Evaluating Alternative Metrics for Dysarthric Speech Recognition

Assessing the Effectiveness of Different Evaluation Metrics in Dysarthric Speech Recognition Systems Across Various Severities

Hung Cuong (Filip) Nguyen Duc
Supervisor(s): Zhengjun Yue, Yuanyuan Zhang
EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Hung Cuong (Filip) Nguyen Duc
Final project course: CSE3000 Research Project
Thesis committee: Zhengjun Yue, Yuanyuan Zhang, Thomas Durieux

Abstract

Dysarthria is a motor speech disorder resulting in slurred or slow speech that can be difficult to understand. This research paper evaluates the effectiveness of various metrics for automatic speech recognition (ASR), such as character error rate (CER), Jaro-Winkler distance, and BERTscore, in assessing performance specifically for dysarthric speech, which is often inadequately measured by the commonly used word error rate (WER). Using the TORGO database, which includes a range of dysarthria severities, we analyze the performance of chosen evaluation metrics with the Whisper and wav2vec 2.0 ASR systems to understand how they reflect the true speech recognition challenges presented by such atypical speech patterns. Our findings reveal that Whisper generally outperforms wav2vec 2.0, particularly in sentence utterances, by effectively managing complex speech patterns and maintaining semantic integrity. The analysis highlights that single-word utterances strongly correlate with dysarthria severity, while sentence utterances show a lesser correlation due to the mitigating effect of additional linguistic context.

Index Terms: automatic speech recognition, evaluation metric, dysarthria

1. Introduction

Dysarthria is a neuromotor speech disorder that can result from conditions like Parkinson's, Alzheimer's, multiple sclerosis, or from traumatic events such as brain injuries or strokes [1]. There are various forms of dysarthria, each distinguished by unique speech traits. While classification typically considers the location of the lesion and the extent of neurological damage, literature often categorizes dysarthria more broadly by severity—measured in terms of speech intelligibility and articulation—using terms such as mild, moderate, and severe [2]. Individuals with dysarthria that is classified as more than mild often face challenges in being understood by both other people and technology. These difficulties are particularly evident in interactions involving automatic speech recognition (ASR) systems.

ASR systems have become pivotal in modern technology, used in everything from virtual assistants to transcription services. Evaluation metrics are used in order to measure how accurate a transcription hypothesis is compared to the reference. The most widely used and accepted metric is word error rate (WER), which is derived from the Levenshtein distance [3]. WER assesses the accuracy of a speech recognition system by counting the errors—substitutions, deletions, and insertions—normalised by the number of words in the reference text [4]. WER can serve well to provide an indication of performance on a word level however it can fail to accurately asses ASR systems handling atypical speech because they do not consider the severity or specific nature of errors [5, 6, 7]. For instance, a person with moderate dysarthria might attempt to say "Please turn on the light", which could be inaccurately transcribed as "Peas turn on light". More severe dysarthric speech might result in a transcription like "Pees turn on the lie", which introduces more phonetic distortions and significantly alters the comprehensibility of the utterance. Both examples would have the same WER score, yet they represent different levels of intelligibility and speech severity. Wang et al. [8] showed that a lower WER doesn't always mean improved accuracy in understanding spoken language. Their findings indicate that transcripts with a reduced WER might align with higher understanding accuracy, highlighting the significance of focusing on understanding as a goal rather than merely decreasing WER. This gap in the effectiveness of WER highlights the importance of adapting ASR technologies to better handle the variability and challenges of atypical speech patterns, such as those presented by dysarthric speech.

Various evaluation metrics have been introduced in the past to address the limitations of WER. Most similar is character error rate (CER), which measures the minimum number of character-level edits required to change the ASR output into the reference text. CER can be particularly useful for languages where character-level errors provide a finer granularity of error analysis than word-level errors. CER is also most commonly used in cases where ASR systems are being evaluated on single-word prompts.

Another evaluation metric that is calcualted using the edit distance of the text is the Jaro-Winkler distance [9]. Originally designed for comparing short strings such as names, the Jaro-Winkler distance is a measure that gives more favor to strings with a common prefix, making it useful for assessing ASR accuracy in contexts where prefixes are predictive of overall speech patterns. Similar to CER, Jaro-Winkler distance is most useful as a metric for evaluating single-word prompts.

All edit distance and n-gram matching based evaluation metrics can suffer from the same problem; they are limited because they only focus on the word/character level accuracy of the hypothesis text. They lack the ability to consider how semantically similar two phrases can be. A metric that does consider semantic context is BERTscore [10]. This metric uses the powerful capabilities of pretrained bidirectional encoder representations from transformers (BERT) contextual embeddings. Unlike traditional metrics that rely on n-gram overlap, which can fail to capture the richness of semantic equivalence, BERTscore computes the similarity between sentences by summing the cosine similarities between their token embeddings. This method allows for a more nuanced understanding of textual similarity by capturing paraphrases, lexical diversity, and changes in syntactic structures, making it robust against traditional evaluation pitfalls and better aligned with human judgement.

The significant variation in dysarthria severity necessitates a nuanced approach to evaluating ASR systems. Current metrics like WER and CER, while effective for typical speech, do not adequately account for the complexities of dysarthric speech patterns, particularly in how they impact intelligibility and articulation across different severities. This oversight presents a critical research gap: there is limited understanding of how well these metrics perform in assessing ASR accuracy for speakers with varying levels of dysarthria. Furthermore, the differentiation between single-word and sentence prompts in testing is crucial. Single-word prompts can help isolate pronunciation and articulation issues, while sentence prompts can better assess the ability of ASR systems to handle linguistic context and syntactic structure, which are often compromised in dysarthric speech.

Addressing this gap, the aim of this paper is to determine **how do various alternative error analysis methods compare in their effectiveness at evaluating ASR system performance across different severities levels of atypical speech?** To that end, we identify which methods provide the most accurate reflections of user experience and comprehension, aligning ASR technology more closely with the needs of individuals with speech disorders. This approach will not only enhance the utility of ASR systems in real-world applications but also contribute to the development of more inclusive technology solutions.

2. Related Works

A recent study by Rugayan et al. [11] introduces and evaluates the aligned semantic distance (ASD) as a metric for assessing the performance of ASR systems. The introduction of ASD marks a significant advancement over traditional metrics like the WER, particularly in terms of measuring the semantic integrity of ASR outputs.

ASD uses dynamic programming to optimally align sequences of token embeddings, thereby calculating the semantic closeness based on the accumulated distance of this alignment. This approach allows ASD to effectively handle variations in sentence length and maintain its robustness, providing a more detailed and semantically meaningful evaluation of ASR accuracy. The implementation of ASD addresses several limitations of WER, including its inability to account for the semantic severity of transcription errors. For instance, while WER would treat all errors equally, ASD distinguishes between errors that have a significant impact on the meaning of the sentence and those that do not.

Rugayan et al.'s research into the ASD offers significant insights for advancing the evaluation of ASR systems, particularly for those dealing with dysarthric speech. The key takeaway from their approach is the emphasis on semantic integrity and contextual understanding, which are critical when evaluating speech that may be highly variable or atypical due to the underlying neurological conditions associated with dysarthria. Their methodology demonstrates how semantic-based metrics can more accurately reflect the real-world effectiveness of ASR systems by focusing on the semantic closeness of the transcribed text to the intended speech.

For our research, examining the efficacy of semantic evaluation metrics like BERTscore across different severities of dysarthria can be particularly enlightening. BERTscore, similar to ASD, leverages the power of contextual embeddings to assess semantic similarity, potentially offering a nuanced understanding of how well an ASR system captures the intended meaning behind speech that might be unclear or distorted due to dysarthria.

3. Methodology

3.1. ASR System

In this study, one of the ASR systems we employ is the Whisper model, a state-of-the-art speech recognition system [12]. Whisper's architecture is designed to be particularly resilient in handling variations in speech, such as accents, dialects, and background noise, making it an exemplary choice for researching ASR performance on atypical dysarthric speech. One of the model's significant advantages is its training on a vast, multilingual dataset, which enhances its capability to accurately transcribe speech even when faced with the complexities of dysarthric speech patterns. Whisper offers five model sizes: tiny, base, small, medium and large. For this study large-v2, the second iteration of the large model, was chosen. This model size has the largest number of parameters at close to 1.5 billion. The large model size is considered state-of-the-art and performs with the greatest accuracy [12]. These attributes make Whisper an ideal ASR system for our research, as it aligns with our goal of exploring speech recognition technology's inclusivity and reliability across a range of dysarthria severities.

Alongside the Whisper model, this study also incorporates the use of wav2vec 2.0, another state-of-the-art ASR technology [13]. The model utilizes a self-supervised learning approach, where the system is trained on raw audio data without the need for manual transcription. A distinctive feature of wav2vec 2.0 is its architecture, which includes a convolutional feature encoder that processes raw audio to produce latent representations, and a transformer that predicts the contextualized representations. These are then fine-tuned with a small amount of labeled data to achieve high levels of accuracy. For our research, we employ the large variant of wav2vec 2.0, which is particularly adept at handling nuanced and complex speech patterns, such as those associated with dysarthric speech. Unlike the Whisper large-v2 which is a multilingual model, the chosen wav2vec 2.0 model is trained only on English language data. This choice is motivated by the model's demonstrated proficiency in discerning subtle differences in speech articulations, making it a valuable asset in assessing ASR performance across different severities of dysarthria.

3.2. Evaluation Metrics

In assessing the performance of ASR systems, especially in handling dysarthric speech, it is crucial to employ metrics that accurately reflect both the literal and contextual correctness of the transcribed text. This section provides a detailed examination of four key evaluation metrics: word error rate (WER), character error rate (CER), BERTscore, and Jaro-Winkler distance.

3.2.1. Word Error Rate

Definition and Calculation:
WER is the conventional metric used to evaluate ASR systems. It quantifies the performance by calculating the ratio of the total number of errors (substitutions, deletions, and insertions) to the number of words in the reference text [4]. The formula for WER is given by:

$$WER = \frac{S + D + I}{N} \tag{1}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the number of words in the reference.

Applications and Limitations:
WER provides a straightforward quantitative assessment but treats all errors equally, lacking sensitivity to the contextual severity of errors. This makes it less ideal for nuanced linguistic analyses, such as in dysarthric speech where different types of errors may impact intelligibility differently.

3.2.2. Character Error Rate

Definition and Calculation:
CER extends the concept of WER to the character level, which can be particularly useful for languages where character-level precision is more indicative of speech recognition accuracy [4]. The CER is defined as:

$$CER = \frac{C_s + C_d + C_i}{C_n} \tag{2}$$

where C_s, C_d, and C_i represent the numbers of character substitutions, deletions, and insertions, respectively, and C_n is the total number of characters in the reference.

Applications and Limitations:
CER offers an analysis at the character level, which can be particularly beneficial for assessing dysarthric speech where slight character changes can significantly impact intelligibility. This

metric can capture the nuances in pronunciation and articulation that are often lost in broader word-level metrics. However, while CER provides finer granularity, it still does not account for semantic changes. This can be a critical oversight in moderate to severe dysarthria, where misrecognitions might lead to completely different meanings, despite minimal character alterations.

3.2.3. BERTscore

Definition and Calculation:
BERTscore leverages the contextual embeddings generated by BERT, a pre-trained deep learning model known for its powerful language understanding capabilities [10]. This metric evaluates the quality of text by computing the cosine similarity between the token embeddings of the hypothesis text generated by the ASR system and the reference text. This method reflects the context of each token in the hypothesis and reference texts, taking into account their semantic and syntactic environment.

The calculation of BERTscore involves the following steps:

1. Tokenization of both the hypothesis and reference texts using BERT's tokenizer.
2. Generation of contextual embeddings for each token in both texts by passing them through a pre-trained BERT model.
3. Calculation of the cosine similarity for each token in the hypothesis with every token in the reference.
4. Identification of the maximum similarity score for each token in the hypothesis, which represents the best semantic match in the reference.
5. Averaging these maximum similarity scores across all tokens in the hypothesis to compute the overall BERTscore.

The formula for BERTscore is formally expressed as:

$$\text{BERTscore} = \frac{1}{|H|} \sum_{h \in H} \max_{r \in R} \cos(h, r) \tag{3}$$

where H and R are the sets of token embeddings for the hypothesis and reference, respectively.

Applications and Limitations:
BERTscore can be advantageous in evaluating dysarthric speech because it assesses semantic similarity between the transcribed and reference texts. This approach is crucial for dysarthria, where speech may be phonetically distorted but still contextually correct. BERTscore's focus on semantic content aligns well with the needs of speakers with varying dysarthria severities, as it can more accurately reflect the intelligibility and clarity of the intended communication. However, the computational intensity of BERTscore and the potential biases inherent in the pre-trained models it relies on can be limitations. Moreover, in cases of severe dysarthria, where speech may be significantly distorted, the semantic analysis might not fully capture the extent of the speech comprehension challenges faced by these individuals.

3.2.4. Jaro-Winkler Distance

Definition and Calculation:
The Jaro-Winkler distance is a string metric for measuring the similarity between two sequences, with an adjustment for common prefixes [9]. The formula is:

$$\text{Jaro-Winkler} = \text{Jaro} + l \times p \times (1 - \text{Jaro}) \tag{4}$$

where l is the length of the common prefix (up to a maximum of 4) and p is a constant scaling factor.

Table 1: *TORGO speakers with their severity level and data about utterance distribution*

Subject	Severity	Word Utterance Count	Sentence Utterance Count
M01	Severe	280 (186)	89 (67)
M02	Severe	293 (107)	92 (40)
M04	Severe	296 (164)	86 (61)
M05	Severe	358 (230)	117 (103)
F01	Severe	94 (78)	20 (16)
M03	Mild	306 (163)	95 (80)
F04	Mild	323 (147)	100 (66)
F03	Moderate	402 (111)	139 (50)
MC01	Typical	786 (220)	284 (105)
MC02	Typical	332 (189)	112 (79)
MC03	Typical	592 (163)	201 (80)
MC04	Typical	725 (188)	262 (100)
FC01	Typical	121 (104)	26 (24)
FC02	Typical	897 (505)	316 (253)
FC03	Typical	695 (108)	261 (58)

Applications and Limitations:
The Jaro-Winkler distance is particularly valuable for its emphasis on phonetic similarities and common prefixes, making it especially useful for single-word analysis in dysarthric speech assessments. This metric enhances the ability to recognize and accurately evaluate words that may start similarly but diverge phonetically due to speech impairments, a common occurrence in dysarthria where motor control deteriorates as the utterance progresses. While highly effective for single-word prompts, which isolate specific pronunciation challenges crucial for tailored therapy and ASR system training, Jaro-Winkler's design focus on short strings limits its applicability to full-sentence evaluations, necessary for understanding contextual ASR performance.

4. Experiments

4.1. Data Description

For this research, we selected the TORGO dysarthric dataset [14] which contains recordings from individuals affected by cerebral palsy and ALS. TORGO includes a balanced cohort of male (M01 - M05) and female (F01, F03, F04) subjects, alongside age- and gender-matched control subjects with typical speech (MC01 - MC04, FC01 - FC03). Importantly, TORGO subjects can be classified in severity levels of atypical speech. Each subject in the database has been evaluated by professional speech-language pathologists, providing detailed assessments of speech-motor functions and labeled data reflecting different severity levels of speech impairment [15]. In Table 1 all TORGO subjects and their severities are listed. This precise labeling is vital for our study, enabling an examination of ASR performance as it relates to the gradations of speech atypicality.

Additionally, the TORGO dataset has a variety of utterances. It includes single-word utterances and sentence utterances. Single-word utterances are crucial for testing phonetic accuracy, while sentence utterances test semantic accuracy, i.e., whether the words together convey the intended meaning. This dual approach helps getting a better understanding of the ASR system's sound recognition and the preservation of meaning,

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

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

which are often challenged in dysarthric speech. Subjects also repeat the same prompt multiple times to reduce variability. Summarised in Table 1, are each subjects' word and sentence utterance count. The number in parenthesis indicates how many of those utterances come from unique prompts. The distribution across different utterance types and severities found within the dataset allows for a comprehensive analysis of ASR system performance.

4.2. Experimental Setup

In our study, we examined the performance of two evaluation metrics, BERTscore and Jaro-Winkler, specifically focusing on their effectiveness across different severity levels of dysarthric speech. We used WER and CER as baselines to determine how well these newer metrics correlate with traditional measures when assessing ASR system accuracy. We hypothesize that BERTscore, a semantic-based metric, will exhibit greater consistency across varying severity levels compared to WER and CER, which do not account for semantic meaning.

The original TORGO dataset is publicly available [14] and includes a rich array of continuous dysarthric speech. In order to improve the quality of data, we used a cleaned TORGO dataset that excludes recordings that are shorter than 25 ms and have incorrectly annotated audio [16]. In addition non-language prompts—such as sounds not forming part of any spoken language—were excluded from the dataset. More details about how the Whisper and wav2vec 2.0 output were matched to the TORGO prompts can be found in the code [1].

The transcription experiment involved running the entire cleaned and preprocessed TORGO dataset through both ASR models, with each audio file individually processed and the output transcriptions collected for further error analysis. For this study we processed all of the TORGO data without using any split since it was not necessary. The experiment was run using the default parameters with the notable exception of the parameter: device='mps'. This is due to the fact that the experiments were executed on an Apple M2 Pro processor with 10 CPU cores and 16 GPU cores. Once the ASR system generated the transcriptions from the dataset we calculated the errors using the selected evaluation metrics.

5. Results and Discussion

5.1. Coefficients

The Pearson correlation coefficient between the severity level and the evaluation metric values are shown in Tables 2 and 3. The data was divided into utterances of a single-word and sentences, which Table 2 and 3 show respectively. The four severity levels (typical, mild, moderate, severe) were given a numerical value (1 - 4) for the correlation calculations. It is important to recognize that higher accuracy in ASR hypotheses leads to increased BERTscore and Jaro-Winkler values. Because of this inverse relationship, the correlation coefficients are negative.

Table 3: *Pearson correlation between speech severity levels and evaluation metrics of sentence utterances. Bold value is the highest absolute correlation coefficient.*

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

The coefficients outlined in Tables 2 and 3 provide a clear illustration of how different evaluation metrics relate to the severity levels of speech impairments across both single-word and sentence utterances. The differences in correlation between single-word and sentence utterances suggest that ASR systems' challenge levels vary with the utterance complexity. Sentence utterances, which are inherently more complex due to longer phrases and increased contextual variability, exhibit statistically significant weaker correlations in all metrics compared to single-word utterances. This distinction highlights the nuanced challenges faced by ASR systems in handling more complex speech patterns under varying levels of impairment.

As can been seen in the bold text of Table 2, the speech severity levels for single-word utterances is most correlated with the Jaro-Winkler distance. This suggests that the metric is particularly sensitive to the phonetic accuracy affected by dysarthric speech impairments. Dysarthric speech often involves distortions, slurring, and other articulation errors that affect phonetic accuracy. Jaro-Winkler, which emphasizes phonetic similarities and differences, particularly in the beginnings of words (through its adjustment for common prefixes), can capture these nuances effectively. Jaro-Winkler distance may therefore provide a more detailed reflection of how speech impairments impact the intelligibility and accuracy of spoken words.

Indicated by the analysis in Table 3, BERTscore exhibits the highest correlation with speech severity levels in sentence-level utterances. This highlights its sensitivity to the semantic integrity of speech. Sentence-level utterances involve syntactic structures and context that is significantly impacted by the articulation of dysarthric speech. BERTscore, leveraging deep learning models to assess semantic similarity, effectively captures the contextual and syntactic nuances, even when phonetic details are compromised. Thus, BERTscore can provide a nuanced reflection of the semantic coherence and comprehensibility of spoken sentences, making it a valuable metric for evaluating ASR systems' performance in maintaining semantic integrity in the presence of speech impairments.

5.2. Utterance Scatter Plots

The relationship between CER and Jaro-Winkler distance for word utterances is shown in Figure 1. The data points cluster distinctly by severity, with 'typical' and 'mild' categories showing higher Jaro-Winkler scores at lower CER values. This clustering indicates better performance in cases with less pronounced speech impairments. For severe dysarthria, the spread of CER and Jaro-Winkler scores is broader, reflecting the variability in how character errors impact word similarity. This

[1] https://github.com/notfilip/research-project

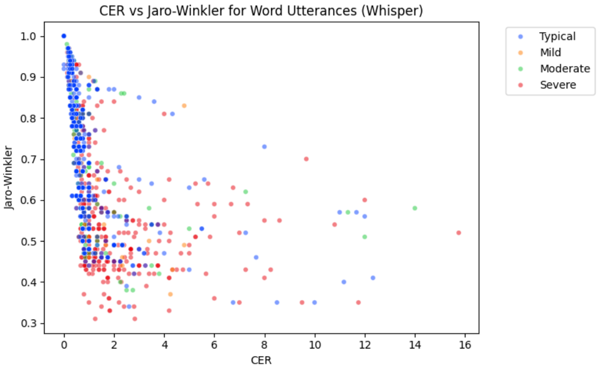

Figure 1: *CER vs. Jaro-Winkler of all word utterances grouped by severity*

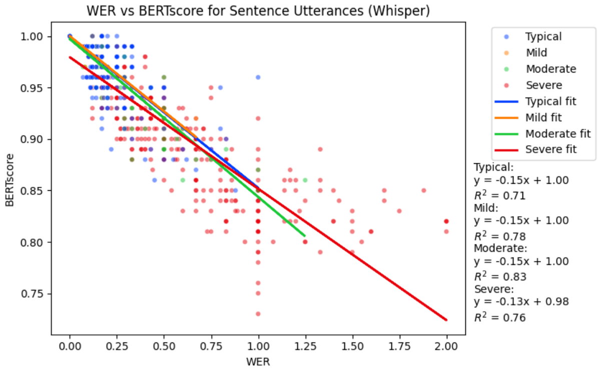

Figure 2: *WER vs. BERTscore of all sentence utterances grouped by severity*

spread suggests a higher unpredictability in speech articulation due to severe dysarthria, affecting character-based recognition accuracy. As CER increases, Jaro-Winkler scores generally decrease, although the relationship is less linear compared to WER vs BERTscore. This pattern suggests that the impact of character errors on perceived word similarity can vary significantly depending on the error's nature and position within words.

In Figure 2 we visualise the the WER and BERTscore values for all sentence utterances. The correlation between WER and BERTscore across four severity categories (typical, mild, moderate, severe) for sentence utterances reveals a clear negative correlation, indicating that higher transcription errors lead to poorer semantic matching. This trend presents the inverse relationship between transcription accuracy and semantic alignment in ASR outputs. The regression lines for each severity category demonstrate variations in slopes and intercepts, with the 'severe' category exhibiting a less steep slope than 'mild' and 'moderate'. This suggests that the impact of increasing WER on BERTscore is less pronounced in severe dysarthric speech, possibly due to the already reduced intelligibility in these cases, which may limit how additional transcription errors affect semantic alignment. The R² values, indicating the strength of correlation, are robust across all categories, with 'moderate' severity showing the highest (0.83). This strong correlation highlights the predictiveness of this model in assessing the relationship between transcription accuracy and semantic alignment across different levels of speech clarity.

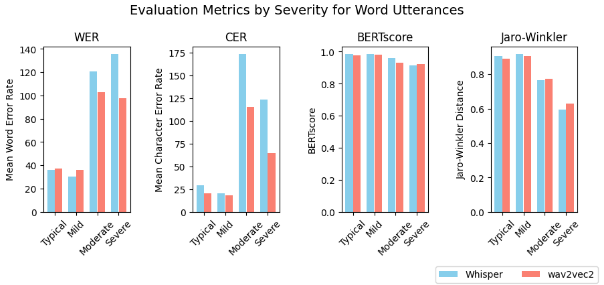

Figure 3: *Results of Evaluation Metrics grouped by subject severity and ASR model for single-word utterances*

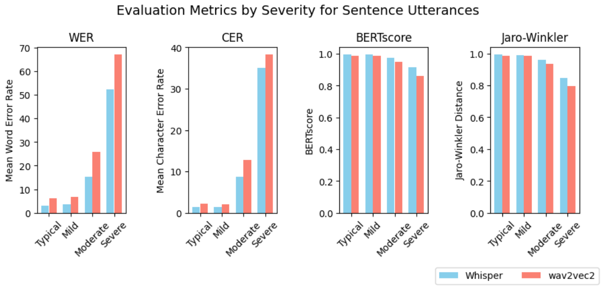

Figure 4: *Results of Evaluation Metrics grouped by subject severity and ASR model for sentence utterances*

5.3. Severity Bar Charts

In Figure 3, we present the metrics for word utterances. Both WER and CER again show that error rates escalate with the severity of speech impairment, with the increase in error rates being more pronounced, especially in the wav2vec 2.0 model which outperforms Whisper. A possible explanation for the lower errors rates of wav2vec 2.0 is that it was trained solely on English language data, while Whisper large-v2 used a multilingual training set.

Comparing BERTscore values across Tables 3 and 4, there is a noticeable drop for severe cases. This indicates that single-word accuracy is crucial for maintaining semantic integrity, more so than in sentences where contextual clues might offer compensatory support. Jaro-Winkler scores also decrease significantly with increased severity in word utterances, with wav2vec 2.0 experiencing a steeper decline compared to Whisper. This highlights greater challenges in maintaining word similarity at higher levels of dysarthria.

In the analysis of sentence utterances shown in Figure 4, both WER and CER metrics exhibit an increase in error rates as the severity of dysarthria intensifies. Notably, the Whisper system consistently shows better performance, indicated by lower error rates, compared to wav2vec 2.0 across all severity levels for both metrics. This suggests that Whisper might be more adept at handling the complexities associated with sentence-level dysarthric speech. The BERTscore across different severities and between the two systems remains relatively stable, with Whisper marginally outperforming wav2vec 2.0. This stability in BERTscore suggests that despite variations in word and character accuracy, the overall semantic content of the sentences is maintained relatively intact. However, the Jaro-Winkler distance demonstrates a noticeable decline as severity increases, particularly in the wav2vec 2.0 model, with a less pronounced decline observed in Whisper, indicating that it maintains closer word-level similarity across severities.

The comparative data from these charts suggests that Whisper generally outperforms wav2vec 2.0 in handling sentence utterances across all evaluated metrics and severity levels, while wav2vec 2.0 outperforms Whisper is most of the metrics for single-word utterances. The relatively stable performance of BERTscore in sentence utterances across both ASR systems indicates that the contextual information present in sentences helps ASR systems preserve semantic meaning even when phonetic or character-level errors are present. However, the significant variation in Jaro-Winkler scores, especially in word utterances, highlights the difficulty ASR systems face in accurately capturing the phonetic content of speech as severity increases.

These findings demonstrate the importance of incorporating robust models into ASR technologies that not only focus on transcription accuracy but also enhance the ability to interpret and reconstruct speech semantically, especially in contexts involving severe speech impairments. The performance differences between Whisper and wav2vec 2.0 further suggest that the choice of ASR model can significantly influence the effectiveness of speech recognition technology in accommodating the variable and often challenging nature of dysarthric speech. This analysis highlights the necessity for ongoing development and refinement of ASR systems to improve their utility and accessibility for individuals with speech disorders.

6. Conclusions and Future Work

This research explored the relationship between various speech impairment severity levels and the performance of ASR systems, examining metrics such as the BERTscore and Jaro-Winkler distance across different speech utterances. The findings indicate that ASR systems encounter distinct challenges when transcribing speech with varying degrees of dysarthria, particularly when it comes to maintaining phonetic accuracy and semantic integrity.

The analysis demonstrated that single-word utterances exhibit stronger correlations with phonetic-based metrics such as Jaro-Winkler, which proved sensitive to articulation errors typical of dysarthric speech. This sensitivity is crucial for recognizing and evaluating the phonetic discrepancies caused by speech impairments, suggesting that Jaro-Winkler is an effective metric for gauging phonetic accuracy in simpler speech forms.

In contrast, sentence-level utterances, which incorporate more complex syntactic structures and contextual variability, showed a higher correlation with semantic-based metrics like BERTscore. This metric effectively captured the semantic coherence of sentences, even when phonetic details were compromised, underscoring its utility in assessing semantic content in more complex speech outputs.

The comparative analysis of ASR models—Whisper and wav2vec 2.0—revealed that while wav2vec 2.0 generally performed better on single-word utterances, Whisper was more effective in handling the complexities of sentence-level utterances. This suggests that the choice of ASR model is critical in achieving optimal performance, particularly in applications involving severe speech impairments.

The findings from this study provide a strong foundation for further research into ASR systems tailored for speech impairments. Several avenues for future work can be outlined based on the insights gained:

1. Exploring Additional Datasets: The use of the TORGO database in this research has provided valuable insights into the challenges and potentials of ASR systems in handling dysarthric speech. Future studies could benefit from incorporating other dysarthric speech datasets to verify the generalizability of the findings across different speech impairments and demographics. Datasets such as the UA-Speech and Nemours could offer new perspectives and more diverse data, potentially revealing unique challenges and opportunities for ASR system refinement.
2. Incorporating More Evaluation Metrics: This study primarily focused on the BERTscore and Jaro-Winkler distance metrics to evaluate semantic integrity and phonetic accuracy. Future research could expand on this by including additional semantic-based and edit distance metrics that assess other aspects of speech recognition quality.
3. Cross-Linguistic Analysis: Investigating the performance of ASR systems on dysarthric speech in different languages could provide insights into linguistic variables affecting speech recognition. Cross-linguistic studies could help in designing more robust ASR systems that are adaptable to a variety of phonetic and syntactic structures present in diverse languages.

7. Responsible Research

In the context of conducting responsible research, it is essential to address several key aspects: data transparency and availability, ethical considerations, and repeatability. Each of these facets contributes to the integrity and impact of the research, especially when dealing with sensitive areas such as speech impairments.

The TORGO database [14], used extensively in this study, is a benchmark in data transparency and availability in research on dysarthric speech. It is publicly accessible, allowing researchers to examine and replicate findings in the realm of speech recognition technologies for individuals with speech disabilities. This transparency not only facilitates broader validation and testing of new ASR systems but also encourages a collaborative approach to advancements in this field.

Using the TORGO dataset adheres to high ethical standards, primarily because the data collection involved comprehensive consent processes, anonymization, and ethical oversight. The dataset includes speech recordings from individuals with cerebral palsy and amyotrophic lateral sclerosis (ALS), alongside data from age- and gender-matched controls. These recordings were obtained under strict ethical guidelines to ensure the dignity and privacy of all participants, making it a safe and respectful resource for conducting speech recognition research.

Repeatability is a cornerstone of robust scientific research. In this study, we have ensured that all experiments are repeatable by making all code used in the analysis and processing of data publicly available [2]. This includes scripts for data cleaning, preprocessing, running the Whisper ASR model, and analyzing the output. By sharing these resources, we aim to enable other researchers to replicate our work, verify our claims, and build upon the foundations we have laid. This open approach not only strengthens the validity of our findings but also enhances the collective capability to develop more inclusive and effective ASR systems for people with speech impairments.

[2] https://github.com/notfilip/research-project

8. References

[1] P. C. Finnerty, "Introduction to communication sciences and disorders," *Topics in Language Disorders*, vol. 16, no. 3, p. 85, 1996.

[2] V. Young and A. Mihailidis, "Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review," *Assistive Technology*, vol. 22, no. 2, p. 99–112, 2010.

[3] V. I. Levenshtein *et al.*, "Binary codes capable of correcting deletions, insertions, and reversals," in *Soviet physics doklady*, vol. 10, no. 8. Soviet Union, 1966, pp. 707–710.

[4] D. Jurafsky and J. H. Martin, *Speech and language processing*, 2nd ed., ser. Prentice Hall series in artificial intelligence. London [u.a.]: Prentice Hall, Pearson Education International, 2009. [Online]. Available: http://aleph.bib.uni-mannheim.de/F/?func=find-b&request=285413791&find_code=020&adjacent=N&local_base=MAN01PUBLIC&x=0&y=0

[5] 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]. Available: https://www.sciencedirect.com/science/article/pii/S0167639307000404

[6] A. Aksënova, D. van Esch, J. Flynn, and P. Golik, "How might we create better benchmarks for speech recognition?" 2021. [Online]. Available: https://aclanthology.org/2021.bppf-1.4/

[7] I. Mccowan, D. Moore, J. Dines, D. Gatica-Perez, M. Flynn, P. Wellner, and H. Bourlard, "On the use of information retrieval measures for speech recognition evaluation," 01 2004.

[8] Y.-Y. Wang, A. Acero, and C. Chelba, "Is word error rate a good indicator for spoken language understanding accuracy," in *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721)*, 2003, pp. 577–582.

[9] Y. Wang, J. Qin, and W. Wang, "Efficient approximate entity matching using jaro-winkler distance," in *Web Information Systems Engineering – WISE 2017*, A. Bouguettaya, Y. Gao, A. Klimenko, L. Chen, X. Zhang, F. Dzerzhinskiy, W. Jia, S. V. Klimenko, and Q. Li, Eds. Cham: Springer International Publishing, 2017, pp. 231–239.

[10] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," 2020.

[11] J. Rugayan, G. Salvi, and T. Svendsen, "Perceptual and Task-Oriented Assessment of a Semantic Metric for ASR Evaluation," in *Proc. INTERSPEECH 2023*, 2023, pp. 2158–2162.

[12] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022.

[13] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020.

[14] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The torgo database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, pp. 523–541, 2011.

[15] K. T. Mengistu and F. Rudzicz, "Adapting acoustic and lexical models to dysarthric speech," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 4924–4927.

[16] Z. Yue, F. Xiong, H. Christensen, and J. Barker, "Exploring appropriate acoustic and language modelling choices for continuous dysarthric speech recognition," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6094–6098.