Preparation of a Free-Running Text Corpus for Maltese Concatenative Speech Synthesis

November 18, 2010

Text-to-speech systems based on concatenative speech synthesis employ the use of databases of recorded utterances which are stringed together to produce speech output. The corpus of recorded speech is segmented into units of concatenation such as individual diphones, and is often read from a training text compiled to provide a high degree of coverage of these basic units. The quality of output speech is highly dependant on the unit coverage of the speech database and in order to provide sufficiently natural speech output, large databases of recorded utterances are often required, spanning several hours. Often the training text is randomly sampled from a large corpus and no optimisations are applied towards the extraction of an optimal sample. Nevertheless, when building a database for an open domain application, recording every possible speech event from a random selection of sentences is practically impossible. In this paper we present a novel search function to maximise diphone coverage when choosing a training text for utterance recording.

The first step consisted of preparing a sufficiently-large text corpus; this was acquired from newspapers, websites, official documents and books written in Maltese. A text cleaning process was applied in order to arrive at a homogeneous corpus; different encodings, formats, abbreviations, semiotic elements present in the text were handled at this stage. A grapheme-to-phoneme conversion algorithm was then applied to the resulting text corpus of just over 33 million words. Since in Maltese the relationship between orthography and phonetics is relatively straightforward (low degree of heterography), a set of context-sensitive rewrite rules is generally sufficient for the phonemic transcription of Maltese text.

Statistical analysis was then performed on the phonetic transcription of the corpus, such as grapheme and diphone frequency counts, giving some interesting results like that 322 diphones (out of a possible 1681) are enough to cover over 90% of all Maltese text. Such statistics is then used to arrive at a much smaller free-text sample that is as representative as much as possible of the main corpus. This involves selecting phonetically-rich text blocks, made up of sentences of regular structure and reasonable length, that should enable the speaker to read them easily and with the expected prosodic patterns, so that naturalness is preserved.

An incremental greedy selection algorithm was developed to generate this free-text sample. The diphone coverage maximisation score used by the selection algorithm is based on diphone position and frequency. In the diphone position score we attempt to capture prosodic variations on each diphone, by matching the diphone position distribution in phrases and words: in phrases by unit position, in words by syllable number. By capturing phrase positions of diphones, we approximate variations due to intonation, while by capturing syllable positions we approximate stress in words.

In order to evaluate the effectiveness of this free-text selection method, a comparison was made with other selection algorithms, both random-based selection methods and against a free-text prepared manually by an expert. The results obtained show that the novel selection algorithm presented in this paper outperforms all the others and arrives at a free-text sample that is highly representative of the main corpus.