R. MOUNIR, R. ALQASEMI, R. DUBEY. Assisting persons with speech impediments using Recurrent Neural Networks. Gerontechnology 2018;17(Suppl):105s; https://doi.org/10.4017/gt.2018.17.s.102.00 **Purpose** This work focuses on the research related to enabling individuals with speech impairment to use speech-to-text software to recognize and dictate their speech. Automatic Speech Recognition (ASR) tends to be a challenging problem for researchers because of the wide range of speech variability. Some of the variabilities include different accents, pronunciations, speeds, volumes, etc. It is very difficult to train an end-to-end speech recognition model on data with speech impediment due to the lack of large enough datasets, and the difficulty of generalizing a speech disorder pattern on all users with speech impediments. This work highlights the different techniques used in deep learning to achieve ASR and how it can be modified to recognize and dictate speech from individuals with speech impediments. **Method** The project is split into three consecutive processes; ASR to phonetic transcription, edit distance and language model. The ASR is the most challenging due to the complexity of the neural network architecture and the preprocessing involved. We apply Mel-Frequency Cepstrum Coefficients (MFCC)[1] to each audio file which results in 13 coefficients for each frame. The labels (text matching the audio) is converted to phonemes using the CMU arpabet phonetic dictionary[2].The Network is trained using the MFCC coefficients as inputs and phonemes' IDs as outputs. The Network architecture implemented is a Bidirectional Recurrent Deep Neural Network (BRDNN)[3], it consists of 2 (one in each direction) LSTM cells with 100 hidden blocks in each direction. The network is made deep by stacking two more layers, which results in a 3 layers network in depth. Two fully connected layers were attached to the output of the recurrent network with 128 hidden units in each. This architecture resulted in a 38.5% Label Error Rate (LER) on the Test set. Levenshtein edit distance[4] is used to generate potential words from phonemes. The language model uses the potential words to generate sentences with the most semantic meaning. The language model is another recurrent neural network model trained on full sentences. The model outputs the probability of a word occurring after a given word or sentence. It is simpler than the main speech recognition model because it is not bidirectional and not as deep. The language model uses beam search decoding to find the best sentences. **Results & Discussion** *Figure 1* shows the number of words found per sentence at every edit distance. subjects having no accent found significantly more words, at an edit distance of one, than subjects with accents. As we increase the edit distance, the words/sentence found increase for all the data points. This concludes that it is recommended to increase the edit distance for data with speech impediment to acquire better results (given a good language model). Increasing the edit distance and beam width can be translated into relying more on the language model than on speech recognition. The idea of relying more on the language model and less on speech recognition seems to be comparable to how the human mind tries to understand distorted audio. We know that the speech must have semantic meaning so we generate words close to what we hear by changing the phonemes and connecting them in a way that makes sense grammatically and logically. The amount of words we generate from each distorted word depends on the severity of distortion that we expect from the audio.

References

1. P. Cryptography. Mel Frequency Cepstral Coefficient (MFCC) tutorial. Available from: practicalcryptography.com
2. C. M. U. The CMU Pronouncing Dictionary. The CMU Pronouncing Dictionary. [Online]. Available from: http://www.speech.cs.cmu.edu/cgi-bin/cmudict
3. Graves A, Jaitly N. Towards End-to-End Speech Recognition with Recurrent Neural Networks
4. Gilleland M. M. P. Software. Levenshtein Distance, in Three Flavors. [Online]

Keywords: speech, neural networks, speech-to-text
Address: University of South Florida, USA;
E: ramy@mail.usf.edu

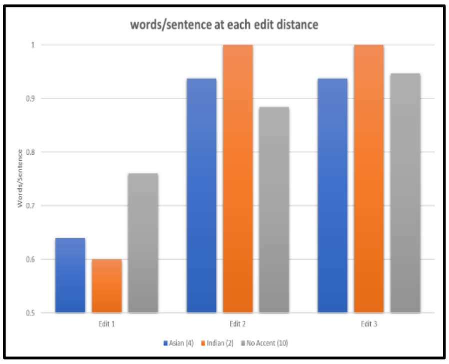

Figure 1. Words/sentence found at each edit distance for different accents