Dynamic Time Warping Based Speech Recognition for Isolated Sinhala Words

P. G. N. Priyadarshani, N. G. J. Dias
Department of Statistics and Computer Science, University of Kelaniya

ABSTRACT

Communication between computer and the human is basically done through keyboard and screen-oriented systems. In the current Sri Lankan context, this restricts the usage of computers to a small fraction of the population, who are both computer literate and conversant with English. Accordingly, the major barrier between the computer and people in Sri Lanka is the language since English is not the mother tongue of most of the people and there is a large proportion of under educated people in rural areas of Sri Lanka. In order to enable a wider proportion of population to benefit from Information Technology, there is a dire need for an interface other than keyboard and screen interface that is widely used at present. The best solution is an efficient speech recognizer so that a natural human-machine interface could be developed to replace the traditional interfaces, such as keyboard and mouse of the computer. Further speech technologies guarantee to be the next generation user interface. For many languages speech recognition applications as well as text to speech synthesis applications have been developed and they have achieved a considerably high precision and applied them in real world applications successfully in developed countries. Even though currently there is no proper speech recognition approach for Sinhala language and the researches in this field in Sri Lanka is still in an infant stage.

Here we investigated the fitness of the dynamic programming technique called Dynamic Time Warping (DTW) algorithm in conjunction with the Mel Frequency Cepstral Coefficients (MFCC) to identify separately pronounced Sinhala words. One of the major difficulties in speech recognition is that although different recordings of the same words includes more or less the same sounds in the same order, the durations of each sub word within the word do not match. Consequently, when recognizing words by matching them with reference templates it gives inaccurate results if there is no temporal alignment. DTW solves this problem by accommodating differences in timing between test words and reference templates.

Converting the sound waves into a parametric representation is a major part of any speech recognition approach and here we have used MFCCs along with their first and second derivatives in time as the feature vector because they have been shown good performance in both speech recognition a well as in speaker recognition than other conventional speech features, In addition the derivatives reflect better dynamic changes of human voice over time. For extracting the features we divide speech signal into equally spaced frames and compute one set of features per frame as the speech signals are not stationary. We developed the reference

templates for each word from one example of that particular word per speaker and matched the test speech against to those reference patterns using DTW approach rather than other methods such as Vector Quantization and Euclidean distance because DTW can successfully deal with test signal and reference templates of the same word having different durations. The local distance measure is the distance between features at a pair of frames while the global distance from beginning of utterance until last pair of frames reflects the similarity between two vectors. Based on that, we could recognize the words that we input from our selected vocabulary.

In most of the systems developed based on DTW for other languages have been used very limited vocabulary for instance ten words but in this work we have used a considerably large vocabulary of 600 words. We obtained the recordings and separated each utterance and made an audio file for each using the software Praat. We developed the program in MATLAB 7.0. For our experiment we used two informants whose native language is Sinhala since we followed speaker dependent approach and tested each speaker separately, it displayed 80.33% overall accuracy.

Keywords: DTW, MFCC, Vector Quantization, Euclidean distance, template.