

Attribution-Based XAI for Dysarthria Assessment: Validation of Acoustic Feature Attributions

Minseo Kim
minseokim@sogang.ac.kr
Department of Artificial
Intelligence, Sogang
University
Seoul, South Korea

Junseok Oh
ohjs@sogang.ac.kr
Department of Computer
Science and Engineering,
Sogang University
Seoul, South Korea

Wonwoo Jeong
jeongwonwoo@sogang.ac.kr
Department of Computer
Science and Engineering,
Sogang University
Seoul, South Korea

Ji-Hwan Kim
kimjihwan@sogang.ac.kr
Department of Computer
Science and Engineering,
Sogang University
Seoul, South Korea

Abstract

Explainable Artificial Intelligence is crucial for transparent clinical decision-making. We propose a novel model that leverages multiple acoustic features from the Diadochokinesis task to assess the severity of dysarthria. The proposed model, integrating three distinct acoustic features, achieves 89.29% accuracy. SHapley Additive exPlanations (SHAP) and Integrated Gradients (IG) are used for feature attributions to identify key factors across severity classes. To further understand the model’s decision process, we categorize features by their contribution to severe-class predictions, capturing each feature’s influence on the model’s tendency to predict a severe outcome. The categorized features are validated against speech-language pathologists’ assessments, resulting in explanation accuracies of 74% and 79% for SHAP and IG, respectively. These findings highlight the potential of attribution-based methods to enhance interpretability and reliability in AI-driven diagnostics.

Keywords

Explainable Artificial Intelligence, Dysarthria Severity Assessment, Shapley Additive exPlanations, Integrated Gradient

1 Introduction

Most machine learning models are inherently complex and non-linear, which makes it difficult to interpret their internal decision-making mechanisms. This complexity has limited their use in critical sectors such as healthcare, finance, and legal systems, where clear interpretability is essential for building trust and making informed decisions [14, 20]. To address this challenge, explainable artificial intelligence (XAI) has emerged as a key research area focused on enhancing model interpretability [1]. By identifying key factors that influence predictions, XAI improves the reliability and practical applicability of these models [1].

XAI methodologies are generally categorized into three categories: model-based, example-based, and attribution-based explanations [14]. Model-based explanations simplify complex models into more interpretable forms. While this simplification enhances understanding, it may also reduce predictive performance [14]. Example-based explanations clarify predictions by presenting representative data instances. However, their effectiveness largely depends on the quality and distribution of the data [14]. Attribution-based methods, which assign importance scores to individual input features, quantify each feature’s contribution to a model’s output. By providing detailed insights into the factors that drive predictions, these methods improve interpretability while maintaining the model’s performance. A central challenge in XAI research remains ensuring

that these explanations accurately reflect the underlying decision-making processes and offer reliable insights to users.

In healthcare, where interpretability is critical, ensuring the reliability of XAI is crucial for adoption in clinical settings. The reliability of model explanations is assessed by comparing them with expert evaluations conducted by Speech-Language Pathologists (SLPs) in clinical decision-making. This study focuses on applying XAI techniques to dysarthria severity assessment and validating the reliability of attribution-based explanations through comparison with SLP evaluations.

Dysarthria is a motor speech disorder resulting from nervous system damage, significantly affecting communication and daily life [5]. Accurate severity assessment is essential for developing effective treatment strategies. Diadochokinesis (DDK), a widely used method for assessing dysarthria, involves the rapid and repetitive pronunciation of syllables such as /pa/, /ta/, /ka/, or /pataka/. It is used to evaluate articulatory speed and coordination, providing insights into phonation, prosody, and respiratory function [5, 6]. In this study, 12 characteristics extracted from DDK were utilized to develop a dysarthria severity assessment model.

Six key characteristics, strongly associated with articulatory speed and regularity, were selected for analysis. These characteristics were grouped into three categories based on attribution scores calculated using Shapley Additive Explanations (SHAP) [13] and Integrated Gradients (IG) [19]. Our study demonstrates the potential of XAI to enhance both the interpretability and reliability of predictive models for dysarthria severity assessment, validated through evaluations from SLPs.

2 Methods

Figure 1 illustrates an example of proposed method for evaluating attribution-based explanations in dysarthria severity assessment by comparing them with SLP evaluations. The dysarthria severity assessment model uses mel-spectrograms and speech signals combined with DDK characteristics as input. Feature attributions for individual DDK characteristics are computed using SHAP and IG based on the model’s predictions and are subsequently categorized into ‘Superior’, ‘Adequate’, and ‘Needs improvement’ groups. These categorizations are validated through comparisons with expert evaluations to ensure the accuracy and reliability of attribution-based explanations.

2.1 Dysarthria severity assessment model based on multi-acoustic feature fusion

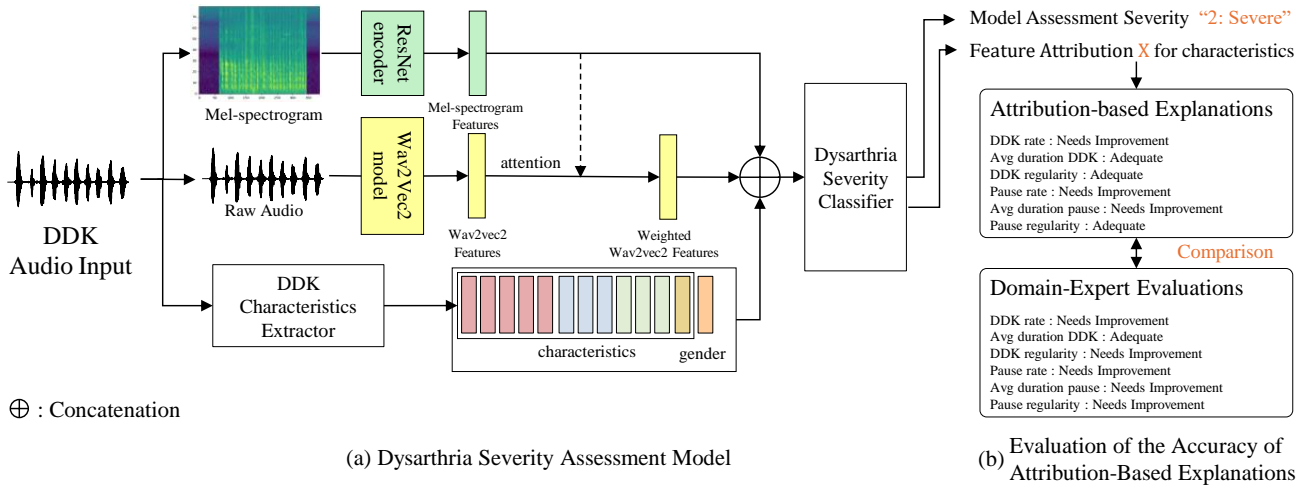


Figure 1: Overall structure of validating the reliability of attribution-based explanations in dysarthria severity assessment through comparisons with expert evaluations. (a) Dysarthria Severity Assessment Model that extracts and fuses multiple acoustic features to predict dysarthria severity. (b) An example illustrating how attribution-based explanations, categorized into *Superior*, *Adequate*, or *Needs improvement*, are evaluated for accuracy against expert assessments of six key dysarthria characteristics.

Table 1: Characteristics evaluated from DDK.

Subsystem	Characteristic	Definition
Phonation	F0 variability [semitones]	Variance of the fundamental frequency in semitones
	F0 variability [Hz]	Variance of the fundamental frequency in Hz
	Avg Energy [dB]	Average of energy
	Energy variability [dB]	Standard deviation of energy
	Max energy [dB]	Maximum value of energy
Prosody	DDK rate [1/s]	Number of syllables per second
	DDK average duration [ms]	Average duration of the syllables
	DDK regularity	Standard deviation of the syllables duration
Respiration	Pause rate [1/s]	Number of pauses per second
	Pause average duration [ms]	Average duration of the pauses
	Pause regularity	Standard deviation of the pauses duration
Intelligibility	Intelligibility	How well a listener can understand the content the speaker intends to deliver

2.1.1 Subsystem characteristics and definitions. Figure 1-(a) illustrates the architecture of the proposed dysarthria severity assessment model, which integrates multiple acoustic features using a joint representation learning approach [10]. Building on models used in emotion recognition [9, 21], our model combines DDK-derived characteristics, mel-spectrograms, and raw audio signals, capturing articulatory and time-frequency patterns for a comprehensive assessment of the severity of dysarthria.

Two types of DDK tasks, Alternating Motion Rate (AMR) and Sequential Motion Rate (SMR), are used to extract key characteristics relevant to dysarthria evaluation [5, 6]. In AMR, the same syllable (e.g., /pa/, /ta/, or /ka/) is rapidly and repetitively pronounced to assess articulatory speed, consistency, and respiratory and phonatory capabilities [5, 6]. In contrast, SMR evaluates the ability to transition quickly between distinct articulatory positions by repeating a sequence of syllables, such as /pataka/ [5, 6]. These tasks provide insights into the coordination, speed, and consistency of articulatory

movements, which are essential for accurately assessing dysarthria severity [5, 6].

Table 1 summarizes the 12 characteristics and their definitions. The characteristics are categorized into four subsystems: phonation, prosody, respiration, and intelligibility. These characteristics were derived from two primary sources. First, the Mayo Clinic rating system [5]—a widely recognized framework for dysarthria evaluation—provided a basis for feature extraction. Second, NeuroSpeech [16], a software tool for automated DDK analysis, was also used to derive these characteristics. The overall structure for characteristic extraction is depicted in Figure 2.

2.1.2 Extraction model architecture. Phonation subsystem characteristics were extracted using Praat software [3], while prosody and respiration characteristics were analyzed with an LSTM-based syllable segmentation model. Intelligibility was assessed using a ResNeXt-based model [15]. The LSTM-based model quantified the rate, duration, and regularity of pronunciation and respiration by

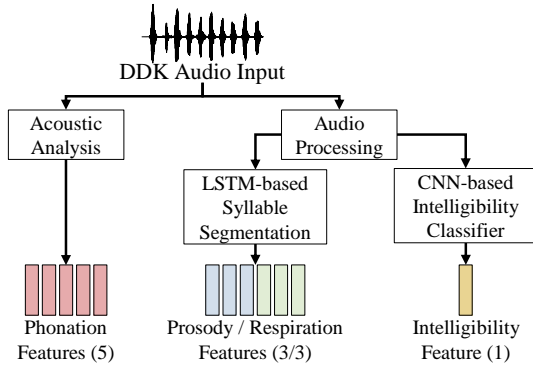


Figure 2: Overall structure for characteristic extraction in automated dysarthria severity assessment. The extracted characteristics are used as one of the inputs to the dysarthria severity assessment model. Parenthetical numbers (e.g., “5”) indicate the number of characteristics in each subsystem.

segmenting audio into speech and non-speech frames. The model consists of 16 LSTM layers and a fully connected (FC) layer. Raw audio signals were converted into spectrograms and fed into the model, which classified each frame as speech or non-speech. Frame-level predictions were aggregated into segment-level results by grouping consecutive frames with identical classifications. Speech segments shorter than 0.07 seconds were classified as silence, and silence segments longer than 0.14 seconds were used to calculate the pause rate. These threshold values (0.07 and 0.14 seconds) were determined based on the best performance observed on the training set. The silence threshold of 0.14 seconds was determined based on previous AMR task research, which found that healthy adults produce syllables at an average rate of approximately 0.143 seconds per syllable [18]. For both prosody and respiration subsystems, three key characteristics were computed: the number of speech and silence segments, as well as the average and standard deviation of segment durations. The 12 characteristics measured from the DDK task, along with gender information, were normalized using min-max scaling to mitigate scale discrepancies. The normalized features were subsequently processed through a FC layer to generate embedding vectors.

Features are extracted from the mel-spectrogram using a ResNet [8] model, capturing averaged characteristics across the frequency and time axes. These features are utilized as embedding vectors. The Wav2Vec 2.0 [2] model processes raw audio signals to generate frame-level representation vectors, which are subsequently used as embedding vectors for raw audio. Previous studies have demonstrated that combining mel-spectrogram and raw audio features enhances the ability to capture both temporal and frequency information. In this study, representation vectors derived from raw audio are fused with feature vectors extracted from the mel-spectrogram through an attention-based mechanism. The fused representations are utilized as input features for the dysarthria severity assessment model.

Three embedding vectors are concatenated into a single vector: a DDK feature vector, a mel-spectrogram vector, and a fused vector that combines the mel-spectrogram and raw audio representations.

This concatenated vector is forwarded to the final layer, which predicts the probabilities of dysarthria severity. A weighted categorical cross-entropy loss function [4] is employed during training to mitigate data imbalance.

2.2 Evaluation of the accuracy of attribution-based explanations

Figure 1-(b) illustrates an example of evaluating the accuracy of attribution-based explanations. The dysarthria severity assessment model utilizes multi-acoustic features as input, which creates challenges for the direct application of XAI algorithms. To address this, a separate model was developed by decoupling the embedding extraction process from the original trained model and isolating the fusion of mel-spectrogram and raw audio embedding vectors prior to their input into the final layer. The embedding vectors for the mel-spectrogram and raw audio were extracted independently from the original model and integrated as inputs into the separated model.

XAI algorithms were applied to the separated model to compute the contributions of 12 characteristics extracted from the DDK task, gender information, and the embedding vectors of the mel-spectrogram and raw audio in assessing dysarthria severity. This study employed SHAP and IG, two widely used algorithms for generating attribution-based explanations.

The dysarthria severity assessment model predicts probabilities for three severity classes: normal, mild-to-moderate, and severe. Regardless of the patient’s severity level, the feature attributions contributing to the severe class were utilized to categorize features into Superior, Adequate, and Needs improvement categories, enhancing the explainability of severity predictions.

Feature importance scores were calculated by scaling the attributions of each feature contributing to the severe class to a range of 0 to 100. These scores were derived from the average attributions for speech samples with normal severity and severe severity. The DDK task, which comprises three AMR tasks (/pa/, /ta/, /ka/) and one SMR task (/pataka/), was used to calculate feature scores separately for each task. The scores were then averaged to obtain the final score. Features with scores of 70 or above were classified as optimal, those with scores of 30 or below as impaired, and the rest as acceptable.

The accuracy of the explanations was validated by comparing the feature classifications generated by XAI with SLP evaluations. For the test dataset, three SLPs assessed six key characteristics of the DDK task—‘DDK rate’, ‘DDK average duration’, ‘DDK regularity’, ‘Pause rate’, ‘Pause average duration’ and ‘Pause regularity’—and categorized them. The agreement between the XAI-based classifications and SLP evaluations was used to assess accuracy. This validation highlights the reliability and practical applicability of XAI-based explanations in dysarthria severity assessment.

3 Experiments

3.1 Dataset

The dataset was collected from 59 healthy controls (HCs; mean age: 22 years) and 314 stroke patients with dysarthria (mean age: 56 years), all of whom were Korean. Each participant completed four tasks—three AMR tasks (repeating the syllables /pa/, /ta/, and

/ka/) and one SMR task (repeating the syllable sequence /pataka/). Recordings were captured using a smartphone at a native sampling frequency of 44.1 kHz and subsequently downsampled to 16 kHz.

The severity of dysarthria in patients ranged from either Mild-to-Moderate to Severe. There are 285 patients with a severity of Mild-to-Moderate and 29 patients with a severity of Severe. Speech intelligibility was rated on a scale from 1 to 5. Speech intelligibility were assessed by SLPs and the HC group was not assessed. The ratings were as follows: 1 for 8 patients, 2 for 14 patients, 3 for 23 patients, 4 for 134 patients, and 5 for 137 patients. Due to the inclusion of personal information of patients, the collected dataset will remain confidential.

3.2 Dysarthria severity assessment model based on multi-acoustic feature fusion

This subsection introduces a model that predicts dysarthria severity by fusing multiple acoustic features

3.2.1 Details. For the syllable segmentation model, speech data from dysarthric patients performing AMR tasks were partitioned into training, validation, and testing sets in an 8:1:1 ratio, stratified by severity levels. The model was trained using the AdamW [12] optimizer with a learning rate of 0.0003 for parameter optimization. To enhance the model’s robustness to noise, random noise was injected into the training data with signal-to-noise ratios of 15, 20, and 25.

For the speech intelligibility prediction model, the dataset was similarly partitioned into training, validation, and testing sets in an 8:1:1 ratio, stratified by speech intelligibility levels. This model was trained using the AdamW optimizer with a learning rate of 0.001. To improve generalization, data augmentation techniques, including SpecAugment [17] and speed perturbation, were applied.

For the dysarthria severity assessment model, the dataset was divided into training, validation, and testing sets following an 8:1:1 ratio, stratified by severity levels. The model was trained using the AdamW optimizer with a learning rate of 0.00003. To evaluate the performance of the proposed multi-acoustic feature fusion model, its results were benchmarked against those of severity assessment models based on LightGBM [11] and ResNet, using the same dataset.

3.2.2 Results. The syllable segmentation model achieved an AUC of 0.99. The correlation between the model’s predictions and the actual number of syllables was 0.94 for patients with mild dysarthria, 0.98 for patients with severe dysarthria, and 0.95 for all patients, indicating a strong correlation across severity levels. The speech intelligibility prediction model achieved a micro accuracy of 72.41% and a macro accuracy of 81.88%.

Table 2: Performance comparison of the proposed multi-acoustic feature fusion model against LightGBM-based and ResNet-based dysarthria severity assessment models.

Model	Micro Accuracy	Macro Accuracy
LightGBM-based	78.12%	82.31%
ResNet-based	83.04%	71.39%
Proposed	88.84%	90.19%

Table 2 summarizes the accuracy of dysarthria severity assessment models on the test set. The LightGBM-based model demonstrated a high macro accuracy, which reflects balanced performance across classes. However, its micro accuracy, representing the accuracy for individual severity levels, was relatively lower at 78.12%. On the other hand, the ResNet-based model achieved higher micro accuracy compared to macro accuracy, indicating strong overall performance. However, its accuracy tended to decline for the most severe dysarthria class. In contrast, the proposed multi-acoustic feature fusion model outperformed the other models, achieving the highest scores with a micro accuracy of 88.84% and a macro accuracy of 90.19%.

3.3 Evaluation of the accuracy of attribution-based explanations

3.3.1 Details. Three SLPs evaluated speech recordings from the test set, where participants performed four DDK tasks (/pa/, /ta/, /ka/, /pataka/). They assessed six key DDK characteristics, grouped into three categories. Inter-annotator reliability was measured using Fleiss’ kappa[7] and accuracy, yielding an average kappa of 0.495 and a mean agreement rate of 0.51 when all three SLPs provided identical evaluations. These findings suggest inconsistencies in evaluation criteria or variations in task interpretation among SLPs. Consequently, we evaluate the accuracy of attribution-based explanations only for cases where all three SLPs agreed in their evaluations.

3.3.2 Results. The table 3 summarizes the accuracy of attribution-based explanations compared to SLP evaluations in cases where all three SLPs provided consistent assessments. Attribution-based explanations were generated using SHAP and IG to compute feature attributions, which were subsequently used to classify six key characteristics into Superior, Adequate, and Needs improvement categories. To evaluate the reliability of the attribution-based explanations, their classification accuracy was compared against the accuracy obtained using only raw feature values. The SHAP-based explanations achieved an average accuracy of 0.74, while the IG-based explanations demonstrated a higher average accuracy of 0.79. Notably, both SHAP- and IG-based explanations outperformed the classification accuracy derived from raw feature values, with improvements of 0.06 and 0.11, respectively.

Table 3: Accuracy comparison between attribution-based explanations and SLP evaluations for six key characteristics. The attribution-based explanations were generated using three methods: characteristic values and feature attributions computed with XAI algorithms (SHAP, IG).

Key features	characteristic values	SHAP	IG
DDK rate	0.81	0.84	0.84
DDK average duration	0.68	0.86	0.76
DDK regularity	0.52	0.6	0.57
Pause rate	0.77	0.88	0.91
Pause average duration	0.56	0.46	0.84
Pause regularity	0.76	0.8	0.83
Average	0.68	0.74	0.79

4 Conclusion

This study presented a novel dysarthria severity assessment model that integrates multiple acoustic features from DDK tasks, achieving a micro accuracy of 88.84% and a macro accuracy of 90.19%. Leveraging attribution-based XAI methods, we categorized six key characteristics into *Superior*, *Adequate*, or *Needs Improvement* by computing their attributions toward the severe class. Compared with SLP evaluations, the SHAP-based explanations achieved an average accuracy of 0.74, while the IG-based explanations achieved 0.79, both exceeding the 0.68 accuracy derived from raw characteristic values alone. These findings highlight the clinical utility of attribution-based explanations in dysarthria assessment. Future research will extend this framework to other dysarthria tasks, such as reading tasks, and explore additional XAI paradigms, including model-based and example-based approaches, to further enhance the generalizability and interpretability of AI-driven clinical assessments.

Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2022-0-00621, RS-2022-II220621, Development of artificial intelligence technology that provides dialog based multi-modal explainability).

References

- [1] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bénézet, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, and Richard Benjamins. 2020. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities And Challenges Toward Responsible AI. *Information fusion* 58 (2020), 82–115.
- [2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*, Vol. 33. 12449–12460.
- [3] Paul Boersma and David Weenink. [n. d.]. Praat: Doing Phonetics by Computer [Computer Program]. <http://www.praat.org/> Version 6.4.27, retrieved 27 January 2025.
- [4] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-Balanced Loss Based on Effective Number of Samples. In *Proc. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9260–9269. doi:10.1109/CVPR.2019.00949
- [5] Frederic L. Darley, Arnold E. Aronson, and Joe R. Brown. 1969. Differential Diagnostic Patterns of Dysarthria. *Journal of Speech and Hearing Research* 12, 2 (June 1969), 246–269. doi:10.1044/jshr.1202.246
- [6] Joseph R. Duffy. 2012. *Motor Speech Disorders: Substrates, Differential Diagnosis, and Management*. Elsevier Health Sciences.
- [7] Joseph L. Fleiss. 1971. Measuring Nominal Scale Agreement among Many Raters. *Psychological Bulletin* 76, 5 (Nov. 1971), 378–382. doi:10.1037/h0031619
- [8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition*. 770–778. doi:10.1109/CVPR.2016.90
- [9] Yurun He, Nobuaki Minematsu, and Daisuke Saito. 2023. Multiple Acoustic Features Speech Emotion Recognition Using Cross-Attention Transformer. In *Proc. 2023 IEEE International Conference on Acoustics, Speech and Signal Processing*. 1–5. doi:10.1109/ICASSP49357.2023.10095777
- [10] Shih-Cheng Huang, Anuj Pareek, Saeed Seyyedi, Imon Banerjee, and Matthew P. Lungren. 2020. Fusion of Medical Imaging and Electronic Health Records Using Deep Learning: A Systematic Review and Implementation Guidelines. *npj Digital Medicine* 3, 1 (Oct. 2020), 1–9. doi:10.1038/s41746-020-00341-z
- [11] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Proc. the 31st International Conference on Neural Information Processing Systems*. 3149–3157.
- [12] Ilya Loshchilov and Frank Hutter. 2018. Decoupled Weight Decay Regularization. In *Proc. International Conference on Learning Representations*.
- [13] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, Vol. 30.
- [14] Aniek F Markus, Jan A Kors, and Peter R Rijnbeek. 2021. The Role of Explainability in Creating Trustworthy Artificial Intelligence for Health Care: a Comprehensive Survey of The Terminology, Design Choices, And Evaluation Strategies. *Journal of Biomedical Informatics* 113 (2021), 103655.
- [15] J. Oh, H. Park, and J. Kim. 2023. Speech Intelligibility Prediction of Dysarthri Using Deep Convolutional Networks. In *Proc. 18th Asia Pacific International Conference on Information Science and Technology*.
- [16] Juan Rafael Orozco-Arroyave, Juan Camilo Vásquez-Correa, Jesús Francisco Vargas-Bonilla, R. Arora, N. Dehak, P. S. Nidadavolu, H. Christensen, F. Rudzicz, M. Yancheva, H. Chinaei, A. Vann, N. Vogler, T. Bocklet, M. Cernak, J. Hannink, and Elmar Nöth. 2018. NeuroSpeech: An Open-Source Software for Parkinson’s Speech Analysis. *Digital Signal Processing* 77 (June 2018), 207–221. doi:10.1016/j.dsp.2017.07.004
- [17] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Proc. Interspeech 2019*. 2613–2617. doi:10.21437/Interspeech.2019-2680
- [18] Kaitlin Schuessler. 2010. *Performance of Alternating Motion Rate (AMR) in Individuals With Parkinson’s Disease Under External And Internal Cueing Conditions*. Master’s thesis. University of Colorado at Boulder.
- [19] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. In *Proc. 34th International Conference on Machine Learning*. 3319–3328.
- [20] Erico Tjoa and Cuntai Guan. 2020. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE transactions on neural networks and learning systems* 32, 11 (2020), 4793–4813.
- [21] Heqing Zou, Yuke Si, Chen Chen, Deepu Rajan, and Eng Siong Chng. 2022. Speech Emotion Recognition with Co-Attention Based Multi-Level Acoustic Information. In *Proc. 2022 IEEE International Conference on Acoustics, Speech and Signal Processing*. 7367–7371. doi:10.1109/ICASSP43922.2022.9747095