Florian Metze
Carnegie Mellon University

Connectionist Temporal Classification for End-to-End Speech Recognition

The performance of automatic speech recognition (ASR) has improved tremendously due to the application of deep neural networks (DNNs). Despite this progress, building a new ASR system remains a challenging task, requiring various resources, multiple training stages and significant expertise. In this talk, I will present an approach that drastically simplifies building acoustic models for the existing weighted finite state transducer (WFST) based decoding approach, and lends itself to end-to-end speech recognition, allowing optimization for arbitrary criteria. Acoustic modeling now involves learning a single recurrent neural network (RNN), which predicts context-independent targets (e.g., syllables, phonemes or characters). The connectionist temporal classification (CTC) objective function marginalizes over all possible alignments between speech frames and label sequences, removing the need for a separate alignment of the training data. We present a generalized decoding approach based on weighted finite-state transducers (WFSTs), which enables the efficient incorporation of lexicons and language models into CTC decoding. Experiments show that this approach achieves state-of-the-art word error rates, while drastically reducing complexity and speeding up decoding when compared to standard hybrid DNN systems.

Florian Metze is an Associate Research Professor at *Carnegie Mellon's University*, and the Associate Director of the *InterACT center*, at *CMU's Language Technologies Institute.* He holds a PhD from *Universität Karlsruhe (TH)*, for a thesis on "Articulatory Features for Conversational Speech Recognition". His current work is centered around speech and multi-media processing with a focus on low resource and multi-lingual speech processing, large-scale multi-media retrieval and summarization, along with recognition of personality or similar meta-data from speech.