RECENT IMPROVEMENTS TO NEURAL NETWORK BASED ACOUSTIC MODELING IN THE EML REAL-TIME TRANSCRIPTION PLATFORM

V. Fischer, O. Ghahabi, S. Kunzmann

EML European Media Laboratory, Berliner Straße 45, D-69120 Heidelberg
fischer@eml.org

Abstract: In this paper we report some recent improvements to DNN/HMM hybrid acoustic modeling for the EML real-time large vocabulary speech recognition system, including the introduction of speaker adaptive long short term memory units (LSTMs) and efficient online decoding with deep bidirectional LSTMs. Based on a thorough latency analysis of our baseline large vocabulary speech recognizer we first abandon multi-pass recognition with fMLLR adapted acoustic features and further simplify decoding by dropping text independent vocal tract length normalization (VTLN) which was identified as a major bottleneck for real time applications. Subsequently, we improve accuracy by a variety of measures that include artificial training data augmentation and the use of additional features derived from an online speaker diarization module currently under development. Moreover, we investigate a hierarchy of feed forward and recurrent neural networks for a further reduction of word error rate. Finally, we demonstrate that established DNN pruning techniques are also applicable to bidirectional LSTMs, resulting in both an appropriate network size and substantial runtime savings. Our experiments are carried out on the publicly available WSJCAM0 corpus. Being simultaneously recorded with both a head-mounted and a desk-mounted microphone, it enables us to study the impact of each of the proposed methods also in case of a channel mismatch between training and test data. The methods described in this paper yield an improvement of up to 15 percent relative to the baseline DNN/HMM acoustic model.