Automatic Dysarthria Severity Assessment using Whisper-extracted Features

Author: Christopher Charlesworth ccharlesworth@tudelft.nl Supervisors: Dr. Zhengjun Yue, Yuanyuan Zhang

Background

Dysarthria

- A speech disorder caused by weakness or incoordination of the muscles necessary for speech
- Commonly caused by neurological disorders like cerebral palsy, strokes or Parkinson's disease
- The severity of dysarthria in a patient greatly impacts their treatment
- Dysarthria severity assessment currently needs to be done by a licensed speech therapist and is a time-consuming process [1]

Whisper

- Multi-lingual and multi-task Automatic Speech Recognition (ASR) model
- Encoder-Decoder architecture [2]
- Dysarthria severity classifiers trained on Whisper embeddings outperform those trained on traditional sound representations [3]

Research Questions

- How do different types of classifiers perform in distinguishing between dysarthria severity levels using Whisper's encodings?
- How do training classifiers on different dysarthria datasets impact their performance?
- How does the inclusion of padded silence in the Whisper embeddings affect the performance of the classifiers?
- How does fine-tuning Whisper to perform dysarthric ASR affect the performance of classifiers trained on its encodings?

Methodology

Classifiers Architectures

- **Traditional RNN**
  - Learns temporal information
  - Has exploding and vanishing gradient problems
- **Long Short-Term Memory (LSTM)**
  - Uses a cell to store important information
  - Trains 3 gates to manage cell memory
- **Bidirectional Long Short-Term Memory (BiLSTM)**
  - Extension of LSTM with data processed in both directions
- **Gated Recurrent Unit (GRU)**
  - Simplified gate structure compared to LSTM
- **Convolutional Neural Network (CNN)**
  - Commonly used in image classification
  - Convolutional layers are used to learn hierarchical features
  - Translationally invariant

[4]

Datasets

- TORGO
  - 3667 subset of the whole dataset
  - Utterances longer than 2.5 seconds
- MSDM
  - Shorter Utterances
  - 61,396 utterances
  - Random resampling of minority classes

Fine-tuned Whisper model

- Developed by Mirella Günther
- Fine-tuned Whisper for dysarthric ASR
- Trained for 2 epochs on TORGO
- Weights were updated using low-rank adaptation

Experiment

All models were trained on:

- TORGO embeddings cut to a timeframe of 375 (including silence)
- TORGO embeddings cut to a timeframe of 125 (no silence)
- TORGO fine-tuned embeddings cut to a timeframe of 125
- MSDM dataset

Model training and optimization techniques

- 90% of data for training, 10% for testing
- 10% of training data used for the validation set
- Patience-based early stopping was used
  - Stop training if validation loss hasn't improved for 13 epochs
- Patience-based learning rate control
  - If validation loss hasn't improved for 3 epochs, then half the learning rate

Results

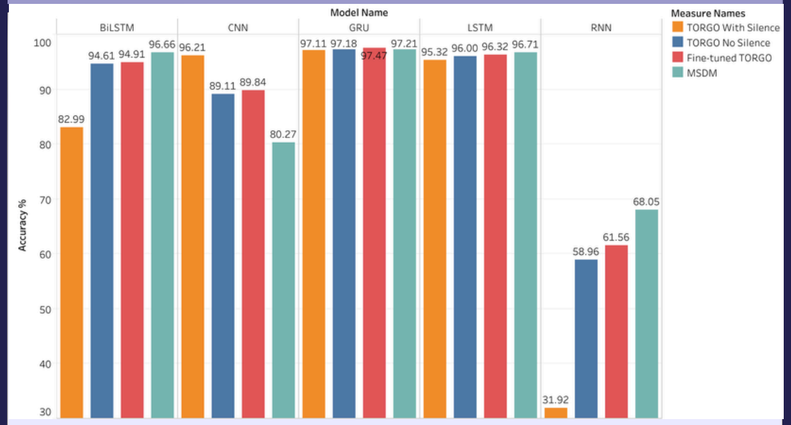

Figure 1: Accuracy of models on all datasets

Discussion

- The GRU had the greatest performance with an accuracy of 97.48% on fine-tuned TORGO embeddings
  - 1.29% improvement on models trained on traditional spectral features [5]
  - Similar performance to the best CNNs trained on Whisper features in previous work [3]
- Removing padded silence improved the performance of RNN variants
  - Fewer gradient issues and shorter temporal dependencies
- The performance of RNN variants was higher on MSDM compared to TORGO
  - MSDM has shorter utterances than TORGO
- All models improved when trained on fine-tuned Whisper embeddings
- BiLSTM performed worse than LSTM

Limitations

- Only a subset of TORGO was utilized due to computational limitations
- The CNN architecture was adapted for MSDM due to their differing utterance lengths
  - Difficult to compare CNN results from TORGO to MSDM
- Batch size, number of layers and other hyperparameters were set using trial and error

Speaker dependency of the models

The training and test data included utterances from all participants meaning models likely learnt specific speaker patterns of the participants. A possible solution would have been to leave one speaker out per severity class for the test and validation sets. This performance would be likely lower but would better represent the models' performance in a real-world setting.

Conclusion

With 97.47% accuracy achieved by the GRU, we have shown that RNN variants trained on Whisper features can outperform CNNs. Next, we found that RNN variants trained on MSDM achieved greater performance than their counterparts trained on TORGO, with the opposite being the case for CNNs. Additionally, processing all utterances to remove padded silence improved performance for RNN variants. Finally, all models trained on fine-tuned Whisper embeddings were found to achieve greater performance than those trained on normal Whisper features.

Future Work

- Train models on normal and fine-tuned Wav2Vec2 embeddings
- Implement ensemble methods to combine multiple different models to improve performance
- Evaluate speaker dependency of models

References

[1] P. Enderby, Chapter 22 - Disorders of communication: dysarthria, ser. Neurological Rehabilitation. Elsevier, Jan. 2013, vol. 110, p. 273–281. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780444529015000228

[2] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision."

[3] S. Rathod, M. Charola, and H. A. Patil, "Noise robust whisper features for dysarthric severity-level classification," in Pattern Recognition and Machine Intelligence, P. Maji, T. Huang, N. R. Pal, S. Chaudhury, and R. K. De, Eds. Cham: Springer Nature Switzerland, 2023, p. 708–715.

[4] R. M. Schmidt, "Recurrent neural networks (rnns): A gentle introduction and overview," no. arXiv:1912.05911, Nov. 2019, arXiv:1912.05911 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1912.05911

[5] A. A. Joshy and R. Rajan, "Automated dysarthria severity classification: A study on acoustic features and deep learning techniques," IEEE transactions on neural systems and rehabilitation engineering: a publication of the IEEE Engineering in Medicine and Biology Society, vol. 30, p. 1147–1157, 2022