Self-Supervised Learning for Classification of Normal *vs.* Dysarthric Speech

Hiya Chaudhari*, Kavya Kumar* and Hemant A. Patil *
* Dhirubhai Ambani University, (Formerly known as DA-IICT)
E-mail: 202101047@dau.ac.in, 202101017@dau.ac.in and hemant_patil@dau.ac.in

***Abstract*—Parkinson's Disease (PD) and Amyotrophic Lateral Sclerosis (ALS) are progressive neurodegenerative disorders that cause motor and non-motor symptoms, including dysarthria. Their early detection is crucial for effective management and improved quality of life. This study explores PD and ALS detection using self-supervised learning (SSL) methods namely, XLS-R, HuBERT, and wav2vec 2.0. We analyze normal *vs.* dysarthric speech using vowel space distribution plots in 2D and 3D, and t-SNE visualizations. BiLSTM and CNN models are employed to classify speech features, comparing XLS-R 300M with other pre-trained models. Our approach achieves an 8.25 % improvement over wave2vec 2.0 on the Italian Parkinson's Dataset and VOC-ALS. Additionally, we assess latency period, noise robustness, and precision for retraining to evaluate the practical applicability of relatively best XLS-R 300M features.**

I. Introduction

Parkinson's Disease (PD) predominately affects the dopamine producing neurons in substantia niagra (which under normal conditions has an inhibitory effect on our basal ganglia). Losing this inhibitory effect results in dopamine deficiency and overactivity of basal ganglia and this eventually manifests as motor symptoms. Amyotropic Lateral Sclerosis (ALS) is a neurological disorder that mainly affects the motor neurons of the central nervous system. As the motor neurons stop their activity, the muscles stop receiving signals and impulses from them. This results in muscle twitch and muscle atrophy. Early symptoms includes difficulty in breathing, chewing, swallowing, and distorted speech. These include chorea (rapid jerky movements), painful contractions of muscle and difficulty in speech. Further, trembling voice, pitch (or fundamental frequency, F_0) changes, and disruptions in speech rhythm are all commonly associated with PD and ALS causes slurred, slow speech, and difficulty with articulation, which is called *dysarthria*. PD is associated with *Hypokinetic Dysarthria*, where the movement of the muscles is relatively less than required, hypo is monotone rigid sounding speech. The causes incorporate injury to the basal ganglia, however, if the cause is unknown, the condition is classified as idiopathic PD. Hypokinetic dysarthria may adversely impact all stages of speech production mechanism, including the respiratory, phonatory, resonatory, articulatory, and prosodic systems. ALS is also known as Lou Gehrig's disease, it is a neurological disorder that affects motor neurons in the brain and spinal cord, which control voluntary muscle movement and breathing. PD and ALS are neurological disorders, however, they are distinct in their causes, symptoms, and affected systems as shown in Fig. 1.

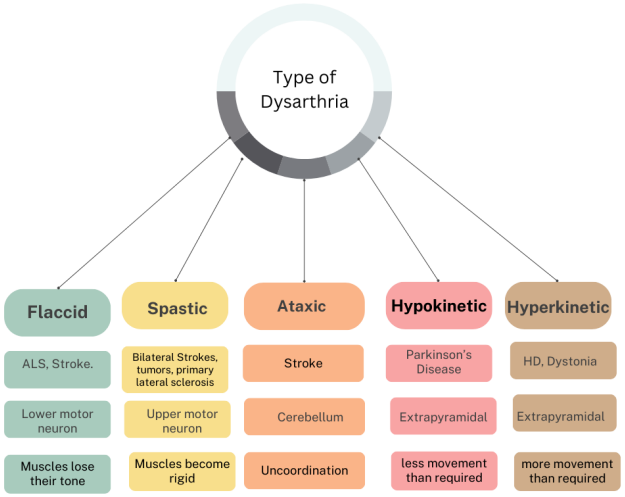

Fig. 1. Taxonomy of Dysarthria and the Articulatory Muscles involved in it.

Early detection of PD and ALS can progress more slowly, symptoms can be better managed, quality of life is improved, and access to innovative medications and treatments is made possible. Patients with both PD and ALS can sustain a higher level of function for a longer amount of time by beginning therapies early. Patients can continue to be more active and involved in their daily lives as a result of the improvements in mobility, cognitive function, and emotional well-being [1]. Former research work includes severity-level classification and speech recognition using XLS-R [2].We are focusing on Parkinson's Disease (PD) and Amyotrophic Lateral Sclerosis (ALS) in this paper, as we have access to well-annotated datasets for both conditions.

This study proposes XLS-R features for classification of normal *vs.* dysarthric speech in PD and ALS.
1) Experimental performance evaluation on three XLS-R models namely 300M, 1B and 2B, wav2vec 2.0, and HuBERT on Italian Parkinson's dataset and VOC-ALS, and a combination of the two.
2) Analysis using t-SNE plots, vowel space distribution and HuBERT heatmap plots for normal *vs.* dysarthric speech using

979-8-3315-7206-8/25/$31.00

an in-house dataset for Dysarthric speech in Indian languages.

3) Considering dysarthria research requires high performance, evaluating model retraining precision is critical. This study documents the experiments conducted for this goal.

4) Experimentation on latency analysis and noise degradation for non-idealistic or real-life conditions for practical system deployement.

II. Related Work

Conventional methods for detecting PD typically involved analyzing acoustic characteristics, such as pitch (or F_0) and tremor.

[3] and [4] proposes whisper features for severity classification of dysarthric speech and Dysarthric ASR. [5] aims at distinguishing between Dysarthric speech and normal speech using Adversarial AutoEncoders. In [6] authors trained an acoustic model was trained with features extracted from Wav2Vec, HuBERT, and the cross-lingual XLSR model for datsets UA-corpus and EasyCall,[7] proposes a cross-lingual classification method for English, Korean, and Tamil that uses both language-independent and language-unique features, [8] is detailed study on dysarthia severity classification using various deep learning architectural choices, namely deep neural network DNN, CNN, and LSTM is carried out. By incorporating multilingual speech data, the cross-lingual extension of wav2vec 2.0, known as XLS-R, has further expanded these capabilities, and improved the robustness of speech disorder detection across various languages and dialects.

This study [9] proposes wav2vec features for emotion recognition task. Even with these developments, further study is still required to solve lingering issues, such as enhancing model performance in a variety of populations and improving detection techniques for real-time use.

III. Methodology

The experimental setup for training and evaluating XLS-R models involves several key steps. First, the data is collected and then preprocessed to ensure consistent audio quality. Depending on the available computational resources and desired performance, an appropriate XLS-R model size (such as 300M, 1B, or 2B parameters) is chosen. GPUs are used to train the models using SSL techniques, where the model learns to predict masked parts of the audio, allowing it to learn from vast amounts of unlabeled speech data. The training process includes optimizing learning rates, batch sizes, and stabilizing training with techniques and gradient clipping. The models are then evaluated on their classification performance using performance metrics, accuracy, and their ability to generalize across different speech pathologies. Fine-tuning may be done on specific pathology to improve accuracy further. After training, the models are optimized for deployment by reducing their size and improving inference speed, making them efficient for real-world applications. This comprehensive setup ensures that XLS-R models are robust and high performing, capable of improving speech pathology identification across diverse acoustical conditions. Paralinguistics involves the non-verbal elements of communication used to convey meaning, such as tone, pitch, volume, speed, and intonation.

Dysarthria can affect paralinguistic features in speech meaning impaired paralinguistic features [10]. Non-lexical speech refers to vocalizations that do not contain recognizable words or linguistic meaning, such as groans, sounds, or noises that are not part of a specific language [11].

In cases of high-severity dysarthria, the speech motor impairments can become so severe that the individual's speech is unintelligible. This may result in vocalizations that lack recognizable words or linguistic structure, effectively rendering the speech non-lexical. Another motivation for explaining XLSR is that since it is multilingual version of wav2vec, it may give comparable classification performance with even training data that is 100 times lesser in duration. This is all the more the case for dysarthric speech, where the training data is extremely lesser in duration because dysarthric individuals are not able to speak even words properly [12].

A. Self-Supervised Learning (SSL)

Due to the lack of labeled data, high annotation costs, and significant speech variability, SSL is critical for classification of dysarthria. SSL uses huge unlabeled datasets to build robust representations, including key phonetic and prosodic aspects that improve classification and severity evaluation. Pretrained models, such as wav2vec 2.0, HuBERT, and XLS-R generalize well to dysarthric speech, improving performance in low-resource environments and allowing effective fine-tuning for enhanced diagnostic and assistive applications. Dysarthric speech is highly variable due to the nature of the motor impairment, and there are frequently significant differences in articulation, intonation, and pacing between speakers. SSL models are trained on large amounts of data that include a wide range of accents, languages, and speaking styles. This exposure to diversity makes the model more robust and adaptable to the variability found in dysarthric speech, even with limited specific training examples. wav2vec 2.0 refers to an unsupervised technique, which segments unlabeled language or speech sounds using self-supervised speech representations, and then employs adversarial training to learn a mapping from these representations to phonemes. It solves a contrastive task defined by a quantization of the jointly learned latent representations, while masking the speech input in the latent space [13]. XLS-R is a self-supervised cross-lingual representation learning model [14]. It is a multilingual wav2vec 2.0 transformer (XLS-R) pretrained on 436K hours of unannotated speech data from 128 languages. The training data comes from various public speech corpora, and the model is fine-tuned for various scenarios.

XLS-R provides models of various sizes to balance performance and computational efficiency: XLS-R 300M, XLS-R 1B, and XLS-R 2B. The 300M model, with approximately 0.3 billion parameters is appropriate for applications with limited computational resources, providing an excellent balance of efficiency and performance. The 1B model, which is trained over approximately 1 billion parameters, improves performance and accuracy, making it ideal for more demanding tasks, where

TABLE I
DISTRIBUTION OF D1 AND D2.

	PD Dataset	ALS Dataset
Healthy	37 (23M/14F)	51 (32M/19F)
Dysarthric	28 (19M/9F)	102 (65M/ 37F)

computational power is available. The largest model, XLS-R 2B, is trained over approximately 2 billion parameters and provides the highest accuracy and robustness. This range of model size enables users to select based on their specific needs and available resources, ensuring optimal performance for various applications in classification of dysarthria.

B. Datasets Used

The datasets being used are Italian Parkinson's Dataset (D1) and VOC-ALS (D2). Italian Parkinson's Dataset. The audio recordings in the [15], (D1) were provided by 28 PD patients between the ages of 40 and 80, 22 healthy individuals between the ages of 60 and 77, and 15 healthy individuals between the ages of 19 and 29. VOC-ALS (D2) is publicly available database of VOiCe signals collected from patients with Amyotrophic Lateral Sclerosis (ALS) and healthy controls while performing various speech tasks [16]. It contains 1224 voice signals recorded from 153 participants, including 51 healthy controls (32 males and 19 females) and 102 ALS patients (65 males and 37 females) with varying degrees of dysarthria.

1) In-House Dataset: We have recorded an In-House dataset for typical and dysarthric patients since its an ongoing project with limited number of subjects we have only used this data in analysis as small data overfits the model giving 100 % accuracy in classification on seen data and not working so efficiently on unlabelled data.

C. Classifiers Used

We have used CNN and Bi-LSTM, which is a deep learning model. Bi-LSTM is composed of LSTM units organized in both forward and backward directions, allowing them to capture data from both past and future states. [17].

D. Performance Metrics

In this study, classification, accuracy, F1-score, and Equal Error Rate (EER are used as the performance metrics utilized to gauge the effectiveness of the system. These metrics facilitated precise measurement of the system's performance. Evaluation accuracy, a metric utilized for assessing the performance of classification models, measures the proportion of correctly predicted instances to the total instances in the evaluation set. It offers a clear indication of the model's capability to accurately predict class labels.

IV. EXPERIMENTAL RESULTS

A. Effect of Different XLS-R Models and Classifiers

We evaluated three XLS-R models, namely, XLS-R 300M, XLS-R 1B, and XLS-R 2B on both datasets using two classifiers, CNN and Bi-LSTM. Fig.5 shows the testing accuracy of the data. It can be observed from the figure that XLS-R 300M yielded the highest relative accuracy. The comparision classification accuracy on other models, namely, wav2vec, HuBERT, and XLS-R on CNN and Bi-LSTM is shown Fig.3. It can be observed that our proposed method of unsupervised learning outperforms other DL model for both sets of classifiers indicating generalizability of proposed XLSR features [18]. The peak highest accuracy achieved by the proposed methodology is found to be 99.98 % and 98.56%. This study used two DNN classifiers, CNN and Bi-LSTM , to offset potential model bias (if any) during performance evaluations. Bi-LSTM outperforms CNN. Since our dataset is small XLS-R 300M performs better than 1B and 2B model as the dataset might overfit the model.

B. Comparison with Baselines

Table III shows comparison with state-of-the-art feature sets MFCC, LFCC, and Whisper [3]. The results show that

TABLE III
COMPARISON WITH BASELINES.

	MFCC	LFCC	Whisper	XLS-R 300M
D1	**95.23**	**94.88**	**98.49**	**99.98**
D2	95.88	97.99	98.01	**98.56**

XLS-R 300M outperforms traditional handcrafted features (MFCC, LFCC) and Whisper embeddings, achieving near-perfect accuracy of 99.98% on D1 and the best performance of 98.56% on D2. Notably, Whisper shows strong results, but the transformer-based XLS-R demonstrates superior generalization across datasets, highlighting its ability to capture rich acoustic and phonetic cues critical for severity-level classification.

C. Analysis of F_1–F_2 Distribution

The plot has formant frequencies F_1 on the y-axis and F_2 on the x-axis and can also help visualize patterns of vowel dispersion, speaker variation, or dialect differences. Insufficient muscle strength in patients with ALS and PD decreases the ability to maintain stable articulatory positions, resulting in greater variability and thus, a more scattered plot. Lower intelligibility was associated with increased overlap between vowels and increased F_1 variability, that is, decreased acoustic stability [19]. Variability in F_1–F_2 values results in a broader and less compact cluster on the F_1–F_2 plot due to inconsistent articulatory control, which causes fluctuations in vowel prediction [19].

The formant frequency is inversely proportional to the length of the vocal tract; in dysarthria the length of the vocal tract shrinks, leading to a higher formant frequency than typical speakers. The corner vowels in Peterson and Barney vowel distribution make up a triangle. The area of the vowel space triangle depends on various factors such as: label=()

1) Gender of the speaker
2) Normal *vs.* Pathological
3) Dialect or Accent
4) Different Languages
5) Severity of Dysarthria

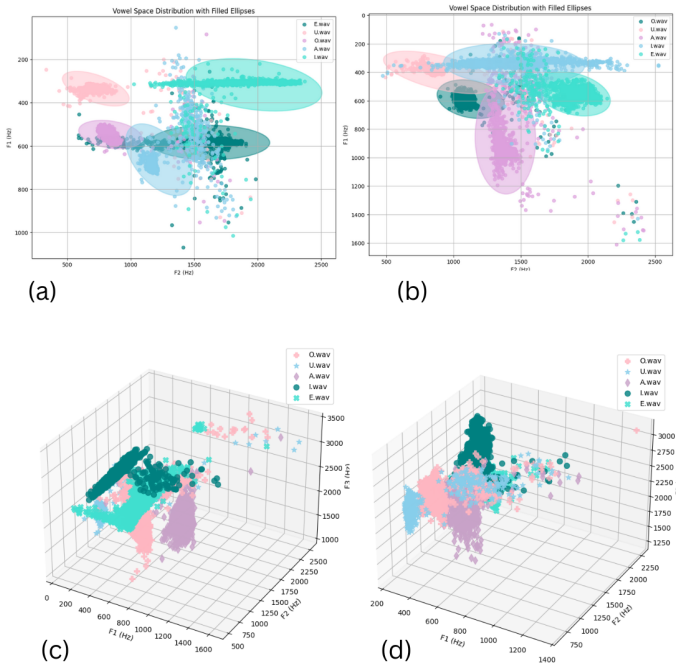

Fig. 2. Vowel Distribution Plot for (a) normal in 2D, (b) dysarthric in 2D, (c) normal in 3D, and (d) dysarthric in 3D.

In dysarthria a smaller vowel space often correlates with lower speech intelligibility since vowel contrast is reduced [19]. A speaker's overall intelligibility is indicated by both the area of the triangle and the degree of overlap within vowels in vowel space. Fig. 2 shows vowel space distribution of a normal speaker *vs.* a dysarthric speaker. In the instance of a dysarthric speaker, we may observe that the area is cut in half. For F_1–F_2 plots, formants are extracted using a standard Linear Prediction (LP)-based method.

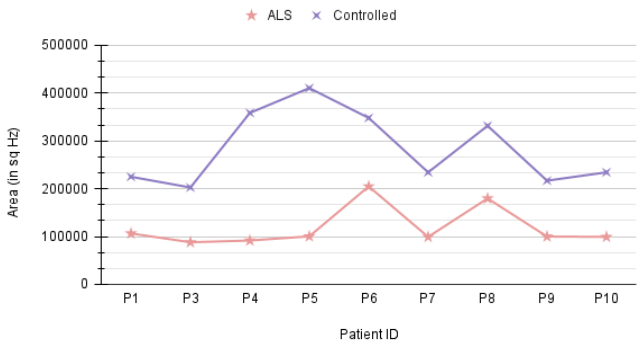

Fig. 3. The figure shows areas of multiple vowel-space triangles for a typical speaker and dysarthric speakers P1–P10, where P1–P6 are from the D2 dataset and P7–P10 are from our in-house dataset.

Fig. 3 presents a line graph of areas of multiple vowel-space triangles of patients and controlled speakers. The energy loss in F_3 and F_4 is more than the loss in F_1 and F_2, leading to a higher 3 dB bandwidth, thus resulting in ambiguity in estimation of F_3 and complicating peak picking [20].

D. Robustness Under Signal Degradation Conditions

Since most existing models for classification of normal *vs.* dysarthic speech have been tested with data collected in controlled settings, this experiment allows us to estimate the correctness of models in real-life application. Fig.4 analyzes how changing the characteristics of the LLM models with babble noise affects their noise robustness for D2. Additionally, we contrasted these characteristics performance with that of HuBERT, wav2vec, XLS-R 2B, and XLS-R 1B. Furthermore, different SNR levels for babble noise, such as -5 dB, 0 dB, 5 dB, and 10 dB, are taken into consideration in order to examine the effects of noise degradation on the proposed features. It is clear that the XLS-R 300M features outperforms the other features, such as, HuBERT, wav2vec, XLS-R 2B, and XLS-R 1B. Additionally, this experiment enables us to gauge the accuracy of models in realistic noisy conditions.

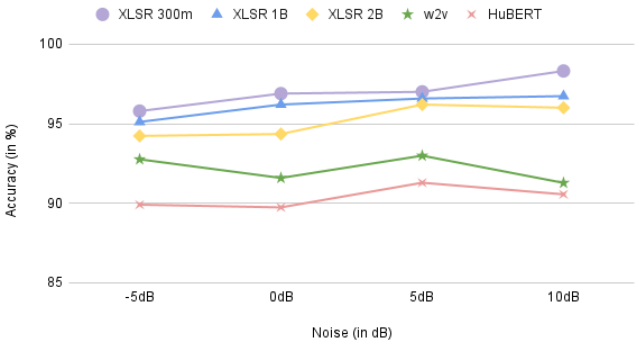

Fig. 4. Noise Robustness of SSL models

E. Analysis of Latency Period

Fig. 5 shows latency periods of several pre-trained speech models, including HuBERT, wav2vec 2.0, XLS-R 1B, XLS-R 2B, and XLS-R 300M for D2, are shown in Fig.5 across a range of necessary frame counts.

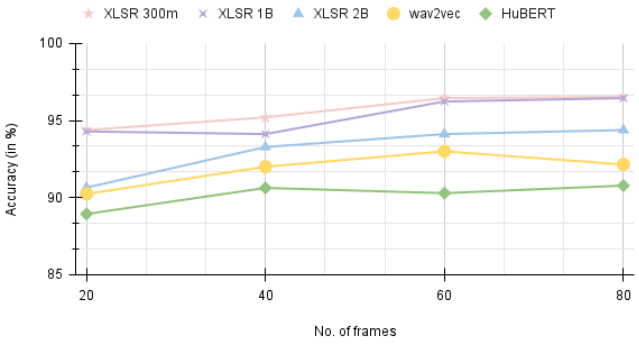

Fig. 5. Analysis of Latency Period.

For all the models, the latency increases as the dimension increases. Lower latency at fewer frames (20 and 40) is exhibited by XLS-R 1B and XLS-R 2B, suggesting that they are more effective for rapid inferences. At mid-range frames, HuBERT and wav2vec 2.0 exhibit moderate latency (60 and

80). All models exhibit increased latency at the highest frames (100), with the XLS-R 300M having the highest latency despite its superior accuracy and F1- score. Therefore, XLS-R 300M is the best option for the highest accuracy, even though XLS-R 1B and XLS-R 2B offer a good balance of efficiency and performance.

F. Analysis of Precision for Retraining

Fig.6 (right) shows the retraining capabilities of the XLS-R 300M model using the D1 Italian Parkinson's dataset. The experiments were performed five times for 100 epochs, with each run achieving a maximum accuracy of 99.98 %. Furthermore, shows that the performance eventually converges to the same value with minimal fluctuations due to model randomness and seed value.

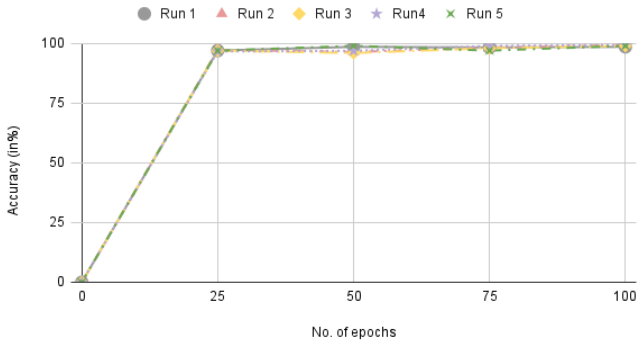

Fig. 6. Precision for Retraining.

G. t-SNE plot

t-SNE aims to capture the similarities between cells in the original dataset by placing the cells close together on a two- or three-dimensional plot. Fig.7 shows the t-SNE plot for XLS-R 300M for D1 (left) and D2 (right). We can examine the raw XLS-R embeddings used to train the acoustic model. In fig.7 we see that XLS-R separates healthy and dysarthric speech but we also see some overlapping due to both datasets being in Italian as this multilingual model forms different clusters for different languages.

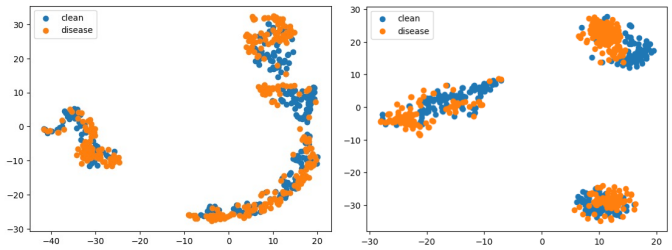

Fig. 7. t-SNE plots of XLS-R feature embeddings: (left) Normal *vs.* PD speech; (right) Normal *vs.* ALS speech.

H. Heatmap

Fig. 8 shows HuBERT heatmap for normal *vs.* dysarthric speech. The plot foe normal speech is structured with periodic

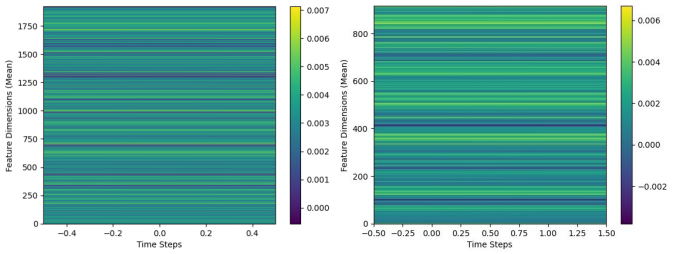

Fig. 8. HuBERT Heatmap for Normal (left)*vs.* Dysarthric Speech (right).

changes, which likely correspond to rhythmic and consistent speech. Dysathric speech has distinct transitions between states more blending and weaker transitions [21] The intensity (brightness) appears slightly different in some areas, hinting at less pronounced phoneme distinction.

V. Summary and Conclusions

This work investigated significance of SSL for the socially relevant task of early detection of PD and ALS . It focused on detecting and identifying pathology in individuals with PD and ALS and classification of normal *vs.* PD and ALS. We compared XLS-R 300M with other DL models, such as wav2vec 2.0, XLS-R 1B, XLS-R 2B, and HuBERT, showed plots corresponding analysis of latency period and noise degradation. We proposed a new optimal system based LLM's. We showed the analysis of normal *vs.* dysarthric speech using vowel space distribution, t-SNE plots and heatmap plots. We studied the effect of babble noise on the proposed XLS-R features and investigated the different dimension of feature vector. Future work involves testing disordered speech data from suspected patients using an in-house dataset and experimenting on various other types of noise (pink noise, white noise, high frequency noise) and feature fusion.

Acknowledgements

The authors sincerely thank the MeitY, for funding this study under project 'BHASHINI', (Grant ID: 11(1)2022-HCC(TDIL)).

References

[1] S. Liang and Y. Gu, "Multi-modal dysarthria severity assessment using dual-branch feature decoupling network and mixed expert framework," in *2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2024, pp. 126–130. DOI: 10.1109/ISCSLP63861.2024.10800159.

[2] F. Javanmardi, S. Tirronen, M. Kodali, S. R. Kadiri, and P. Alku, "Wav2vec-based detection and severity level classification of dysarthria from speech," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[3] S. Rathod, M. Charola, A. Vora, Y. Jogi, and H. A. Patil, "Whisper features for dysarthric severity-level classification," in *INTERSPEECH 2023*, 2023, pp. 1523–1527. DOI: 10.21437/Interspeech.2023-1891.

[4] J. Bhatt, H. Patel, and H. A. Patil, "Noise robust whisper features for dysarthric automatic speech recognition," in *The Speaker and Language Recognition Workshop (Odyssey 2024)*, 2024, pp. 217–224. DOI: 10.21437/odyssey.2024-31.

[5] V. Kanchana Devi, R. Sreenivas, E. Umamaheshwari, and N. Bacanin, "Adversarial auto-encoders based model for classification of speech dysarthria," in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2024, pp. 1–7. DOI: 10.1109/ICCCNT61001.2024.10724410.

[6] A. Hernandez, P. A. Pérez-Toro, E. Noeth, J. R. Orozco-Arroyave, A. Maier, and S. H. Yang, "Cross-lingual self-supervised speech representations for improved dysarthric speech recognition," in *Interspeech 2022*, 2022, pp. 51–55. DOI: 10.21437/Interspeech.2022-10674.

[7] E. J. Yeo, K. Choi, S. Kim, and M. Chung, "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)*, 2022, pp. 566–574. DOI: 10.23919/APSIPAASC55919.2022.9980124.

[8] A. A. Joshy and R. Rajan, "Automated dysarthria severity classification using deep learning frameworks," in *2020 28th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 116–120. DOI: 10.23919/Eusipco47968.2020.9287741.

[9] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," in *Interspeech 2021*, 2021, pp. 3400–3404. DOI: 10.21437/Interspeech.2021-703.

[10] J. Qi and H. Van hamme, "A study on model training strategies for speaker-independent and vocabulary-mismatched dysarthric speech recognition," *Applied Sciences*, vol. 15, no. 4, 2025, ISSN: 2076-3417. DOI: 10.3390/app15042006. [Online]. Available: https://www.mdpi.com/2076-3417/15/4/2006.

[11] Y. Jiao, M. Tu, V. Berisha, and J. Liss, "Simulating dysarthric speech for training data augmentation in clinical speech applications," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6009–6013. DOI: 10.1109/ICASSP.2018.8462290.

[12] I.-T. Hsieh and C.-H. Wu, "Dysarthric speech recognition using curriculum learning and articulatory feature embedding," Sep. 2024, pp. 1300–1304. DOI: 10.21437/Interspeech.2024-444.

[13] X. Xu, Y. Kang, S. Cao, B. Lin, and L. Ma, "Explore wav2vec 2.0 for mispronunciation detection," in *Interspeech 2021*, 2021, pp. 4428–4432. DOI: 10.21437/Interspeech.2021-777.

[14] A. Babu, C. Wang, A. Tjandra, *et al.*, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," in *Interspeech 2022*, 2022, pp. 2278–2282. DOI: 10.21437/Interspeech.2022-143.

[15] G. Dimauro and F. Girardi, *Italian parkinson's voice and speech*, 2019. DOI: 10.21227/aw6b-tg17. [Online]. Available: https://dx.doi.org/10.21227/aw6b-tg17.

[16] R. Dubbioso, M. Spisto, L. Verde, *et al.*, "Voice signals database of als patients with different dysarthria severity and healthy controls," *Scientific Data*, vol. 11, no. 1, p. 800, Jul. 2024. DOI: 10.1038/s41597-024-03597-2.

[17] K. Vora, D. Padalia, D. Mehta, and D. Sharma, "Hybrid cnn-lstm network to detect dysarthria using mel-frequency cepstral coefficients," in *2022 5th International Conference on Advances in Science and Technology (ICAST)*, 2022, pp. 615–621. DOI: 10.1109/ICAST55766.2022.10039514.

[18] S. Leivaditi, T. Matsushima, M. Coler, S. Nayak, and V. Verkhodanova, "Fine-tuning strategies for dutch dysarthric speech recognition: Evaluating the impact of healthy, disease-specific, and speaker-specific data," in *Interspeech 2024*, 2024, pp. 1295–1299. DOI: 10.21437/Interspeech.2024-1231.

[19] H. Kim, M. Hasegawa-Johnson, and A. Perlman, "Vowel contrast and speech intelligibility in dysarthria," *Folia phoniatrica et logopaedica : official organ of the International Association of Logopedics and Phoniatrics (IALP)*, vol. 63, pp. 187–94, Oct. 2010. DOI: 10.1159/000318881.

[20] X. Liu, X. Du, J. Liu, *et al.*, *Automatic assessment of dysarthria using audio-visual vowel graph attention network*, 2024. arXiv: 2405.03254 [eess.AS]. [Online]. Available: https://arxiv.org/abs/2405.03254.

[21] Y. Kheir, A. Ali, and S. Chowdhury, *Speech representation analysis based on inter- and intra-model similarities*, Jun. 2024. DOI: 10.48550/arXiv.2406.16099.