Beyond Binary Detection: Multi-Etiology Dysarthria Classification with Pre-trained Speech Models

Zihan Zhong*, Qianli Wang*, Satwinder Singh*, Clarion Mendes†,
Mark Hasegawa-Johnson†, Waleed Abdulla*, and Seyed Reza Shahamiri*
*University of Auckland, New Zealand
E-mail: {zzho680, qwan121}@aucklanduni.ac.nz, {satwinder.singh, w.abdulla, reza.shahamiri}@auckland.ac.nz
†University of Illinois Urbana-Champaign, USA
E-mail: {cmendes2, jhasegaw}@illinois.edu

***Abstract*—Automated classification of dysarthria's etiology and distinguishing it from typical speech is critical for informing the differential diagnosis and subsequent targeted treatment. We present the first study to tackle this comprehensive diagnostic task on the newly released Speech Accessibility Project (SAP) corpus. SAP is the largest and most etiologically diverse dysarthric-speech corpus to date, comprising speakers with amyotrophic lateral sclerosis, cerebral palsy, Down syndrome, Parkinson's disease, stroke. We applied linear probing to assess features from three widely used pre-trained speech models on the five-etiology classification task. HuBERT features achieved the highest performance (AUC 0.95), followed closely by Wav2vec2 (AUC 0.94), while the supervised Whisper model fell behind (AUC 0.55). We further explored a multi-task learning (MTL) framework based on Whisper in which a six-class classification (five etiologies plus typical speech) was trained jointly with an auxiliary automatic speech recognition task to improve classification performance. MTL boosted Whisper's AUC from 0.55 to 0.93. The results indicate that self-supervised models are superior feature extractors and validate MTL as a powerful technique to enhance supervised models. This study is a step toward objective, data-driven tools that assist speech-language pathologists during diagnosis and patients in receiving optimal treatment.**

I. Introduction

Dysarthria is a motor-speech impairment arising from central or peripheral nervous system damage that compromises the neuromuscular control required for accurate speech production [1]. The resulting impairment disrupts function across the core speech subsystems and is a common symptom in conditions like amyotrophic lateral sclerosis (ALS), cerebral palsy (CP), Parkinson's disease (PD) [2]. Importantly, different etiologies are often associated with distinct dysarthria subtypes. As shown in a seminal work by Darley et al., each neurological disorder may present a unique cluster of perceptual speech characteristics that correspond to specific neuromuscular deficits [3]. These etiology-specific acoustic-phonetic patterns capture disorder variability and are one of the targets for automated analysis.

Because a wide range of neurological conditions can cause dysarthria, accurately classifying its etiology is clinically valuable for differential diagnosis and subsequent targeted treatment [4]. However, current clinical assessment methods face significant challenges in delivering timely and effective diagnoses. While objective instrumental methods such as neuroimaging and kinematic analysis are available, these methods' cost, equipment demands, and specialized training requirements make them impractical for routine clinical deployment, limiting widespread adoption [5].

Machine learning (ML) has emerged as a promising solution to the limitations of medical devices assessment. Early research centered on feature engineering. These studies focused on extracting a robust set of handcrafted acoustic features based on expert knowledge of speech science to target speech tasks [6]. These feature sets were often refined using selection techniques to identify the most informative subset and reduce overfitting before being fed into conventional classifiers like Support Vector Machines (SVMs) [7]. However, this approach relies heavily on selected features. These handcrafted descriptors target specific, predefined properties of the speech signal [8] and are difficult to transfer to other assessment dimensions. Consequently, system performance is highly sensitive to the chosen features and does not generalize well to new tasks. This challenge motivated a shift toward deep-learning methods that are more robust for automatically learning hierarchical, task-relevant representations.

Deep learning (DL) instead learns complex, hierarchical feature representations directly from less-processed data like spectrograms without the need for manual feature engineering [9]. A study combined Convolutional Neural Networks (CNNs) with temporal models and applied novel architectures like patch-based CNNs to exploit raw waveforms [10]. The CNN system performance distinguished ataxic from hypokinetic dysarthria with an AUC of 0.95, even surpassing medical residents' diagnostic accuracy. Later, the field gained increased attention in self-supervised learning (SSL) and transfer learning. State-of-the-art (STOA) systems on binary dysarthric/healthy detection show that features from large pretrained models exceed 95% accuracy on English dysarthric speech using SSL features [11], and can achieve even 99% accuracy on a Tamil dysarthria corpus with a fine-tuned HuBERT model [12]. Farhad et al. further showed that pretrained features outperform traditional handcrafted sets. HuBERT features improved dysarthria detection by 25 % over MFCC baselines [11]. The pre-trained speech models have huge potential as the feature

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

extractor for speech assessment tasks for dysarthric speech.

Multi-task learning (MTL) has recently gained substantial scholarly attention [13]. By sharing layers across related tasks, MTL encourages more robust speech representations. For instance, Xu et al. [14] jointly trained a CNN to detect dysarthria while predicting four clinical speech impairment scores. Their MTL model achieved a high detection accuracy of around 95% while providing interpretable indicators, illustrating the effectiveness of MTL. Similarly, Xiong et al. [15] improved detection accuracy by adding speaker-embedding learning as an auxiliary objective. These studies indicate that rich acoustic embeddings learned via multi-task learning MTL can jointly support detection and clinical assessment.

Although binary dysarthria detection performs well, its coarse diagnostic granularity limits clinical utility. A more clinically impactful task is the comprehensive differential diagnosis, which aims not only to identify a speaker's underlying neurological condition but also to distinguish disordered speech from typical speech. This step is crucial for automated differential diagnosis yet has long been hindered by limited, etiologically diverse data.

The recent release of the Speech Accessibility Project, as detailed in Section II-A, now makes tackling this challenge feasible. Therefore, in this study, we aim to investigate the six-class etiology classification task. To our knowledge, this is the first multi-etiology classification study on SAP. We used linear probing to benchmark three STOA pretrained speech models as feature extractors on the five-etiology sub-task: HuBERT [16], Wav2Vec 2.0 [17], and Whisper [18]. In addition, we propose an MTL framework that leverages automatic speech recognition (ASR) as an auxiliary objective to learn more robust representations for the full six-class diagnostic task. The main contributions of this work are threefold:

- We systematically compared and analyzed the effectiveness of features extracted by pretrained speech models for five-class etiology classification sub-task on SAP-1008 using simple linear probing.
- We propose a novel multi-task learning framework that jointly optimizes six-class differential diagnosis (five etiologies plus typical speech) and ASR, demonstrating that this approach significantly improves classification performance for the Whisper model.
- Our results highlight that models pre-trained with SSL (Wav2vec2, Hubert) are superior feature extractors and validate the potential of MTL to enhance the performance of encoder-decoder architectures further for complex, multi-class etiology classification.

II. Materials

A. Datasets

This study utilizes two speech corpora. LibriSpeech [19] was used to pre-train the models on typical speech, whereas the Speech Accessibility Project (SAP) [20] corpus serves as the primary dataset for our classification experiments.

B. LibriSpeech

LibriSpeech is a large-scale English audio dataset with audiobook recordings from 2,484 typical speakers. Most recordings are in American English. It provides a robust foundation for pre-training speech models.

C. Speech Accessibility Project

SAP is an ongoing large-scale initiative that collects speech from individuals with diverse impairments to advance speech technology. The project is led by the Beckman Institute at the University of Illinois Urbana-Champaign and supported by leading technology companies, including Amazon, Apple, Google, Meta, and Microsoft. SAP is available upon request. In this study, we used the October 8, 2024 release, denoted SAP-1008. Because the project targets ASR improvement rather than medical diagnosis, etiology labels rely on participant self-reports with an experienced speech-language pathologist verifies the presence of a speech impairment only. SAP-1008 release is the largest publicly available corpus of etiologically diverse dysarthric speech. It is more than 20 times larger than prior datasets, such as UA-Speech [21] and TORGO [22], that were widely used in the field. As Figure 1 shows, it also covers a broader etiology distribution.

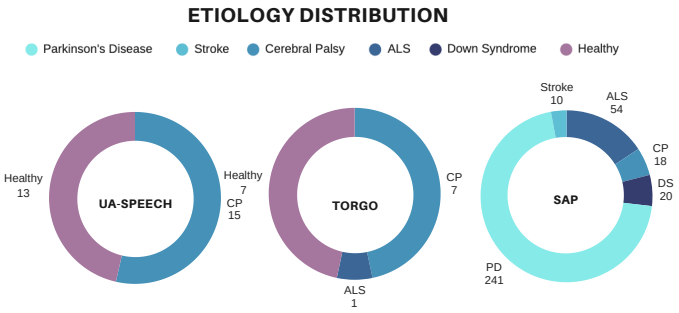

Fig. 1. Etiology distribution comparison between atypical speech dataset [20]

TABLE I
STATISTICS OF SAP'S TRAIN AND DEV SET BY ETIOLOGY

Etiology	Train		dev	
	Speakers	Hours	Speakers	Hours
Parkinson's disease	241	205.1	35	31.2
ALS	54	29.1	8	4.3
Down syndrome	20	10.0	3	1.8
Cerebral palsy	18	14.1	4	3.1
Stroke	10	5.7	1	0.7
Total	343	264.0	51	41.1

SAP is structured for speaker-independent tasks with predefined train, dev, and test sets. It contains a train set of 253 hours (343 speakers) and a dev set of 41.1 hours (51 speakers). Test sets remain reserved for official SAP challenges and are not publicly available. Table I summarises speaker counts and durations for the train and dev splits. Notably, SAP's etiology distribution exhibits a data imbalance. Initial recruitment focused on speakers with PD, while enrollment

for other etiologies began later and is still ongoing. Consequently, only a subset of non-PD speakers appears in SAP-1008. Future releases are expected to be more balanced. To comply with ethical regulations, SAP does not provide speaker-level demographic data such as age, gender, or intelligibility. Consequently, our analysis does not investigate the influence of these factors.

D. Pre-trained Speech Models

We evaluate three STOA speech models as feature extractors for etiology classification. These models represent three distinct pre-training paradigms. HuBERT [16] and Wav2Vec 2.0 [17] are already popular for dysarthria detection. In addition, we investigate Whisper [18], a STOA model in many speech-related tasks such as ASR, but it is rarely studied for dysarthria assessment. Whisper's encoder-decoder architecture makes it better suited for our proposed MTL framework (Section III-D) than the encoder-only HuBERT and Wav2Vec 2.0 models. Key characteristics of each model are summarised below.

a) Whisper: is a supervised model with an encoder-decoder architecture pre-trained on a massive, diverse, multilingual dataset (680,000 hours). We compare two variants: Whisper-medium.en (769M parameters) and Whisper-large-v3 (1.55B parameters). Its extensive training enables robust generalization, and its architecture is uniquely suitable for our proposed multi-task learning framework.

b) Wav2Vec 2.0: is a self-supervised model pre-trained on approximately 53,000 hours of unlabeled audio from sources such as CommonVoice data. Its contrastive learning approach, which learns representations by predicting masked segments of the raw audio waveform from a set of distractors, produces rich acoustic representations. We use wav2vec2-large-960h with 317 million parameters.

c) HuBERT: is pre-trained on approximately 60,000 hours of unlabeled LibriLight data and employs a two-stage SSL strategy. By first deriving hidden acoustic units through unsupervised clustering and then using these pseudo-labels in a masked prediction task. HuBERT learns robust representations that capture subtle phonetic and acoustic variations, making it a leading model for feature extraction in speech analysis. We use hubert-large-ls960-ft, which also has 317 million parameters.

III. Methodology

A. Data Preprocessing

All recordings were resampled to 16 kHz. Utterances longer than 30 seconds were excluded to accommodate the context window limitations of the Whisper model.

Text normalization was performed using the script provided by the SAP team for ASR tasks in MTL to ensure consistency for evaluation. This procedure first cleaned the transcripts by removing transcriber comments in brackets and speech disfluencies in parentheses, while mapping uncertain segments to a special `<UNK>` token. The script then expanded common acronyms into their constituent letters (e.g., "FBI" becomes "F B I") and converted numbers into words. Finally, the text was converted to uppercase, all punctuation except for internal apostrophes was removed, and whitespace was standardized to produce the final clean transcript.

B. Experimental Setup

We loaded pre-trained speech models from the official Hugging Face Hub—wav2vec2-large-960h, hubert-large-ls960-ft, Whisper-medium.en, and Whisper-large-v3—all of which were pre-trained on the 960 h LibriSpeech dataset [23].

For the initial benchmarking of pre-trained models, we addressed the 5-class etiology classification task using only the SAP corpus. Because the test set of SAP-1008 is reserved, we used the original dev set as the test set. Accordingly, the SAP training set was split at the speaker level to create our training and validation sets. We held out 10% of the original SAP training set as a dev set. The remaining 90% of the data served as the training set for this experiment.

For our 6-class experiment, we constructed the 'Typical' speech class using data from the LibriSpeech corpus, as the SAP dataset lacks an age- and sex-matched control group. We added the train-clean-100 subset to our training set and dev-clean to our validation set. Finally, 6-class performance was evaluated using two test sets: the SAP dev set for the etiology classes and the LibriSpeech test-clean subset for the 'Typical' class. This setup enabled us to assess performance across all six categories.

C. Five-Etiology Classification Using Linear Probing

Prior work shows that robust feature extractors can achieve strong classification performance even when paired with a simple linear model [24]. Accordingly, we conducted a linear probing experiment to evaluate the discriminative power of features from various pre-trained speech models.

We evaluated three models' checkpoints loaded from Hugging Face, as detailed in Section II-A. For each utterance, we aggregated its sequence of final hidden states, $\mathbf{H} = \{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_T\}$, into a single fixed-size embedding $\mathbf{z}$ using temporal mean pooling, calculated as

$$\mathbf{z} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{h}_t. \tag{1}$$

The resulting vector $\mathbf{z} \in \mathbf{R}^D$, where D is the feature dimension, served as the input to a classifier as shown in Figure 2. This probing classifier consists of a single linear layer that was trained on the frozen features to produce a vector of logits $\mathbf{y}$ for the five speaker etiologies:

$$\mathbf{y} = W\mathbf{z} + \mathbf{b}, \tag{2}$$

where W and $\mathbf{b}$ are the learnable parameters. The model was trained using the Adam optimizer ($\text{lr} = 1 \times 10^{-3}$) to minimize the standard cross-entropy loss. The loss function operates directly on the output logits, ensuring the probe's architecture remains strictly linear.

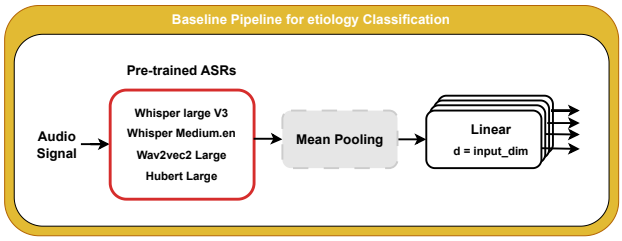

Fig. 2. Baseline pipelines for etiology classification

D. Six-Class Diagnosis Using Multi-task Learning

Previous studies show that MTL can learn more robust shared features by jointly training related tasks [13]. In this experiment, we jointly trained six-class etiology classification with an auxiliary ASR objective. The shared encoder is expected to learn features useful for both tasks, thereby improving classification accuracy. We adopted the Whisper model as the backbone for this framework because its encoder–decoder design already supports ASR, whereas HuBERT and Wav2Vec 2.0 are encoder-only.

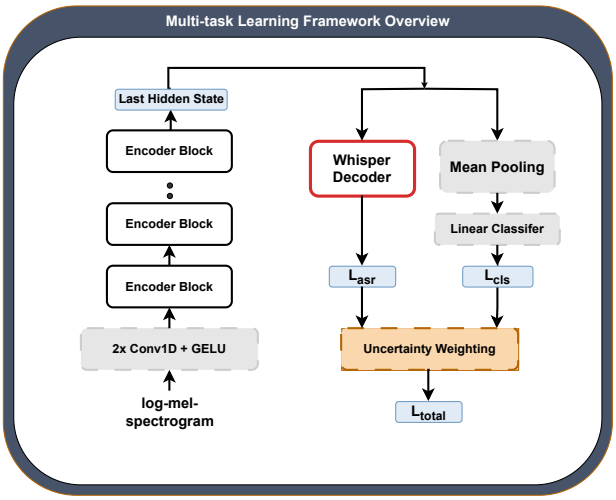

Fig. 3. Multi-task learning framework combining ASR (auxiliary) and six-class classification.

Our MTL model extends the standard Whisper architecture by adding a lightweight classification head on top of its encoder (see Fig. 3). The original Whisper decoder operates on the encoder's outputs to perform the ASR task, generating transcriptions. Meanwhile, we aggregated the final hidden states from the encoder using temporal mean pooling and fed the resulting vector into a linear layer, which is the same as our linear probing experiments.

During training, the encoder, decoder, and classification head were optimized by minimizing a composite loss, a weighted sum of the standard Cross-Entropy losses for the ASR task ($\mathcal{L}_{\text{ASR}}$) and the etiology classification task ($\mathcal{L}_{\text{CLS}}$). The total loss is defined as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ASR}} + \lambda \cdot \mathcal{L}_{\text{CLS}}, \tag{3}$$

where λ is a hyperparameter that balances the contribution of the etiology classification task relative to the primary ASR task. This *etiology_weight* (λ) was tuned as a hyperparameter during our experiments.

E. Baseline

To the best of our knowledge, no prior work has performed multi-etiology classification on the SAP dataset. Consequently, there is no established external benchmark for this task. Therefore, our evaluation is based on an internal comparison between the feature representations derived from HuBERT, Wav2vec2, and Whisper model. Furthermore, to validate the effectiveness of our proposed MTL framework, we compare the performance of the MTL-trained Whisper model against its identical counterpart in linear probing experiment, which serves as the direct baseline.

F. Evaluation Metrics

The performance of the etiology classifier is evaluated using standard classification metrics, including class-wise precision, recall, and F1-score. We also report the area under the curve (AUC) [25] to assess the classifier's discriminative power. For the auxiliary ASR task, performance is evaluated using the Word Error Rate (WER), the standard metric for this task.

IV. Results & Discussion

A. Five-Etiology Classification

Table II and Figure 4 summarize the performance of the three pretrained models on the five-etiology task. Our results indicate that self-supervised models outperform the supervised Whisper variants.

TABLE II
AVERAGE CLASSIFICATION METRICS FOR FIVE ETIOLOGIES

Model	Accuracy	Precision	Recall	F1-Score	AUC
Whisper-m	74.2%	0.58	0.74	0.64	0.55
Whisper-l	74.1%	0.56	0.74	0.64	0.55
HuBERT-l	**82.5%**	**0.85**	**0.82**	**0.83**	**0.95**
Wav2vec2-l	79.9%	0.80	0.80	0.79	0.94

Results indicate HuBERT-Large as the optimal model that achieved an overall accuracy of 82.5% and an average AUC of 0.95. Wav2vec2-Large also demonstrated strong performance, with an accuracy of 79.9% and an average AUC of 0.94. The superiority of HuBERT is consistent with previous findings on dysarthria detection tasks [12]. These results suggest that the representations learned through self-supervised objectives are highly effective at capturing the subtle, discriminative acoustic markers that differentiate between dysarthria etiologies.

In contrast, both Whisper variants attained an accuracy of approximately 74% and showed notably weaker discriminative capability, yielding an average AUC of only 0.55. These models particularly struggled to distinguish between classes such as PD and Down syndrome (DS).

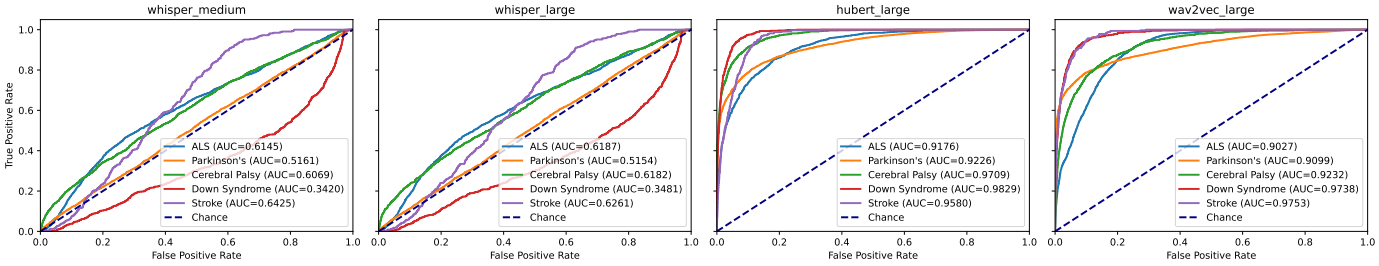

Fig. 4. ROC AUC curves for the four pre-trained models.

Whisper's under-performance is likely due to its different training objectives. It was pre-trained under a weakly supervised paradigm on paired audio-transcript data to predict the next token. This encourages its encoder to emphasize high-level lexical and semantic context rather than the fine-grained articulatory cues critical for distinguishing pathologies [18]. By contrast, self-supervised models like Wav2vec2 and HuBERT learn to reconstruct masked acoustic units without textual supervision for detailed, low-level phonetic structure that is more informative for this task [16], [17].

B. Six-Class Classification

The proposed MTL framework significantly enhanced Whisper-medium.en model's discriminative power by jointly training the model on the six-class classification and ASR. Figure 5 shows that we boosted the average AUC from a baseline of 0.55 to 0.93. This demonstrates that forcing the model to learn robust phonetic representations for the ASR task provides a good regularization effect, enabling the shared encoder to extract much more effective features for six-class classification. As a measure of auxiliary task performance, our approach achieved a WER of 17.8% on the combined test sets.

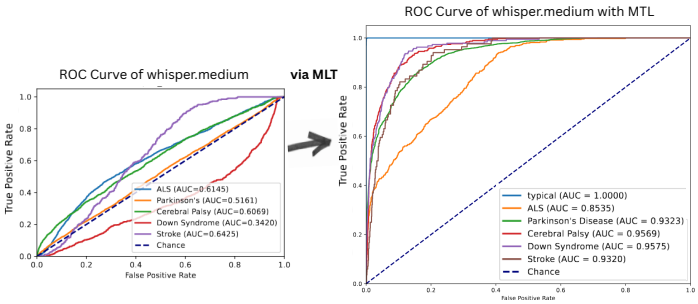

Fig. 5. ROC AUC curves for Whisper-medium.en with MTL fine-tuning

Table III summarizes the class-wise performance of the MTL framework. The model achieved near-perfect performance for the "Typical" speech class and a high F1 score of 0.90 for PD. The superior performance of PD compared to other etiologies is likely due to it being the majority class in the SAP dataset, providing more training data for the model to learn its characteristics. On the other hand, performance was considerably lower in the minority classes such as Stroke, ALS, and Cerebral Palsy. While a worse performance was expected, given the significant data imbalance in the training set, the low precision and recall scores underscore the ongoing challenge of training robust classifiers with data imbalance.

TABLE III
PER-ETIOLOGY CLASSIFICATION RESULTS FOR THE MTL FRAMEWORK

Etiology	Accuracy	Precision	Recall	F1-Score
Typical	1.00	0.99	1.00	0.99
ALS	0.43	0.55	0.42	0.47
PD	0.93	0.88	0.92	0.90
CP	0.44	0.79	0.39	0.52
DS	0.55	0.47	0.56	0.51
Stroke	0.39	0.19	0.48	0.27

V. Limitation & Future Work

We acknowledge several limitations of this study. First, we did not explicitly address the significant class imbalance. Future work could use forthcoming balanced SAP releases or apply SMOTE techniques [26] such as oversampling to handle the data imbalance. Second, our evaluation relied on internal comparisons because SAP is newly released, and this work has no baseline yet. It would be valuable to perform a cross-dataset evaluation to assess model robustness. Third, the MTL framework was applied only to Whisper. Future research could adapt this framework to HuBERT and Wav2vec2 to investigate whether their strong encoders can be enhanced even further.

VI. Conclusion

This study is the first to tackle multi-etiology dysarthria classification on SAP-1008. Our linear probing experiments demonstrated that self-supervised models, particularly HuBERT, serve as superior feature extractors for this task compared to supervised models like Whisper. Furthermore, by jointly optimizing for etiology classification and ASR via an MTL framework, we significantly elevated the performance of the Whisper model to be competitive with the top-performing self-supervised models. This work establishes a critical performance baseline and validates powerful new methods for future research in automated dysarthria assessment.

References

[1] J. R. Duffy, *Motor Speech Disorders: Substrates, Differential Diagnosis, and Management*, 4th ed. Mosby, 1995.

[2] H. P. Rowe *et al.*, "Characterizing dysarthria diversity for automatic speech recognition: A tutorial from the clinical perspective," *Frontiers in Computer Science*, vol. 4, 2022. DOI: 10.3389/fcomp.2022.770210.

[3] F. L. Darley, A. E. Aronson, and J. R. Brown, "Differential diagnostic patterns of dysarthria," *Journal of Speech and Hearing Research*, vol. 12, no. 2, pp. 246–269, 1969. DOI: 10.1044/jshr.1202.246.

[4] K. M. Yorkston, D. R. Beukelman, E. A. Strand, and M. Hakel, *Management of Motor Speech Disorders in Children and Adults*, 3rd. Pro-Ed, 2010.

[5] C. Estes and A. Johnson, "Practical considerations for instrumental acoustic and aerodynamic assessment of voice: Discussion points from an open forum of clinicians," *Perspectives of the ASHA Special Interest Groups*, vol. 8, pp. 1354–1362, 2023. DOI: 10.1044/2023_PERSP-23-00039.

[6] A. Tsanas *et al.*, "Novel speech signal processing algorithms for high-accuracy classification of parkinson's disease," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 5, pp. 1264–1271, 2012. DOI: 10.1109/TBME.2012.2183367.

[7] I. Guyon and A. Elisseeff, "An introduction of variable and feature selection," *J. Machine Learning Research Special Issue on Variable and Feature Selection*, vol. 3, pp. 1157–1182, 2003. DOI: 10.1162/153244303322753616.

[8] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012. DOI: 10.1109/MSP.2012.2205597.

[9] C. Bhat and H. Strik, "Speech technology for automatic recognition and assessment of dysarthric speech: An overview," *Journal of Speech, Language, and Hearing Research*, vol. 68, no. 2, pp. 547–577, 2025. DOI: 10.1044/2024_JSLHR-23-00740.

[10] J. Song *et al.*, "Detection and differentiation of ataxic and hypokinetic dysarthria in cerebellar ataxia and parkinsonian disorders via wave splitting and integrating neural networks," *PLoS One*, vol. 17, no. 6, e0268337, 2022. DOI: 10.1371/journal.pone.0268337.

[11] F. Javanmardi, S. R. Kadiri, and P. Alku, "Pre-trained models for detection and severity level classification of dysarthria from speech," *Speech Communication*, vol. 158, p. 103 047, 2024. DOI: 10.1016/j.specom.2024.103047.

[12] B. Sanjay, M. K. Priyadharshini, P. Vijayalakshmi, and T. Nagarajan, "Severity classification and dysarthric speech detection using self-supervised representations," in *Proceedings of the 21st ICON*, 2024, pp. 621–628. [Online]. Available: https://aclanthology.org/2024.icon-1.74/.

[13] S. Vandenhende *et al.*, "Multi-task learning with deep neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. DOI: 10.1109/tpami.2021.3054719.

[14] L. Xu, J. Liss, and V. Berisha, "Dysarthria detection based on a deep learning model with a clinically-interpretable layer," *JASA Express Letters*, vol. 3, no. 1, 2023. DOI: 10.1121/10.0016833.

[15] Y. Xiong, V. Berisha, J. Liss, and C. Chakrabarti, "Improving speech-based dysarthria detection using multi-task learning with gradient projection," *Interspeech 2022*, pp. 902–906, 2024. DOI: 10.21437/interspeech.2024-1563.

[16] W.-N. Hsu *et al.*, "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. DOI: 10.1109/TASLP.2021.3122891.

[17] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Neural Information Processing Systems*, pp. 12 449–12 460, 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf.

[18] A. Radford *et al.*, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the ICML*, vol. 202, 2023, pp. 28 492–28 518.

[19] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, 2015, pp. 5206–5210. DOI: 10.1109/ICASSP.2015.7178964.

[20] M. Hasegawa-Johnson *et al.*, "Community-supported shared infrastructure in support of speech accessibility," *Journal of Speech, Language, and Hearing Research*, pp. 1–14, 2024. DOI: 10.1044/2024_jslhr-24-00122.

[21] H. Kim *et al.*, "Dysarthric speech database for universal access research," in *Proceedings of Interspeech*, 2008. DOI: 10.21437/interspeech.2008-480.

[22] 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, no. 4, pp. 523–541, 2011. DOI: 10.1007/s10579-011-9145-0.

[23] T. Wolf *et al.*, *Huggingface's transformers: State-of-the-art natural language processing*, 2020. [Online]. Available: https://arxiv.org/abs/1910.03771.

[24] D. Chakrabarti, "Robust linear classification from limited training data," *Machine Learning*, vol. 111, no. 4, p. 1621, 2022. DOI: 10.1007/s10994-021-06093-5.

[25] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern Recognit.*, vol. 30, pp. 1145–1159, 1997.

[26] W. Chen *et al.*, "A survey on imbalanced learning: Latest research, applications and future directions," *Artificial Intelligence Review*, vol. 57, no. 6, May 2024, Art. no. 137. DOI: 10.1007/s10462-024-10759-6.