Efficient Algorithm for Transcription of Spontaneous Speech
Ami Moyal
ACLP – Afeka Center for Language Processing,
Afeka College of Engineering, Tel Aviv, Israel
E-mail: amim@afeka.ac.il

Transcription of spontaneous speech is a major requirement for various applications. One such application, Voice Mail (VM) transcription, is the primary focus of this project. The main business motivation behind the transcription of VM messages is to offer users the option of receiving VM messages via SMS (Short Textual Message) rather than dialing into the VM system and listening to the messages.

An analysis of VM content shows that messages contain all of the characteristics of spontaneous speech: very large vocabulary, hesitations, interjections, deleted words, truncated syllables, unstructured speech and a large number of names. In the case of VM speech, the number of names used is even more substantial.

The project focuses on spontaneous speech transcription using a multi-stage speech recognition engine. The first stage converts speech to phonemes using a phoneme recognizer. The second stage identifies the most probable hypothesized words in the phoneme string. The third stage composes the sequence of words using previous stage word list output.

This sequential structure differs from classical approaches to Large Vocabulary Continuous Speech Recognition (LVCSR) mainly due to the inclusion of the second stage. By identifying the most probable words, the vocabulary is substantially reduced, thus decreasing recognition time and improving performance at the third stage.

The main target of the project is to enable the development of a product that can perform VM transcription using a huge representative lexicon of words required for the task.

The project includes several parallel paths including: definition and construction of a large lexicon of words including a large relative ratio of names; algorithmic development of efficient hypothesis calculations using anchors; algorithmic development of an efficient phoneme similarity measure; and testing the multi-stage speech recognition engine on a database of real VM messages.

The design goal of the lexicon is for at least 90% coverage of target words. Common words were gathered from several spontaneous speech databases (including voice messages) and were cross-referenced with textual sources from the internet. The list includes hesitations, fillers and colloquial vocabulary common to spontaneous speech. The lexicon will include multi-transcriptions for both common words and names. A lexicon comprising 100,000 entries has already been compiled.

At the algorithmic level, the main goal is to develop an algorithm that will significantly reduce search complexity with minimal effect on recognition performance. An exhaustive search of words in a string of recognized phonemes would assume that any word from the lexicon can occur at any location on a given phoneme string. This means that the number of potential hypotheses is substantial, particularly when using a very large lexicon. The approach proposed by this project allows a long string of phonemes to be effectively matched against a large lexicon by relying on both linguistic and statistically based anchor points for formulating word hypotheses.

There are three main issues in the second stage responsible for the vocabulary reduction phase: hypothesis creation, distance measurement and N best word selection. At present, a common method used for distance measures is the Levenshtein distance - a textual distance measure. The Levenshtein measure is a dynamic programming based method with an O(n^2) complexity – which is extremely problematic when dealing with very large lexicons. The goal of this research is to identify a more efficient distance measure that will reduce computation time without adversely affecting the quality of the results.

For testing purposes, IBM’s Voicemail I and Voicemail II were licensed from LDC (Linguistic Data Consortium).

Current results show a decrease of up to 90% in the computation of the second stage compared to exhaustive search, while word coverage in the N best output words decreases only by 2%. At present, the overall multi-stage transcription engine is being compared to a classical LVCSR engine.

This research is part of grant (#41914) provided by the Chief Scientist of the Israeli Ministry of Commerce and is being carried out as part of the Magneton program which encourages the transfer of research results from academic institutions to industrial companies – In this case Afeka Academic College of Engineering and SpeechModules LTD.