Generalization Problem in ASR Acoustic Model Training and Adaptation

Sadaoki Furui
Department of Computer Science, Tokyo Institute of Technology

Since speech is highly variable, even if we have a fairly large-scale database, we cannot avoid the data sparseness problem in constructing automatic speech recognition (ASR) systems. How to train and adapt statistical models using limited amounts of data is one of the most important research issues in ASR. This paper summarizes major techniques that have been proposed to solve the generalization problem in acoustic model training and adaptation, that is, how to achieve high recognition accuracy for new utterances. One of the common approaches is controlling the degree of freedom in model training and adaptation. The techniques can be classified by whether a priori knowledge of speech obtained by a speech database such as those spoken by many speakers is used or not. Another approach is maximizing "margins" between training samples and the decision boundaries. Many of these techniques have also been combined and extended to further improve performance. Although many useful techniques have been developed, we still do not have a golden standard that can be applied to any kind of speech variation and any condition of the speech data available for training and adaptation.

It's Not You, It's Me: Automatically Extracting Social Meaning from Speed Dates

Dan Jurafsky
Department of Linguistics and, by courtesy, Department of Computer Science Stanford University

Automatically detecting human social intentions from spoken conversation is an important task for social computing and for dialogue systems. We describe a system for detecting elements of interactional style: whether a speaker is awkward, friendly, or Participants rated themelves and each other for these elements of style. Using rich dialogue, lexical, and prosodic features, we are able to detect flirtatious, awkward, and friendly styles in noisy natural conversational data with above 70% accuracy, significantly outperforming not only the baseline but also, for flirtation, outperforming the human interlocutors. We find that features like rate of speech, pitch range, energy, and the use of questions help detect flirtation, collaborative conversational style (laughter, questions, collaborative completions) help in detecting friendliness, and disfluencies help in detecting awkwardness. In analyzing why our system outperforms humans, we show that humans are very poor perceivers of flirtatiousness in this task, and instead often project their own intended behavior onto their interlocutors. This talk describes joint work with Dan McFarland (School of Education) and Rajesh Ranganath (Computer Science Department).