Speech Recognition with Segmental Conditional Random Fields

Geoffrey Zweig and Patrick Nguyen
Senior Members: Dirk van Compernolle, Kris Demuynck, Les Atlas, Fei Sha

The goal of this workshop group is to advance the state-of-the-art in core speech recognition by developing new kinds of informative segment level features for use in a Segmental Conditional Random Field (SCRF). In the recently proposed SCRF approach [Zweig & Nguyen 2009], we generalize Conditional Random Fields to operate at the segment level, rather than at the traditional frame level. Key to our approach, each segment is labeled directly with a word. To simplify somewhat, denote a specific segmentation by **q** and specific word sequence by **w**. We compute the conditional probability of the word sequence using a set of features f_i which are combined in a log-linear model with weights λ_i:

$$P(\vec{w} \mid \vec{o}) = \frac{1}{Z} \sum_{\vec{q}} \exp\left(\sum_i \lambda_i f_i(\vec{w}, \vec{o}, \vec{q})\right)$$

While we refer the reader to [Zweig & Nguyen ASRU 2009] for the full details of this approach, intuitively the features which are extracted for a segment each measure some form of consistency between the underlying audio and the word hypothesis for the segment. In previous work, we have used features based on the detection of phoneme and multi-phone units in the audio input. Specifically, one feature is the edit distance between the observed (detected) phoneme sequence in a segment, and that expected based on the hypothesis.

We believe the use of SCRFs has several advantages which can be explored in the workshop:

1) By operating at the segment level, we can use long-range features that are otherwise difficult to represent (for example, the edit distance between expected and observed detector events)
2) Again by operating at the segment level, we can use long-span template-based features that are potentially more robust to noise than single frame-level features
3) The model supports segment and label-dependent signal processing, which is a relatively unexplored and powerful area
4) The method is inherently discriminative in nature, and we are able to jointly train the acoustic and language models in a discriminative way

To support a research effort in these areas, we have compiled an internationally recognized multi-disciplinary team of academic and industrial researchers. Professor Van Compernolle and Dr. Demuynck at Leuven University in Belgium have done groundbreaking work on template based ASR [Wachter et al. 2007, Demange & Compernolle 2009], which will form the foundation of one set of features. Another line of research will revolve around the use of coherent modulation features, pioneered by signal processing expert Prof. Les Atlas at the University of Washington. Professor Fei Sha at the University of Southern California will contribute deep knowledge of machine learning. Drs. Zweig and Nguyen at Microsoft Research have developed the theory of segmental CRFs in speech recognition, and released a toolkit to implement it. Dr. Zweig will lead the workshop. To provide further scientific background for this work, we anticipate visits by Malcolm Slaney and Shihab Shamma to discuss their relevant work, and input from Chief Scientist Ken Church at the JHU COE. In concluding, we would like to note that this proposal complements the research being conducted at JHU itself, and Professors Hermansky and Andreou are anticipated to contribute.