3.20 A tool for automatic derivation of phone transitions for the creation of a diphone database for Sinhala text to speech synthesis

K.H.Kumara and N.G.J. Dias
Department of Statistics and Computer Science, University of Kelaniya

ABSTRACT

Since the conventional user interfaces such as keyboard and monitors restrict the usage of computers, there is a dire need for an interface other than keyboard and screen-interface that is widely in use at present. Speech technologies promise to be the next generation user interfaces. In general, two technologies for processing speech are needed. One is speech recognition, and the other is speech synthesis. 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. TTS can be defined as "the production of speech by machines, by way of the automatic phonetization of the sentences to utter". Before a synthesizer can produce an utterance, several steps have to be completed. First, the right segments/units have to be selected. The units usually used are diphones, half-syllables, and triphones etc. Many synthesizers use diphones as their basic units of concatenation. A diphone is the transition between two speech sounds, obtained from natural speech. Creating a diphone database, which contains all the sound transitions in the target language, is critical in diphone TTS synthesis.

Diphone Studio, developed by MBROLA research team at the Facult´e Polytechnique de Mons (Belgium), is a software tool for developing and maintaining a set of diphones that can be used in Text to Speech synthesis. Before Diphone Studio can be used for recording of new set of diphones, a data file must be created, according to the following format:

<Left> <Right> <Wavefile> <Utterance> <Start> <Boundary> <End>

where <Left> <Right> is the name of the diphone, <Wavefile> is the name of the corresponding wave file, <Utterance> is the utterance from which the diphones is extracted, <Start> is the start point of the diphone segment, <end> is the end point of the diphone segment and <Boundary> is the transition point of the diphone.

A sample of a created .dat file is as follows:

```
!16000
e     k     w1.wav      <eka>       7171  8382  9555
k     e     w3.wav      <kelaya>    7419  7836  8762
..    ..     ……        <……..>      ……  ……. ……
```

At present this is done by manually and it is a very tedious as well as a time consuming task.

Therefore, we developed a tool that can create this .dat file for Sinhala Language by listing a set of possible diphones automatically. Once a text (single word, a sentence or set of sentences) is input to this tool, it is capable of creating/updating .dat file including all the possible phone transitions, already not in the .dat file, form the above text. However, at this stage input text should be in its authographic form and in the development of this tool allophone variations have not taken into consideration. In order to test the efficiency of this tool we followed the following method. Using the above tool we created a .dat file for a selected test dataset. Then with a help of linguistic specialist we derived the possible phone transitions of above test data set by manually without considering the allophone variations. Then both results were unified. Finally, we observed that the unification rate was 100%. In future, it is need to integrate an intelligent text preprocessor with this tool in order to enhance this tool for unrestricted input text which may contain numbers, abbreviations, special formats and formatting characters.