# Summary: End-to-End Speech Recognition in English and Mandarin

# A simple alternative to hand-engineered traditional speech recognition systems is deep learning with an end-to-end training of the model. Deep learning eliminates many processes often necessary for the state-of-art ASR models. The paper highlights how porting from one language to another in deep learning is significantly easier as compared to traditional speech recognition systems since the latter system has to take care of language-specific developments. The only change the authors had to make to their model when porting from English to Mandarin was to change the number of output layers which depend on the number of characters in each language. The goal of the paper was achieving human-level performance in speech recognition. It turned out this model outperforms human performance at times. This paper sites three things that proved to be very crucial for the success of their model - extensive experimentation on model architectures, really large datasets and use of high computation GPUs.

# In recent times, most ASR systems have gone the deep neural network path using both convolutional as well as recurrent neural networks. The authors claim that Recurrent neural networks used along with CTC loss function work very well for this end-to-end model. The model used in this paper consists of few convolutional layers followed by many recurrent layers and a fully connected layer ending with a softmax layer used with CTC loss function. Inputs for this model are log-spectrograms and output is alphabets obtained using a specialized beam search. More amount of data leads to increase in recurrent layers making gradient descent technique ineffective to some extent. The authors chose to use an alternate method called Batch Normalization, which not only decreased convergence time but also improved generalization error. In addition, a sequence-wise normalization was used which normalizes only the non-recurrent connections as opposed to normalizing all non-linearity which turned out to ineffective. It turned out that even Batch Normalization with CTC proved to be unstable due to exploding and volatile gradients when long transcriptions were encountered. The authors introduced a new technique which deals with this problem by simply training the model taking the length of words into account (shorter ones coming before during the first epoch only). Also, to capture variable length utterances more accurately the model uses few convolutional layers specifically 2D (time and frequency) instead of fully connected layers which do not do a good job in comparison. Bidirectional RNNs perform well in ASR as compared to unidirectional models, but since they are difficult to deploy, a special lookahead convolution is used to give some future context in unidirectional models. During deployment in an online setting, and while using bidirectional models a technique called Batch Dispatch is used that gives a stream of data to the network in batches increasing latency crucial for bidirectional models to read the future context.

# One important factor for the success of this model was using a large amount of labeled data for both English and Mandarin. Additionally, bootstrapping large datasets and data augmentation achieved by adding noise in each epoch were used to increase the size of labeled data. The authors noticed that increasing the data available for training further decreases the Word Error Rate (WER) in both noisy and regular datasets.

# Training speed being a crucial factor, a high-performance computing infrastructure was used where important routines were carefully optimized and memory allocators were used. Training was done in parallel fashion on various GPUs each using a local copy of the model and then exchanging the calculated gradient. In addition, highly optimized kernels were used. The authors acknowledge how model and data parallelization proved to be a good catalyst in improving speed.

The authors found out that their English model outperformed human workers in speech recognition on 3 out of 4 test sets. This model outperformed humans for various accents too except the Indian accent. Also, this model performed exceptionally well in recognizing noisy data as compared to humans. Similar results were achieved with Mandarin model too.