



An Investigation to Identify Optimal Setup for Automated Assessment of Dysarthric Intelligibility using Deep Learning Technologies

Kyle Hall¹ · Andy Huang¹ · Seyed Reza Shahamiri¹

Received: 16 January 2022 / Accepted: 24 June 2022 / Published online: 7 July 2022
© The Author(s) 2022

Abstract

Recent advances in deep learning have provided an opportunity to improve and automate dysarthria intelligibility assessment, offering a cost-effective, accessible, and less subjective way to assess dysarthric speakers. However, reviewing previous literature in the area determines that the generalization of results on new dysarthric patients was not measured properly or incomplete among the previous studies that yielded very high accuracies due to the gaps in the adopted evaluation methodologies. This is of particular importance as any practical and clinical application of intelligibility assessment approaches must reliably generalize on new patients; otherwise, the clinicians cannot accept the assessment results provided by the system deploying the approach. In this paper, after these gaps are explained, we report on our extensive investigation to propose a deep learning-based dysarthric intelligibility assessment optimal setup. Then, we explain different evaluation strategies that were applied to thoroughly verify how the optimal setup performs with new speakers and across different classes of speech intelligibility. Finally, a comparative study was conducted, benchmarking the performance of our proposed optimal setup against the state of the art by adopting similar strategies previous studies employed. Results indicate an average of 78.2% classification accuracy for unforeseen low intelligibility speakers, 40.6% for moderate intelligibility speakers, and 40.4% for high intelligibility speakers. Furthermore, we noticed a high variance of classification accuracies among individual speakers. Finally, our proposed optimal setup delivered an average of 97.19% classification accuracy when adopting a similar evaluation strategy used by the previous studies.

Keywords Dysarthria · Speech intelligibility · Evaluation · Deep learning

Introduction

Dysarthria is a motor speech disorder caused by weakness or lack of control of the muscles used for articulation [1]. This impairment can be caused by physical injuries that damage the neural cells in the brain, stroke, etc., and is a symptom of neurological disorders such as Parkinson's disease and cerebral palsy [2]. The lack of articulatory control

and muscle weakness associated with dysarthria affect the speaking rate, varied amplitude and pitch dynamics, and articulation, all of which result in unintelligible speech that is more difficult to understand due to imprecise articulation of phonemes or abnormal variations in speech [3]. Intelligibility can be thought of as the quality of speech indicating how well healthy speakers can understand the spoken words [4]; dysarthric individuals may have reduced intelligibility, and variation in dysarthria severity can lead to speech of differing degrees of intelligibility that can significantly reduce the patient's communication abilities. Low speech intelligibility can severely impact dysarthric individuals' life and independence. A study has shown that over 47% of dysarthric sufferers reported often repeating what they say as healthy speakers find it difficult to understand them [5]. Speech and Language Pathologists (SLPs) are required to know the severity of dysarthria to assess the progression in the underlying cause of the impairment [6], which helps them to design effective treatment plans, exercises, and

✉ Seyed Reza Shahamiri
admin@rezanet.com

Kyle Hall
khal745@aucklanduni.ac.nz

Andy Huang
ahua049@aucklanduni.ac.nz

¹ Department of Electrical, Computer, and Software Engineering, Faculty of Engineering, The University of Auckland, Building 405, Level 6, Room 669, 3 Grafton Road, Auckland, New Zealand

recovery sessions, as well as monitor the effectiveness of the therapy [7].

Usually, dysarthric intelligibility assessments are performed clinically by trained SLPs, who estimate and classify the intelligibility based on the number of words they can understand. However, this auditory-based clinical assessment can be subjective [8], and studies have shown these assessments can be influenced by factors such as assessor familiarity and different use of rating scales between assessors [9]. Such factors introduce errors and make these assessments difficult to administer and unreliable, leading to ineffective therapy. Additionally, SLPs need to perform the assessments manually, making them expensive and inaccessible for individuals with limited access to SLPs and speech clinics [10]. Hence, automated intelligibility assessment approaches could offer the potential for an inexpensive, simple, and less subjective way to classify dysarthric speakers and reduce the likelihood of error in the process. Additionally, this may improve access to such assessments and help provide a framework to facilitate remote treatment, recovery plans, and rehabilitation of dysarthric speakers. While traditional computation methods may be insufficient to accurately capture and learn complex patterns associated with human speech, deep learning algorithms are better equipped to offer an avenue to produce a model with sufficient predictive abilities to meaningfully improve dysarthric intelligibility assessments.

Automatic dysarthria intelligibility assessment using machine learning technologies is typically done in two stages. In the first stage, a computational model is built (i.e., trained) based on the historical speech samples and their respective speech intelligibility class label. Once the model is trained, it can map speech samples from speakers with unknown intelligibility levels to a certain class by finding and analyzing similarities among speakers' acoustic features to those used during training without any human intervention. As such, the two primary components of such systems are the feature extraction method presenting acoustic features and the machine learning classifier used to learn and map the features to classes of speech intelligibility. Outputs could be formulated in intelligibility classes, such as low or high intelligibility, and could assist clinicians by providing an interpretable classification of the speaker's intelligibility level.

Studying the literature shows few attempts to leverage machine learning algorithms to automate intelligibility assessment. Nevertheless, as explained in [11], despite high accuracies reported, some previous studies did not properly verify the generalizability of their methods. For example, most studies verified their approaches against the same speaker(s) whose speech samples were used during the training stage, which means the speaker information was leaked in the testing data resulting in the model tending to perform speaker recognition [12] instead of learning intelligibility patterns associated with dysarthria. Additionally, some studies did

not apply cross-validation or test the predictive capability of their models by including all dysarthric subjects in the dataset they considered. This flaw in evaluation strategy is important as we noticed significant fluctuations in the predictive performance across the dysarthric speakers from the same intelligibility class, indicating that any results obtained from one speaker per intelligibility class are not a reliable indication of model performance. Likewise, only reporting the average accuracies without the breakdown of the results per subject or intelligibility class may not be a solid indication of the model performance as higher average accuracies can be achieved if there are more speech samples from intelligible test subjects than severe dysarthria in the testing data.

This paper first presents an overview of the state-of-the-art automatic dysarthric intelligibility assessment approaches, highlighting the gaps in the evaluation methodologies adopted in the previous studies. Next, we report on our extensive investigation in which we conducted multiple sets of experiments to propose a deep learning-based dysarthric intelligibility assessment optimal setup that recommends feature extraction approaches and other parameters future researchers and practitioners need to consider in order to design such a system. Then, we employed different evaluation strategies to thoroughly verify how the optimal setup performs with the 16 dysarthric subjects in the UA-Speech corpus [13] and across different intelligibility classes. Finally, we conducted further experiments to perform a comparative study benchmarking the performance of our proposed optimal setup against the state of the art by adopting similar strategies previous studies employed to verify their models. Our objectives in this study were to address the gaps in evaluation strategy remained by the previous studies and provide a robust indication of intelligibility assessment generalization on new speakers.

Related Work

Based on the data used to build and classify speech models, automatic dysarthria intelligibility classification approaches can be classified into two categories: reference-free approaches, which do not require any prior knowledge of healthy speech, or reference-based approaches that utilize healthy speech signals when measuring intelligibility [11]. To put it differently, reference-free approaches rely on extracting features that are directly correlated with speech intelligibility, whereas reference-based approaches use healthy (normal) speech to make a comparison and measure the ratio of deviations from normal speech, then relate the measurement to intelligibility classes. The rest of this section provides examples of both approaches to intelligibility assessment.

Reference-Free Approaches

Reference-free approaches are more commonly seen in the literature; hence, we can see a larger variety of machine learning algorithms and input modalities have been explored. Some of these studies focused on finding the best acoustic features and others on the machine learning algorithms. For example, in [14], acoustic features, such as the kurtosis of the linear prediction residual and the range of the fundamental frequency, were manually extracted to investigate their correlation with intelligibility and resulted in creating a “composite measure” using six such features. The authors then concatenated these features and built a regression classifier achieving Pearson correlation coefficient (PC) of 0.97 when predicting intelligibility scores of UA-Speech subjects. This work presented intelligibility assessment by calculating an intelligibility score instead of perceiving intelligibility assessment as a multi-class classification task in which the classifier maps the subject’s intelligibility to different classes of speech intelligibility such as mild, moderate, high, and severe intelligibility. An example of the multi-class approach is the study conducted in [15] where artificial neural networks with a single hidden layer were employed to perform intelligibility classification on UA-Speech corpora, achieving a 96.4% classification accuracy using audio descriptors related to the timbre of musical instruments as input features.

Nonetheless, deep learning algorithms have shown to deliver state-of-the-art performances when dealing with unstructured data such as speech in comparison to shallow algorithms. In terms of deep learning algorithms to perform reference-free intelligibility assessment, we can refer to [16], where multiple standard deep learning architectures were built and evaluated on UA-Speech and TORGO [17] corpora. In particular, a fully connected dense neural network, a convolutional neural network (CNN), and a long short-term memory network (LSTM) were considered in this study. The dysarthric subjects in TORGO were classified into three classes of intelligibility (high, moderate, and low) and four classes for UA-Speech subjects (high, moderate, low, and very low). Speech features were presented by mel-frequency cepstral coefficients (MFCCs) in all experiments. The authors also provided experimental results obtained from support vector machines (SVMs). The overall classification accuracies of both the dense and CNN models outperformed the SVM on both TORGO and UA-Speech experiments concluding the superiority of the deep learning architectures. However, the LSTM did not perform as well as the other models since, according to the study, the temporal information learned by the LSTM exposed to the simpler common words during training was insufficient to identify the severity level of the complex uncommon words in the test set.

Nevertheless, this hypothesis could have been explored by using the common words for both training and testing.

When it comes to mining visual data and learning patterns presented in images and videos, CNNs have been the most successful algorithm due to their abilities in learning translation-invariant patterns. Similarly, CNNs have been successfully used in speech modeling tasks where acoustic features were presented visually. In this context, CNNs have been applied for dysarthria intelligibility classification where dysarthric speech features were presented as spectrograms, for example, in [18]. A more comprehensive example is [5], in which a cross-modal framework including both video and acoustic data was proposed. In this framework, MFCCs were considered to present the acoustic features, and a Facial Action Coding system was applied to track facial muscle movement. The architecture used was a multi-input deep learning architecture with two CNN streams, one CNN for each input type. The framework obtained the highest recorded accuracy of over 99% on UA-Speech. Another example is [19], where short speech segments (less than 1 s) were fed to a standard CNN and a deeper CNN with residual connections. With the maximum of 1-s-long speech segments, the study reported that the residual CNN achieved 86.63% classification accuracy compared to 64.35% accuracy obtained from the standard CNN.

Reference-Based Approaches

Automatic speech recognition (ASR) technologies play a key role in reference-based automated approaches. Such approaches rely on the assumption that if the ASR model is trained only on healthy speech, then it should perform poorly on dysarthric speech, and that the performance deteriorates as the intelligibility of the speech decreases. Hence, ASR-based approaches measure speech intelligibility by computing the word or alphabet accuracy ratio obtained from an ASR system. A notable study here is [20] in which Mozilla DeepSpeech speech-to-alphabet system was used to assess the intelligibility of dysarthric speech. In this work, the impaired speech was given to DeepSpeech, and then intelligibility estimates were provided by calculating the cost of transforming the output string from the ASR to the ground truth. The study also compared different subsets of words from the UA-Speech database, and it was found that measuring speech intelligibility using some specific words can be more effective than other words. Estimations obtained from this approach using the specific set of words achieved a 0.98 Pearson correlation coefficient value. Another example is [20], which used an acoustic model based on i-vectors and SVMs trained on English healthy speech and common words from UA-Speech, then evaluated on the dysarthric uncommon words from UA-Speech. The model achieved accuracies up to 0.91 PC.

Limitations of the Previous Studies

Table 1 summarizes the existing intelligibility assessment approaches on the UA-Speech corpus. The studies cited in this table classified the severity of dysarthria based on speech intelligibility into severe dysarthria (very low intelligibility < 25%), high dysarthria (low intelligibility between 25 and 50%), moderate dysarthria (intelligibility between 50 and 75%), and mild dysarthria (high intelligibility > 75%) based on the intelligibility of the subjects reported in the dataset. The table also explains the evaluation strategies adopted in those studies, particularly whether unforeseen subjects were considered during evaluation. UA-Speech corpus is explained in detail in the next section.

As shown in Table 1, previous studies reported very high classification accuracies, but the evaluation approaches adopted did not properly measure the generalizability of the models. For example, the highest performing model achieved an accuracy of 99.6% for their CNN-based uni-model approach using only speech data [7]. Nevertheless, the authors evaluated their model on two high (93% and 90.4%) intelligibility subjects, one moderate (62%), one low (29%), and only one very low (7.4%) intelligibility subject in which the authors handpicked his speech samples. This means it was likely their results were affected by the number

of higher intelligibility test subjects (60% of the test subjects had intelligibility of more than 62%), and the authors did not provide the breakdown of the results per subject. Moreover, there was no explanation why the speech samples of the only very low intelligibility subject were manually selected, and not all audio samples provided in the dataset for this subject were not considered. Thus, the top result may not imply how the model performs with very low intelligibility dysarthric subjects. They also did not test the system for all UA-Speech speakers. Nonetheless, the results reported in this paper are still significant.

Another limitation we noticed in the rest of the works cited in Table 1 was the evaluation data contained speech samples of the same subjects used for training the models; i.e., the models were evaluated on the training speakers. Using the test speaker's data during training may prompt the models to conduct closed-set speaker recognition instead of speech intelligibility assessment. In particular, these studies mostly split the evaluation datasets based on the type of utterances instead of the speakers; a common approach when UA-Speech subjects were considered was using the common 155 words provided by the dataset as the training data, with the 300 uncommon words of the same speaker for the testing data. This provides evidence to the notion that reported intelligibility assessment models may focus more on who

Table 1 Summary of highest performing dysarthria intelligibility classification models on UA-Speech

Paper	Model	Acoustic features	Evaluation methodology	Mean classification accuracy	Comments
Bhat et al. 2017 [15]	ANN	Audio features designed for spectral characterization of musical instruments	Trained and Tested on the same ten subjects	96.4%	Trained and tested across 2812 digits and command utterances from blocks 1 and 2 70% of data used for training and 15% for validation and 15% for testing
Tong et al. 2020 [7]	CNN	MFCC and recorded video of speakers	Trained on ten subjects, tested on five different subjects	99.6%	Held out five manually selected subjects for testing purposes but used only 1/3 of “very low” intelligibility speaker’s data in testing (selecting a single block instead of all three available). Likewise, the testing dataset was imbalanced with more test samples from higher intelligibility speakers, which may have affected the average accuracy provided
Gupta et al. 2021 [19]	ResNet	Spectrogram	Tested and trained on the same eight subjects	98.9%	Subjects were manually selected and tested with 10% of the entire dataset
Joshy and Rajan 2021 [16]	SVM	MFCC	Tested and trained on the same 15 subjects	82.9%	Uncommon words used for testing (4500 utterances), the rest of the data for training (6975 utterances)
	DNN			93.6%	
	CNN			93.2%	
	LSTM			75.1%	

the speaker was and not necessarily their speech intelligibility. As an illustration, the study that utilized residual CNNs [20] achieved 98.9% accuracy but considered only eight UA-Speech dysarthric subjects for training and the same speakers for testing with 90:10 train/test split across all speakers' data.

Consequently, it is inconclusive how well these models perform when given speech data from dysarthric speakers not seen during training. Using test speakers whose utterances were not considered during training delivers a more robust estimation of the model's applicability in real scenarios. This is of significant importance when the models are deployed in clinical practice since they must classify new patients reliably.

Finally, none of the previous studies in Table 1 considered all UA-Speech subjects, nor verified their models against all the subjects. Given the scarcity of dysarthric data is a significant issue [21], not using all available data does not seem justified. Our study addresses all these limitations and provides realistic measurements to indicate generalizability best.

Dysarthric Participants

The UA-Speech [13] corpus was used for all the experiments conducted in this study. The corpus consists of 16 dysarthric speakers whose data can be used, each speaking 455 unique words, with three repetitions (aka blocks) of each word except the uncommon words. A 7-channel microphone array was used to record the utterances, generating seven audio files for each utterance in each block. The prompt words consist of 150 common words such as 10 digits, 26 radio alphabet letters, and 19 basic computer commands such as “delete” and “enter” and 100 common words selected from the most common words in the Brown corpus of written English such as “it,” “is,” and “you”; and 300 uncommon words selected from Project Gutenberg in which children's novels were digitized. These uncommon words are a set of more complex words selected to maximize biphone diversity.

Speech intelligibility of the subjects, provided by the corpus, was measured by having five American-English speakers aged 18–40 years old attempt to transcribe 255 spoken words for a given dysarthric subject. The listeners rated each word between 0 and 2 indicating how certain they were about their transcribe choice, in which 0 was not certain and 2 very certain. Then, UA-Speech authors calculated the ratio of correct responses per listener, and the average percentage of correct responses became the intelligibility rating for the subject. To assess the reliability of assessments among the listeners, an examination of the repeated words was conducted to check whether the repeated words were the same. For the words marked very certain, the authors reported 91.64% average

agreement among the listeners. For the words marked otherwise, the transcriptions provided by the listeners were either identical or of the same phonological similarity. This indicated that the listeners did not randomly transcribe the unintelligible words.

Based on the perceptual intelligibility scores provided in the dataset, we classified UA-Speech subjects into three classes of intelligibility: low intelligibility (less than 33% intelligibility), mild or moderate intelligibility (33–66% intelligibility), and high intelligibility (more than 66% intelligibility). We have considered speech samples provided by all 16 UA-Speech subjects shown in Table 2 in our experiments.

Experiments and Results

In this study, we conducted multiple sets of experiments to identify the optimal setup to conduct dysarthric intelligibility assessment and then verify the optimal setup against all dysarthric subjects in Table 2 and deliver per-speaker results in addition to results per intelligibility class.

We considered speech features presented as both MFCCs and spectrograms for the optimal setup identification since they both delivered significant results in the previous studies. While MFCCs have been widely studied in the literature, spectrograms do not appear to be thoroughly investigated in the context of intelligibility assessment, although they have outperformed other feature extraction approaches in ASR tasks [18]. We experimented with different configurations by selecting different MFCC parameters and spectrogram setups explained below.

Our evaluation method differed from those shown in Table 1 by using held-out speakers to measure how well our models generalize to unseen speakers. In particular, no speech sample from the testing speakers was among the training set in any of the experiment sets explained below. Additionally, accuracy was measured in all experiments.

Additionally, in the remaining of this paper, we refer to *configurations* (confs) as experiments with different MFCC feature extraction parameters, such as different number of MFCC features and frame lengths. On the other hand, *setups* refer to different spectrogram experiments to identify optimal utterance length and whether common UA-Speech words were included. We opted to select different terms (configurations vs setups) to indicate that the latter was not experiments with different spectrogram configurations.

Experiment Set 1: Identifying the Best MFCC-Based Feature Set

In experiment sets 1 and 2, all models were tested on speakers M01 (low intelligibility), M05 (moderate intelligibility), and M09 (high intelligibility) utterances while the rest of the

Table 2 The UA-Speech participants

Dysarthric subjects	Age	Perceptual intelligibility score	Dysarthria diagnosis	Intelligibility class
<i>M04</i>	> 18	2	Spastic	Low
<i>F03</i>	51	6	Spastic	Low
<i>M12</i>	19	7	Mixed	Low
<i>M01</i>	> 18	10	Spastic	Low
<i>M07</i>	58	28	Spastic	Low
<i>F02</i>	30	29	Spastic	Low
<i>M06</i>	18	39	Spastic	Mid
<i>M16</i>	Not provided	43	Spastic	Mid
<i>M05</i>	21	58	Spastic	Mid
<i>M11</i>	48	62	Athetoid	Mid
<i>F04</i>	18	62	Athetoid (or mixed)	Mid
<i>M09</i>	19	86	Spastic	High
<i>M14</i>	40	90	Spastic	High
<i>M10</i>	21	93	Not provided	High
<i>M08</i>	28	95	Spastic	High
<i>F05</i>	22	95	Spastic	High

utterances from the other 13 speakers were used for training. All available words from UA-Speech were utilized in training and testing, but only M5 microphone data was used. The number of audio samples per participant was highly varied concerning the available microphone data, but the M5 microphone had the highest number of collected utterances across all microphones. Table 3 summarizes the speakers used in training and evaluating the models during set 1 and 2 experiments.

In order to design the network architecture for this set of experiments, we employed Keras Tuner [22] applying Hyperband [23] to automatically identify the CNN configuration and architecture shown in Fig. 1. This was done by configuring a search space with different number of neurons in each layer, convolutional filters, batch sizes, dropout values, optimizers, etc. Then, the hyperband optimization algorithm employed an adaptive approach to allocate resources paired with early stopping, trained a large number of models with different architectures and hyperparameters defined in the search space in a few epochs, and proceeded with further training with the models that delivered the best performances. This process iterated until the best performing models were identified and reported.

Identifying the best MFCC configuration, we explored different MFCC frame and sliding window rates, and whether

to include the MFCC delta coefficients. For each setup, we retrained the CNN with the training subjects of Table 3 and measured its performance on the testing subjects. This process was repeated ten times in each configuration while the testing accuracies were recorded and averaged. The results of this experiment are presented in Table 4. As can be seen, the best performance was achieved when the first 13 MFCCs (i.e., mel cepstrum with 12 coefficients plus the energy information) were used with relatively long frames of 256 ms sliding each 128 ms. The inclusion of MFCC first and second derivatives degraded the model performance. Likewise, results demonstrate that longer frames achieved higher performance. Although this is contrary to typical usage of MFCCs in healthy speech processing when short frames of 20–30 ms are commonly used, it was not unexpected and is consistent with our previous findings for dysarthric ASR [24].

Experiment Set 2: Identify the Best Spectrogram Setup

In the second set of experiments, mel-spectrograms were created using a decibel scale on the frequency axis to emphasize the frequency ranges relevant to human speech. Librosa mel-spectrograms [25] with default parameters were used to generate the spectrograms. The default parameters were FFT window length of 2048, 512 samples between successive frames (hop length), and window function was set to Hann.

Generally, the utterance length of the high intelligibility speakers in the dataset was shorter than 5 s, while low intelligibility speakers had a longer average length, typically exceeding 7 s. A trade-off between not using excessively longer utterances that were mostly empty while still long enough to capture low

Table 3 Experiment 1 and 2 train/test dysarthric speakers

Intelligibility category	Test subjects	Training subjects
Low	M01	F02, F03, M01, M04, M07, M12
Moderate (mid)	M05	F04, M05, M06, M1, M16
High	M09	F05, M08, M09, M10, M14

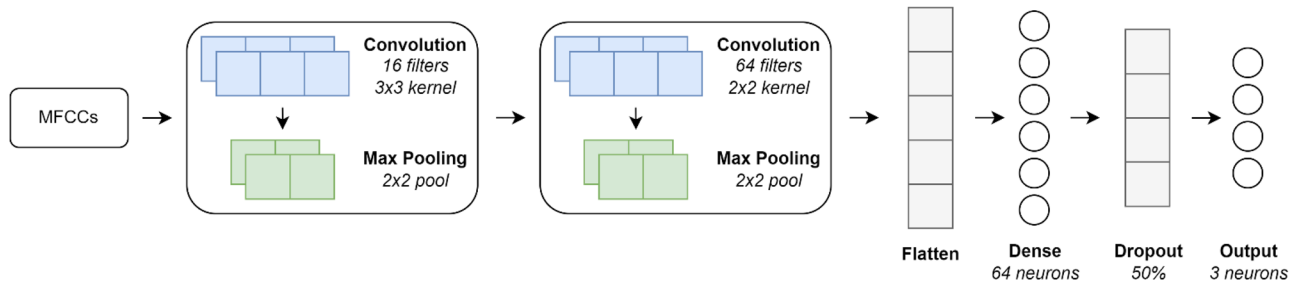


Fig. 1 Experiment set 1 CNN architecture

intelligibility speech was difficult to determine. Hence, we consider the optimal length of the utterance as one of the parameters to investigate during this experiment set. To do so, we conducted multiple experiments retraining the model using 5-, 7-, and 10-s audio lengths. Audio files that were longer than 2 s over the predetermined audio length were excluded from both training and evaluation to ensure consistency among the generated spectrograms. From our observations, spectrograms from longer UA-Speech utterances mostly consist of silence or background noise and did not include useful speech data.

Moreover, in our early experiments, we noticed that models trained only on uncommon UA-Speech words performed better — this seemed counterintuitive to exclude data, especially when the dysarthria acoustic data is scarce, but the difference in results was significant enough to be a consideration. However, it was unclear whether changing the spectrogram length and the inclusion/exclusion of the common words would impact the model performance. Hence, we built several models trained on all words and only the uncommon words to investigate this effect. Overall, six setups were configured and evaluated for this experimental set using audio length values of 5, 7, and 10 s, and the two different word selection schemes (uncommon words only or all words). The same three subjects from Table 3 were considered for model evaluation.

Similar to experiment 1, Keras Tuner was used to identify the best CNN architecture depicted in Fig. 2. The spectrograms' resolution was 128×157 pixels for experiments with 5-s audios,

128×219 pixels for 7-s audio experiments, and 128×313 pixels for 10-s audio experiments.

In each experiment, we retrained the model five times for 30 epochs with early stopping if the validation loss did not decrease in at least five epochs. Adam optimizer with an initial learning rate of 0.0003 and a batch size of 16 was used. Table 5 shows the mean classification accuracy of the models under each of the audio length and training word setups and the maximum classification accuracies achieved.

Based on the results obtained, the best performing setup was the 5-s spectrogram length using only the uncommon words that achieved the maximum of 81% accuracy on the three unseen speakers. Interestingly, the models trained only on the uncommon words outperformed all those trained on both the common and uncommon words. It appears that the addition of the common words may not be beneficial in an intelligibility classification setting. We assume that the common words may not contain enough discriminatory information for the model to learn intelligibility assessment. On the other hand, there may be more stark differences between speakers of different intelligibility classes when using only uncommon words, which may help the model generalize more successfully. Given that setup #1 delivered the best results, it can be concluded that 5-s audios uttering uncommon words are the optimal setup to conduct dysarthria intelligibility classification, considering generalizability to unseen speakers, when features are presented via spectrograms.

Table 4 Experiment set 1 results

MFCC configurations	Frame length	Sliding window	Coefficients	Average accuracy
Conf #1	25 ms	10 ms	13MFCCs	45.2%
Conf #2	64 ms	32 ms	13 MFCCs	52.5%
Conf #3	64 ms	16 ms	13 MFCCs	49.2%
Conf #4	128 ms	64 ms	13 MFCCs	57.8%
Conf #5	256 ms	128 ms	13 MFCCs	63.1%
Conf #6	256 ms	128 ms	13 MFCCs + delta MFCCs	56.4%
Conf #7	256 ms	128 ms	13 MFCCs + delta MFCCs + delta-delta MFCCs	51.7%

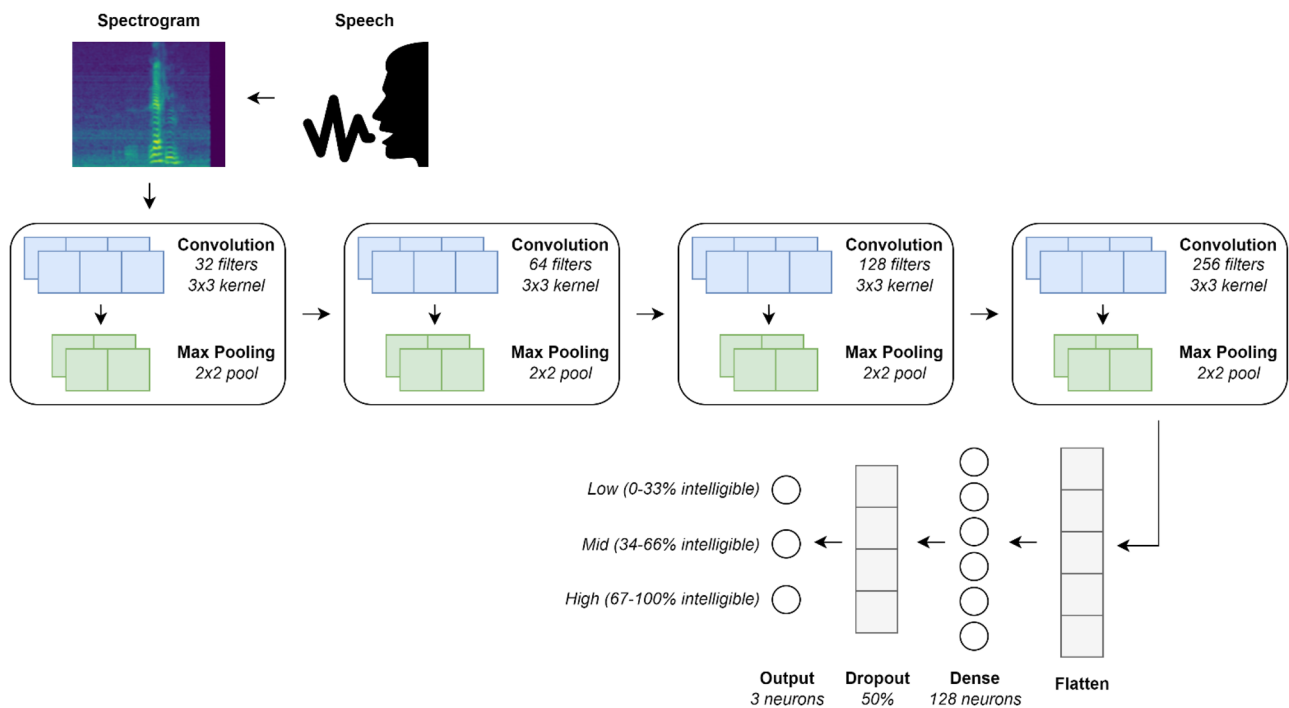


Fig. 2 Experiment 2 CNN architecture

As for the optimal length, the longer utterances appear to reduce performance, at least when using uncommon words only, which is consistent with previous findings reported in the literature [26]. The 5-s length performs the best on average and has the highest maximum accuracy. However, one additional consideration is that the shorter audio length reduces the number of utterances used in the training since some low intelligibility utterances may get excluded. Nonetheless, this can be remedied by trimming longer utterances to 5 s. However, this requires an automated audio segmentation technique in case the content of the audio files is skewed towards the beginning or end of the file.

Figure 3 shows the confusion matrix delivered by setup #1 model in terms of performance per intelligibility class. It can be noted that the misclassifications were not spread evenly among the three intelligibility classes. Particularly, the low intelligibility speaker was most frequently misclassified

(34%), in which the model considered his speech samples as high intelligibility 22% of the time. Moderate intelligibility was most accurately classified (90%), followed by high intelligibility with an 80% classification rate. Similar to the low intelligibility misclassification rate, the high intelligibility subject was misclassified more often as low intelligibility (15%) than moderate (4%).

Experiment Set 3: Verification of the Optimal Setup with All UA-Speech Dysarthric Speakers

Between experiment sets 1 and 2, the CNN trained with spectrograms of 5-s utterances with uncommon words delivered a 12.05% better average classification rate over the best MFCC configuration for M01, M05, and M09 dysarthric subjects, and, as such, was selected as the optimal setup. The next step was to confirm how this setup performs for

Table 5 Experiment set 2 results

Experiment setup	Utterance length	Word selection	Mean accuracy	Max accuracy
Setup #1	5 s	Uncommon only	75.15%	81.00%
Setup #2	5 s	All words	57.78%	64.28%
Setup #3	7 s	Uncommon only	73.92%	78.52%
Setup #4	7 s	All words	61.23%	63.60%
Setup #5	10 s	Uncommon only	70.39%	77.26%
Setup #6	10 s	All words	57.56%	61.43%

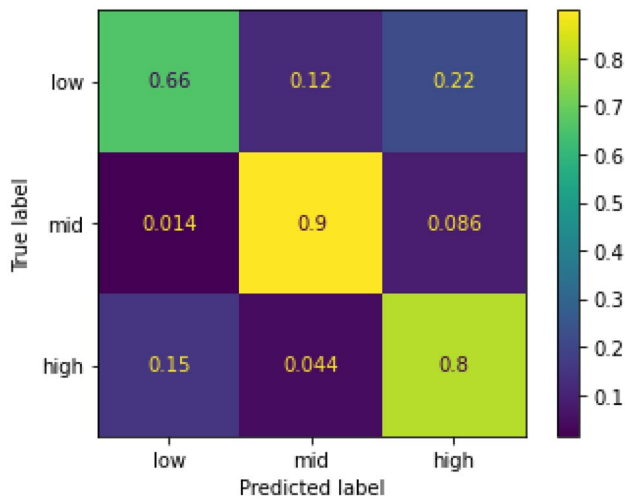
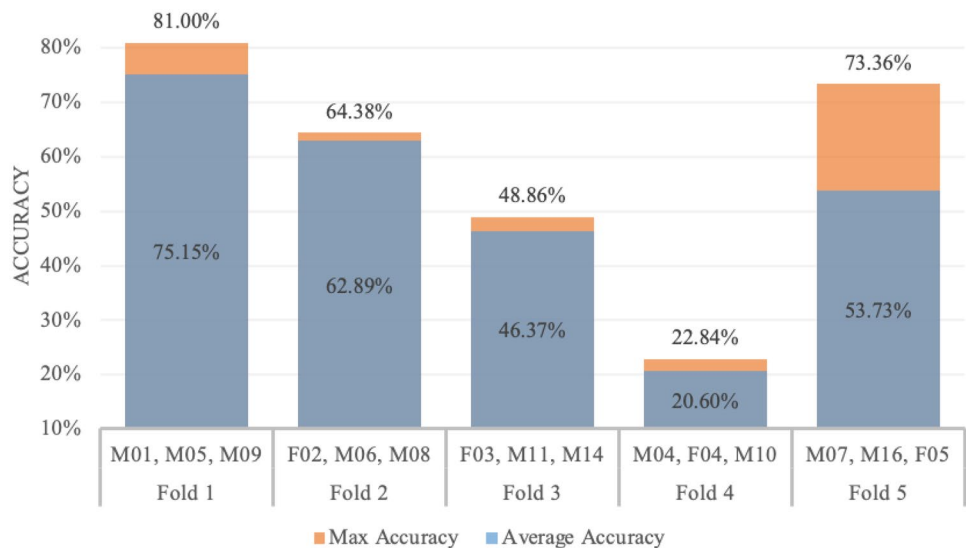


Fig. 3 Confusion matrix for spectrogram setup #1 (5-s audio length, uncommon words only) tested on speakers M01 (low), M05 (moderate), and M09 (high)

the remaining participants, which was done by conducting two more sets of experiments.

In the first set, 5-fold cross-validation was employed where different combinations of held-out speakers were used for evaluation purposes. In each fold, utterances of different speakers were entirely held-out during training. For example, we held out F02, M06, and M08 utterances in the second fold, trained the model with the rest of the speakers’ data, and then tested the model on the speakers mentioned above. In the next fold, we selected F03, M11, and M14 for testing and the rest of the speakers for training. Each fold contained testing speakers from all three dysarthric intelligibility classes and used the optimal 5-s utterances of only uncommon words during training and evaluation with the CNN shown in Fig. 2. Additionally, the CNN in each fold

Fig. 4 5-fold cross-validation results



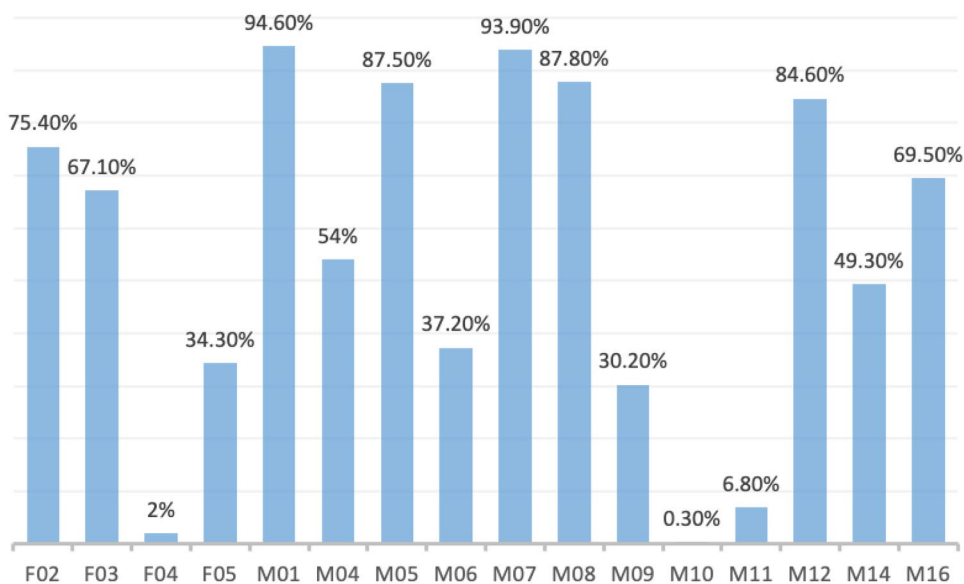
was retrained five times, as explained in the “[Experiment Set 1: Identifying the Best MFCC-Based Feature Set](#)” section. It is pertinent to note that the first fold results are from Table 5. Figure 4 depicts the results of these experiments.

In the second set of these experiments, we conducted leave-one-speaker-out approach [27] with 16 folds to verify the performance against each individual UA-Speech speaker. This was done by holding out a speaker in each fold for testing and training the CNN with the remaining 15 speakers based on the optimal setup. The process was repeated in the next fold with another speaker selected to held-out. These experiments were important to identify if particular speakers were skewing the performance. The same setup of 30 epochs with early stopping if validation loss does not decrease was used in each fold. The results of these experiments are illustrated in Fig. 5.

Discussion

Across different experiments we conducted, varying results were obtained. For example, each of the alternative sets of test speakers in the “[Experiment Set 3: Verification of the Optimal Setup with All UA-Speech Dysarthric Speakers](#)” section and Fig. 4 led to worse performance compared to the initial set of speakers in the first fold, which suggests the results from one speaker are not necessarily generalizable to the rest of the speakers of the same intelligibility class. Likewise, a large variation in results can be seen in the results shown by Fig. 5, ranging from as low as 0.3% classification accuracy up to 94.60%. This can also be seen from Fig. 6, where the results obtained from each speaker is plotted against the mean (56.66%) and standard deviation (32%). Only 56% of the observations are within one standard deviation above or below the mean, suggesting that the results do not follow normal distribution.

Fig. 5 16-fold leave-one speaker-out classification accuracy results



As shown in Fig. 6, speakers such as M01 and M07 were classified with over 90% accuracy, but few speakers such as F04, M10, and M11 with less than 10% classification accuracy, which indicates the model significantly misclassified the utterances from these speakers. Interestingly, M09 utterances were not classified well when trained on all other speakers, despite the 80% accuracy reported in Fig. 3. On the other hand, M01 and M05 were classified with more than 87% accuracy when evaluated individually.

In general, it appears that the low intelligibility speakers were identified more correctly on average than the other

intelligibility classes during the 16-fold experiments. Table 6 displays the average accuracy per intelligibility class based on the results obtained from the 16-fold experiments. We can see that the low intelligibility speakers appear to have the highest classification accuracy on average in this setting, even with the shorter 5-s audio length cut-off. The other two intelligibility classes seem more difficult to classify unseen speakers. However, the mid intelligibility speaker M05 and the high intelligibility speaker M08 had utterances classified more than 87% of the time correctly — there was not a universally poor classification of mid and high intelligibility

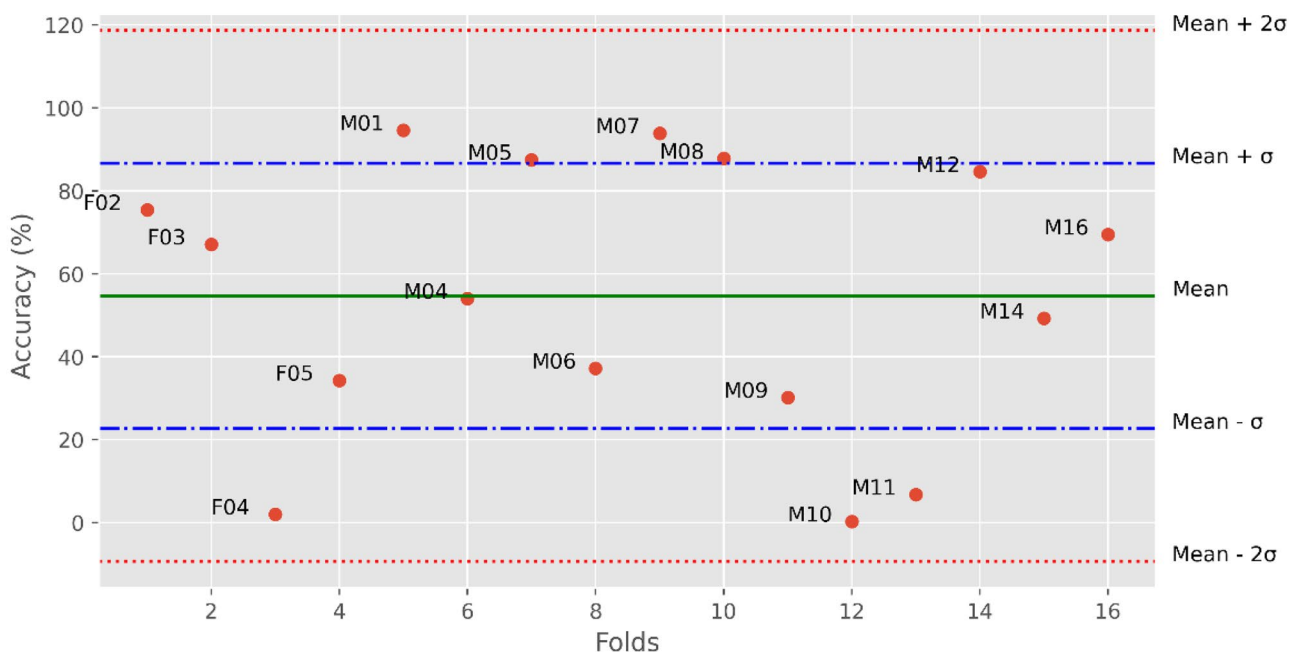


Fig. 6 16-fold leave-one-speaker-out results analysis

Table 6 Intelligibility class average results (16-fold experiments)

Intelligibility class	Speakers	Average accuracy
Low intelligibility (less than 33%)	F02, F03, M01, M04, M07, M12	78.2%
Moderate intelligibility (between 33 and 66%)	F04, M05, M06, M11, M16	40.6%
High intelligibility (more than 66%)	F05, M08, M09, M10, M14	40.4%

speakers during the 16-fold experiments. While the high degree of variation makes definitive conclusions difficult to draw, perhaps the low intelligibility speakers are more dissimilar from the other intelligibility classes and hence are easier to distinguish generally. Additionally, the mid and high intelligibility classes have only one female speaker each, explaining their suboptimal results. These results show that further investigation into the characteristics of speakers like M10 and M11 may be useful to explore difficulties in dysarthric intelligibility modeling not generalizing as well to unseen speakers.

Comparison with the State of the Art

Performance inconsistencies are obvious from the results obtained by evaluation on completely unseen speakers. We did not see the same level of performance achieved by the studies reported in Table 1, although for some speakers, the results are comparable with the state of the art. Nevertheless, because of the held-out strategy we adopted, a direct comparison with those studies that did not consider unseen speakers is not informative. To verify how our identified optimal setup compares with those reported in the literature, we have conducted a final set of experiments adopting a similar evaluation strategy considered in [19]. This would provide more confidence that the presence of unseen speakers is responsible for the variations in performance, not the optimal setup.

In [19], 90% of the utterances from four male speakers (M01, M05, M07, M09) and four female speakers (F02, F03, F04, F05) were used for training, and the remaining 10% of the same participants' utterances for testing. The data contained all 455 distinct words, including both common and uncommon words. In our comparative study, we decided to use all available speakers and the uncommon words in training and evaluation based on the proposed optimal setup. We also used an 80:20 train/test split across all speakers' data to increase the number of testing utterances. Additionally, the previous performances reported in the “[Experiments and Results](#)” section were obtained by perceiving the intelligibility classification as a three-class problem, but the baseline paper divided dysarthric intelligibility into four classes. As such, we conducted two comparative experiments based on both three-class and four-class tasks. We obtained 97.83% accuracy for the three-class and 97.19% for the four-class

task compared to the maximum 98.90% accuracy reported in [19].

Observing the results of the 3-class setting, it is much higher than the 81% accuracy obtained from unseen speakers. We believe this validates our concerns that using the training speakers for evaluation leaks speaker information, prompting the model to lean towards speaker recognition rather than intelligibility assessment. Additionally, this suggests that the variations in performance compared to the state-of-the-art results are due to the more difficult evaluation procedure and not because of a suboptimal setup.

Conclusions

In this paper, we reported on multiple sets of experiments we conducted to assess how well deep learning-based automatic dysarthric intelligibility assessment models generalize to new speakers, and to identify the optimal setup indicating which acoustic features perform better, different parameters such as utterance length, etc. Studying the literature identified that the state-of-the-art results may not properly reflect actual performances despite the very high accuracy reported, which was confirmed based on the results we obtained.

To find the optimal setup, we conducted 13 experiments investigating different MFCC configurations and spectrogram setups as input modalities to establish a high-performing dysarthria intelligibility assessment model when evaluated on speakers that the model was not exposed to. These experiments delivered a maximum of 81% classification accuracy on three unseen dysarthric speakers from the UA-Speech corpus, each presenting a different intelligibility class. The optimal setup was found to be a 5-s audio length and focusing on the uncommon words from the UA-Speech vocabulary using a CNN architecture we identified via hyperband tuning. While MFCCs are more commonly used in the literature, spectrograms outperformed them by delivering around 12% better classification accuracy in our experiments.

Once the optimal setup was identified, we proceeded with more experiments to investigate the setup's generalization performance on all available UA-Speech speakers via a 5-fold cross validation and 16-fold leave-one-out cross validation procedures; each employed a different strategy in dividing the train and test speakers. The results obtained from these

experiments indicated varying accuracies ranging from 0.30 to 94.60% for different speakers. No test speaker utterance was leaked in the training set in all of these experiments, ensuring test speakers remained completely unseen.

In order to measure how the optimal setup performs in comparison to the state of the art, we conducted two more comparative experiments adopting a similar strategy commonly used in the literature in which the test utterances were selected from all available speakers that their speech samples were provided to the model during training. The optimal setup achieved over 97% accuracy, which is as high as those reported in the literature. Nonetheless, this evaluation strategy does not indicate how well the model performs in clinical settings where new speakers need to be assessed.

The high variance in the results obtained from experiments with unforeseen speakers highlights the need for more investigations with more dysarthric subjects. Nevertheless, this is difficult in practice, especially given the communication challenges dysarthric speakers face and the fatigue they endure when speaking for longer periods. Another possibility is a cross-database evaluation to extend the amount of data available for training and testing — although the means to standardize and preprocess the speech across different databases may require significant effort and poses extra challenges. Another issue is the strategy applied to assess the subjects' speech intelligibility in different dysarthric corpora, which may not be consistent and result in further performance degradation and variance.

Another strategy that we are currently pursuing is to design a dysarthria assessment system that deploys the optimal setup to help SLPs conduct intelligibility assessments. The system will capture further speech data and have a feedback loop from SLPs when the system's prediction is incorrect to learn from the new data and SLPs' assessment.

Acknowledgements The authors would like to thank Dr. Catherine Watson for her help during the initial steps of this project.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Declarations

Ethics Approval This article does not contain any studies with human participants or animals performed by any of the authors. Publicly available datasets were used.

Conflict of Interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are

included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Kooi-Van Es M, et al. Dysphagia and dysarthria in children with neuromuscular diseases, a prevalence study. *J Neuromuscul Dis.* 2020;7(3). <https://doi.org/10.3233/JND-190436>.
2. Dyoniputri H, Afiahayati. A hybrid convolutional neural network and support vector machine for dysarthria speech classification. *Int J Innov Comput Inf Control.* 2021;17(1). <https://doi.org/10.24507/ijicic.17.01.111>.
3. Mendoza Ramos V, Pauly C, van den Steen L, Hernandez-Diaz Huici ME, de Bodt M, van Nuffelen G. Effect of boost articulation therapy (BArT) on intelligibility in adults with dysarthria. *Int J Lang Commun Disord.* 2021;56(2). <https://doi.org/10.1111/1460-6984.12595>.
4. Reithofer K. Intelligibility in English as a lingua franca - the interpreters' perspective. *J Engl Ling Fr.* 2020;9(2). <https://doi.org/10.1515/jelf-2020-2037>.
5. Hartelius L, Elmberg M, Holm R, Löfberg AS, Nikolaidis S. Living with dysarthria: evaluation of a self-report questionnaire. *Folia Phoniatr Logop.* 2008;60(1). <https://doi.org/10.1159/000111799>.
6. Tripathi A, Bhosale S, Koppurapu SK. Automatic speaker independent dysarthric speech intelligibility assessment system. *Comput Speech Lang.* 2021;69. <https://doi.org/10.1016/j.csl.2021.101213>.
7. Tong H, Sharifzadeh H, McLoughlin I. Automatic assessment of dysarthric severity level using audio-video cross-modal approach in deep learning. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH.* 2020;2020:4786–4790. <https://doi.org/10.21437/Interspeech.2020-1997>.
8. Schölderle T, Haas E, Ziegler W. Age norms for auditory-perceptual neurophonetic parameters: a prerequisite for the assessment of childhood dysarthria. *J Speech Lang Hear Res.* 2020;63(4). https://doi.org/10.1044/2020_JSLHR-19-00114.
9. Kent RD. Hearing and believing: some limits to the auditory-perceptual assessment of speech and voice disorders. *Am J Speech Lang Pathol.* 1996;5(3). <https://doi.org/10.1044/1058-0360.0503.07>.
10. Bhat C, Strik H. Automatic assessment of sentence-level dysarthria intelligibility using BLSTM. *IEEE J Select Topics Signal Process.* 2020;14(2). <https://doi.org/10.1109/JSTSP.2020.2967652>.
11. Huang A, Hall K, Watson C, Shahamiri SR. A review of automated intelligibility assessment for dysarthric speakers. 2021 11th International Conference on Speech Technology and Human-Computer Dialogue, *SpeD.* 2021. pp. 19–24. <https://doi.org/10.1109/SPED53181.2021.9587400>.
12. Tirumala SS, Shahamiri SR, Garhwal AS, Wang R. Speaker identification features extraction methods: a systematic review. *Expert Syst Appl.* 2017;90:250–71. <https://doi.org/10.1016/j.eswa.2017.08.015>.
13. Kim H, et al. Dysarthric speech database for universal access research. In: *INTER SPEECH 2008 - 9th Annual Conference of the International Speech Communication Association.* 2008. p. 1741–1744.
14. Falk TH, Chan WY, Shein F. Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective

- measurement of dysarthric word intelligibility. *Speech Commun.* 2012;54(5). <https://doi.org/10.1016/j.specom.2011.03.007>.
15. Bhat F, Vachhani B, Koppurapu SK. Automatic assessment of dysarthria severity level using audio descriptors. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings.* 2017. pp. 5070–5074. <https://doi.org/10.1109/ICASSP.2017.7953122>.
 16. Joshy AA, Rajan R. Automated dysarthria severity classification using deep learning frameworks. In: *European Signal Processing Conference.* 2021, vol. 2021-January. <https://doi.org/10.23919/Eusipco47968.2020.9287741>.
 17. Rudzicz F, Namasivayam AK, Wolff T. The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Lang Resour Eval.* 2012;46(4):523–41.
 18. Shahamiri SR. Speech vision: an end-to-end deep learning-based dysarthric automatic speech recognition system. *IEEE Trans Neural Syst Rehabil Eng.* 2021;29:852–61. <https://doi.org/10.1109/TNSRE.2021.3076778>.
 19. Gupta S, et al. Residual Neural Network precisely quantifies dysarthria severity-level based on short-duration speech segments. *Neural Netw.* 2021;139:105–17. <https://doi.org/10.1016/j.neunet.2021.02.008>.
 20. Martínez D, Green P, Christensen H. Dysarthria intelligibility assessment in a factor analysis total variability space. In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH.* 2013. pp. 2133–2137. <https://doi.org/10.21437/interspeech.2013-505>.
 21. Hu A, Phadnis D, Shahamiri SR. Generating synthetic dysarthric speech to overcome dysarthria acoustic data scarcity. *J Ambient Intell Humaniz Comput.* 2021. <https://doi.org/10.1007/s12652-021-03542-w>.
 22. O'Malley T, Bursztein E, Long J, Chollet F. KerasTuner. 2019.
 23. Li L, Jamieson K, DeSalvo G, Rostamizadeh A, Talwalkar A. Hyperband: a novel bandit-based approach to hyperparameter optimization. *J Mach Learn Res.* 2018;18.
 24. Shahamiri SR, Binti Salim SS. Artificial neural networks as speech recognisers for dysarthric speech: identifying the best-performing set of MFCC parameters and studying a speaker-independent approach. *Adv Eng Inf.* 2014;28(1). <https://doi.org/10.1016/j.aei.2014.01.001>.
 25. McFee B, et al. Librosa: audio and music signal analysis in Python. 2015. <https://doi.org/10.25080/majora-7b98e3ed-003>.
 26. Barreto SDS, Ortiz KZ. Speech intelligibility in dysarthrias: influence of utterance length. *Folia Phoniatr Logop.* 2020;72(3):202–10. <https://doi.org/10.1159/000497178>.
 27. Wong TT. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recogn.* 2015;48(9). <https://doi.org/10.1016/j.patcog.2015.03.009>.
- Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.