TOWARDS AUTOMATIC DYSARTHRIC SPEECH RECOGNITION

L. Mendez[1], C. Leyva[1], V. Vega[1], M. A. Umana[2], J. M. Tua[1], W. Aponte,[2] R. Rosado[1], N. G. Santiago[2]
Computer Science and Engr. Dept.[1]
Electrical and Computer Engr. Dept.[2]
University of Puerto Rico – Mayaguez, Mayaguez, Puerto Rico, 00681

Keywords: Speech recognition, dysarthria, seq2seq, knowledge distillation, end-to-end

A motor speech disorder is the damage or interruption of connections between parts of the brain in charge of speech motor planning or muscle tone (Motor Speech Disorders, 2021). Dysarthria is a motor speech disorder, and it is normally caused by medical conditions that affect the muscles used to produce speech. These muscles are damaged, paralyzed, or weakened, causing the person to have no control over it and make it difficult for them to pronounce words (Dysarthria, n.d.). However, these articulatory errors are not random, unlike in the case of apraxia (Kim, et al., 2008). "The location and severity of the brain lesion can determine the acoustics of the sound waves" (Motor Speech Disorders, 2021). Symptoms, such as slurred speech, varied rate of speech, among other acoustic anomalies, vary based on location and severity, therefore distinct acoustic outputs are due to the location and severity. These characteristics make it unique and that's the reason that traditional Automatic Speech Recognition (ASR) cannot identify their speech.

In the past, there have been studies of using Hidden Markov Models (HMM) with transition-interpolation to determine bidirectional probability from a sequential input. This model gave the best results when the speaker had severe dysarthria, suggesting that the severity of the dysarthria is not a sufficient indicator for the performance of the models (Vardhan Sharma & Hasegawa-Johnson, 2010). Another architecture previously proposed by Shor, et al. was a Recurrent Neural Network Transducer (RNN-T), which was originally trained with non-dysarthric speech and fine-tuned with dysarthric speech (Shor, et al., 2019). More recently, there was the implementation that uses the Transformer in a Multi-Task Learning Architecture (Ding, Sun, & Zhao, 2021) which improves the features learning by separating the domain into multiple tasks within the same domain (Caruana, 1997). In their proposed approach they used an attention-based architecture along with a Connectionist Temporal Classification (CTC) to divide the task into the sub-domains (Ding, Sun, & Zhao, 2021). The methods described are computationally intensive and do not incorporate information about the structure of the language that may reduce errors as well as computation time.

The proposed approach is to use an end-to-end architecture for natural language processing, along with Knowledge Distillation (KD) to develop a model that understands dysarthric speech. Through this process, an initial model is trained to understand non-dysarthric speech, and then used as the Teacher to train another model in a Response-Base Knowledge environment (Gou, Baosheng, Maybank, & Tao, 2021). In this environment the second model (known as the student) is optimized to interpret the input data of dysarthric speech and return a similar output as the non-dysarthric speech; both inputs have the same label.

The data for non-dysarthric speech is obtained from the LJ Speech Dataset composed of 13,100 audio clips and containing a total of 13,821 distinct words. For the dysarthric data,

we are going to be using the UA Speech Dataset that consists of 765 isolated words per speaker for a total of 19 speakers with cerebral palsy (Ito & Johnson, 2017). For the latter dataset, each speaker was classified into four categories; very low, low, mid, and high intelligibility (Kim, et al., 2008). This dataset provides the waveforms and spectrograms which can be used to train our model. Both datasets are labeled, and this data is saved in their own CSV which includes what's being said in the audio and the sample to which audio that label represents.

This environment was originally proposed for cross-modal knowledge distillation (Wang, et al., 2020). Their proposed method to develop a text-to-speech autoencoder for dysarthric speech reconstruction achieved a Word Error Rate (WER) between 9.33% and 48.65%, depending on the intensity of the dysarthria. Granted their problem really was transforming the dysarthric speech into non-standard, through Voice Conversion; we believe that using knowledge distillation can be used for Automatic Speech Recognition.

Currently, for the Knowledge Distillation Environment, we have been studying both Recurrent Auto-Encoders using Long-Short Term Memory (LSTM) and Transformer architectures. The latter one uses a multi-head attention layer along with positional encoding to interpret the relationship between the values within the sequence (Vaswani, et al., 2017). A model that became state-of-the-art when it surpassed previous models in the WMT 2014 English-to-German and WMT 2014 English-to-French translation tasks. On the other hand, we have the LSTMs, which is a type of RNN that uses a series of gates to control the information that gets passed on to the recurrent state by forgetting some information and only keeping the relevant one, thus minimizing the issue of Vanishing Gradient Descent (Hochreiter & Schmidhuber, 1997). Both architectures have shown great results with Natural Language Processing and other Sequential Problems and follow the encoder-decoder structure that is needed for the proposed approach.

References

Caruana, R. (1997). Multitask Learning. *Machine Learning*, 41–75. doi:https://doi.org/10.1023/A:1007379606734

Ding, C., Sun, S., & Zhao, J. (2021). Multi-Task Transformer with Input Feature Reconstruction for Dysarthric Speech Recognition. *International Conference on Acoustics, Speech and Signal Processing* (pp. 7318-7322). IEEE.

Dysarthria. (n.d.). Retrieved 09 18, 2020, from Speech and Hearing Sciences - University of Washington: https://sphsc.washington.edu/dysarthria

Gou, J., Baosheng , Y., Maybank, S. J., & Tao, D. (2021). Knowledge Distillation: A Survey. *Int J Comput Vis*, 129.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Comput*, 1735–1780. doi:https://doi.org/10.1162/neco.1997.9.8.1735

Ito , K., & Johnson, L. (2017). *The LJ Speech Dataset*. Retrieved 09 2020, from Keithito: https://keithito.com/LJ-Speech-Dataset

Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J., Watkin, K., & Frame, S. (2008). Dysarthric speech database for universal access research. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 1741-1744.

Motor Speech Disorders. (2021). Retrieved from Midwestern University Clinics: https://www.mwuclinics.com/illinois/services/specialty/speech-language/motor-speech-disorders

Shor, J., Emanuel, D., Tuval, Tuval, O., Brenner, M., Cattiau, J., . . . Matias, Y. (2019, 09). Personalizing ASR for Dysarthric and Accented Speech with Limited Data. doi:10.21437/interspeech.2019-1427

Vardhan Sharma, H., & Hasegawa-Johnson, M. (2010). State-Transition Interpolation and MAP Adaptation for HMM-based. *Proceedings of the NAACL HLT 2010 workshop on speech and language processing for assistive technologies.*, 72-79.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention Is All You Need. *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 6000–6010). Long Beach, California, USA: Curran Associates Inc.

Wang, D., Yu, J., Wu, X., Liu, S., Sun, L., Liu, X., & Meng, H. (2020). End-to-End Voice Conversion Via Cross-Modal Knowledge Destillation. *ICASSP*. Spain: IEEE.