## [Personalizing ASR for Dysarthric and Accented Speech with Limited Data | Technical Approach for Google's Euphonia](https://arxiv.org/abs/1907.13511) (2019)

* Focused on two types of non-standard speech
  + Speech from people with amyotrophic lateral sclerosis (ALS) (Slurred speech to incomprehensible)
  + Heavy accented speech
* Challenges faced
  + "Difficult to find enough speakers to train state-of-art model"
  + "Individuals within a group like ALS or a particular accent can have different ways of speaking."
* How did they overcome these challenges?
  + Fine tuning a model which is trained on 1000s of hours on normal speech
  + "Gets around sub-group heterogeneity by training personalized speech"
* What model was used/currently in use?
  + Google's ASR model with standard speech corpus used RNN-Transducers. The model for euphonia was fine-tuned on 1st,2nd and 3rd layers of RNN-T.

For LAS architectures - Fine Tuned the entire network.

* Results
  + “TTS can be successfully used for data augmentation in dysarthric ASR”
  + “this method cannot be applied to unseen speakers because the synthetic speech on its own is not of sufficient quality”
* Demonstration
  + Good performance on average dysarthric ALS speakers, large improvements in very dysarthric speakers.
  + Better performance on ALS and accented speech than just training on RNN-T encoder

## [Parrotron: An End-to-End Speech-to-Speech Conversion Model](https://arxiv.org/abs/1904.04169)

* This model is used to normalize speech from any speaker regardless of their accent, background noise into the voice of a single target speaker with a fixed accent.
* It has been shown in this paper that this model can be adapted to normalize highly atypical speech from a deaf speaker.
* Model Architecture
  + End to end sequence to sequence model architecture that takes an input source speech and generates/synthesizes target speech as output
  + The model maps an input spectrogram directly to another spectrogram without any intermediate representation.
* This model is also used in speech separation where it can separate the loudest voice in a mixture of overlapping speech.