ABSTRACT

Reading is one of the prime ways to acquire knowledge. This era has been named as ‘Information era’, wherein every human always wants to acquire some information. People acquire knowledge through information they receive. When there is no or little time to read, systems which can aid a person to ‘speak out’ a text to be read, comes to picture. There exist systems that automatically extract data from the text and to generate text out of a given data. It would be convenient when text is replaced by speech. Speech Processing systems have come into existence where information is handled in terms of speech. Recent technological developments in the field of Computational Linguistics have been drastically influenced various speech recognition and synthesis systems. Speech synthesis plays a vital role in many assisting and aiding systems. Dedicated Text to Speech Synthesis systems are available for the Tamil Language. The synthesized speech needs to be more natural and human-like. Therefore the synthesizers need to concentrate upon appropriate speech units to be employed. We have selected syllables and diphones as the two suitable speech units, to be used at variable places of articulation. Explicit Prosody Modeling is also required to bring naturalness to the synthetic speech. These speech units are stored in a speech database, which is known as a Speech Corpus. Our research work concentrates on designing a Dual Database for a Tamil Text to Synthesis System. One database consists of syllable units and the other is a diphone corpus. Necessary prosody information is annotated to those units stored in the databases. Therefore, it is called as an annotated corpus. An annotated corpus consists of more than one entry for each speech unit, based on the Prosody information annotated to them. A clustering technique was also adapted to organize and speed up the process.

The method of synthesis adapted in the proposed synthesizer is the Concatenative synthesis technique. This methodology concatenates individual speech units to form a complete speech. When text is inputted to the system, it is divided into strings of text units. Each text unit is mapped with the corresponding speech unit in the speech corpus. Appropriate entry of the speech unit is selected from the speech corpus, based on the Prosody requirements. These selected speech units are concatenated to form a complete speech. The key aspects of prosody such as pitch, duration and intensity are modelled and evaluated. Still, there were some overheads concerning to the context switching of databases. The processing time taken for synthesizing each speech unit increased because of this context switching overhead. The proposed synthesizer produced consistent good results regarding the key aspects of Prosody, resulting in a human-like speech.