A PUBLIC DOMAIN SPEECH-TO-TEXT SYSTEM

N. Deshmukh, A. Ganapathiraju, J. Hamaker, J. Picone

Institute for Signal and Information Processing
Mississippi State University, Mississippi State, MS 39762
{deshmukh, ganapath, hamaker, picone}@isip.msstate.edu

M. Ordowski

Department of Defense
Ft. Meade, MD 20755
mordow@afterlife.ncsc.mil

ABSTRACT

The lack of freely available state-of-the-art Speech-to-Text (STT) software has been a major hindrance to the development of new audio information processing technology. The high cost of the infrastructure required to conduct state-of-the-art speech recognition research prevents many small research groups from evaluating new ideas on large-scale tasks. The Institute for Signal and Information Processing (ISIP) has been committed to providing the research community with free software tools for digital information processing via the Internet to facilitate worldwide synergistic development of speech recognition technology.

In this paper, we present the core components of an available state-of-the-art Speech-to-Text system: an acoustic processor which converts the speech signal into a sequence of feature vectors; a training module which estimates the parameters for a Hidden Markov Model; a linguistic processor which predicts the next word given a sequence of previously recognized words; and a search engine which finds the most probable word sequence given a set of feature vectors.

By far, the most important component of a Speech-to-Text system is the search engine or decoder. The decoder was designed to be modular and extensible in order to be able to handle a wide variety speech recognition problems (connected digits, studio-quality read speech and spontaneous telephone conversations) in a transparent fashion. The process of moving from a well defined task to a less rigorously defined recognition problem (Spontaneous Speech Recognition, i.e. Switchboard) requires the decoder to have a sophisticated control structure. Hence very few good decoders exist and the best decoders are always considered proprietary.

The ISIP decoder has the capability to compile network grammars, efficiently decode n-gram language models, generate and rescore lattices, generate N-best lists, and perform forced alignments. The decoder is based on a hierarchical Viterbi, breadth-first search tree which will support cross-word triphone acoustic models. The decoder uses lexical trees to represent the pronunciations of all words. The decoder uses beam pruning at the state, phone and word levels and limits the number of active model instances per frame to prevent the evaluation of low-scoring hypothesis. A benchmark evaluation (which does not include MLLR or vocal tract normalization) conducted on a subset of the Switchboard corpus yielded a WER of 46.1% at 30xRT. This is competitive with commercially available Speech-to-Text systems.

The ISIP Speech-to-Text system currently produces mel-frequency scaled cepstral coefficients and is capable of estimating the mixture densities using Viterbi training. The design of the acoustic processor will allow other feature sets to be easily incorporated into the ISIP Speech-to-Text system.

Finally, some experimental results of the complete system will be presented in this paper. To obtain further information of the ISIP Speech-to-Text system the following URL is available: *http://WWW.ISIP.MsState.Edu/resources/technology/projects/speech_recognition/*.