Acoustic Modeling for Multi-Language, Multi-Style, Multi-Channel Speech Recognition

Mark Hasegawa-Johnson, University of Illinois at Urbana-Champaign

Automatic speech recognizers exist for a few dozen of the most widely spoken languages in the world. In most of those languages, speech recognizers exist only for read speech, spoken with a standard accent in a quiet environment. Porting to a new language, dialect, or speaking style requires, typically, a large amount of labeled speech data. Data requirements can be reduced somewhat if acoustic models for the new recognizer can be ported or adapted from the acoustic models learned in other languages, dialects, and speaking styles. This talk will describe the methods used to port acoustic models among component languages in a multi-language, multi-style, multi-channel speech recognizer currently under development at the University of Illinois, designed and tested using open-source software packages published by institutions around the world. The core technology is a detailed context description. Every word in every training file is broken into phones using standard prononunciation dictionaries (many of which are open source), and each phone is annotated with a long list of context variables automatically derived from the orthographic transcription and from the waveform itself: phonetic context (what phones are to the left and right), syllable position, stress, word position, sentence position, sentence mood, part of speech, speaker gender, speaker category, voice quality, speaking style, speaking rate, language, channel type and background noise. Standard normalization and adaptation methods are used to reduce the acoustic importance of channel, background noise, and speaker identity. Context variables are then organized into a tree structure, for each root phone, according to their acoustic importance. Context variables that don't affect the way a phone is pronounced are discarded as irrelevant; context variables that affect pronunciation are used to divide the training examples into different context-dependent allophones. During recognition of speech in a known language, all context variables are uniquely specified by either the audio signal or the candidate word string, hence no extra ambiguity is introduced by this type of fine-grained context encoding. In order to port the recognizer to a new language, candidate acoustic models for each phoneme are proposed on the basis of acoustic similarity, or (if there are no acoustic data available) on the basis of linguistic similarity. The set of models selected in this way can be applied to speech files recorded in the new language, with or without a dictionary, with or without any acoustic training data, in order to generate a lattice of candidate phone transcriptions suitable for information retrieval applications. Similar methods have been used by teams in Europe to port software among the European languages, and similar methods are being developed by several sites participating in the Singaporean Star Challenge information retrieval task; methods used in the Illinois recognizer differ from the state of the art primarily in the variety and selection of context features. An open debate in the field is the degree to which mismatch among the acoustic units in different languages, or in different speaking styles, can be compensated using targeted feature transformation methods. Research at the 2006 Johns Hopkins workshop demonstrated, for example, that speech recognition error rates in English may be reduced using a hybrid neural network-Bayesian network architecture, in which the auditory spectrum is augmented by estimated phonological distinctive features

prior to speech recognition. Neural network adaptation methods, or Bayesian network adaptation methods, might be able to rapidly transform such a system from one language to another, thus reducing error rates using an extremely small amount of labeled training data in the new language.