ACHIEVING HUMAN PARITY IN CONVERSATIONAL SPEECH RECOGNITION

Summary

This paper compares professional transcribers – people who listen to recorded speech and type out what they hear – with the Microsoft Speech Recognition system and based on the results conclude that modern automated systems have reached human parity. The NIST 2000 Test set, which is a part of an ongoing Speaker Recognition Evaluation conducted by the National Institute of Standards and Technology, is used for this research. The NIST 2000 Test set uses data collected from the Switchboard corpus, which has recorded speech of people talking about different assigned topics, and data from CallHome, which are records of family members and friends having open ended conversation. The speech recognition system uses a number of convolutional and Long Short-Term Memory (LSTM) acoustic model architectures. An acoustic model is a file that contains statistical representations of each of the distinct sounds that makes up a word. Novel spatial smoothing, which means that data points are averaged with their neighbours, is used alongside Lattice free MMI acoustic training, which is a method of training neural network acoustic models, for the system. Multiple recurrent neural network language modelling approaches and a systematic use of system combination has also been used.

Key Points

- Both the Speech recognition system and the professional transcriber were given the exact same audio segment. A single channel of audio is given do that is much easier to identify words.

- Performance of the system aligned almost exactly with the transcriber. The result was an error rate of 5.9% and 5.8% for the Switchboard portion and 11.3% and 11% for the professional transcriber and speech recognition system respectively.

- The human and machine errors are almost the same except in terms of backchannel words and hesitations where the system seems to mix them up while people don't. Backchannel words are words like "Uh huh" and "hmm" which a listener uses to interject words to the speaker.

Questions

1. Will Speech Recognition ever be perfect?

2. Will the speech recognition system work as efficiently in everyday conditions where the background noise levels varies a lot?

3. Can the problem of the system failing to distinguish between backchannel words and hesitations be fixed?