



Increasing the Precision of Dysarthric Speech Intelligibility and Severity Level Estimate

Mohammad Soleymanpour¹(✉), Michael T. Johnson¹, and Jeffrey Berry²

¹ University of Kentucky, Lexington, KY 40506, USA

{m.soleymanpour, mike.johnson}@uky.edu

² Marquette University, Milwaukee, WI 53201, USA

jeffrey.berry@marquette.edu

Abstract. Dysarthria is a speech disorder often characterized by slow speech with reduced intelligibility. Automated assessment of the severity-level and intelligibility of dysarthric speech can improve the efficiency and reliability of clinical assessment as well as benefit automatic speech recognition systems (ASR). However, in order to evaluate them, there are not sentence-level severity and intelligibility label. We only have access to speaker-per-level severity and intelligibility labels. This is a problem as dysarthric talkers might be able to produce some intelligible utterances due to frequent use and short utterances. Therefore, label based analysis might not be very accurate. To address this problem, we explore methods to estimate the severity-level and speech intelligibility in dysarthria given discrete speaker-level labeling in the training set. To accomplish this, we propose a machine learning based method using one-dimensional Convolutional Neural Networks (1-D CNN). The TORGO dataset is used to test the performance of the proposed method, with the UASpeech dataset used for Transfer learning (TL). To evaluate, an Averaged Ranking Score (ARS) and intelligibility probability distribution are used. Our findings demonstrate that the proposed method can assess speakers based on severity-level and intelligibility to provide a more granular analysis of factors underlying speech intelligibility deficits associated with dysarthria.

Keywords: Dysarthria · Intelligibility assessment · Dysarthric speech severity · Convolutional Neural Network · Transfer learning

1 Introduction

Dysarthria is motor speech disorder, often caused by traumatic injury or neurological disfunction, that decreases speech intelligibility through slow or uncoordinated control of speech production muscles [1]. People with moderate and severe levels of dysarthria may be less able to communicate with others through speech due to poor intelligibility [2].

Dysarthria severity-level is conventionally assessed clinically using subjective assessments of neuromuscular function during both speech and non-speech tasks. Standardized testing procedures, such as the Frenchay Dysarthria Assessment (FDA) [3] and

the Speech Intelligibility Test (SIT) [4], find common clinical use and prescribe methods for the auditory-perceptual assessment of speech intelligibility [5, 6]. These tests are often time-consuming to implement clinically and some approaches suffer from a lack of intra-rater reliability, due to the subjective nature of these tools [7]. Automated assessment of dysarthria severity-level and speech intelligibility could improve both the efficiency and reliability of clinical assessment. This has led researchers to investigate systems to automatically evaluate these dimensions in dysarthria.

Prior research has investigated automatic assessment of dysarthria severity level and speech intelligibility [8–10]. Automatic Speech Recognition based models have been applied to evaluate dysarthric speech intelligibility [10–12]. K. Gurugubelli et al. have proposed perceptually enhanced single frequency cepstral coefficients (PE-SFCC) as a new perceptually feature representation to assess dysarthric speech [13]. A non-linguistic method of dysarthria severity level has also been presented using audio descriptor, traditional musical-related features [14].

Since the suprasegmental characteristics such as pause occurrence, pause and phonemes duration, speaking rate and f0 decline and overall energy degradation vary across the dysarthric talkers with different degrees of severity and typical talkers, we aim to assess sentence-level dysarthria severity [15–21]. Sentence-level dysarthria severity has been done using Bidirectional Long Short-term Memory BLSTM (BLSTM), in which each sentence is classified into intelligible and non-intelligible groups [22]. Another research [23] has investigated using different DNN frameworks such as CNN and long short-term memory network (LSTM) with MFCC feature to classify dysarthria. In [24], sentence-level features are proposed to capture abnormal variation in the prosodic, voice quality and pronunciation aspects of pathological speech. A final intelligibility decision is made using feature-level fusions and subsystem fusion.

One of the problems in building automatic assessment models is the lack of severity-level and intelligibility labels for individual spoken utterances. Existing dysarthria datasets typically contain only severity-level and intelligibility labels per each speaker. This assumes that all sentences spoken by a speaker have the same degree of dysarthria. However, there is often a varying level of intelligibility in reality. This problem motivated us to use a regression approach to estimate a continuously-valued level of intelligibility.

In this work, we propose using a CNN-based model to automatically analyze dysarthria severity-level and speech intelligibility. Studies shows that one dimensional CNN would perform better over 2-D CNN with limited one-dimensional data [25]. The main dataset used here is TORGO, described in more detail in Sect. 3.1. The features used to represent speech are Mel Frequency Cepstral Coefficients (MFCCs) due to its potential to capture the global spectral envelope characteristics of speech and results of previous studies [23, 24]. Initially, we train the model with four groups of dysarthria severity levels. After this, the model is trained based on speech intelligibility labels. Unlike most of previous works, we use a regression approach to estimate a continuously-valued level of intelligibility rather than applying a simple classification structure. We believe that this approach will enable a more granular assessment of speech, which may be more informative to clinicians.

2 Methodology

We propose a new approach to automatically estimate dysarthria severity and speech intelligibility at a finer-grained level than that given by the dataset labels.

2.1 Model and Experiments

A one-dimensional CNN-based model is used in the proposed approach. Figure 1 shows the model applied for both tasks, containing three 1D-CNN layers, each followed by dropout and maxpooling layers. After the last convolutional layer, two fully connected layers are added for dysarthric severity-level analysis. However, only one fully connected layer is used in the intelligibility detection task. The convolutional layers attempt to capture the local characteristics, while the maxpooling layers reduce the dimensionality. Dropout is also used to avoid overfitting.

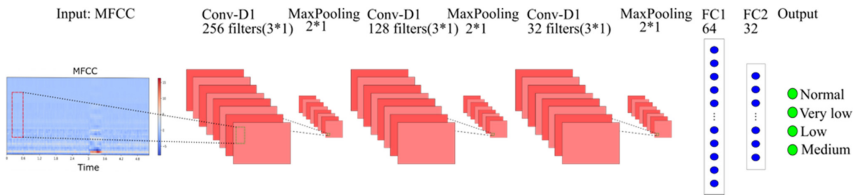


Fig. 1. Block diagram of the proposed architecture.

CNN-based models generally need a large amount of data to capture the varieties between groups. Transfer Learning (TL) is applied to reduce the effect of speaker variability and better learn the spectral features. In addition, since we are using a leave-one-speaker-out classification procedure, training is likely sensitive to the groups with a small number of individuals, in particular the group with only two people (one male and one female). To apply TL, the model is first trained on the UASpeech dataset and then the first three convolutional layers are saved when the model approaches optimal performance. These saved layers are used as initial layers to train the model on the TORGO dataset.

To evaluate, we used the Averaged Ranking Score (ARS) metric as an estimate of dysarthria severity for an individual utterance. For each sentence in the test set, four probabilities were generated to show the probability of the given sentence for each severity level. The final severity level was estimated as the weighted mean from these probabilities, using numeric values 1, 2, 3, and 4 for Normal, Very Low, Low, and Medium dysarthria levels. For example, if the model for a sentence generates the probability of 0.19, 0.15, 0.20, 0.46 for the four classes, respectively, the ranking score would be calculated as follows:

$$ARS = 1 \times 0.19 + 2 \times 0.15 + 3 \times 0.20 + 4 \times 0.46 = 2.93. \tag{1}$$

With this approach, an overall dysarthria severity-level can be obtained for each sentence in the range between 1 to 4. This can be interpreted on a continuous scale with 1 indicating normal and 4 indicating medium severity dysarthria. The average ranking score for each unseen speaker can then be computed across all utterances, allowing us to estimate both the average severity-level of that speaker and variance across utterances.

To estimate overall intelligibility on a per-speaker basis, the posterior probabilities from the intelligibility classifier for each of a speaker's utterances can be used to create a probability distribution for that speaker. The mean of the distribution can be used as an indicator the speaker's overall intelligibility, while the variance can provide information about the consistency of intelligibility.

3 Experimental Setup

We implemented three experiments to evaluate the effectiveness of the proposed method. In the first experiment, the dysarthria estimation model is trained based on four categories of dysarthric speech severity, including Normal, Very Low, Low, and Medium. Before training the model on TORGO, the model was trained on UASpeech.

In the second experiment, we excluded the normal category of speech and only used the dysarthric speech contained in TORGO. Because the categories of normal speech and very low dysarthric speech are quite similar, this allowed us to better distinguish the severity level of dysarthric speech in mild cases. The experimental setup and evaluation were the same as the previous method except for the number of classes. The categories of Very Low, Low, and Medium speech were used with the same ranking factors of 2, 3 and 4, respectively, as used in the first experiment.

The third experiment focused on estimating overall speaker intelligibility from the results of a binary intelligibility classification task. All speech was divided into two groups, intelligible and non-intelligible, which were used to train the model for binary speech intelligibility detection. This model was then used to generate the posterior intelligibility probabilities for individual utterances in the dataset, and the distribution of intelligibility probabilities across utterances from each speaker were used to assess the speaker's overall intelligibility profile.

For both dysarthria severity detection and speech intelligibility, the leave-one-speaker-out cross-validation procedure is applied. Before training, one speaker was kept out for test as unseen speaker and the remaining were used to train the model. 39 MFCC features were extracted for a window of 25 ms with 10 ms overlap. Utterances are zero-padded to the maximum length of training data. For training the model, all words and sentences were exploited whereas only sentences were used for testing. In addition, both words and sentences in UASpeech were used to train the initial TL model.

As described previously, three convolutional layers along with fully connected layers construct the main part of the model. The convolutional layers contain 256, 128 and 32 filters respectively with a kernel length of 3. Each of the convolutional layers is followed by a maxpooling of size 2×1 . The coefficient of the dropout layer is 20 percent. The number of neurons in the fully connected layers are 64 and 32, respectively, for the severity detection task and 32 for the one connected layer in the intelligibility task. The optimizer algorithm is Adam with a small learning rate of 0.0001. The number of outputs is four for the dysarthria severity detection and two for intelligibility detection.

3.1 Dataset

The main dataset used in this work is TORGO [26], containing 8 dysarthric speakers and 7 normal speakers. This dataset consists of non-word, short words, restricted and non-restricted sentences. Two types of microphones were used in this dataset, a head-mounted microphone as well as an array of 8 microphones placed approximately 61cm from each speaker. Dysarthric speakers are categorized into three dysarthria severity levels, Very Low, Low, and Medium and into two groups for intelligibility, intelligible and non-intelligible. The standardized Frenchay Dysarthria Assessment by a speech-language pathologist was applied to investigate the motor functions of each subject [26].

The UA-Speech dataset is used for Transfer Learning. This dataset includes speech recordings of 15 dysarthric speakers and control speakers. Each speaker was asked to read utterances containing 10 digits, 26 radio alphabet letters, computer commands, common words from the Brown corpus of written English, and uncommon words from children's novels selected to maximize phone-sequence diversity. All participants produced the same 765 words in citation form, 455 of them unique. Speech was recorded with an eight-channel microphone array at a sampling rate of 48 kHz, but in this experiment only one channel is used. Speakers are categorized in four groups of very low, low, middle and high by five native English listeners for each speaker [27].

4 Results and Discussion

The ARS results for each unseen speaker for the first and second experiments are shown in Table 1. For the first experiment each speaker ranged between 1 (normal) and 4 (medium severity).

Results for the dysarthria severity estimation indicate ARS severity rankings which were ordered in severity and mostly in the expected range. Although the ARS among the normal speakers was lower than those of the very low severity dysarthria group, this was by a small margin with most speakers in the normal category having an ARS close to 2 rather than 1, as might be anticipated. To see the difference between these two groups, Fig. 2 depicts a box plot of dysarthria severity levels. It can be observed that although the mean values of the ARS are similar between the normal and very low category speakers, there is a significant greater variance for the talkers with dysarthria across individual utterances, indicating that talkers with very low severity-level produced some utterances ranking as high as medium severity talkers.

Table 1. Averaged ranking score for the first experiment.

Severity Level	Intelligibility Category	Speaker ID	ARS	
			Exp. 1 4-levels	Exp. 2 3-levels
Normal	Intelligible	FC01	1.98	
		FC02	1.95	
		FC03	1.89	
		MC01	1.98	
		MC02	1.97	
		MC03	1.77	
		MC04	2.12	
Very low	Unintelligible	F03	2.26	2.63
		F04	2.52	2.46
		M03	2.21	2.60
Low	Unintelligible	F01	2.97	3.20
		M05	3.85	3.91
Medium	Unintelligible	M01	3.81	3.91
		M02	3.72	3.70
		M04	3.5	3.72

The last column in Table 1 shows the scores for the second experiment which estimated severity for only the dysarthric speech. The results for most speakers align with their labeled severity level; however, the M05 speaker is labeled as having a “low” severity level but the severity estimation for both of the experiments suggests a more severe level, on par with the “medium” speakers.

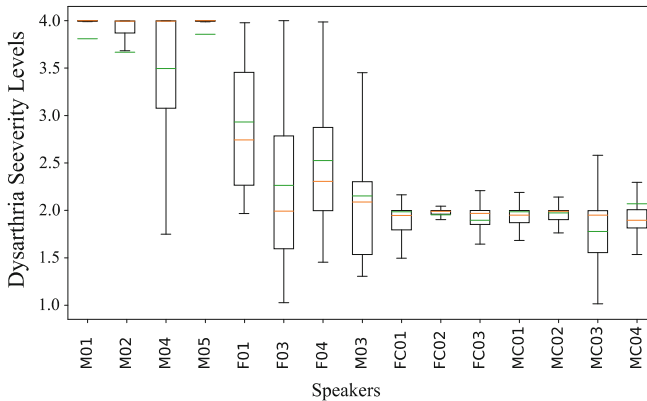


Fig. 2. Box plot of dysarthric severity level for the first experiment. The red line shows the median of the ranking scores, the dashed green line shows the mean (or averaged ranking score), and the box indicates the 25th to 75th percentile range. (Color figure online)

Figure 3 shows the results of the second experiment. This allows visualization of the relative severities as well as the variance within individual utterances. Comparing this to Fig. 2, excluding normal speech from training gives more precise severity estimates and less variation.

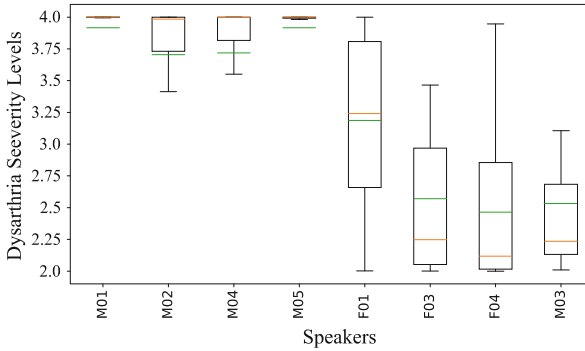


Fig. 3. Box plot of dysarthric severity level, with normal speech excluded. The red line shows the median of the ranking scores, the dashed green line shows the mean (or averaged ranking score), and the box indicates the 25th to 75th percentile range. (Color figure online)

In the third experiment, we analyzed the intelligibility probability distribution across individual utterances. Figure 4 shows the histogram of intelligibility probabilities calculated on individual utterances for select speakers with a bin-size of 0.05. The difference in the mean values of intelligibility is clear between the intelligible and unintelligible groups.

Moreover, with the “intelligible” speakers, “normal” talkers have almost no low-intelligibility utterances but “very low” severity speakers have numerous occurrences of such utterances. There are also notable differences in the distribution patterns across speakers. The extent of this variation suggests the possibility that an utterance-by-utterance assessment of intelligibility variance could be clinically useful, insofar as it could be used as a basis for a phonetic level characterization of the sound contrasts contributing to the intelligibility deficits [28].

To the best of our knowledge, this is the first work to continuously assess dysuria severity level and intelligibility, so there is not a direct way to compare these results with the findings of other works reported in classification metrics. For instance, Bhat et al. [22] have reported an average accuracy of 98.2 percent using BLSTM with transfer learning and balance data. Joshy et al. [23] reported the classification accuracy of 96.1 for TORGO dataset. As we mentioned in introduction section, existing dysarthria datasets like TORGO and UA-Speech contain only severity-level and intelligibility labels per each speaker, lacking severity-level and intelligibility labels for individual spoken utterances. This assumes that all sentences spoken by a speaker have the same degree of dysarthria which is not always correct in reality. Therefore, the classification metrics reported in these papers are based on this assumption.

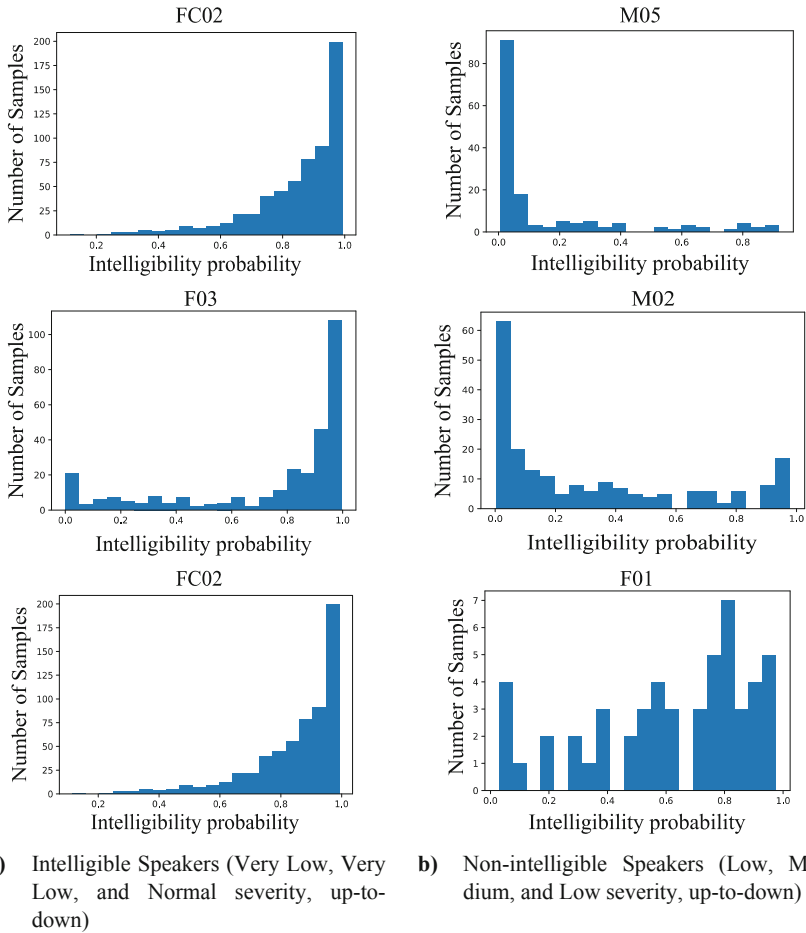


Fig. 4. The intelligibility probability histogram for each unseen speaker with a bin-size of 0.05.

5 Conclusion

This paper describes an automatic assessment of per-utterance dysarthria severity-level and speech intelligibility of individual speakers using a 1D-CNN-based model with Transfer learning. The models were trained with discrete dysarthria severity-level and speech intelligibility labels per speaker but used weighted probabilities of the discrete categories across individual utterances and speakers to estimate continuously-valued severity and intelligibility assessment metrics. Our findings demonstrate substantial variations across utterances and speakers for multiple dysarthria severity-levels and support the idea that this type of approach could be an effective tool to support objective clinical assessment of dysarthria.

Acknowledgments. This work was supported by National Institutes of Health under NIDCD R15 DC017296-01.

References

1. Duffy, J.R.: Motor speech disorders E-Book: Substrates, differential diagnosis, and management. Elsevier Health Sciences (2019)
2. Mitchell, C., et al.: Interventions for dysarthria due to stroke and other adult-acquired, non-progressive brain injury. *Cochrane Database Syst. Rev.* **1**, CD002088–CD002088 (2007)
3. Enderby, P.: Frenchay dysarthria assessment. *Br. J. Disord. Commun.* **15**(3), 165–173 (1980)
4. Dorsey, M., et al.: Speech intelligibility test for windows. Lincoln, NE: Institute for Rehabilitation Science and Engineering at Madonna Rehabilitation Hospital (2007)
5. Freed, D.: Motor speech disorders: diagnosis and treatment. Nelson Education (2011)
6. Hijikata, N., et al.: Assessment of dysarthria with Frenchay dysarthria assessment (FDA-2) in patients with Duchenne muscular dystrophy. *Disabil. Rehabil.*, 1–8 (2020)
7. Kent, R.D.: Hearing and believing: some limits to the auditory-perceptual assessment of speech and voice disorders. *Am. J. Speech Lang. Pathol.* **5**(3), 7–23 (1996)
8. Berisha, V., Utianski, R., Liss, J.: Towards a clinical tool for automatic intelligibility assessment. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2825–2828 (2013)
9. Kim, M.J., Kim, H.: Automatic assessment of dysarthric speech intelligibility based on selected phonetic quality features. In: *International Conference on Computers for Handicapped Persons*, pp. 447–450 (2012)
10. Hummel, R., Chan, W.-Y., Falk, T.H.: Spectral features for automatic blind intelligibility estimation of spastic dysarthric speech. In: *Twelfth Annual Conference of the International Speech Communication Association* (2011)
11. Ferrier, L., et al.: Dysarthric speakers' intelligibility and speech characteristics in relation to computer speech recognition. *Augment. Altern. Commun.* **11**(3), 165–175 (1995)
12. Martínez, D., et al.: Intelligibility assessment and speech recognizer word accuracy rate prediction for dysarthric speakers in a factor analysis subspace. *ACM Transactions on Accessible Computing (TACCESS)* **6**(3), 1–21 (2015)
13. Gurugubelli, K., Vuppala, A.K.: Perceptually enhanced single frequency filtering for dysarthric speech detection and intelligibility assessment. In: *International Conference on Acoustics, Speech and Signal Processing*, pp. 6410–6414 (2019)
14. Bhat, C., Vachhani, B., Kopparapu, S.K.: Automatic assessment of dysarthria severity level using audio descriptors. In: *International Conference on Acoustics, Speech and Signal Processing*, pp. 5070–5074 (2017)
15. Looze, C.D., et al.: Pitch declination and reset as a function of utterance duration in conversational speech data. In: *Sixteenth Annual Conference of the International Speech Communication Association* (2015)
16. Teodorescu, H.-N.: Pitch analysis of dysarthria helps differentiating between dysarthria mechanisms. *Bull. Integr. Psychiatry* **84**(1), 89–95 (2019)
17. Feenaughty, L., et al.: Speech and pause characteristics in multiple sclerosis: a preliminary study of speakers with high and low neuropsychological test performance. *Clin. Linguist. Phon.* **27**(2), 134–151 (2013)
18. Allison, K.M., Yunusova, Y., Green, J.R.: Shorter sentence length maximizes intelligibility and speech motor performance in persons with dysarthria due to amyotrophic lateral sclerosis. *Am. J. Speech Lang. Pathol.* **28**(1), 96–107 (2019)
19. Patel, R.: Prosodic control in severe dysarthria. *J. Speech Lang. Hear. Res.* **45**, 858–878 (2002)
20. Bunton, K., et al.: Perceptuo-acoustic assessment of prosodic impairment in dysarthria. *Clin. Linguist. Phon.* **14**(1), 13–24 (2000)

21. Bigi, B., et al.: A syllable-based analysis of speech temporal organization: a comparison between speaking styles in dysarthric and healthy populations. In: Sixteenth Annual Conference of the International Speech Communication Association, vol. 1, pp. 2977–2981 (2015)
22. Bhat, C., Strik, H.: Automatic assessment of sentence-level dysarthria intelligibility using BLSTM. *J. Sel. Top. Sign. Process.* **14**(2), 322–330 (2020)
23. Joshy, A.A., Rajan, R.: Automated dysarthria severity classification using deep learning frameworks. In: European Signal Processing Conference, pp. 116–120 (2021)
24. Kim, J., et al.: Automatic intelligibility classification of sentence-level pathological speech. *Comput. Speech Lang.* **29**(1), 132–144 (2015)
25. Kiranyaz, S., et al.: 1D convolutional neural networks and applications: a survey. In: *Mechanical Systems and Signal Processing*, vol. 151, p. 107398 (2021)
26. Rudzicz, F., Namasivayam, A.K., Wolff, T.: The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Lang. Resour. Eval.* **46**(4), 523–541 (2012)
27. Kim, H., et al.: Dysarthric speech database for universal access research. In: Ninth Annual Conference of the International Speech Communication Association (2008)
28. Kent, R.D., et al.: Toward phonetic intelligibility testing in dysarthria. *J. Speech Hear. Disord.* **54**(4), 482–499 (1989)