

Fusing Heterogeneous Speech Tasks for Automated Dysarthria Severity Classification

Jaeyeong Lee
lee19492@sogang.ac.kr
Sogang University
Seoul, Korea

Hyung-Min Park
hpark@sogang.ac.kr
Sogang University
Seoul, Korea

Abstract

We present an automated system for assessing dysarthria severity to reduce reliance on time- and resource-intensive clinical ratings. From Maximum Phonation Time (MPT) and paragraph-reading speech, the model uses Mel spectrograms, raw waveforms, and acoustic features as input. Task-specific representations are fused via an attention module, and a final ensemble predicts three severity levels. On a self-collected Korean dataset, the ensemble reaches 72.44% macro-accuracy, outperforming MPT-only (46.49%) and paragraph-only (67.64%) baselines. These results show that fusing heterogeneous speech tasks provides complementary information for dysarthria severity classification.

CCS Concepts

• **Social and professional topics** → **People with disabilities**; • **Applied computing** → *Health care information systems*.

Keywords

Dysarthria, classification, fusion, attention, ensemble

1 Introduction

Dysarthria is a motor speech disorder caused by abnormalities in the brain and peripheral nerves due to conditions such as neurodegenerative diseases like Parkinson’s disease or stroke. It results in paralysis, weakness, or incoordination of speech muscles during various production processes, including respiration, phonation, resonance, articulation, and prosody [2, 4]. Currently, dysarthria diagnosis involves neurologists and speech-language pathologists assessing severity levels by evaluating the patient’s pronunciation during tasks such as Maximum Phonation Time (MPT) [1], diadochokinetic (DDK) [2, 4], word reading, and paragraph reading [10].

However, this diagnostic method relies on the individual experience and subjectivity of the evaluator and has limitations in requiring significant time and labor for diagnosis. Therefore, there is a need for an automated assessment system for the severity of dysarthria to provide unbiased diagnostic results to patients more quickly and to enable continuous monitoring of the patient’s condition and effective treatment.

Recently, numerous studies have utilized deep learning for classifying dysarthria severity. Previous research has focused on enhancing classification performance by utilizing entire datasets or comparing model performance by dividing data according to specific criteria [9, 21]. However, research on the potential synergistic effects when combining different speech tasks has been relatively scarce.

This paper proposes an automated severity assessment system that jointly utilizes MPT and paragraph-reading speech data from both healthy individuals and patients with dysarthria. Specifically, we aim to measure the performance of models processing each speech task individually and validate synergistic effects through an ensemble model that fuses features extracted from both tasks.

2 Method

2.1 Dataset

This study utilized a self-collected Korean dysarthria dataset. Speech data was recorded in a quiet environment (below 50 dB) using either a smartphone recording app or a dedicated application (repeech)[5], with the microphone placed about 30cm from the mouth. A total of 315 participants performed both MPT and paragraph-reading tasks. Each speaker was labeled as Healthy, Mild-to-Moderate, and Severe by a neurosurgeon based on National Institute of Health Stroke Scale (NIHSS) [13].

The MPT data for the vowels /a/, /i/ and /u/ were collected by instructing the patient to sustain each vowel as much as possible[19].

Paragraph reading data were collected by recording the patients as they read the passage ‘Gaeul’ (Autumn), a standardized Korean paragraph reading[8]. This 369-syllable passage is commonly used for diagnosing communication disorders and dysarthria as it includes all Korean consonants and vowels based on their frequency of occurrence. While the paragraph-reading speech data for training was divided into six segments, the test set uses raw, unsegmented data. The final dataset used for analysis comprised 2,553 audio files, totaling 8 hours and 51 minutes in length.

The entire dataset was split by speaker at an 8:1:1 ratio (train: validation: test). Since the number of patients in the severe class is small, splitting at same ratio would result in insufficient data in the validation and test sets makes it difficult to guarantee statistical reliability for the model’s classification performance. Therefore, we aimed to maintain the 80:10:10 ratio for the total dataset volume while ensuring speakers belonging to each class were evenly distributed across the sets to secure statistical reliability. Table 1 shows the final number of speakers and data points in each split set.

Audio data were recorded at 44.1kHz sampling rate, and down-sampled to 16kHz for model input.

2.2 Feature extraction

Guided by the Mayo Clinic classification system[3] for dysarthria, we extracted a set of acoustic features covering phonation, articulation, and prosody. For the MPT audio files, 19 acoustic features were extracted using Praat software [15]. Gender and age information was added to account for potential variation in feature distribution, resulting in a total of 21 features used.

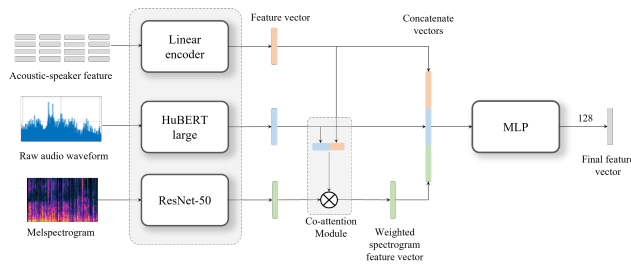
Table 1: Training Set Division

| Sets | Speakers | # of samples | hours | Severity | | | Gender | | Age | | | | | | | |
|-------|----------|--------------|--------|----------|----------|--------|--------|--------|-----|-----|-----|-----|-----|-----|-----|--|
| | | | | Healthy | Mild-Mod | Severe | Male | Female | 20s | 30s | 40s | 50s | 60s | 70s | 80s | |
| Train | 216 | 1928 | 5h 59m | 136 | 57 | 23 | 110 | 106 | 39 | 5 | 7 | 50 | 63 | 50 | 2 | |
| Valid | 46 | 413 | 1h 21m | 20 | 18 | 8 | 22 | 24 | 4 | 4 | 6 | 10 | 10 | 10 | 2 | |
| Test | 53 | 212 | 1h 30m | 22 | 20 | 11 | 27 | 26 | 7 | 6 | 6 | 10 | 10 | 10 | 4 | |
| Total | 315 | 2553 | 8h 51m | 178 | 95 | 42 | 159 | 156 | 50 | 15 | 19 | 70 | 83 | 70 | 8 | |

For paragraph-reading audio files, in addition to the 21 acoustic features and gender/age information, Character Error Rate (CER) value extracted using the Whisper large-v2 [16] model was included to measure pronunciation clarity, resulting in a total of 24 features. Table 2 summarizes the acoustic features extracted from MPT and paragraph reading task.

2.3 Model architecture

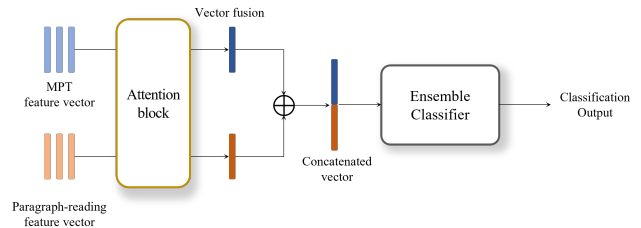
The severity classification system proposed in this study consists of a multi-input model for feature extraction and an ensemble model that fuses the extracted features to perform the final classification.

**Figure 1: Architecture of Multi-Input feature extractor.**

2.3.1 Multi-input feature extractor. Figure 1 shows the architecture of Multi-Input feature extractor model. The Multi-Input model is a multi-modal deep learning architecture that takes three types of information from the audio signal as input: Mel spectrogram, raw waveform, and acoustic features. Each input is processed through separate sub-networks based on ResNet[6], HuBERT[7], and Multi-Layer Perceptrons (MLP), and the extracted features are hierarchically combined.

In this study, this multi-input model is utilized not as a final classifier but as a feature extractor that represents the information of each speech task as a compressed vector. To achieve this, the 128-dimensional vector immediately preceding the final classification is used as the feature vector for that input.

2.3.2 Attention-based ensemble classifier. Figure 2 shows the architecture of Attention-based ensemble classifier. The multi-input model generates a 128-dimensional feature vector from each audio file. Tasks composed of multiple audio sources, such as MPT, are combined into a single representative vector via an attention mechanism. For the paragraph reading task, vectors extracted from each segmented sentence are also reconstructed into a single representative vector via attention. Finally, a 128-dimensional vectors

**Figure 2: Architecture of Ensemble classification model.**

representing the MPT task and the paragraph reading task are concatenated to generate a 256-dimensional fusion vector. This vector is input into an Ensemble classifier with an MLP structure, which outputs the final logit values for the three severity classes. Actual training occurs in the attention layer and the Ensemble classifier.

2.3.3 Co-Attention applied ensemble classifier. The ensemble classifier model uses simple concatenation for fusion. This static approach combines the 128-dim MPT vector and the 128-dim Paragraph Reading vector but has limitations in modeling their complex inter-modal dependencies. To address this, we replace concatenation with a co-attention mechanism [20].

Calculate an affinity matrix, allowing the representation of one task to be weighted by the information from the other. This process generates a final, context-aware fused vector which is then input into the MLP classifier. We hypothesize that this adaptive fusion improves performance by prioritizing relevant information across tasks, which we validate in the Results section.

3 Experiment

3.1 Basic Setting

The training process for the multi-input feature extraction model is as follows.

The MPT model utilized 21 acoustic features, and the speaker's recorded audio was segmented into clips up to 25 seconds for input to the HuBERT model (average total phonation time: 12.44 seconds; zero-padded when under 25 seconds). Training was conducted for 200 epochs with a learning rate of $1e-3$. The optimizer used was Adam[12], and the loss function was Cross-Entropy loss[14].

For the paragraph reading model, recordings from single speaker were segmented by sentence. Each sentence was then cut into segments of up to 15 seconds in length for model input (zero-padded if shorter than 15 seconds). Training was conducted for 200 epochs with a learning rate of $1e-4$ and a batch size of 16.

Table 2: Speech acoustic features extracted for MPT and paragraph reading task

| Tasks | Speech subsystem | Feature | Explanation for Each Feature |
|----------------------|--|----------------------------|--|
| Paragraph Reading | Articulation | Avg CER | Average Character Error Rate (CER) calculated across sentences in a paragraph |
| | Prosody | Speaking rate | Rate of spoken syllables per second, including pauses |
| | | Articulation rate | Rate of spoken syllables per second, excluding pauses |
| | | Avg syllable duration | Average duration of syllables in the speech signal |
| | | Phonation ratio | Ratio of speaking time to total time |
| | | Pause frequency ratio | Ratio of the number of pauses to the total speech duration |
| Pause duration ratio | Ratio of the number of pauses to the total speech duration | | |
| MPT | Phonation | Tilt | Spectral slope of the speech signal |
| | Breath control | Duration | Total duration of actual voice vocalization |
| Both Included | Phonation | Pitch(mean, std, min, max) | Fundamental frequency(F0) and its standard deviation, minimum, and maximum value |
| | | Intensity (mean, std, max) | Speech signal’s intensity and its standard deviation and maximum value |
| | | Jitter (mean, DDP, PPQ) | Measures of pitch variation; DDP (Difference of Differences of Periods) and PPQ (Pitch Period Perturbation Quotient) |
| | Articulation | Shimmer (mean, APQ) | Measures of amplitude variation; APQ (Amplitude Perturbation Quotient) |
| | | Harmonics-to-Noise Ratio | Ratio of harmonic sound to noise in the voice |
| | Speaker metadata | Age | Age of speaker |
| | Gender | Gender of speaker | |

The ensemble model was trained using the feature vectors extracted for each speaker as input. The train/valid/test sets were applied identically here as well. Training was conducted for 200 epochs with a learning rate of $5e-6$ and a batch size of 1.

3.2 LIME Analysis

To understand which input representations (raw audio, Mel spectrograms, or acoustic features) most influence our composite ensemble model’s predictions, we employ the LIME framework [17]. We selected LIME because its model-agnostic nature is essential for reliably analyzing a complex, multi-input "black-box" model like ours.

4 Results

4.1 Classification Result

The classification performances of each model are summarized in Table 3. The standalone model using only MPT data showed poor performance with a macro-accuracy of 46.49%. A standalone model using only paragraph reading data achieved a macro accuracy of 67.64%, demonstrating a higher classification accuracy than the MPT model, although this performance level has limitations for real-world applications.

In contrast, the proposed ensemble model in this study, which fuses feature vectors from both tasks for classification, achieved

Table 3: Classification Performance of Each Model

| Tasks | Macro Acc. | Micro Acc. |
|------------------------|---------------|---------------|
| MPT Only | 46.49% | 56.8% |
| paragraph Reading Only | 67.64% | 69.08% |
| Ensemble-Concatenation | 70.96% | 75.47% |
| Ensemble-Co-Attention | 72.44% | 75.47% |

a macro accuracy of 70.96%, showing a significant performance improvement compared to each standalone model. Moreover, Co-Attention applied ensemble model achieved highest macro accuracy of 72.44%. This Result shows that the co-attention method has better information representation than concatenation, which confirms our hypothesis. The class-specific classification performance of the proposed ensemble model is shown in Table 4 and the confusion matrix, Figure 3.

When measuring the Precision and F1-score for the best model, the 'Severe' class achieved the highest precision and 'Healthy' class achieved the highest F1-score. The class with the worst classification performance is Mild-to-Moderate, and when compared with the confusion matrix at the same time, An analysis of the confusion matrix indicates it is difficult to classify with the Healthy class. This trend is not only between the Healthy and Mild-to-Moderate

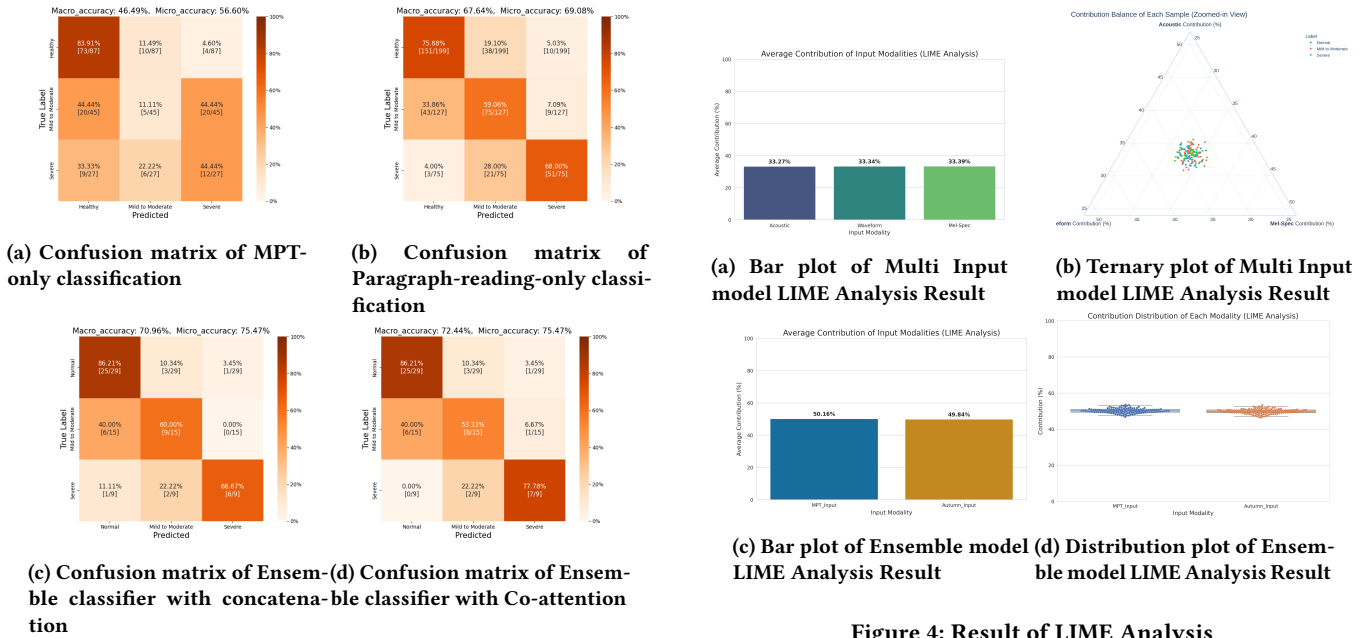


Figure 4: Result of LIME Analysis

Figure 3: Comparison between MPT only, paragraph-reading only, MPT&Paragraph-reading Ensemble.

Table 4: Detailed Ensemble Classifier Result

| | Precision | F1-score |
|------------------|--------------|-------------|
| Healthy | 76.47% | 0.83 |
| Mild to Moderate | 63.64% | 0.53 |
| Severe | 87.5% | 0.82 |
| Macro Accuracy | 72.44% | 0.73 |
| Micro Accuracy | 75.47% | 0.74 |

class, but also between the Mild-to-Moderate and Severe class. This suggests that the acoustic boundaries between neighboring classes may not be sufficiently distinct.

4.2 LIME Analysis Result

As illustrated in 4a, the LIME analysis of multi-input model reveals a highly uniform distribution of average contributions from the three input modalities: Acoustic (33.27%), Waveform (33.34%), and Mel-Spec (33.39%). The ternary plot in 4b confirms that this trend is not merely an artifact of averaging, as all individual samples are densely clustered near the center of the plot, irrespective of their true severity labels.

A similar trend is observed in the LIME analysis of the Ensemble Classifier model. As shown in 4c, the inputs that influence the final classification result are MPT at 50.16% and paragraph reading at 49.84%, indicating a negligible difference. 4d further visualizes this distribution, revealing that the two distributions are highly similar.

This indicates that the model is not biased toward any single modality, but instead leverages all information sources in a balanced manner to make its classifications. This finding provides strong

validation for the multi-modal approach proposed in this study. Had the model learned a "shortcut" (i.e., relying on Acoustic features mostly), the contributions would have been significantly distorted.

Instead, the results demonstrate that the multi-input-based mechanism effectively integrates all three sources of information and utilizes them equitably in making the final decision. In essence, this shows that the model is operating in a truly hybrid fashion, as intended by its design.

5 Conclusion

This study demonstrated that fusing features from different speech tasks, namely MPT and paragraph reading, significantly improves the accuracy of automated dysarthria severity classification compared to using either task alone.

However, analysis of the proposed model's confusion matrix revealed challenges in distinguishing between adjacent severity classes. Specifically, a notable number of Mild-to-Moderate cases were misclassified as Healthy, and some Severe cases were misclassified as Mild to Moderate. This indicates that the acoustic boundary between two severity classes is ambiguous, implying the potential need for alternative approaches or additional data to clearly distinguish them.

Based on this study, we propose the following follow-up research. First, it is necessary to evaluate the performance of an extended model utilizing all four tasks: MPT, paragraph reading, word reading, and DDK data. Second, cross-validating the generalization performance of the proposed model using public datasets like TORGO[18] and UAspeech[11] is necessary. Finally, research is required to enhance the feature expressiveness of the MPT task, which showed relatively low performance, thereby improving the overall model's performance.

6 Acknowledgement

This work was supported by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2022-0-00621, RS-2022- IIT20621, Development of artificial intelligence technology that provides dialog-based multi-modal explainability).

References

- [1] Ben Barsties v. Latoszek, Christopher R. Watts, Katharina Schwan, and Svetlana Hetjens. 2023. The maximum phonation time as marker for voice treatment efficacy: A network meta-analysis. *Clinical Otolaryngology* 48, 2 (2023), 130–138. arXiv:<https://onlinelibrary.wiley.com/doi/pdf/10.1111/coa.14019> doi:10.1111/coa.14019
- [2] 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
- [3] Frederic L. Darley, Arnold E. Aronson, and Joe R. Brown. 1969. Differential diagnostic patterns of dysarthria. *Journal of Speech and Hearing Research* 12, 2 (1969), 246–269. arXiv:<https://pubs.asha.org/doi/pdf/10.1044/jshr.1202.246> doi:10.1044/jshr.1202.246
- [4] Joseph R Duffy. 2019. *Motor speech disorders*.
- [5] HAIL. [n. d.]. <https://hail.co.kr/repeechkr>
- [6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs.CV] <https://arxiv.org/abs/1512.03385>
- [7] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhota, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. arXiv:2106.07447 [cs.CL] <https://arxiv.org/abs/2106.07447>
- [8] Kim Hyang Hui. 1996. Perceptual, Acoustical, and Physiological Tools in Ataxic Dysarthria Management: A Case Report. *Proceedings of the KSPS conference* (1996), 9–22. <https://koreascience.kr/article/CFKO199613842057217.page>
- [9] Biswajit Karan and Sitanshu Sekhar Sahu. 2021. An improved framework for Parkinson’s disease prediction using Variational Mode Decomposition-Hilbert spectrum of speech signal. *Biocybernetics and Biomedical Engineering* 41, 2 (April 2021), 717–732. doi:10.1016/j.bbe.2021.04.014
- [10] HyangHee Kim. 2005. Dysarthria evaluation. *Communication Sciences & Disorders* (2005), 23–28.
- [11] Heejin Kim, Mark Hasegawa-Johnson, Adrienne Perlman, Jon Gunderson, Thomas S. Huang, Kenneth Watkin, and Simone Frame. 2008. Dysarthric speech database for universal access research. In *Interspeech 2008*. 1741–1744. doi:10.21437/Interspeech.2008-480
- [12] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs.LG] <https://arxiv.org/abs/1412.6980>
- [13] Li Khim Kwah and Joanna Diong. 2014. National Institutes of Health Stroke Scale (NIHSS). *Journal of Physiotherapy* 60, 1 (2014), 61. doi:10.1016/j.jphys.2013.12.012
- [14] Anqi Mao, Mehryar Mohri, and Yutao Zhong. 2023. Cross-Entropy Loss Functions: Theoretical Analysis and Applications. arXiv:2304.07288 [cs.LG] <https://arxiv.org/abs/2304.07288>
- [15] Boersma P. 2001. Praat, a System for Doing Phonetics by Computer. *Glott. Int.* 5, 9 (2001), 341–345.
- [16] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*. PMLR, 28492–28518.
- [17] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. arXiv:1602.04938 [cs.LG] <https://arxiv.org/abs/1602.04938>
- [18] Frank Rudzicz, Aravind Namasivayam, and Talya Wolff. 2010. The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation* 46 (01 2010), 1–19. doi:10.1007/s10579-011-9145-0
- [19] M. J. Shin, J. O. Kim, S. B. Lee, and S. Y. Lee. 2010. *Speech mechanism screening test*. Hakjisa.
- [20] Caiming Xiong, Victor Zhong, and Richard Socher. 2018. Dynamic Coattention Networks For Question Answering. arXiv:1611.01604 [cs.CL] <https://arxiv.org/abs/1611.01604>
- [21] Eun Jung Yeo, Kwanghee Choi, Sunhee Kim, and Minhwa Chung. 2022. Cross-lingual Dysarthria Severity Classification for English, Korean, and Tamil. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. 566–574. doi:10.23919/APSIPAASC55919.2022.9980124 ISSN: 2640-0103.

Received 30 September 2025; revised 19 October 2025; revised 13 November 2025