PennState
College of Medicine

Dysarthric speech recognition of anesthesia-affected patient speech

Patrick Kudo, MAI, Elie Sarraf, M.D., C.M.

Artificial Intelligence, PennState Great Valley, Anesthesiology, PennState College of Medicine

Introduction

Dysarthric speech recognition (DSR) models have been sought by researchers to identify medical conditions which affect vocal articulation. An effective DSR model would assist clinical staff with non-invasive monitoring of patients emerging from anesthesia. Deep learning models have been used to classify dysarthric speech audio waveform features, as convolutional neural network (CNN) and residual neural network (ResNet) models have been found to be particularly effective with such tasks [1][2]. In our work, we demonstrate a functioning DSR model and a practical Python toolbox which may aide fellow researchers in developing future DSR models [3].

Methods

After obtaining Institutional Review Board approval, 48 patients were voluntarily recorded in the ambulatory endoscopy unit at the Penn State Health's University Physician Center. The subjects were recorded speaking a key word ("Pennsylvania") repeatedly prior to anesthesia, when emerging from anesthesia in the recovery area, and prior to discharge from the unit. After segmenting each patient's spoken utterance from their recorded audio files, all utterances were converted to spectrograms using the short-time Fourier transform (STFT) and then converted once more to input tensors for a total of 727 usable samples (see Table 1). CNN and ResNet models were then used to classify tensors with severity-level and binary (healthy/dysarthric) approaches. The binary classification approach assumes that speech articulation prior to anesthesia and at discharge is generally equivalent, while the severity-level approach assumes that the speech articulation in those groups are not equivalent.

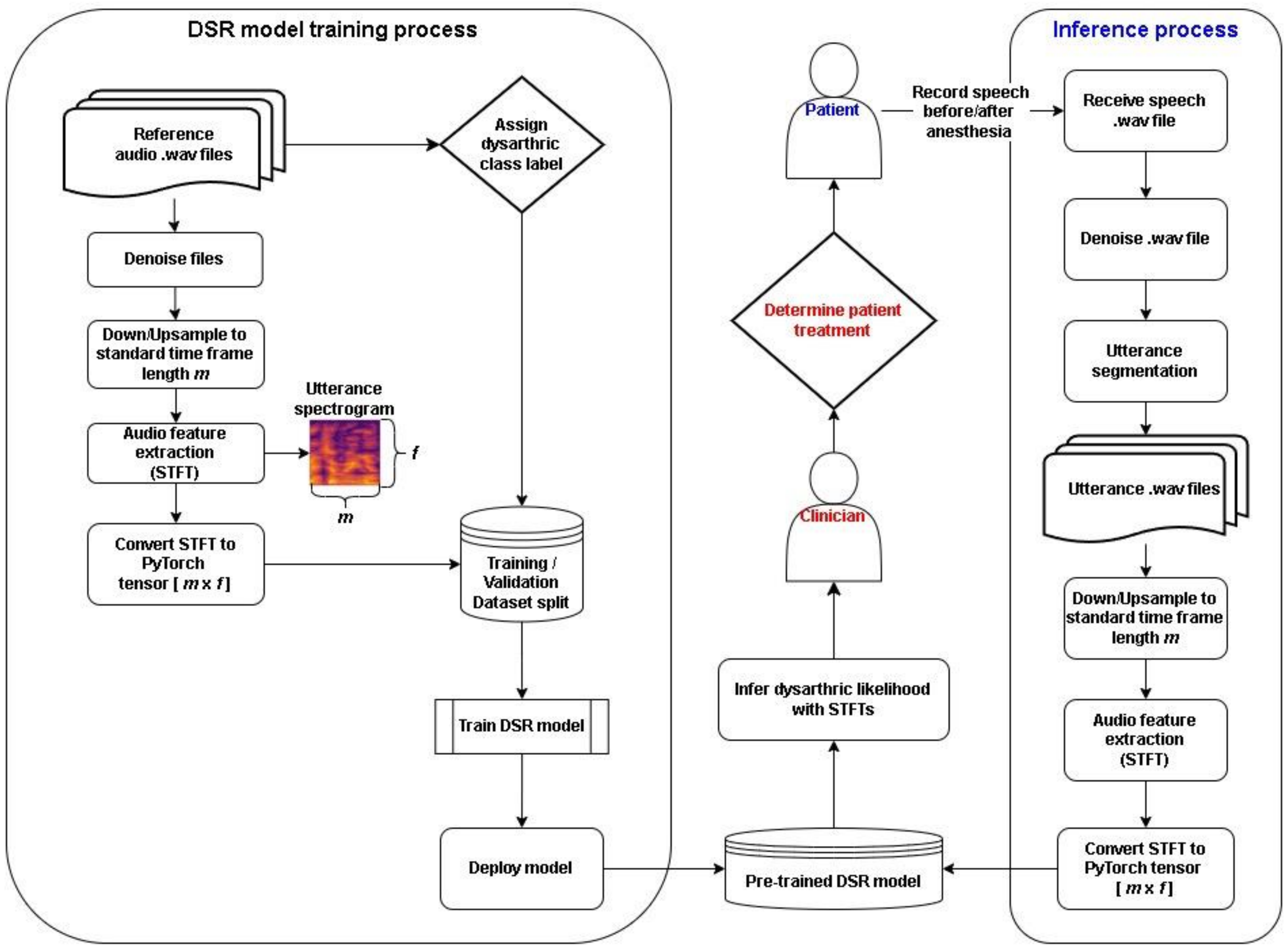
Results

Both model architectures were found to overfit the training dataset and provided poor results with the severity-level classification approach, while the ResNet-18 model achieved modest yet promising results with 85% overall accuracy and macro F1-score of 0.75 on the test set with the binary approach. Summary results are presented in Table 2.

Conclusions

Further work in this area may focus on identifying articulation changes within speaker samples as pairwise distance-based CNN models have shown promise in other dysarthric corpuses [4]. Lastly, we acknowledge that well-known benchmark dysarthric speech datasets such as TORGO and UASpeech provide additional opportunities for greater evaluation of these model architectures as those corpuses contain a variety of speech samples [5][6].

Automatic Dysarthric Speech Recognition System



DSR model training process

Reference audio .wav files

Assign dysarthric class label

Denoise files

Down/Upsample to standard time frame length m

Audio feature extraction (STFT)

Utterance spectrogram

Convert STFT to PyTorch tensor [$m \times f$]

Training / Validation Dataset split

Train DSR model

Deploy model

Pre-trained DSR model

Patient

Record speech before/after anesthesia

Determine patient treatment

Clinician

Infer dysarthric likelihood with STFTs

Inference process

Receive speech .wav file

Denoise .wav file

Utterance segmentation

Utterance .wav files

Down/Upsample to standard time frame length m

Audio feature extraction (STFT)

Convert STFT to PyTorch tensor [$m \times f$]

Figure 1. Flowchart diagram demonstrating intended application of DSR system.

Table 1. Utterance dataset statistics

Patient Status	Average Length (sec)	Min Length (sec)	Max Length (sec)	n
pre	0.84 ± 0.13	0.51	1.32	459
post	0.93 ± 0.19	0.54	1.59	133
dc	0.91 ± 0.13	0.63	1.29	135

Table 2. Held out (20%) test set results

Architecture	Classes	Accuracy %	Macro F1
CNN	2	82	0.45
ResNet		**85**	**0.75**
CNN	3	63	0.34
ResNet		68	0.55

References

[1] Vásquez-Correa, J.C., et al. "Convolutional Neural Network to Model Articulation Impairments in Patients with Parkinson's Disease." Interspeech 2017, 20 Aug. 2017, https://doi.org/10.21437/interspeech.2017-1078. Accessed 20 Dec. 2021.

[2] Gupta, S. et al., "Residual Neural Network precisely quantifies dysarthria severity-level based on short-duration speech segments," Neural Networks, vol. 139, pp. 105–117, Jul. 2021, doi: 10.1016/j.neunet.2021.02.008.

[3] Kudo, P. "GitHub - PatrickKudo/psudsr: PSU Dysarthric Speech Recognition (psudsr)," GitHub, 2024. https://github.com/PatrickKudo/psudsr.

[4] Janbakhshi, P., Kodrasi, I., & Bourlard, H. (2021). Automatic dysarthric speech detection exploiting pairwise distance-based convolutional neural networks. Ithaca: Cornell University Library, arXiv.org. https://doi.org/10.48550/arxiv.2011.07545

[5] "UASpeech – Statistical Speech Technology Group." https://speechtechnology.web.illinois.edu/uaspeech/

[6] "The TORGO database." https://www.cs.toronto.edu/~complingweb/data/TORGO/torgo.html