Investigating the Performance of MIKNN for Objective Speech Intelligibility Assessment of Dysarthric Speech

Kruthika Reddy Kowkuntla

Supervisor(s): Jorge Martinez Castaneda, Dimme de Groot

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
January 26, 2025

Name of the student: Kruthika Reddy Kowkuntla
Final project course: CSE3000 Research Project
Thesis committee: Jorge Martinez Castaneda, Dimme de Groot, Przemyslaw Pawelczak

Abstract

Assessing speech intelligibility for individuals with dysarthria is critical for understanding the severity of motor speech disorders and evaluating speech therapy interventions. Traditional subjective assessments, while effective, are resource-intensive and prone to bias, which highlights the need for reliable objective measures. This study investigates the applicability of MIKNN (Mutual Information with K-Nearest Neighbors) as an objective speech intelligibility measure for dysarthric speech, by comparing objective intelligibility scores with subjective ratings. Unlike its proven effectiveness with neurotypical speech, the performance of objective measures on atypical speech, such as dysarthria, remains under-explored. The study compares MIKNN with state-of-the-art measures, including P-STOI and P-ESTOI, using the UA-Speech dataset. Key challenges addressed include adapting MIKNN to handle the temporal and spectral variability inherent in dysarthric speech. The results demonstrate that while MIKNN offers promising correlations with subjective scores, it is outperformed by P-STOI and P-ESTOI.

Index Terms: MIKNN, dysarthria, dynamic time warping, P-STOI, P-ESTOI

1 INTRODUCTION

Dysarthria is a motor speech disorder that is caused by damage to the central or peripheral nervous system [1]. It commonly occurs in individuals with neurological conditions such as Cerebral Palsy (CP), Brain injury or Amyotrophic lateral sclerosis (ALS). This results in affecting the respiration, phonation, resonance, articulation and prosody [1] during speech production, leading to articulation deficiencies, vowel distortions, reduced loudness variation, hypernasality or syllabification [2], [3], ultimately reducing the speech intelligibility. The disorder occurs in various forms, including spastic dysarthria (characterized by poor control of volume of speech and slow rate of speech), athetoid dysarthria (characterized by involuntary, irregular movements), and mixed dysarthria (a combination of features from multiple types) [4]. These variations in speech patterns introduce considerable complexity when trying to automatically evaluate its intelligibility .

Identifying speech disorders like dysarthria is crucial in clinical practice because these conditions often are early indicators of neurodegenerative diseases [5]. Tracking speech characteristics like intelligibility and severity over time is vital for monitoring progression of the disease and evaluating the impact of speech therapy [6]. The standard approach to measure the intelligibility of dysarthric speech is based on subjective listening tests, in which human evaluators assess the speech to determine how well it is understood. However, this method is resource intensive and time consuming and susceptible to subjective biases, such as the availability of contextual cues [7]. These evaluations can be challenging, particularly for mild impairments, as distinguishing subtle clinical features can be difficult even for experts. Furthermore, there is often overlap in the clinical characteristics of various speech disorders, which complicates accurate diagnosis. Non-expert clinicians may find these assessments particularly challenging, and even among experts, the reliability can be low [8]. Therefore, the need for automatic assessment for dysarthric speech intelligibility arose.

In recent years, the use of machine learning and deep learning in research on dysarthria has grown significantly, with particular emphasis on automatic pathological speech detection. Classic machine learning approaches rely on machine learning models like SVMs, random forests, and Gaussian mixture models, using features such as spectral, cepstral, and articulatory characteristics. While these approaches demonstrated promising results on small datasets, their generalization to larger, diverse populations was limited due to biases in demographics, recording conditions, and language [9]. Deep learning models, including CNNs and LSTMs, have significantly improved performance by leveraging advanced feature representations. Multi-task learning and self-supervised models like wav2vec 2.0 have shown state-of-the-art results, outperforming classical methods. However, challenges such as language dependency and lack of clinical interpretability remain barriers to universal application [3], [10].

Building on this progress in detecting pathological speech, the focus has also been on assessing intelligibility —a critical factor for understanding the severity and progression of speech impairments. Methods for automatically assessing the intelligibility of pathological speech are generally divided into two main categories: blind and non-blind approaches. Blind approaches aim to evaluate the intelligibility of pathological speech without typical speech data as reference. These methods often focus on analyzing acoustic features such as jitter, shimmer, fundamental frequency, and formant frequencies, which are thought to have a strong connection to speech intelligibility. Non-blind approaches use the intelligible speech data from typical speakers to assess the intelligibility of pathological speech. These methods commonly utilize features derived from automatic speech recognition (ASR) systems trained on extensive datasets of typical speech. Despite the potential of this technology, its integration into standard clinical practice remains limited, and even ‘methodologically rigorous‘ approaches often yield disappointing results [3].

While blind and non-blind assessment methods focus on developing new measures, the adaptation of existing intrusive objective speech intelligibility measures (OIMs) from the speech enhancement domain remains largely underexplored. Similar to non-blind approaches, the intrusive OIMs require typical speech data for reference. Recent research [11] demonstrated promising results in assessing dysarthric speech using the Short-Time Objective Intelligibility (STOI) measure, a well-established intrusive OIM. This success has inspired further exploration into adapting any other intrusive OIM for evaluating dysarthric speech in this paper.

Notable examples of intrusive OIMs include the Speech Intelligibility Index (SII), which evaluates the signal-to-noise ratio across frequency bands, and the Speech Transmission Index (STI), which measures modulation transfer functions of the speech signal. However, many OIMs have significant limitations. Such as their reliance on long-term statistics, which fail to capture the fine-grained temporal distortions often present in dysarthric speech and their inability to account for the non-linear dependencies commonly observed in pathological speech [12]. STOI and ESTOI addressed this limitation by this limitation by incorporating short-term temporal resolution [13]. Among the existing measures, MIKNN (Mutual Information with K-Nearest Neighbors) stands out as theoretically closest to STOI, offering a high-performing alternative to assess dysarthric speech.

MIKNN is built based on information theory [14] and uses the same speech representation as STOI but takes a different approach to quantify distortion. Instead of relying on the short-time correlation, MIKNN calculates the mutual information between the clean and distorted temporal envelopes using a non-parametric k-nearest neighbor estimator. This shift allows MIKNN to capture both linear and non-linear dependencies between signals, addressing a key limitation of correlation-based methods. The development of MIKNN was supported by the TaalPOST and KjemsITFS datasets, which were also used during the development of STOI [13]. These datasets are widely regarded as benchmarks for evaluating OIMs. Their inclusion of diverse range of distortions, ranging from stationary and fluctuating noise to reverberant conditions, ensures robust evaluation across varied real-world scenarios. The shared use of these datasets highlights key alignments between MIKNN and STOI. Both metrics were developed and validated on identical data sources, enabling a direct comparison of their performance. Additionally, the mutual reliance on temporal envelope representations reinforces the methodological similarity, while MIKNN's information-theoretic framework offers a more flexible and theoretically grounded approach to evaluating intelligibility.

This exploration led to the formulation of the research question: **Investigating the Performance of MIKNN for Objective Speech Intelligibility Assessment of Dysarthric Speech**. By applying MIKNN, this research aims to evaluate its capability to handle dysarthric speech by comparing its objective intelligibility scores with subjective scores obtained from listening tests. Correlation analysis is conducted to determine the relationship between MIKNN's predictions and subjective intelligibility scores. Finally, MIKNN's performance is compared to state-of-the-art measures, such as STOI and ESTOI, to assess its relative effectiveness in the pathological speech domain.

The remainder of this paper is organized as follows. Section 2 provides a detailed overview of the related work in OIMS and pathological speech, highlighting key methodologies and state-of-the-art measures that will be used for the comparative analysis of MIKNN. Section 3 outlines the overview of MIKNN, detailing how it uses information theory and knn method. Section 4 explains the methodology used in this study, including the personal contributions to the methodology, the preprocessing steps for the data and the intelligibility assessment using MIKNN. Section 4 describes the experimental setup and results, including the dataset used, evaluation metrics, mapping function employed, and presents the results and comparative analysis, evaluating MIKNN's performance against state-of-the-art measures like STOI and ESTOI. Section 6 provides a discussion of the findings, exploring their implications, limitations, and potential areas for improvement. Finally, Section 7 concludes the paper by summarizing the key contributions and outcomes of this study.

2 RELATED WORKS

The study in [11] is the primary work evaluating the performance of a standard OIM for pathological speech intelligibility assessment. It proposes a novel approach that uses STOI and ESTOI measures, adapted for pathological speech as P-STOI and P-ESTOI. Their method addresses the limitations of traditional intelligibility measures by employing dynamic time warping (DTW) to align pathological speech signals with reference representations. The authors also proposed the new approach of constructing reference representations from 'multiple' healthy speakers. These reference representations are generated on an utterance-specific basis, using DTW-based clustering and averaging across healthy speaker templates in the one-third octave band domain, ensuring that the reference signal captures the key characteristics of intelligible speech. P-STOI quantifies temporal distortions, while P-ESTOI incorporates spectral correlations, providing a comprehensive analysis of intelligibility.

Experimental evaluation across English (CP) and French (ALS) datasets in [11] demonstrated that P-STOI and P-ESTOI high Pearson correlation coefficients (as high as 0.95) with subjective intelligibility scores, outperforming several state-of-the-art feature-based approaches that were introduced in [15]. The work in [11] represents a significant advancement in automatic pathological speech intelligibility assessment by providing a robust framework for evaluating intelligibility in pathological contexts. The proposed methodology aligns closely with the research focus of this paper, on enhancing the generalization of objective intelligibility metrics to atypical speech populations.

Other existing state-of-the-art objective intelligibility measures were proposed in [15]:

- Linear Prediction Residual Kurtosis(K_{LP}): evaluates the shape of the residual signal obtained from linear prediction (LP) analysis by calculating its kurtosis. Lower kurtosis values suggest irregular, noise-like excitation patterns, which are often linked to severe speech disorders, while higher values are typical of normal speech.
- Standard deviation of the zeroth-order delta coefficient (σ_{Δ}): captures variations in short-term speech dynamics by analyzing the standard deviation of delta cepstral coefficients, which represent changes in spectral features over time. It provides insights into how smoothly energy and spectral features change within a speech signal, which can be disrupted in dysarthric speech.
- Voicing percentage ($\%V$): calculates the proportion of voiced speech segments compared to the total utterance

duration. It reflects prosodic characteristics, as changes in voicing are a key feature of intelligibility in dysarthric speech.

- Fundamental frequency range (Δ_{f_0}): assesses the range of fundamental frequency (f_0) variations, which are essential for prosody and naturalness in speech. Reduced pitch range or monotonic pitch patterns are common in individuals with dysarthria
- Low-to-high modulation energy ratio ($LHMR$): analyzes the modulation spectrum of the speech signal by comparing energy in low-frequency bands (below 4 Hz) to that in higher-frequency bands. It captures rhythmic aspects of speech, with deviations from typical modulation patterns indicating potential intelligibility issues.

The P-STOI study [11] conducted a comparative analysis of P-STOI and P-ESTOI to performances K_{LP}, σ_Δ, $\%V$, Δ_{f_0}, and $LHMR$ from [15]. In this study, we will add MIKNN to the comparative analysis.

3 OVERVIEW OF MIKNN

The MIKNN (Mutual Information with K-Nearest Neighbours) is a speech intelligibility measure based on mutual information(MI). It was proposed in [14], where the prediction of speech intelligibility based on information theory was investigated. MI is an information-theoretic measure that quantifies the statistical dependency between two random variables, extending beyond linear relationships typically captured by correlation or signal-to-noise ratio (SNR). In the context of speech intelligibility, MIKNN evaluates the dependency between the temporal envelopes of clean and processed speech in specific subband domains, making it particularly suitable for assessing nonlinear distortions. The KNN method for MI estimation was formalized in [14], using properties of k-nearest-neighbor distances to estimate joint and marginal entropy distributions.
The choice of the k parameter (defining the neighborhood size) significantly impacts the accuracy of the MI estimation. A balance between sampling errors and systematic estimation error is taken into account for choosing k, to make the error as small as possible.
The MI between two continuous random variables X and Y is calculated as:

$$I(X,Y) = \int_X \int_Y p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) dx\, dy \qquad (1)$$

Here, p(x,y) represents the joint probability density function of X and Y, while p(x) and p(y) are their marginal densities.
Input speech signals (clean and processed) are resampled to a uniform sampling rate, typically 10 kHz, to capture a relevant frequency range for intelligibility. The signals are divided into overlapping frames, each windowed with a Hann function and zero-padded to ensure a consistent Fourier Transform length. Discrete Fourier Transform (DFT) is done on each frame and the DFT bins are grouped into one-third octave bands. These bands represent the temporal envelopes of the clean and processed speech signals, denoted by X (clean) and Y (degraded).
The intelligibility score of each one-third octave band is measured by means of MI.

$$d_j = \hat{I}(\Phi_{x_j}; \Phi_{y_j}), \qquad (2)$$

where d_j is the intermediate intelligibility measure of the jth octave band, $\hat{I}(.)$ is the estimated MI calculated using (1) and Φ_x and Φ_y are the vectors of temporal envelopes of the clean and processed signals. After calculating d_j for all bands using (2), the final intelligibility score (d_{raw}) is obtained by averaging the MI values across all bands.
The unit of d_{raw} is nats. A normalized intelligibility score (%) is computed by dividing the intelligibility score between the test signal and the reference signal with the intelligibility score obtained between the signal and itself. The implementation of MIKNN algorithm used in this paper was obtained from the website of the developer mentioned in [14].

4 METHODOLOGY

This study adopts an extended approach to assessing pathological speech intelligibility, heavily inspired by the study in [11] which was also discussed in section 2 of this paper. The primary objective of following a methodology similar to that in [11] is to enable a fair and consistent comparative analysis of the performance of MIKNN against the established measures (P-STOI and P-ESTOI). This section contains the contributions made to this research method, and how the intelligibility assessment using MIKNN was carried out.

4.1 Contributions to the Methodology

Several key contributions were made to adapt the MIKNN algorithm for pathological speech intelligibility assessment. While the original MIKNN algorithm focuses on estimating MI using a k-nearest neighbor (KNN) approach, this work specifically utilizes the KNN-based MI estimation as its core and introduces the following advancements:

1. Reorganization of algorithmic steps for better alignment with P-STOI: Based on the recognition of similarities in data representation between MIKNN and P-STOI , the MIKNN framework was adapted to align its data processing approach with that of P-STOI. This ensures compatibility with P-STOI's efficient handling of pathological speech data while retaining MIKNN's mutual information-based foundation.
2. Integration of dynamic time warping (DTW): To account for temporal variability in pathological speech, DTW was incorporated into the adapted MIKNN framework. DTW dynamically aligns the time frames of the reference and test signals, enhancing the algorithm's ability to address temporal mismatches before MI estimation.
3. Simplification of the framework: Elements of the original MIKNN framework that were not directly relevant to the goals of this study were excluded. The 'silent frame removal' was omitted as it can significantly reduce the duration of already short signals, potentially leaving insufficient data for meaningful analysis. Silent frames, characterized by minimal amplitude or energy, naturally

contribute less to DTW's distance metric, thereby minimizing their impact. This inherent handling of silent frames by DTW makes explicit removal redundant.

This restructuring was implemented to enhance the MIKNN's suitability for assessing speech intelligibility in this specific context, laying the foundation for preprocessing and reference signal construction, as described in subsequent sections.

4.2 Data preprocessing

To assess pathological speech intelligibility, the input data underwent a series of preprocessing steps to prepare the representations for further analysis. These steps include extracting the time-frequency (TF) representation, mapping it to a 1/3 octave band scale, and applying DTW for temporal alignment.

The speech signals were first converted into a TF representation using a short-time Fourier transform (STFT). This process involves segmenting the audio signal into overlapping time windows and computing the Fourier transform for each segment. The resulting TF representation captures the spectral energy distribution across time, with each frame representing a snapshot of the frequency content over a short interval.

To achieve a meaningful acoustic representation, the TF representation was mapped onto a 1/3 octave band scale. This mapping compresses the frequency spectrum into bands that approximate the human auditory system's sensitivity to sound frequencies. Specifically, the energy in each band was computed by summing the squared magnitudes of the TF bins corresponding to that band. This transformation reduces dimensionality and also make it well-suited for intelligibility analysis.

DTW was employed to align the 1/3 octave band representations of speech signals from different speakers. DTW minimizes the distance between two sequences by non-linearly aligning their frames, allowing for differences in speaking rates and durations. A simple Euclidean distance metric was used as the local cost function for alignment. This step ensured that the pathological speech signals were temporally aligned with the reference signals, enabling frame-by-frame comparisons for intelligibility evaluation.

4.3 Reference Signal Construction

Since MIKNN requires comparison between dysarthric speech and a clean reference, we construct an utterance-dependent reference representation: For each utterance being evaluated, a healthy speaker is chosen at random from the pool of available 13 healthy speakers in the dataset. The one-third octave band representation of the selected healthy speaker (lets denote it with X) is aligned to the representations of all other healthy speakers using DTW. This process ensures that frames from different speakers are mapped to corresponding points in time, even if their speaking rates differ. For every frame in X, we extract all frames that DTW maps to it from the representations of the other speakers. The collected frames for each point in X are averaged to create a single reference frame that represents the corresponding segment. The complete reference template for the utterance is formed by concatenating all the averaged reference frames. This ensures that the reference template has the same length as the initial selected speaker's representation X. In this way, a reference template was created for each word (utterance) provided in the dataset. By including all healthy speakers, the reference captures a broader range of speech patterns, accounting for natural variations in pronunciation, speaking rate, and acoustic characteristics. This ensures that the reference template is not biased toward the any single healthy speaker, making it more robust and representative of general "healthy" speech.

4.4 Intelligibility assessment

1. Compute the time-frequency representation of the signals.
2. Apply 1/3 octave band decomposition to break the speech signals into 15 frequency bands. This representation captures the speech signal in frequency bands that align with human auditory perception.
3. Use DTW to align the one-third octave band representation of the test utterance with the reference template. Use Euclidean distances as the local scoring metric during alignment to match corresponding frames effectively.
4. Compute the normalized intelligibility score using the k-nearest neighbour method proposed in the MIKNN algorithm for the estimation of MI between the aligned dysarthric speech signal and the aligned reference template.
5. The objective intelligibility score (d) for each subject is considered as the mean of the scores computed in step 3 of all utterances.

5 EXPERIMENTAL SETUP AND RESULTS

5.1 Database

For this research, the publicly available 'Dysarthric Speech Database for Universal Access Research (UA-Speech)' made by University of Illinois [16] was used. The speech data was recorded using an 8-channel microphone array, and sampled at 16kHz. For the objective intelligibility assessment, the recordings of the 5th channel were used. This choice was made to maintain consistency with the P-STOI experiments in [11]. The normalized version of the speech data was utilized, where the files were scaled to fully utilize the dynamic range, ensuring consistent amplitude levels and improving the quality of intelligibility evaluations.

Each participant read 765 utterances in total, with 455 distinct words, including three repetitions of 155 words for training and testing, and 300 uncommon words to enhance phonetic diversity. Subjective listening tests were performed to assess the speech intelligibility and obtain score for each pathological speaker. Based on the mean subjective intelligibility scores, each speaker was then classified into one of four intelligibility categories: very low (0– 25%), low (26–50%), mid (51–75%) and high (76–100%). Recordings from 13 age-matched healthy speakers were provided in the database, and

are considered for clean reference speech signals in the experiments in this paper . Ten spastic dysarthric speakers (7 males, 3 females) are the subjects; Table 1 shows their subjective intelligibility scores.

Speaker	Age	Intelligibility (%)	Category
M01	18	15	Very low
M04	18	2	Very low
M05	21	58	Mid
M07	58	28	Low
M08	28	93	High
M14	40	90.4	High
M16	Unreported	43	Low
F02	30	29	Low
F03	51	6	Very low
F05	22	95	High

Table 1: Demographics of the ten spastic dysarthric speakers

5.2 Value of k-nearest parameter

The k-value used in these experiments was set to 10, which is the minimum recommended value by the original authors of the MIKNN algorithm. Using $k = 10$ ensures a balance between statistical reliability and computational efficiency. It is important to note that selecting a larger k-value is not advisable in this context due to the short length of the utterances in the dataset. Choosing larger k could reduce the sensitivity of the mutual information estimation, especially when the data size is limited. This makes $k = 10$ an optimal choice for these experiments.

5.3 Mapping

The evaluation of intelligibility scores in this study is grounded in comparing the normalized objective score (d) that is produced from Section 4, with the subjective scores. It is important to note that d quantifies the intelligibility of speech in an objective manner, representing the percentage of information retained in the signal. In contrast, subjective measures, such as the word correct score (WCS), capture human perception and understanding of speech. This distinction highlights the need for additional processing to align objective metrics with subjective evaluations. In [14], it is explicitly mentioned that d is not directly equivalent to the WCS and suggests use of a mapping function.

Accordingly, a logistic mapping function is used, consistent with prior literature, [17] and [18]. This approach has also been explicitly used with MIKNN intelligibility scores in [12]. The mapping function, expressed as:

$$f(d) = \frac{100}{1 + e^{a(d-b)}} \tag{3}$$

serves as a tool to quantify the strength of the relationship between the d and the subjective intelligibility scores. Here, a is the slope and b is the midpoint, and these parameters are optimized to minimize the mean squared error between p and $f(d)$.

Without the mapping function, the correlation coefficients that will be used in the subsequent analysis, might fail to capture the non-linear relationship between objective and subjective measures, resulting in misleading performance evaluations.

5.4 Evaluation metrics

- Pearson's Correlation Coefficient (R): quantifies the strength and direction of the linear relationship between the predicted intelligibility scores and subjective listening test results.
- Spearman's Rank Correlation Coefficient (R_s): assesses the monotonic relationship between predicted and subjective scores, making it suitable for non-linear but monotonic dependencies.
- p-values: indicate the statistical significance of the correlation coefficients, representing the probability that the observed correlations occurred by chance. A smaller p-value ($p < 0.05$) indicates higher confidence in the validity of the correlation.

5.5 Comparative analysis of MIKNN with state-of-the-art measures

Figure 1 illustrates that MIKNN exhibits a moderate linear relationship between its mapped scores and the subjective scores. This indicates that MIKNN captures intelligibility to some extent. However, its performance is notably weaker compared to P-STOI and P-ESTOI. The gentler slope of MIKNN's regression line, combined with the greater dispersion of its scatter points from the line, reflects higher variability in its predictions and reduced consistency in aligning with subjective intelligibility ratings.

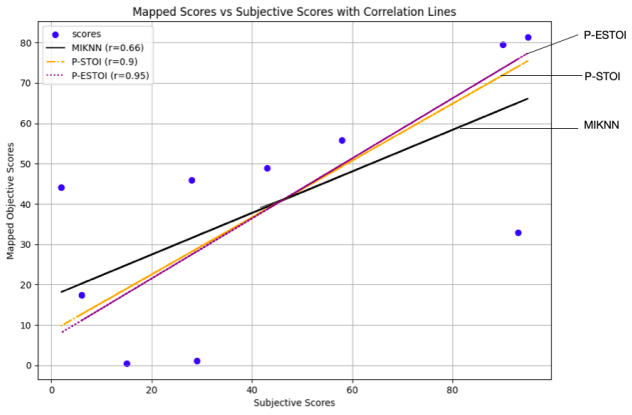

Figure 1: Scatter plot of the intelligibility scores. The vertical axis is the 'mapped objective intelligibility scores' and horizontal axis is the 'subjective intelligiblity scores'. The respective Pearson correlation coefficients (r) for each measure are shown. Each line represents the linear fit corresponding to the respective measure, with the slope and alignment of the lines indicating the strength of correlation between the subjective and objective scores.

Table 2 reports the performance of the MIKNN, along with other state-of-the-art measures (that were discussed in section 2) on the UASpeech dataset. MIKNN values were computed as part of this study. The values for P-STOI, P-ESTOI, K_{LP}, σ_{Δ}, $\%V$, Δ_{f_0}, and $LHMR$ were taken from prior studies in [11] and [15]. The same parameters used for the time-frequency (TF) analysis in P-STOI and P-ESTOI were applied in this study. Specifically, a Hamming window of 32 ms with a 50% overlap was employed.

Measures	R	p	R_s	P
MIKNN	**0.66**	0.0389	0.62	0.0537
P-STOI	**0.90**	5E-04	**0.82**	7E-03
P-ESTOI	**0.95**	4.3E-5	**0.91**	2E-04
K_{LP}	0.41	0.23	0.42	0.23
σ_{Δ}	0.45	0.2	0.51	0.13
%V	-0.40	0.25	-0.58	0.08
Δ_{f0}	**-0.70**	0.02	-0.61	0.06
LHMR	-0.55	0.09	-0.54	0.10

Table 2: Performance of MIKNN and other state-of-the-art measures. Bold text indicates significant correlations ($p < 0.05$). R indicates the Pearson correlation coefficient and R_s indicates Spearman correlation coefficient.

While MIKNN is outperformed by P-STOI and P-ESTOI, it is important to note that it demonstrates a significantly stronger correlation with subjective scores compared to K_{LP}, σ_{Δ}, $\%V$, Δ_{f_0}, and $LHMR$. This highlights its potential as a promising metric for intelligibility assessment.
Moreover, this opens up the possibility of developing an enhanced measure that builds upon MIKNN's approach to assess dysarthric speech. A similar methodology has been demonstrated by [15], where a composite measure combining K_{LP}, σ_{Δ}, $\%V$, Δ_{f_0}, and $LHMR$ was shown to composite measure was shown to be a reliable indicator of dysarthric word intelligibility [15]. Then adapting such a strategy for MIKNN could further enhance its utility in this domain.

6 DISCUSSION

This study has presented the results of evaluating MIKNN for objective speech intelligibility assessment of dysarthric speech. Several aspects of the methodology and results deserve further discussion, particularly regarding experimental conditions, dataset constraints, and factors affecting the performance of MIKNN.
The length of the utterances used for MIKNN plays a critical role in the reliability of intelligibility scores. Short utterances may fail to provide sufficient temporal and spectral information for accurate alignment and intelligibility estimation, which could lead to less robust correlations with subjective scores. Ensuring that future datasets include longer utterances across multiple utterances from the same speaker could address this issue. The idea of aggregating results was omitted because [11] mentioned that repeated frames can affect intelligibility. This limitation was particularly relevant here due to the already limited number of frames in the a speech representation and the reliance of the KNN method on sufficient and diverse neighbor data for accurate mutual information estimation.
A major limitation of this study is the small size of the dataset used. The UA-Speech database, while valuable, includes only 16 spastic dysarthric speakers, out of which some subjects' data is unavailable. This restricts the generalizability of the findings, particularly across other types of dysarthria, such as athetoid or mixed dysarthria. Moreover, the lack of publicly available, diverse datasets for dysarthric speech remains a significant barrier to advancing research in this field [3].
Another factor that could influence the performance of MIKNN is the construction of utterance representations. This study employed a reference-based approach, where the utterance representations were derived from a pool of healthy speakers. Selecting the initial reference represenation might affect the computation of the intelligibility scores. To investigate this, the process needs to be repeated with different initial reference represenations [11].
While the experiments in this study used $k = 10$ based on the MIKNN's restriction to not have the $k < 10$, it is worth noting that varying k could significantly affect the results. The limitation to not conduct experiments with various k-values is due to the short length of the utterances. Futher experiments should systematically explore the impact of different k-values on both Pearson and Spearman correlation coefficients, which may help optimize MIKNN for different speech datasets.
Future work should focus on addressing the methodological constraints discussed here. Exploring variations in k-values, ensuring sufficient utterance lengths, expanding dataset size and diversity, and refining reference template construction are the key steps.

7 RESPONSIBLE RESEARCH

The database [16] utilized for this study has been specifically developed for advanced research in automatic speech recognition (ASR) for people with neuromotor disabilities. Ethical considerations were an integral part of the creation and use of this dataset. The 19 participants with Cerebral Palsy were recruited through rehabilitation centers and personal networks, ensuring their voluntary participation. To protect participant privacy, personally identifiable information was excluded from the dataset. The dataset files were only made available on request. The intended use of the corpus and the research institution/university name and institutional email address was asked before the access was made available. Informed consent was obtained prior to recording and participants were informed of the purpose of the study and how their data would be used in research. Participants were required to provide explicit consent for their data to be shared; those who declined are not included in the dataset's distribution. For example, subject M06 did not approve the redistribution of his data, hence his data was no longer in the dataset that can be accessed now.
The dataset represents a diverse group of speakers in terms of

age, gender, and severity of dysarthria. This inclusivity ensures that ASR systems trained on the dataset are more representative of real-world scenarios. However, it is acknowledged that biases could still arise due to the specific focus on cerebral palsy, which may limit generalizability to other forms of dysarthria.

8 CONCLUSIONS

This study investigates the performance of MIKNN for the objective assessment of dysarthric speech intelligibility. It addresses the research gap of using existing intrusive objective speech intelligibility measures from the speech enhancement domain for dysarthric speech assessment. By utilizing mutual information, MIKNN captures both linear and nonlinear dependencies in speech signals, offering a theoretically grounded approach to intelligibility evaluation. The research compares MIKNN to P-STOI or P-ESTOI and other state-of-the-art metrics, analyzing its correlation with subjective intelligibility scores. Key methods of this work include adapting MIKNN for pathological speech, integrating dynamic time warping for temporal alignment, and optimizing the preprocessing pipeline for intelligibility evaluation, using the UA-Speech database. The findings revealed that while MIKNN does not outperform P-STOI or P-ESTOI, it demonstrates stronger correlations than several other feature-based measures K_{LP}, σ_{Δ}, $\%V$, Δ_{f_0}, and $LHMR$. MIKNN shows significant correlation with subjective scores which highlights its potential as a complementary tool for dysarthric speech intelligibility assessment.

References

[1] P. Enderby. "Disorders of communication: Dysarthria". In *Handbook of Clinical Neurology*, volume 110, pages 273–281. Elsevier, 2013.

[2] J. R. Duffy. "Motor speech disorders". In *Handbook of Clinical Neurology*. Elsevier, 1995.

[3] Md. Sahidullah S. A. Sheikh and T. Kodrasi. "Deep Learning for Pathological Speech: A Survey". 2025.

[4] P. Enderby. "Frenchay Dysarthria Assessment". In *International Journal of Language Communication Disorders*, 2007.

[5] D. C. Atkins J. M. Tracy, Y. Özkanca and R. H. Ghomi. "Investigating voice as a biomarker: Deep phenotyping methods for early detection of Parkinson's disease". volume 104, 2020.

[6] N. Müller J. S. Damico and M. J. Ball. "The Handbook of Language and Speech Disorders". In *Handbook of Clinical Neurology*. Wiley Online Library, 2010.

[7] N. Miller S. Robson V. Thompson S. Landa, L. Pennington and N. Steen. "Association between objective measurement of the speech intelligibility of young people with dysarthria and listener ratings of ease of understanding". volume 16 no.4, page 408–416, 2014.

[8] P. Maat M. Aldenhoven A. Algra S. Fonville, H. B. van der Worp and J. van Gijn. "Accuracy and inter-observer variation in the classification of dysarthria from speech recordings". volume 255 no.4, page 1545–1548, 2008.

[9] E. Moore J. Laures-Gore S. Russell S. Gillespie, Y.-Y. Logan and R. Patel. "Cross-database models for the classification of dysarthria presence". volume 255 no.4, page 3127–3131, 2017.

[10] V. Aharonson P. Harar Z. Galaz S. Rapcsak J. R. Orozco-Arroyave L. Brabenec D. Kovac, J. Mekyska and I. Rektorova. "Exploring digital speech biomarkers of hypokinetic dysarthria in a multilingual cohort". In *Biomedical Signal Processing and Control*, volume 88, 2024.

[11] H. Bourlard P. Janbakhshi, I. Kodrasi. "Pathological Speech Intelligibility Assessment Based on the Short-time Objective Intelligibility Measure". In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6405–6409, 2019.

[12] W. B. Kleijn S. Van Kuyk and R. C. Hendriks. "An Evaluation of Intrusive Instrumental Intelligibility Metrics". In *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, volume 26, page 2153–2166, 2018.

[13] R. Heusdens C. H. Taal, R. C. Hendriks and J. Jensen. "An algorithm for intelligibility prediction of time–frequency weighted noisy speech". In *IEEE Transactions on Speech and Audio Processing*, volume 19, no. 7, page 2125– 2136, 2011.

[14] J. Taghia and R. Martin. "Objective Intelligibility Measures Based on Mutual Information for Speech Subjected to Speech Enhancement Processing". volume 22, pages 6–16, 2014.

[15] T. H. Falk, W.-Y. Chan, and F. Shein. "Characterization of Atypical Vocal Source Excitation, Temporal Dynamics and Prosody for Objective Measurement of Dysarthric Word Intelligibility". *Speech Communication*, 54:622–631, 2012.

[16] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T.S. Huang, K.Watkin, and S.Frame. "Dysarthric speech database for universal access research". In *Interspeech 2008*, pages 1741–1744, 2008.

[17] B. Schwerin and K. Paliwal. "An improved speech transmission index for intelligibility prediction". volume 65, pages 9–19, 2014.

[18] R. Heusdens C. H. Taal, R. C. Hendriks and J. Jensen. "An algorithm for intelligibility prediction of time–frequency weighted noisy speech". volume 19, pages 2125–2136, 2011.