



Mitigating Overfitting During Speech Foundation Model Fine-tuning: Applications to Dysarthric Speech Detection

Yan Xiong¹, Visar Berisha^{1,2}, Julie Liss², Chaitali Chakrabarti¹

¹School of Electrical, Computer and Energy Engineering, Arizona State University, USA

²College of Health Solutions, Arizona State University, USA

Yan.Xiong.1@asu.edu, visar@asu.edu, julie.liss@asu.edu, chaitali@asu.edu

Abstract

Speech foundation models have shown significant success in various speech-processing applications. However, fine-tuning these models on dysarthric speech is challenging due to overfitting caused by limited dataset sizes. This work proposes a modified multitask learning framework to mitigate overfitting in foundation model fine-tuning. Specifically, we train the model on a more complex task along with the task of interest and use gradient projection to preserve beneficial updates while resolving conflicts. We demonstrate that using automatic speech recognition as the main task and dysarthria detection as the auxiliary task improves model robustness and dysarthria detection performance. The proposed method¹ reduces overfitting and improves in-corpus and cross-corpus detection accuracy by 5.4% to 13.4% compared to standard multi-task learning. These findings highlight the importance of structured multitask training for enhancing foundation model adaptability.

Index Terms: Clinical Acoustic Analysis, Dysarthria Detection, Speech Foundation Model Tuning, Multi-task Learning

1. Introduction

In recent years, speech foundation models have rapidly evolved, demonstrating remarkable performance in several speech-related tasks. These models are based on transformer architectures [1] and trained on large-scale datasets to generate general-purpose speech representations. The speech foundation models can be fine-tuned for specific applications, where their transfer learning capability minimizes the need for extensive labeled data, enhancing efficiency and adaptability across different languages and domains [2]. Models such as wav2vec [3, 4], HuBERT [5], Whisper [6], and Conformer [7] have driven significant advancements in the field, enabling widespread adoption in real-world applications like voice-activated assistants and real-time language translation. As these models continue to evolve, their application domains have expanded to include clinical tasks such as dysarthric speech analysis.

Dysarthria is a motor speech disorder caused by impaired neuromuscular control, which significantly affects speech clarity and intelligibility [8]. Machine learning and deep learning methods have been used for dysarthric speech recognition, dysarthria assessment, dysarthria-related features for disease identification, etc. While multiple dysarthria speech datasets have been collected for clinical analysis purposes [9, 10, 11], the size of such datasets is relatively small compared to other speech applications. The lack of data has been shown to increase the risk of overfitting when using machine learning and deep learning models for clinical analysis, leading to less mean-

ingful tuning and poor generalization, which ultimately results in unreliable performance during deployment [12, 13]. Speech foundation models reduce such risk, as they are pre-trained with large-scale datasets and maintain robustness when properly tuned. These models have been effectively used for dysarthric speech recognition [14], severity level classification [15], disease detection based on dysarthria analysis [16], speech intelligibility assessing [17], etc. However, these existing approaches use speech foundation models pre-trained on normal speech samples to generate feature embeddings. Fine-tuning these models with dysarthric speech data could enhance their ability to capture the unique characteristics of dysarthric speech, leading to improved performance in related downstream tasks.

As dysarthric speech datasets become larger, tuning speech foundation models with dysarthric speech is now feasible. With dysarthric speech fine-tuning, the foundation model captures acoustic features and patterns that are not present in normal speech. However, the tuning task is challenging as the model runs the risk of overfitting [18]. Due to the mismatch between the analytic flexibility of the speech foundation model and the complexity of the downstream task, the model captures unrelated patterns in the training data rather than learning the essential task-specific features, leading to poor generalization.

In this work, we propose a training scheme that utilizes multitask learning (MTL) and task-specific gradient projection (TGP) [19] to mitigate model overfitting when tuning speech foundation models for dysarthric speech detection. MTL has been shown to help mitigate overfitting by leveraging shared representations across related tasks, which regularizes the model and prevents it from memorizing task-specific noise. However, conventional MTL performs poorly when tasks are very different, leading to the gradients from different tasks significantly diverging in direction or amplitude. MTL with TGP mitigates the problem by using the main task for gradient computation and capturing the contribution of the auxiliary task by projecting the main task gradient to the normal plane of the auxiliary task gradient when the two gradients have a large divergence. We propose a TGP MTL training scheme to fine-tune a wav2vec 2.0 model for dysarthria detection. Our approach treats Automatic Speech Recognition (ASR) as the primary task, leveraging its higher complexity to align with the capabilities of the speech foundation model. Dysarthria detection serves as an auxiliary task to reduce overfitting risks introduced by its strong supervision. Evaluation loss curves demonstrate that our method achieves better convergence than both single-task learning and standard MTL. We assess the pre-trained model's dysarthria detection accuracy on a local dysarthric speech dataset and out-of-domain public datasets. Results show that our TGP MTL-trained model improves sample-level accuracy by 5.4% to 13.4% compared to standard MTL.

¹<https://github.com/BearockXY/TGP-SFM-Tuning-for-Dysarthria-Detection>

2. Methodology

2.1. Mitigating overfitting with Multitask Learning

Fine-tuning speech foundation models on dysarthric speech data is challenging due to the high risk of overfitting caused by the limited size and variability of available datasets. MTL provides an effective solution by introducing additional supervision, improving generalization, and preventing the model from memorizing irrelevant and task-specific features. ASR is particularly well-suited for speech foundation model tuning because it aligns naturally with the capabilities of speech foundation models, which are pre-trained to extract robust speech representations. Besides, ASR can also contribute to dysarthria detection, as impaired articulation is a cardinal feature shared by dysarthria. The proposed method maintains generalizable speech features by integrating ASR into the MTL framework for dysarthria detection while adapting effectively to dysarthric speech, where ASR serves as a complementary task to dysarthria detection.

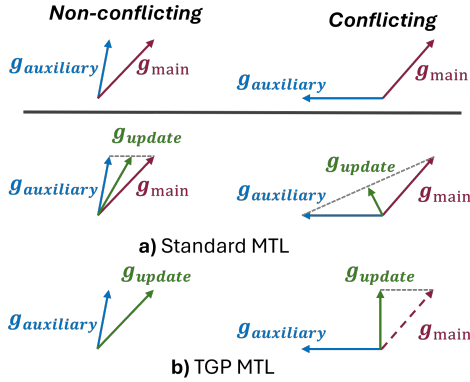


Figure 1: (a) *Standard MTL*: summing the gradients of each task. (b) *TGP MTL*: Projecting the main task gradient to the normal plane of the auxiliary task when two tasks conflict.

Standard MTL that uses the weighted sum of objective functions of the sub-tasks suffers from gradient discrepancy. When the gradients from different tasks diverge, the MTL becomes less effective as conflicting updates can hinder optimization convergence, reduce learning efficiency, and lead to sub-optimal task performance by prioritizing one objective over others or causing instability in parameter updates. In this work, we build upon the task-specific gradient projection (TGP) method in [19] to train the MTL model. TGP MTL considers one of the tasks as the main task and the other as the auxiliary task. As shown in Figure 1, TGP MTL uses the cosine similarity to decide whether the gradients from the two tasks are conflicting. In non-conflicting cases, TGP MTL uses the gradient from the main task to update the model weight. When there is a conflict, TGP MTL projects the main task gradient to the normal plane of the auxiliary task, removing the conflict. We consider both ASR and dysarthria detection as the main task in our evaluation.

2.2. Multi-task Learning with Speech Foundation Model

Our multi-task learning model is shown in Figure 2. The model uses a wav2vec 2.0 model to extract embedded feature vectors from the raw speech samples. Two independent heads are used for ASR and dysarthria detection. Both heads are MLP models with one hidden layer and include one dropout layer for regularization. The ASR head generates predicted transcription, and the classification head generates the dysarthria detection predictions. The input to the model is raw speech waveform

$x \in R^{L \times 1}$, where L is the sample length. The wav2vec 2.0 model (W) generates the embedded feature $h \in R^{T \times M}$, where T and M are the time and feature dimension, respectively.

$$h = W(x), x \in R^{L \times 1}, h \in R^{T \times M} \quad (1)$$

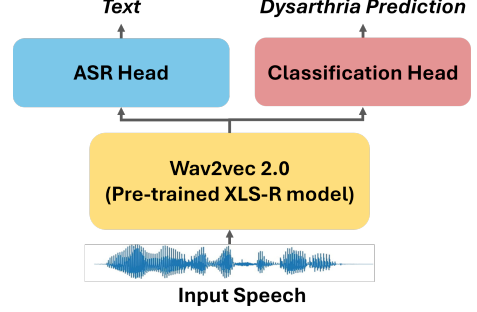


Figure 2: *Multi-task Learning Model with Wav2vec 2.0*

The dysarthria detection head (H_{dd}) and ASR head (H_{asr}) take the embedded feature h as input and generate the corresponding prediction for each task, p_{dd} and p_{asr} , where p_{dd} is the binary classification likelihood and p_{asr} is the $T \times V$ predicted word embedding where V is the vocabulary size.

$$p_{dd} = H_{dd}(h), p_{dd} \in R^{2 \times 1} \quad (2)$$

$$p_{asr} = H_{asr}(h), p_{asr} \in R^{T \times V} \quad (3)$$

We use Cross-Entropy (CE) loss for dysarthria detection and Connectionist Temporal Classification (CTC) [20] loss for ASR. For the baseline standard MTL scheme, the objective function is the weighted sum of the two losses:

$$L_{std-MTL} = L_{CE} + \alpha L_{CTC} \quad (4)$$

$$= CE(l_{dys}, p_{dd}) + \alpha CTC(l_{asr}, p_{asr}), \quad (5)$$

where l_{dys} and l_{asr} are the ground-truth labels of dysarthria detection and ASR, and factor α is tuned using the development set. In this case, the gradients that are used to update the model are calculated as shown in Figure 1 a):

$$g_{stdMTL} = \partial(L_{CE} + \alpha L_{CTC}) \quad (6)$$

$$= \partial L_{CE} + \alpha \partial(L_{CTC}). \quad (7)$$

The gradient from TGP MTL is calculated as is shown in Figure 1 b):

$$g_{tgpMTL} = \begin{cases} \partial L_{main}, & \text{if } \text{Sim}(\partial L_{main}, \partial L_{aux}) \geq 0 \\ \text{Proj}(\partial L_{main}, \partial L_{aux}), & \text{else} \end{cases} \quad (8)$$

where $\text{Proj}(g_1, g_2)$ refers to the projection computation that projects g_1 to the normal plane of g_2 . L_{main} is the main task loss function and L_{aux} is the auxiliary task loss function. $\text{Sim}(g_1, g_2)$ is the similarity function to determine whether two gradients are conflicting. In this work, we use the cosine similarity function to determine the gradient similarity.

2.3. Speech Foundation Model Tuning

The tuning process of speech foundation models is inspired by the way human listeners adapt to dysarthric speech. When encountering impaired speech caused by dysarthria, a listener does not start from scratch but rather builds upon their pre-existing knowledge of normal speech patterns. By listening to dysarthric

speech accompanied by transcripts, the listener progressively refines their ability to recognize and interpret the altered phonetic and prosodic characteristics associated with the disorder. This adaptation process involves leveraging prior linguistic and acoustic knowledge and incorporating new knowledge specific to dysarthric speech, resulting in an improved capability of understanding dysarthric speech.

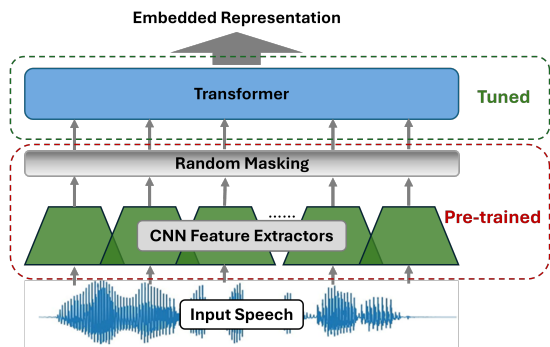


Figure 3: *Partially Tuning Wav2vec 2.0 using Dysarthric Speech*

To simulate this process, we partially fine-tune the wav2vec 2.0 speech foundation model, as illustrated in Figure 3. During fine-tuning, the pre-trained representations serve as a robust starting point, and the model is gradually adapted to dysarthric speech data through supervised learning. To maintain the general speech representations acquired from large-scale training, we freeze the pre-trained CNN feature extractor in the wav2vec 2.0 model. To adapt the model to dysarthric speech, the context transformer model is tuned to better capture the unique acoustic characteristics of dysarthric speech.

3. Evaluation

3.1. Dataset and Training Setup

We use a locally collected dysarthric speech dataset, which includes speech samples from 187 participants with dysarthria, aged 57-78, and 157 control participants, aged 33-80. It covers various dysarthria subtypes, including hypokinetic, ataxic, flaccid-spastic, and non-specific forms. The dysarthric speech recordings include sentence reading, scene description, phrase reading, etc. Four speaking styles—habitual, clear, loud, and slow—were elicited from participants to capture diverse acoustic profiles. In total, the dataset contains approximately 110,000 speech samples with a total length of ~ 200 hours.

The dataset is divided into training, evaluation, and testing sets in a speaker-independent way, i.e., the samples in three sets are from three different groups of speakers. The gender of the speakers is balanced in all three sets. The model is trained with a training set for 50 epochs, and the hyper-parameters are tuned with the evaluation set. After the model is tuned, the testing set is used to report the performance.

The proposed MTL model takes raw speech waveform as input. The samples have a sample rate of 16 kHz. Amplitude normalization is applied to normalize the amplitude of the input waveform between -1 and 1.

3.2. Evaluation Loss Curve Study

We use the evaluation loss curve (classification loss as a function of epoch number) to show the extent of overfitting in foundation model training. Figure 4 plots the evaluation loss curve on the evaluation set for the various STL and MTL schemes.

The STL baseline shows severe overfitting, where the evaluation loss shows an increasing trend during training. The use of standard MTL mitigates the overfitting and slows down the evaluation loss increase at later epochs but still shows an overall increase in evaluation loss as the training proceeds.

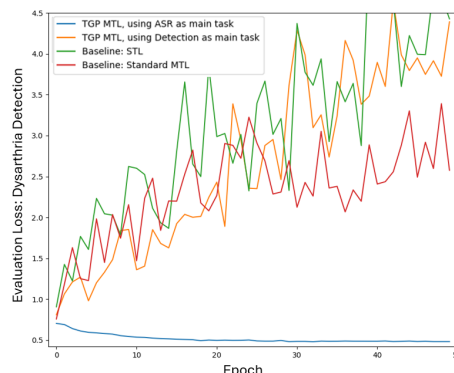


Figure 4: *Evaluation Loss Curve of Different Training Schemes*

To evaluate TGP MTL performance, we consider the main task to be either dysarthria detection or ASR. When using dysarthria detection as the main task, the evaluation loss shows a trend similar to the STL baseline. The loss increases and is in a similar range after 50 epochs of training. In comparison, using ASR as the main task shows a steady and decreasing trend in evaluation loss. With more training epochs, the evaluation loss keeps decreasing and finally converges. This shows that the use of TGP MTL with ASR as the main task efficiently mitigates the overfitting of the dysarthria detection task when tuning the wav2vec 2.0 model.

3.3. Dysarthria Detection Performance on Local Data

First, we test the model with dysarthric speech samples from a local dataset where the speech samples are from the same dysarthria speech dataset that is used for training and evaluation. Both dysarthria and control speakers are different from those used in the training and evaluation sets. The sample-level accuracy of different training schemes is shown in Table 1. On the local dysarthria speech testing set, STL and Standard MTL yield comparable performance, indicating that conventional MTL does not significantly contribute to single-task training for dysarthria detection tasks. TGP MTL with dysarthria detection as the main task performs the worst among all methods, suggesting that the use of ASR as the auxiliary task does not help with the regularization. The model trained using TGP MTL with ASR as the main task achieves the highest accuracy of 75.18%, surpassing all other training strategies. Thus, TGP MTL with ASR as the main task mitigates the overfitting problem and enhances the model's ability to capture useful acoustic characteristics of dysarthric speech.

To evaluate model robustness to disease types, a testing set that consists of speech samples from Amyotrophic Lateral Sclerosis (ALS) patients is used. The samples are from the local dysarthria speech dataset; however, no dysarthric speech from patients with ALS was used in the training set. The ALS speakers in this corpus show more severe dysarthria symptoms, and, as a result, all four models show higher accuracy in dysarthria detection. TGP MTL with ASR as the main task again demonstrates the highest performance, achieving an accuracy of 77.14%. This suggests that TGP MTL with ASR as the main task helps the model better capture the acoustic features that correspond to dysarthria and show robustness to the type of

Data Source	Evaluation on Local Data		Cross-corpus Testing		
	Training	Dysarthria	ALS (etiology not included in training)	TORGO	UA Speech
STL		67.83%	74.64%	60.68%	60.63%
Standard MTL		66.29%	75.71%	60.73%	61.21%
TGP MTL, detection as main task		64.55%	76.79%	71.54%	63.55%
TGP MTL, ASR as main task		75.18%	77.14%	74.18%	66.70%

Table 1: Performance comparison of different models on local and cross-corpus testing datasets.

disease that causes dysarthria.

3.4. Cross-corpus Evaluation with Public Dysarthria Speech Datasets

To further validate the model robustness, we evaluate the performance of the pre-trained models using two widely used public dysarthria speech datasets, namely, TORGO dataset [10] and UA speech dataset [9]. Achieving good performance on cross-corpus evaluation is challenging. It requires the model to capture dysarthria-related features from speech samples while ignoring the differences in recording conditions. For both corpora, the pre-trained model is not tuned with samples from the target corpus to avoid over-optimistic estimation caused by limited number of speakers.

The cross-corpus evaluation results on TORGO and UA Speech datasets show a similar trend as the local data evaluation. STL and Standard MTL models perform notably worse, with accuracies around 60-61%, highlighting their limitations in handling cross-corpus variability. TGP MTL with detection as the main task shows an accuracy of 71.54% on TORGO and 63.55% on UA Speech, suggesting that detection-oriented MTL provides more robustness compared to standard MTL. Among all methods, TGP MTL with ASR as the main task achieves the highest accuracy for both datasets, reaching 74.18% on TORGO and 66.70% on UA Speech, demonstrating its adaptability to unseen dysarthric speech collected from different sources. These results collectively indicate that ASR-prioritized TGP MTL enhances model generalization across different dysarthric speech corpus. Across all methods, the performance of the UA Speech dataset is lower. This is likely because the samples from the UA speech dataset consist of individual keyword recordings from dysarthria patients. The keywords include digits, radio alphabet letters, computer commands, and common/uncommon words. In comparison, all models in this paper are trained with longer speech recordings that include sentences, keyword combinations, and narrative descriptions. This mismatch likely causes lower performance. Nevertheless, TGP MTL with ASR as the main task still shows improvement compared to other baselines.

4. Discussion

The evaluation results show the benefit of using TGP MTL. The performance difference between the different choices of the main tasks indicates that choosing the right complementary task is important in speech foundation model tuning.

Cross-entropy (CE) loss has been widely used in classification and detection tasks. It provides strong supervision by directly optimizing class probabilities, making it highly effective for classification tasks where precise one-to-one label assignments are required. Our results show that even with heavy regularization (dropout and random masking), the CE loss is prone to overfitting due to the strong supervision at each training instance. This makes it easier for models to memorize training data or focus on confounding features for separation, especially in low-data or imbalanced scenarios. Unlike sequence-

level losses like CTC, which introduce alignment uncertainty and distribute learning across multiple valid paths, CE loss enforces strict label assignments. Such assignments can lead to sharper decision boundaries that may not correspond to the target tasks and reflect other confounding factors in the data that are correlated with the target labels. As is shown in Figure 4, when CE loss is applied in STL or standard MTL, the evaluation loss shows an increasing trend, indicating that the learned decision boundary is not effective for the evaluation set.

Compared to the CE loss, the CTC loss helps prevent overfitting by introducing alignment flexibility and sequence-level supervision, allowing the model to learn from multiple valid paths rather than memorizing exact input-output mappings. CTC dynamically aligns input sequences with target outputs, spreading gradient updates across different possible alignments. This implicit regularization discourages the model from overfitting to specific training examples, especially in tasks like speech recognition, where input sequences vary in length and noise is present. We know *a priori* that the features learned by the ASR task are also likely to be useful for dysarthria detection since articulation imprecision is an important feature in dysarthria.

From the data utilization perspective, the CTC loss utilizes the full input context rather than focusing on isolated features, thus achieving better generalization to unseen data. With ASR as the main task, TGP MTL integrates the effect of CE loss during training in a less aggressive style. When the two gradients do not conflict, the ASR gradient is likely to contribute to both tasks. When the two gradients disagree, the modification is made on the main task gradient to minimize the negative effect on the auxiliary task while still aligning with the main task gradient. In this way, the training makes full use of the ASR task to prevent overfitting and teaches the model how to distinguish between dysarthric speech and normal speech on the side. This leads to a decreasing and converging evaluation curve, as is shown in Figure 4.

5. Conclusion

This paper presents an MTL framework with task-specific gradient projections to mitigate overfitting in fine-tuning speech foundation models for dysarthric speech detection. By utilizing TGP MTL and making ASR the main task, our approach achieved significant performance improvement even with a limited-size dysarthric speech dataset. Experimental results demonstrate substantial improvements in model generalization and robustness across both local and cross-corpus datasets. Our findings emphasize the role of structured multitask training in optimizing speech foundation models for specialized applications. Future work includes extending this approach to more diverse speech disorders and exploring adaptive gradient projection strategies for dynamic task weighting.

6. Acknowledgements

This work is funded in part by NIH NIDCD grants R01DC006859-11 and R21DC019475 grants.

7. References

- [1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 5998–6008.
- [2] C. Yi, J. Wang, N. Cheng, S. Zhou, and B. Xu, "Transfer ability of monolingual wav2vec2.0 for low-resource speech recognition," in *2021 international joint conference on neural networks (IJCNN)*. IEEE, 2021, pp. 1–6.
- [3] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," 2019.
- [4] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," vol. 33, 2020, pp. 12 449–12 460.
- [5] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.
- [6] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision." PMLR, 2023, pp. 28 492–28 518.
- [7] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Revolution-augmented transformer for speech recognition," 2020.
- [8] P. Enderby, "Disorders of communication: dysarthria," *Handbook of clinical neurology*, vol. 110, pp. 273–281, 2013.
- [9] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. R. Gunderson, T. S. Huang, K. L. Watkin, and S. Frame, "Dysarthric speech database for universal access research," in *Interspeech*, vol. 2008, 2008, pp. 1741–1744.
- [10] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The torgo database of acoustic and articulatory speech from speakers with dysarthria," *Language resources and evaluation*, vol. 46, pp. 523–541, 2012.
- [11] R. Dubbioso, M. Spisto, L. Verde, V. V. Iuzzolino, G. Senerchia, E. Salvatore, G. De Pietro, I. De Falco, and G. Sannino, "Voice signals database of als patients with different dysarthria severity and healthy controls," *Scientific Data*, vol. 11, no. 1, p. 800, 2024.
- [12] C. Flint, M. Cearns, N. Opel, R. Redlich, D. M. Mehler, D. Emden, N. R. Winter, R. Leenings, S. B. Eickhoff, T. Kircher *et al.*, "Systematic misestimation of machine learning performance in neuroimaging studies of depression," *Neuropsychopharmacology*, vol. 46, no. 8, pp. 1510–1517, 2021.
- [13] B. A. Yawer, J. Liss, and V. Berisha, "Reliability and validity of a widely-available ai tool for assessment of stress based on speech," *Scientific reports*, vol. 13, no. 1, p. 20224, 2023.
- [14] S. Hu, X. Xie, M. Geng, Z. Jin, J. Deng, G. Li, Y. Wang, M. Cui, T. Wang, H. Meng *et al.*, "Self-supervised asr models and features for dysarthric and elderly speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.
- [15] 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.
- [16] O. Klempir and R. Krupicka, "Analyzing wav2vec embedding in parkinson's disease speech: A study on cross-database classification and regression tasks," *medRxiv*, pp. 2024–04, 2024.
- [17] T. Smolik, R. Krupicka, and O. Klempir, "Assessing speech intelligibility and severity level in parkinson's disease using wav2vec 2.0," in *2024 47th International Conference on Telecommunications and Signal Processing (TSP)*. IEEE, 2024, pp. 231–234.
- [18] E. J. Yeo, K. Choi, S. Kim, and M. Chung, "Automatic severity classification of dysarthric speech by using self-supervised model with multi-task learning," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [19] Y. Xiong, V. Berisha, J. Liss, and C. Chakrabarti, "Improving speech-based dysarthria detection using multi-task learning with gradient projection," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTER-SPEECH*, 2024, pp. 902–906.
- [20] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.