Overcoming data limitations in deep learning-based dysarthric speech processing

Neuromuscular disorders such as Parkinson's disease, amyotrophic lateral sclerosis (ALS), Friedrich ataxia, and Huntington's disease often have a pronounced impact on speech. They precipitate a breakdown in fine motor control that often gives rise to dysarthria—a slurred, slow, monotone, raspy, nasal or uneven sort of speech that is often less intelligible than healthy speech [1]. Due to the precision of muscular control required to quickly and accurately render speech phonemes, and the subconscious nature of the discrete muscle movements that compose speech, speech symptoms are often the first to manifest under these conditions. Thus, dysarthric speech is an appealing topic for study, serving as *a rich source of biomarkers which may be used to diagnose neurological disease and track its progression* [2].

Some such biomarkers, such as speaking rate, are easy to objectively assess in a clinical setting. Anyone can use a stopwatch to time a patient reading a passage aloud. However, many useful biomarkers—such as the slurred, raspy, or nasal qualities—are perceptually defined, making them much more difficult to assess objectively. Hypernasality, a lack of ability to properly regulate airflow between the oral and nasal cavities, has been demonstrated to be an early indicator of ALS and Huntington's disease, for example [3], making its estimation desirable in clinical applications. The gold standard in assessing hypernasality in speech is the opinion score of a trained speech pathologist; yet even these expert clinicians have been shown susceptible to errors [4]. Perhaps more problematic is the need for a skilled speech clinician to provide the rating, making such assessment costly and difficult to scale to broader use in neurological clinical settings.

These limitations in human perceptual assessment of speech biomarkers drive interest in their automated objective assessment. Many of the pioneering techniques for neurological speech symptom estimation have been derived from work in automatic speech recognition (ASR). For example, in hypernasality assessment, from early spectral analysis methods, based on formant amplitudes and bandwidths, to more sophisticated techniques utilizing Gaussian mixture models (GMM) or support vector machines (SVM) processing Mel-frequency cepstral coefficients, or perceptual linear prediction features, the state of the art has consistently been built atop features and models employed in ASR [5]. However, since the advent of deep learning (DL) and neural network techniques in ASR, transfer of ASR insights to nasality assessment has slowed [6], with classical techniques such as hidden Markov acoustic model (HMM) features outperforming DL approaches (**Figure 1**). This change in insight transfer is primarily due to data availability and "the curse of dimensionality".

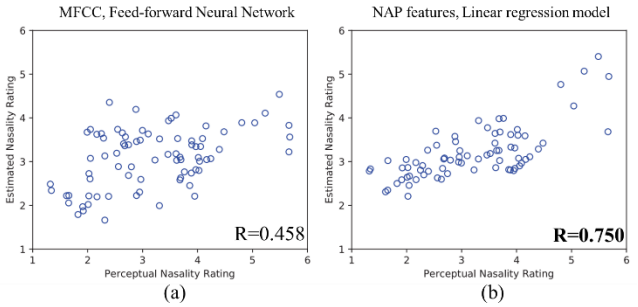

Figure 1 Performance of two models predicting hypernasality scores from speech, a state-of-the-art DL neural model (a) and a simple linear regression model using our GMM-HMM acoustic model-based NAP features (b).

Collecting speech data for ASR is resource intensive, and dysarthric speakers represent a small fraction of the population. As a result, there is only a limited amount of disordered speech data available for model training. While classical ASR methods, with their compartmental nature, hand-engineered features, and relatively small amounts of learnable free parameters, aren't seriously limited by small datasets, deep learning-based techniques are, as they are optimized end-to-end as a single integrated pipeline of thousands to millions of learnable parameters. My plan, put simply, is to pursue a solution to this problem. In other DL application areas, **transfer learning** is a broadly applied technique to overcome this issue. Even in some speech processing tasks, transfer learning

is recognized as a way to bring model "knowledge" learned from larger outside datasets into a problem for which data is sparse [7]. However, there is little analogous work in speech demonstrating generalized, transferrable representations learned from large speech datasets, like there is in text understanding [8] and image processing.

I propose building domain-informed DL models that overcome the problem of data sparsity in dysarthric speech by leveraging specialist understanding of the physical mechanisms that underlie the speech patterns, and by working to further adapt existing transfer learning techniques such as representation learning to dysarthric speech. Furthermore, I plan to apply these techniques to neural ASR more broadly, to ensure that burgeoning voice interface technologies will be accessible to those with neuromuscular disease.

I will draw on much of my prior research experience to accomplish this task. My previous project focusing on hypernasality assessment [5], applied this same approach to specialist knowledge, using the tools of classical ASR rather than DL. Furthermore, in my manuscript currently under review for an Automatic Assessment of Health Disorders from Voice, Speech and Language Processing IEEE JSTSP issue, I directly studied the limitations of current hypernasality estimating neural networks (**Figure 1**). In my research work performed as an applied science intern with Amazon I worked directly on neural speech modeling, representation learning, and end-to-end neural ASR and natural language interpretation from speech, enabling me to replace the old tools of classical ASR I worked with before with those of DL.

Success will be measured in terms of speech biomarker estimation performance, and accuracy of ASR systems for dysarthric speakers. Success might allow me to go back to **Figure 1** and add a subplot (c), depicting a neural model outperforming the current state of the art in classical ASR-based features for hypernasality prediction. Success might enable the employment of newly learned dysarthric speech representations to a DL-based neural ASR model and improve recognition performance speakers with Parkinson's disease or ALS.

Intellectual Merit: Application-oriented research in deep learning has often led to generalized advancements that apply in other domains, it is likely that this is the case with this problem; this work could potentially lead to insights that improve neural network performance on a variety of other problems. Furthermore, the targeted topic is a natural next step in neural speech processing.

Broader Impacts: The primary motivation behind this work is a broader impact: <u>improved well-being of people with neuromuscular disease</u>, through improved detection and tracking of the early stages of disease using innovative, improved neural network-based metrics. Better neurological classification metrics from speech will *improve well-being for individuals with early onset disease*, allowing us to noninvasively and persistently track the progression of their disease. The secondary goal of improved neural ASR performance for the same population will drive more <u>full participation of people with speech symptom-manifesting neurological disease in the benefits of neural ASR-based technologies</u>; improving the performance of speech interfaces for such individuals will <u>ensure that persons with disabilities may fully participate in the benefits</u> they bring.

References: [1] F.L. Darley, et al. "Differential diagnostic patterns of dysarthria." Journal of speech and hearing research (1969) [2] B.T. Harel, et al. "Acoustic characteristics of Parkinsonian speech: a potential biomarker of early disease progression and treatment." Journal of Neurolinguistics (2004) [3] M. Novotny, et al. "Hypernasality associated with basal ganglia dysfunction: evidence from Parkinson's disease and Huntington's disease," (2016)
[4] J. Duffy, "Motor Speech Disorders: Substrates, Differential Diagnosis, and Management." Mosby, (1995)
[5] M. Saxon, J. Liss, V. Berisha, "Objective Measures of Plosive Nasalization in Hypernasal Speech", ICASSP, (2019) [6] M. Tu, et al. "Interpretable Objective Assessment of Dysarthric Speech Based on Deep Neural Networks." *INTERSPEECH*. 2017. [7] L. Lugosch, et al. "Speech Model Pre-training for End-to-End Spoken Language Understanding." *INTERSPEECH*, 2019. [8] J. Howard, S. Ruder. "Universal language model fine-tuning for text classification." (2018).