

Investigators: Bilmes, Byrd, Jurafsky, Kirchoff, Manning, Narayanan □
Agency: MURI-Office of Naval Research □
Dates: June 2005-2010

Project Summary

Computer recognition of speech is a crucial application for the Department of Defense and a key challenge for the scientific and engineering goals of our nation. Current speech-to-text technology has attained impressive performance on constrained data, primarily through painstaking engineering and the use of massive amounts of training data. The existing technology, however, still falls short of human performance in many situations, including accented speech, varying channel and noise conditions, or mismatches in training and testing conditions. To overcome these limitations, we must depart from the current dominant Hidden Markov Model (HMM) paradigm.

We introduce a radically new, unified approach to speech-to-text which replaces the HMM at all levels yet is firmly grounded in a powerful statistical machine learning framework. It is motivated by insights into human speech perception, specifically recent studies of human word recognition in everyday conversations. These indicate that large improvements in speech-to-text under diverse conditions cannot be obtained by changing only isolated system components. Rather, human robustness derives from flexibly combining high-level and low-level knowledge sources, and from integrating these in highly context-dependent ways and over long time spans.

We incorporate these insights by introducing a strategy termed **attention-shift decoding**: speech is not processed sequentially, but hierarchically, in accordance with time-varying assessments of the salience/relevance of different signal regions, and using sophisticated and appropriately combined knowledge sources. First, acoustic-prosodic information pre-categorizes speech into reliable and unreliable regions, and provides prominence and boundary information. Then, word hypotheses are generated for reliable regions while partial analyses are generated for unreliable regions. The latter are subsequently filled in using novel **discriminative word classifiers** which combine acoustic, prosodic and contextual information (obtained by dedicated extraction modules) to directly choose between competing word hypotheses. This process is constrained by **tightly coupled speech and language processing**, in the form of a probabilistic **spoken-language parser** that uses prosodic and disfluency information and is integrated into our language model. The development of this architecture will be supported by psycholinguistic **human lexical access** experiments on the one hand, and by speech production data on the other. We will collect and distribute a new **20-hour MRI/EMA articulatory corpus** of spontaneous conversational speech to facilitate the analysis of coarticulation and reduction phenomena, and for co-training acoustic models.

Our approach will be implemented using the formal mathematical infrastructure of statistical **dynamic graphical models** (DGMs), in which diverse and powerful ideas may be expressed in simple and computationally efficient ways. DGMs can efficiently accommodate highly structured domains, multiple knowledge sources, and missing observations. We will express all of our models and knowledge integrations methods **discriminatively** in this paradigm, and will moreover use DGMs to build novel **asynchronous & multi-resolution** models.

We have assembled an outstanding team of six researchers at three universities: **University of Washington, Stanford University, and University of Southern California**. Our team includes computer scientists, electrical engineers, linguists, and psycholinguists, with expertise in natural language processing, linguistics/phonetics, speech recognition/machine-learning and high-performance computing. Our laboratories have a history of exploring new, creative directions in speech and language processing, and of high-quality graduate and undergraduate student training, which will form an integral part of this project.