Evaluating Alternative Metrics for Dysarthric Speech Recognition

H.C. (Filip) Nguyen Duc
Email: H.C.NguyenDuc@student.tudelft.nl
Supervisors: Zhenjun Yue, Yuanyuan Zhang
EEMCS, Delft University of Technology, The Netherlands

1 Background

Dysarthria is a motor speech disorder resulting in slurred or slow speech that can be difficult to understand. **Automatic speech recogniction** (ASR) systems have trouble transcribing dysarthric speech.

Word error rate (WER) and character error rate (CER) can provide an indication of performance on a word/character level however they can fail to accurately asses ASR systems handling atypical speech because they do not consider the severity or specific nature of errors [1].

Reference: Please turn on the light

Moderate dysarthria: Peas turn on light

Severe dysarthria: Pees turn on the lie

Example of how varying levels of dyarthria can be transcribed by an ASR system

Research gap: there is limited understanding of how well metrics perform in assessing ASR accuracy for speakers with varying levels of dysarthria

2 Research Question

How do various alternative error analysis methods compare in their effectiveness at evaluating ASR system performance across different severities levels of atypical speech?

References

[1] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens, "Automatic speech recognition and speech variability: A review," Speech Communication, vol. 49, no. 10, pp. 763–786, 2007, intrinsic Speech Variations. [Online].

3 Methodology

State-of-the-art ASR systems

- **Whisper** employs a transformer architecture that processes audio input to perform multilingual speech recognition, translation, and generation, delivering high accuracy across a diverse set of languages
- **Wav2vec 2.0** uses a self-supervised learning approach within a transformer-based architecture to convert raw audio into contextualized representations, enhancing the model's performance on speech recognition tasks without relying heavily on labeled data

Dysarthric Dataset

- **TORGO** database provides dysarthic data that is split by utterances and by subject severity

Evaluation Metrics

1. **Word Error Rate (WER)** - edit distance based metric on word level
2. **Character Error Rate (CER)** - edit distance based metric on character level
3. **Jaro–Winkler Distance** - scores strings based on their matching characters and the order in which they appear
4. **BERTscore** - uses contextual embeddings from BERT model to compute similarity focusing on semantic accuracy

4 Experimental Setup

- Whisper large-v2 and wav2vec 2.0 large
- Running TORGO on Whisper and wav2vec 2.0 for single word and sentence utterances
- Cleaned and processed out from ASR system to match TORGO prompts
- Pearson correlation between each evalaution metric and subejct severity

5 Results

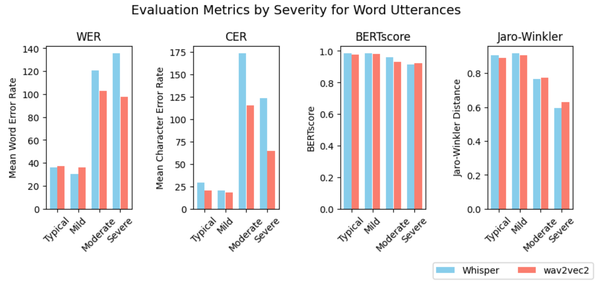

Figure 1: Evaluation metric errors grouped by severity and model for word utterances

Metric	Correlation Coeff.	P-Value
WER	0.927	<0.001
CER	0.830	<0.001
BERTscore	-0.928	<0.001
Jaro-Winkler	**-0.942**	<0.001

Table 1: Pearson correlation between speech severity levels and evaluation metrics of single-word utterances. Bold value is the highest absolute correlation coefficient.

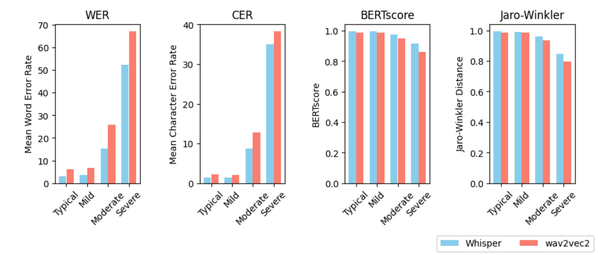

Figure 2: Evaluation metrics errors grouped by severity and model for sentence utterances

Metric	Correlation Coeff.	P-Value
WER	0.865	<0.001
CER	0.851	<0.001
BERTscore	**-0.866**	<0.001
Jaro-Winkler	-0.864	<0.001

Table 2: Pearson correlation between speech severity levels and evaluation metrics of sentence utterances. Bold value is the highest absolute correlation coefficient

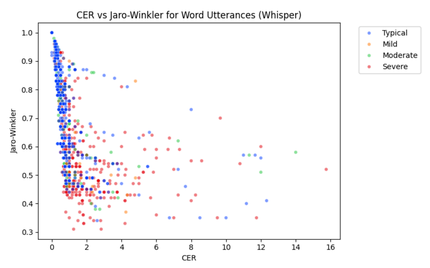

Figure 3: CER vs. Jaro-Winkler of all word utterances grouped by severity

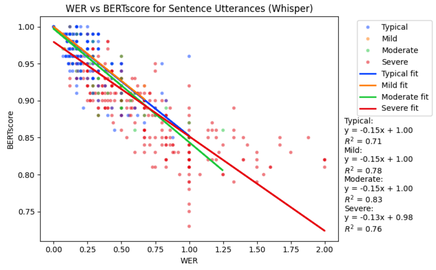

Figure 4: WER vs. BERTscore of all sentence utterances grouped by severity

6 Conclusions

- Single-word utterances showed strong correlations with phonetic metrics like Jaro-Winkler, sensitive to articulation errors,
- Sentence-level utterances correlated better with semantic metrics like BERTscore, effectively capturing semantic coherence
- The comparison between wav2vec 2.0 and Whisper revealed that wav2vec 2.0 excels in single-word utterances, whereas Whisper better handles the complexities of sentence-level utterances