration from Angry to Sarcastic. No regular pattern in Median, Mean, Standard Deviation, Minimum,

and Maximum values of the Pitch parameter. Regarding the pulses, we computed the Number of pulses, Number of periods, Mean period, Standard deviation of period for each of the above sound files and we observed that there is no regular pattern in the parameter Pulses. For voicing parameter we computed the Fraction of locally unvoiced frames, Number of voice breaks and Degree of voice breaks. However for this parameter there were not regular patterns too. Then we computed the Harmonicity values as Mean autocorrelation, Mean noise-to-harmonics ratio, Mean harmonics-to-noise ratio and found that there is no regular pattern. After computing the mean-energy intensity of each sentences, we found that there is an incremental pattern in the Intensity by concerning the order Angry, Sarcastic and Sadness. Finally we computed the formant values as First formant, First Bandwidth, Second Formant, Second Bandwidth, Third formant, Third Bandwidth, fourth formant and forth bandwidth and found that there is no regular pattern in different formant parameters.

Although there are no regular patterns in most of the above speech parameters, in order to develop a more natural sounding speech synthesizer, however these parameters should be annotated with basic pronunciation computed from the authograpich text in speech synthesis. Therefore in future we hope to develop more generic probabilistic models based on this analysis to model above speech parameters for Sinhala speech synthesis.