Title: Automatic speech analysis for dysarthria recognition and ALS diagnosis

G027 (s2693044, s2716530, s2748897)

Abstract

Amyotrophic Lateral Sclerosis (ALS) is a severe neurodegenerative illness that affects motor and speech functions, often presenting dysarthria — impaired vocal articulation due to weakened muscular control — as an early symptom. This research examines the potential of audiometric biomarkers to uncover ALS' progression exploring both Machine Learning (ML) and Artificial Intelligence (AI) techniques.

Two data sources are used: the TORGO database, which contains a wide range of dysarthric and healthy voice samples, and the VOC-ALS dataset, comprising structured clinical and acoustic information on ALS patients and control individuals, together with the corresponding ensemble of scripted audio recordings.

The primary objective is to merge the results of these studies and build a detection system able to identify dysarthria and, eventually, distinguish the cases caused by sclerosis. A state-of-the-art deep speech representation strategy known as Wav2Vec 2.0 is fine-tuned on the TORGO instances to recognise patterns in vocal deterioration and spot dysarthric voices. Afterwards, different AI methods are used on the ALS-labelled dataset to create a disease-specific detector based on predetermined vocalisations. This approach has the potential to support early diagnosis and monitoring in clinical and remote settings.

1. Introduction

The fatal, adult-onset neurodegenerative disease known as Amyotrophic Lateral Sclerosis (ALS) primarily damages the motor neurones controlling voluntary muscles. These include the lower motor nerves, extending from the spinal cord to the skeletal muscles (Brown & Al-Chalabi, 2017), and the upper motor nerves, which start in the cerebral cortex and reach the spinal cord. Degeneration of these neurones causes progressive muscle weakness and loss of coordination, eventually leading to full paralysis. Although early stages usually spare sensory, psychological, and autonomic functions, recent studies indicate that a subset of patients may also show minor cognitive or behavioural deficits, especially those associated with frontotemporal dementia (Hardiman et al., 2017). Ultimately, the gradual weakening of diaphragm and intercostal muscles causes most people to die from respiratory failure.

ALS occurs slightly more in men than in women and affects about 2 to 3 individually per 100,000 annually (Brown & Al-Chalabi, 2017). Though patient progression rates differ greatly, the median survival period following diagnosis is usually 2 to 5 years, with the average age of start between 55 and 70 years. (Talbott et al., 2016).

Gradual bulbar dysfunction (i.e., problems with speech, chewing, and swallowing) is a classic sign of ALS. This is caused by a degeneration of the lower motor neurones in the brainstem, specifically within the *medulla oblongata* — or bulbar region — which innervates the muscles of face, tongue, mouth, and larynx. Among bulbar-related symptoms, dysarthria is one of the earliest and most frequent, affecting up to 80% of ALS patients over the course of the disease (Yunusova et al., 2016). The reason is a weakness and impaired coordination of the articulatory musculature; ALS-provoked dysarthria is indeed recognisable by inaccurate speech, hypernasal pitch, diminished loudness, and strained or breathy voice quality (Darley et al., 1969). Compared to spinal-onset ALS, which starts with limb weakening (Chiò et al., 2009), the bulbar variant is linked with faster disease progression and poorer prognosis. In he latter ALS case, hindered articulation is therefore not only representing a functional communication loss, but might also act as a signal of more aggressive disease subtypes.

Speech is an especially sensitive indicator of neuromuscular health because it relies on the rapid, finely coordinated activity of multiple muscle groups — including the tongue, lips, jaw, larynx, and respiratory system ones. By consequence, even minor disruptions in neurone signalling can significantly impact speech acoustics and articulation (Yunusova et al., 2016). These alterations generally precede gross motor symptoms, and their recognition is lowly dependent on patient prejudice or physician subjectivity, making them potential markers for objective, continuous disease monitoring (Green et al., 2013). Most importantly, vocal samples can be collected non-invasively using accessible tools such as microphones or smartphones, offering a low-cost, scalable, and reproducible solution for both clinical and home-based screening. This approach is especially valuable for tracking ALS in situations where access to clinical facilities is limited or regular visits are not feasible. Thus, recognising and measuring speech impairments can become crucial for early diagnosis, classification of ALS subtypes, and better studying the disease progression.

2. Objectives

The primary aim of this research is to investigate the viability of using Machine Learning (ML) techniques to uncover and track ALS-induced dysarthria. In particular, the study seeks to enhance early recognition of bulbar-onset ALS and monitor disease progression by analysing speech characteristics. Building upon previous studies in automated vocal analysis (Green et al., 2013), the final objective is to produce an advance in the field of automated diagnosis, by integrating two complementary datasets (see Section 3) to construct a comprehensive detection pipeline.

Wav2Vec 2.0 — a self-supervised Deep Learning (DL) model for speech representation (Baevski et al., 2020) — is applied to classify dysarthric speech in ALS patients, using a diverse set of clinically annotated samples for training and validation (see Section 5.1). While this architecture has been successfully applied in general speech recognition and some pathologies-related studies (wen Yang et al., 2021; Bar et al., 2022), its applicability to ALS-caused dysarthria remains unexplored. By leveraging this advanced feature extraction, the model is expected to achieve an improved disease detection performance. In fact, empowering vocal analysis through a properly trained Deep Neural Network (DNN) might allow to distinguish dysarthric speech even with relatively short, unstructured, and noisy input recordings (Rudzicz et al., 2012), offering a robust and preventive tool to assess early voice impairment.

In parallel, this research complements symptom recognition through generalised audio-embeddings with a more composite clinical analysis. This is done by applying AI models to a series of specifically prescribed vocalisation, using a set of predefined audiometric features extracted from the recordings (see Section 5.2). So, a more structured and interpretable disease classification approach is explored, leveraging the acquisition of clinically recognised vocal biomarkers and experts' annotations.

By comparing and potentially combining these two methodologies, the study aims to develop a dual-model pipeline capable of both generalised data-driven symptoms detection and domain-informed diagnosis. The conclusive aim is thus to evaluate the feasibility of a scalable, automated system for speech-based ALS monitoring, considering the results obtained from the produced classifiers in terms of accuracy (see Section 5). Given that speech is a non-invasive and easily accessible biomarker, a thorough ML-driven assessment tool could greatly benefit early disease recognition and patient stratification (Green et al., 2013). These advancements would not only assist clinical professionals in refining diagnostic workflows but also provide a cost-effective alternative for home-based severity progression tracking (Helleman et al., 2020). In summary, this work aims to bridge the gap between state-of-the-art speech recognition techniques and neurodegenerative disease monitoring, with a particular focus on advancing computational pathology detection for ALS and its revealing symptom.

3. Data

Two publicly available data sources are promising for the just described examination of speech as a biomarker for neurological diseases: the TORGO database (Rudzicz et al., 2012) and the VOC-ALS dataset (et al., 2024).

3.1. TORGO

The "TORGO database of acoustic and articulatory speech from speakers with dysarthria" contains 2,000 vocal instances, distributed evenly among dysarthric males, dysarthric females, non-dysarthric males, and non-dysarthric females. Originally created to advance research in Automatic Speech Recognition (ASR) for individuals with vocal impairments, the dataset includes hindered speech samples stemming from various conditions, not limited to ALS (e.g., Cerebral Palsy). Rather than identifying the specific medical cause, it classifies speakers broadly as either dysarthric or non-dysarthric. As such, the labelling focuses on distinguishing impaired vs healthy speech only, without considering the underlying pathological condition. In this setting, a speech-recognition Deep Neural Network is unquestionably the best option to handle a symptoms detection task, given the information's low level of structure, which consists of recordings of variegated vocalisations (e.g., phoneme repetitions, short words, restricted and unrestricted sentences) from a group of subjects categorised solely by gender and healthiness (Rudzicz et al., 2012).

3.2. VOC-ALS

Group	Female	Male	Total (F+M)
ALS patients	37 *24.2%*	65 *42.5%*	102 *66.7%*
Healthy Controls	19 *12.4%*	32 *20.9%*	51 *33.3%*
Total (ALS + HC)	56 *36.6%*	97 *63.4%*	153 *100%*

Table 1. Distribution of VOC-ALS experiment's subjects.

The "VOiCe signals acquired in Amyotrophic Lateral Sclerosis patients and healthy controls" (VOC-ALS) dataset includes clinical annotations alongside acoustically derived features extracted from vocal recordings of ALS patients and healthy controls (et al., 2024). Participants in the experiment were instructed to perform multiple specific vocal tasks: repeating syllables (i.e., */ka/*, */pa/*, */ta/*), and sustaining vowel phonations. The dataset includes a total of 1,224 audio samples from 153 Italian-speaking individuals — average age: 63 — comprising 102 ALS patients with varying levels of dysarthria severity and 51 healthy controls (see Table 1). These recordings were used to determine the following acoustic features (et al., 2024).

Fundamental Frequency (F_0) (Bäckström et al., 2022) Defined as the lowest frequency of a periodic signal, in the context of human speech it represents the frequency at which the vocal cords vibrate to produce sound. Higher F_0 are heard as higher in pitch and vice versa.

Subsequently, the mean fundamental frequency ($\overline{F_0}$) (Andrade et al., 2020) would represent the average pitch of a speaker over a given vocal segment, calculated as the arithmetic mean of all measured F_0 values:

$$\overline{F_0} = \frac{1}{N}\sum_{i=1}^{N} F_{0,i} \; ; \tag{1}$$

where:

- $F_{0,i}$ is the fundamental frequency at frame i,
- N is the total number of frames considered.

The standard deviation of fundamental frequency (σ_{F_0}) (Andrade et al., 2020) quantifies instead the spread of F_0 values around the mean, being an important measure of prosodic variation in speech by reflecting how much the pitch fluctuates over time. It is calculated as:

$$\sigma_{F_0} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(F_{0,i} - \overline{F_0})^2} \; ; \tag{2}$$

where:

- $F_{0,i}$ is the fundamental frequency at frame i,
- $\overline{F_0}$ is the mean fundamental frequency,
- N is the total number of frames.

Harmonics-to-Noise Ratio (HNR) (Fernandes et al., 2018)
The Harmonics-to-Noise Ratio quantifies the relationship between the periodic and aperiodic (i.e., harmonic and noise, respectively) components of a speech signal. It is expressed in decibels (dB) and serves as indicator of voice quality. An higher HNR indicate a cleaner, harmonic-dominated and regular voice, such that lower values are typically associated with vocal disorders. It is defined as:

$$\text{HNR} = 10 \cdot \log_{10}\left(\frac{r'_x(\tau_{\max})}{1 - r'_x(\tau_{\max})}\right) \; ; \tag{3}$$

where:

- $r'_x(\tau_{\max})$ is the normalised autocorrelation value at the first local maximum (excluding $\tau = 0$),
- $\tau_{\max}$ corresponds to the time lag at which the first significant periodic peak occurs.

Local jitter (Jitt) (Teixeira et al., 2013)
The local jitter represents the average absolute time difference between two consecutive periods (i.e., jitta), normalised by the average period. A Jitt beyond 1.04% may indicate the presence of pathologies. It is computed as:

$$\text{Jitt} = \frac{\text{jitta}}{\frac{1}{N}\sum_{i=1}^{N} T_i} \text{, with} \quad \text{jitta} = \frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i-1}| \; ; \tag{4}$$

where:

- T_i is the duration in seconds for each period,
- N is the number of periods.

Local shimmer (Shim) (Teixeira et al., 2013)
The shimmer measures the average absolute variation in Amplitude (A) between two consecutive voice periods, normalised by the average A. It is used as an indicator of vocal stability too, as values above 3.81% may suggest a clinical condition. It is calculated as:

$$\text{Shim} = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|A_i - A_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \; ; \tag{5}$$

where:

- A_i is the peak amplitude of the i^{th} pitch period,
- N is the number of periods.

The dataset also includes a cumulative ALS Functional Rating Scale Revised (**ALSFRS-R**) score (Cedarbaum et al., 1999) for each subject, consisting of the state-of-the-art clinical assessment of dysarthria severity, based on a perceptual rating scale spanning from 48 (indicating health) to 0. The disease progression rate is thus computable as (Kimura et al., 2006):

$$\text{Progr. Rate} = \frac{48 - \text{ALSFRS-R}_{\text{\,at examination time}}}{\text{Disease Duration}_{\text{\,in months}}} \; . \tag{6}$$

This Functional Rating Scale is a commonly used clinical metric, but it relies heavily on a subjective assessment of patients' symptoms. It is calculated by adding up the responses to a 12-item neurologist-administered questionnaire, gauging the impairment levels of specific functions (e.g., speech, salivation, swallowing, respiration), each scored on a 0-to-4 scale. Therefore, this metric lacks sensitivity to detect objective subclinical changes in the bulbar motor system and reliably classify patients based on problematicness (Dubbioso et al., 2024). Accurately predicting functional decline is essential for timely clinical intervention, as early decisions regarding communication support and palliative care have the greatest impact when made proactively. In this context, voice-based diagnosis has been highlighted as a promising approach due to its low cost, ease of use, and potential for early detection and continuous monitoring — factors that are critical for improving ALS prognosis (Green et al., 2013).
In fact, the dataset at issue is very suitable for training an AI model able to either act a classification or even a regression on the ALS presence and level, based on a set of raw audio files containing predetermined vocal performances through the extraction of salient acoustic characteristics (Schindler & Gullì, 2012). To unambiguously quantify the disease, a target *Severity* score was additionally allocated to each patient, obtained exactly reverting the ALSFRS-R value on its scale; thus, 0 would be awarded to the healthy patient, while 48 would reflect the deepest level of ALS:

$$\textit{Severity} = 48 - \text{ALSFRS-R} \; . \tag{7}$$

4. Experiments

Two parallel experimental procedures were carried out (as outlined in Figure 1): the primary analysis investigated the potential of Wav2Vec 2.0 for dysarthria detection based on raw speech recordings, using the TORGO dataset; the secondary one focused on AI-predicting ALS severity starting from patients' audio data structured as in the VOC-ALS study, through determined pre-computed acoustic features.

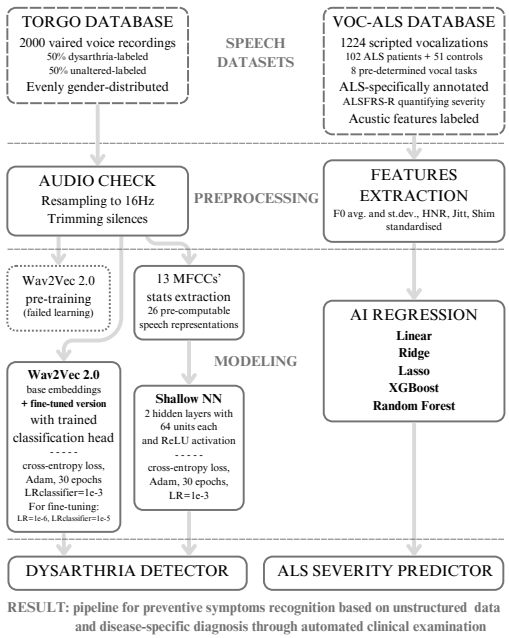


Figure 1. Methodology flowchart.

4.1. Wav2Vec 2.0 for dysarthria detection

Being a powerful self-supervised speech representation model, Wav2Vec 2.0 was used to extract meaningful voice embeddings from the TORGO's audio files, which were then used in a deep learning-based classification pipeline to detect the presence of dysarthria (Baevski et al., 2020). The approach aimed to create an automated method for detecting the relevant symptoms based on a virtually small set of unscripted voice recordings of an individual.

To ensure compatibility and uniformity, some preprocessing steps were applied to the dataset: all audio files were resampled to a standard rate of 16 kHz and trimmed to remove leading and trailing silences, reducing irrelevant acoustic noise. For the classification task, the file paths were standardised into consistent recording-pointing variables, ensuring accurate class-coded referencing within the dataset directory. A validation check was then performed to confirm the existence and usability of each file for training purposes, revealing that five instances were unreadable (i.e., too short for the model to be loaded) and therefore to be excluded from further analysis. The dysarthria labels were encoded into binary format, with *1* representing dysarthric speech and *0* denoting healthy controls' recordings.

Initially, a pre-training experiment was attempted, aiming to apply Wave2Vec 2.0 from inception on the dataset in question for preliminary learning. This allowed to gauge the model's capacity of generating more task-representative embeddings based only on this clinically-biased recording types — possibly characterised by short duration and symptoms-related sounds pronounced. This strategy was inspired by prior research that illustrated the capacity of self-supervised models to effectively adapt to limited-label domains, such as pathological speech (Schneider et al., 2019). The hypothesis was that, if successfully learning, an intrinsically specific model could outperform the publicly available similar architectures pre-trained for general-purpose on wider and diverse corpora. In order to further improve data consistency, all audio files were converted to mono and normalized to reduce loudness variations, in addition to resampling. The majority of instances were left available for the learning task, as approximately 2% of the samples only were reserved for validation due to the limited dataset size. The pre-training was launched on GPU cluster, utilizing a slightly modified version of the public fairseq Wav2Vec 2.0 repository (Facebook AI Research, 2021), and terminated prematurely after the interim outputs were sufficient to assess this approach's efficacy.

Metric	Train	Valid
accuracy	0.359	0.605
loss	6.658	3.242
perplexity	126.5	3.6
tokens p.b.	229.1	205.5
sentences p.b.	3.4	3.2
num updates	133613	
prob perplexity	639.976	
Learning Rate	3.62×10^{-4}	
Gradient Norm	0.001	

Table 2. Wave2Vec TORGO-based pre-training stats, epoch 234.

By analysing the values of the last faced pre-training epoch (see Table 2), it was indeed noticeable that the key indicators (i.e., 35.9% accuracy, 6.658 loss) highlighted a quite poor performance in learning robust speech representations, with the validation scores expressing a slightly better generalisation on unseen data. The fundamental explanation for a similarly high perplexity is a failure of the method's application on such a small dataset: the model scarcely predicted the probability distributions over tokens, demonstrating difficulties in uncovering dysarthric speech based on the relevant brief ~2K audios only. This matches the prior research findings highlighting Wav2Vec's struggle in training from inception on relatively small datasets (Schneider et al., 2019). Moreover, the Learning Rate and Gradient Norm at the relevant step indicated small coefficients updates, suggesting that the training had plateaued; the loss value was indeed not lowering along epochs anymore, confirming a stuck optimisation (Smith, 2017).

In summary, the findings demonstrated the necessity of using an already pre-trained Wav2Vec 2.0 model, which could then be fine-tuned for the dysarthria detection task (Riviere et al., 2020). A robust speech feature extraction was needed as a base to subsequently reach the set objective, while trying to directly focus the training on an insufficient dataset did not allow to consistently capture any essential phonetic and acoustic variation. The resulting performance also suggested that even the application of data augmentation strategies would have been unlikely to compensate for the dataset's inherent limitations. These observations are consistent with prior work in speech representation learning (Baevski et al., 2020), which emphasises the importance of transfer learning and self-supervised pre-training in handling domain-specific tasks with scarce labelled data.

To retain the desirable generalisation performance, the base **Wav2Vec 2.0** pre-trained on 960 hours of speech (Facebook AI Research, 2021) was **fine-tuned** on the TORGO dataset. This aimed to obtain task-relevant vocal representations while avoiding the instability observed when learning from scratch. Audio files were loaded and converted into waveforms, which were then passed through the model to extract embeddings from the final hidden state. These embeddings were mean-pooled across the time dimension, resulting in a 768-dimensional feature vector per sample. The dataset was split for model evaluation into training, validation and test subsets (80%, 10% and 10% of the instances respectively), with a balanced distribution of dysarthric and unaltered speech samples. The model was trained for 30 epochs using cross-entropy loss and the Adam optimizer.

In order to have a direct comparison of the obtained architecture to a **frozen-Wave2Vec**-based dysarthria detector, another model version inspired by previous applications (wen Yang et al., 2021) was trained freezing all the voice-embedding layers, allowing learning only for the appended classification head, which would therefore take as input unchanged embeddings from *wav2vec2-base-960h*. The latter type of model has demonstrated good performance in low-resource speech tasks and has recently been applied to clinical audio domains (Baevski et al., 2020).

An additional ML-based classifier was implemented using the Mel-Frequency Cepstral Coefficients (**MFCCs**) — traditional representations in speech processing to capture the power spectrum of audio signals in a perceptually meaningful way (Davis & Mermelstein, 1980). In particular, 13 MFCCs were extracted from the recordings, and their mean and standard deviation computed across time, producing a 26-dimensional feature vector per sample. The architecture chosen consisted of a shallow feedforward **Neural Net** with two hidden layers of 64 units each, using ReLU activations and a final linear output layer for binary classification, as in literature-relevant approaches (Eyben et al., 2016). The data loading was enacted in mini-batches to reduce memory overhead, and the training based on cross-entropy loss minimization through Adam optimizer over 30 epochs.

4.2. AI-fostered ALS diagnosis

Given the highly structured nature of the clinically annotated VOC-ALS dataset, various standard AI methods were evaluated for the task of regressing the *Severity* score, using a set of predefined acoustic features (see Section 3.2) extracted from the audio recordings (Dubbioso et al., 2024). However, the automatic computation of these audio variables was tested too, by adapting the code published for the underlying research (Sannino, 2024) in order to ensure full reproducibility of the present paper results, making the repository associated with this analysis suitable for ALS detection starting directly from a properly scripted set of raw audio recordings.

Specifically, for each of the eight main vocal tasks prescribed to participants (i.e., sustaining the five vowels and repeating the syllables */ka/*, */pa/* and */ta/*), five metrics were considered and used as input for the following models: $\overline{F_0}$, σ_{F_0}, HNR, Shim, Jitt; for a total of 40 features. To ensure consistent scaling, these predictors were standardised to zero mean and unit variance (Ioffe & Szegedy, 2015; Eyben et al., 2016). The dataset was split for model evaluation into training, validation and test subsets (80%, 10% and 10% of the instances, respectively) with a balanced distribution of ALS patients and healthy controls.

Linear Regression was used as a baseline model; while highly interpretable, its performance is deemed limited with non-linear underlying relationship patterns between the features and the target variable (Seber & Lee, 2012) .

Ridge Regression — regularised version of linear regression — introduces L2 penalty to reduce the impact of multicollinearity, avoiding overfits (Hoerl & Kennard, 1970).

Lasso Regression adds an L1 penalty instead, not only regularising but also performing feature selection by shrinking some coefficients to exactly zero (Tibshirani, 1996).

XGBoost Regressor is a gradient boosting framework optimised for speed and performance. It builds trees sequentially, where each new tree corrects the errors of its predecessors and provides robust feature importance scores based on gain and frequency (Chen & Guestrin, 2016).

Random Forest Regressor is an ensemble learning technique that builds multiple decision trees and averages their predictions, effectively capturing non-linear relationships. It also provides an inherent feature importance evaluation by measuring how much each feature contributes to reduce the prediction error (i.e., Impurity) across the trees (Breiman, 2001). This model was also cross-validated to ensure reliable performance and generalisation across the dataset (Hastie et al., 2009). Feature importance was then assessed to identify which features had the greatest impact on predicting ALS severity, based on their contribution to reducing the loss during training.

5. Results

5.1. Wav2Vec 2.0 for dysarthria detection

Considering the progression of the **fine-tuned Wav2Vec** model (see Figures 2 and 3), the training accuracy curve exhibits a smooth, upward trajectory, increasing steadily from 50% to over 95% across 30 epochs. Most importantly, the validation accuracy closely follows this trend, showing no significant divergence — a strong indicator of robust generalisation. The corresponding loss curves further support this conclusion: both training and validation values consistently decrease, with validation loss dropping from 0.65 to below 0.20 by epoch 25 and flattening thereafter. There are no sharp spikes or irregularities, suggesting that the model is learning effectively and stably. This training behaviour is characteristic of well-regularised transfer learning, where the pre-trained feature extractor is effectively fine-tuned to identify task-specific patterns (Riviere et al., 2020).
Notably, despite more coefficients were optimizable in the above model, the two different and lower LRs (i.e., 10^{-6} and 10^{-5} for embedder and classifier respectively) ensured the learning power at inception to be comparable with the frozen model's one (LR of 10^{-3}). Moreover, higher LRs were attempted for the former model, making its already huge architecture producing a less stable learning. By contrast, tuning the frozen-embeddings detector's LR produced results similar to the presented ones: actual regularization forms might be necessary for performance improvement.

At equal epochs' span, the **frozen Wav2Vec** model (see Figures 4 and 5) displays signs of overfitting and under-adaptation. While training accuracy increases steadily to ~90%, the validation accuracy plateaus early around epoch 10 and oscillates between 78% and 83%, never surpassing 84%. This suggests that the classifier head is memorising training examples but failing to capture generalisable speech patterns from the frozen pre-determined embeddings. The loss curves further reveal this issue: training values decrease rapidly, but validation ones plateau above 0.30 with minor fluctuations, indicating no continued learning on unseen data. This instability highlights the limitations of using Wav2Vec 2.0 without fine-tuning, particularly in clinical settings where subtle speech abnormalities must be captured for accurate classification (Tripathi et al., 2020).

The **MFCCs-based NN** (see Figures 6 and 7) shows a decoupled performance too: training accuracy rises above 98%, but validation values peak near 94% with visible oscillation. Similarly, training loss drops, while validation one stabilises around 0.40 after epoch 20. This behaviour suggests that although MFCCs provide useful descriptors in general, they lack sufficient specificity and dimensionality for nuanced dysarthria detection. Thus, the NN classificator overfit them soon, pushing to derive significant patterns and reach an higher accuracy on the train instances only. Previous studies found indeed that MFCCs may not match the representational richness of task-optimised deep embeddings (Dubbioso et al., 2024; Schindler & Gullì, 2012).

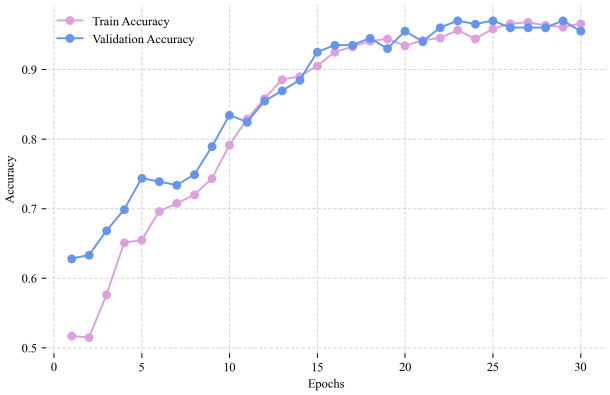


Figure 2. Fine-tuned Wav2vec 2.0 classifier's Accuracy curves.

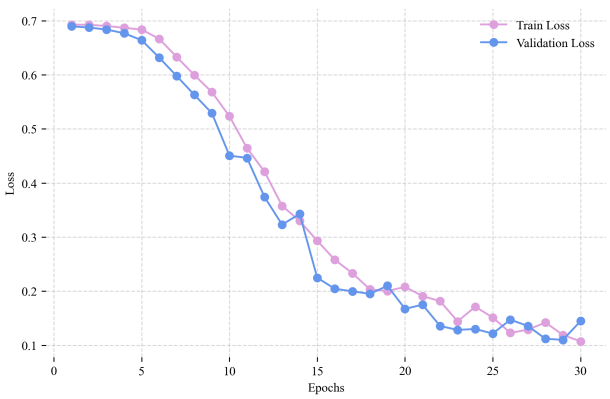


Figure 3. Fine-tuned Wav2vec 2.0 classifier's Loss curves.

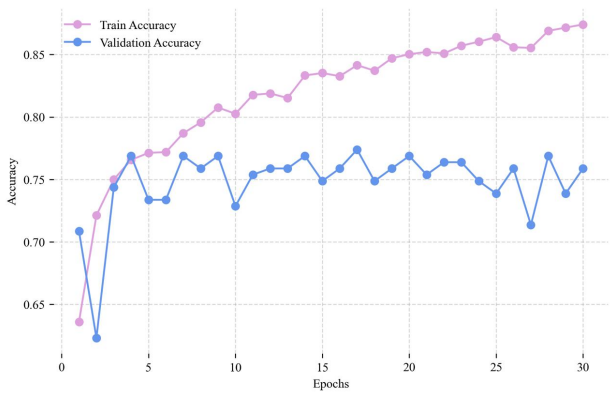


Figure 4. Wav2vec 2.0 classifier's Accuracy curves.

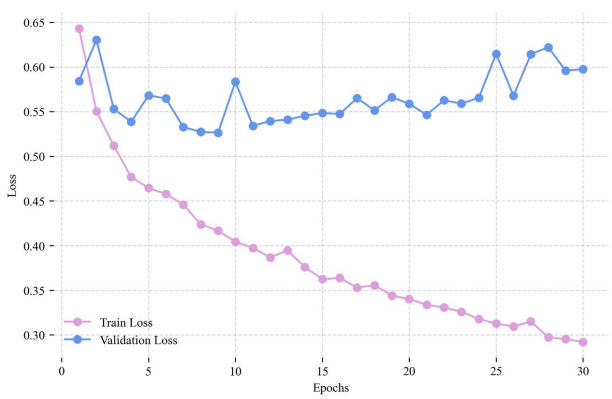


Figure 5. Wav2vec 2.0 classifier's Loss curves.

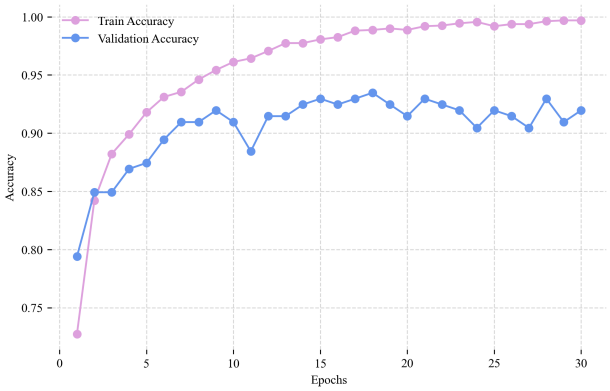


Figure 6. MFCCs-based NN classifier's Accuracy curves.

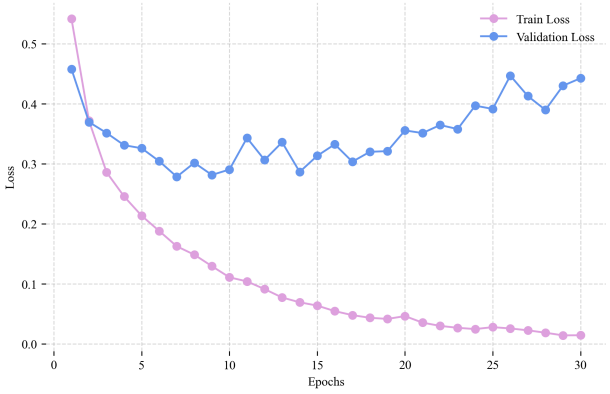


Figure 7. MFCCs-based NN classifier's Loss curves.

Considering test performances at epoch 30 (see Table 3), the fine-tuned Wav2Vec 2.0 classificator achieves an F1-score of 0.97, outperforming both the frozen variant and the MFCCs-based model. The latter reaches an F1 of about 0.95, yet its higher validation loss and oscillating accuracy indicate slightly weaker robustness; this supports previous evidence suggesting that while handcrafted features are effective, deep learnt embeddings better capture complex, task-specific vocal patterns in neurodegenerative diseases (Dubbioso et al., 2024).

Model	**Precision**	**Recall**	**F1-score**
Fine-tuned Wav2Vec	0.94	0.99	0.97
Frozen Wav2Vec	0.80	0.82	0.81
MFCCs-based NN	0.94	0.95	0.95

Table 3. Dysarthria detectors' performances on test set, epoch 30.

Predicted ⇒	Fine-tuned Wav2Vec		Frozen Wav2Vec		MFCCs-based Neural Net	
	D	**H**	**D**	**H**	**D**	**H**
Dysarthria	94	6	80	20	94	6
Healthy	1	99	18	82	5	95

Table 4. Dysarthria detectors' test confusion matrices, epoch 30.

Further insight comes from the confusion matrices over the test instances (see Table 4): whereas the frozen model and MFCCs-based NN show more wrong predictions, the fine-tuned Wav2Vec 2.0 misclassifies only 7 of the 200 samples. The latter detector shows the best performance particularly in uncovering symptoms' absence, displaying a single false positive (i.e., recall of 99%). Despite in clinical diagnosis contexts the precision is definitely crucial (e.g., missing early signs of dysarthria can delay ALS intervention) (Yunusova et al., 2016), also accurately recognizing negatives is fundamental to keep automated screening tools trusted and prescribe further examinations to actual ill subjects (Green et al., 2013). On the other hand, the MFCCs-based NN produces 5 false positives, meaning that its non-negligible accuracy disadvantage and higher loss are lowly related with the performance on truly dysarthric samples. Finally, the frozen Wav2Vec 2.0 model exhibits a much higher number of misclassifications (19%), struggling to distinguish the instances and demonstrating its limited adaptability due to a missing feature-level specialization.

5.2. AI-fostered ALS diagnosis

Analysing the results of the *Severity*-score prediction (see Table 5), the Random Forest Regressor is highlighted as best performing overall, with a MAE of ~6.8 and a MSE of ~57.1 on the test set, indicating the most coherently low predictions' distancing from the target. These values are promising, given the already discussed (see Section 3.2) subjectivity and limited resolution of the ALSFRS-R scale (Green et al., 2013). Several studies have indeed suggested that ALS-level estimation, even when not exact, offers valuable clinical stratification (Schindler & Gullì, 2012). Linear, Ridge, and Lasso regression models show poorer generalisation, likely due to their inability to capture non-linear relationships in the acoustic feature space; XGBoost performs exceptionally well on training data, but significantly underperforms on validation and test sets, demonstrating an even higher overfitting of the relatively small dataset.

Model	**Metric**	**Train**	**Val**	**Test**
Linear Regression	MAE	5.504	8.037	9.528
	MSE	46.603	122.087	170.694
Ridge Regression	MAE	5.615	6.888	8.570
	MSE	47.173	93.473	136.781
Lasso Regression	MAE	5.961	6.156	7.694
	MSE	51.386	70.079	110.249
XGBoost Regressor	MAE	0.002	6.579	7.372
	MSE	0.000	71.909	70.570
Random Forest R.	MAE	3.984	5.459	6.834
	MSE	20.999	44.825	57.128

Table 5. Severity-regressors' performances across data subsets.

Random Forest's stability is confirmed via cross-validation (see Table 6), where MAE averages ~7.3 and RMSE ~8.8, Although modest, the R^2 of 0.125 is within range of previous work and indicates potential for disease's progression tracking, especially if *Severity* is divided into coarse bins (e.g., mild, moderate, severe) rather than treated as an exact score (Green et al., 2013; Dubbioso et al., 2024).

Metric	Value
Mean Absolute Error (MAE)	7.281
Mean Squared Error (MSE)	77.023
Root Mean Squared Error (RMSE)	8.776
R^2 Score	0.125

Table 6. Random Forest model's cross-validation results.

To interpret the Random Forest's behaviour, features' importances were extracted (see Appendix A), revealing the prominence of a varied set of fundamental frequency, Jitt and Shim variables for the vowel vocalizations. In fact, their acoustic consistency and physiological load make them especially sensitive to early neuromotor deterioration, enabling models to detect ALS severity variations even in the absence of visible physical symptoms. Syllables-repeating samples' σ_{F_0} and the HNR for the */i/* phonation were of salient predictive power too. Overall, this distribution mirrors the clinically-established scheme of markers for bulbar dysfunction and muscular control loss (Yunusova et al., 2016; Schindler & Gullì, 2012), reinforcing the relevance of VOC-ALS study's pre-determined vocal tasks.

6. Conclusions

This project explored how speech analysis can contribute to the early detection of dysarthria and the ALS progression monitoring through Machine Learning. By combining insights from two datasets — TORGO, focusing on general dysarthric speech, and VOC-ALS, featuring structured clinical annotations — the present study developed two main approaches: a main examination of raw voice recordings using Wav2Vec 2.0, and a secondary analysis based on pre-computed acoustic biomarkers specific to ALS severity.

Pre-trained on general speech and further adapted to TORGO dataset, the fine-tuned Wav2Vec 2.0 demonstrated high reliability for **ML dysarthric speech detection**, reaching a 96.5% accuracy and an F1-score above 0.96. The latter model thus outperformed both its non-specialized counterpart and a standard NN-classificator applied on a sector-relevant but non-specialized MFCCs-based representation, with particular success in reducing the false positives predictions. This confirms the feasibility of a top-to-bottom detection system operating directly on unstructured audio files without manual feature engineering (Baevski et al., 2020), demonstrating significant ability in detecting dysarthria even in short and noisy recordings. Therefore, this type of system would be especially suitable for generalised screening and preventive symptoms spotting, potentially acting as an early alarm to prompt further clinical investigations (Yunusova et al., 2016; Green et al., 2013). Notably, early attempts to pre-train Wav2Vec on the relatively limited clinical dataset failed to produce effective representations, confirming that large-scale pretraining remains essential for deep speech models, leaning on self-supervised learning paired with subsequent fine-tuning (Schneider et al., 2019).

In parallel, a set of structured features (i.e., $\overline{F_0}$, σ_{F_0}, HNR, Jitt and Shim) were used to train several disease-specific **regression models targeting ALS severity**. Among them, the Random Forest performed best, achieving a test MAE of 6.83 and a cross-validation R^2 of 0.125, consistent for the task. To ground the predictions in a clinical context, the ALS level was measured by an ALSFRS-R-derived *Severity* score, so a *Progression Rate* (see Equation 6) is derivable to quantify the disease advancement. These results support Random Forest models' ability to capture underlying non-linear relationships between vocal features and sclerosis' magnitude, making them a strong candidate for automated, lightweight and explainable ALS monitoring (Dubbioso et al., 2024; Schindler & Gullì, 2012; Hastie et al., 2009).

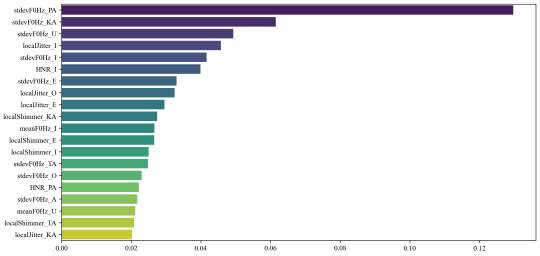
6.1. Future work

This study adds to the expanding field of speech-based digital biomarkers for neurological disorders, building on previous work that highlights vocal impairments as early signs of motor neuron degeneration (Green et al., 2013; Yunusova et al., 2016). Tools like Wav2Vec (Bar et al., 2022; Baevski et al., 2020) and MFCCs-based models (Eyben et al., 2016) have already shown promise in detecting speech impairments; by combining raw audio analysis and structured acoustic features, this approach strikes a balance between cutting-edge ML and clinical interpretability.

Developing a hybrid system — using deep, flexible embeddings alongside experts-outlined, interpretable features — is a possible path forward. Including longitudinal speech data so that models can track voice changes over time is another key goal. This time dimension could assist in capturing illness progression at earlier stages, potentially even before symptoms are visible to patients or doctors (Luz, 2017). Longitudinal modelling also provides personalized health monitoring, increasingly stressed in neurodegenerative disease treatment. Increasing the demographic variety of the training data sets is also very important. Recent evaluations of commercial speech recognition systems reveal that training data often underrepresents diverse demographic groups, leading to significant performance discrepancies across accents, age groups, and genders (Fu et al., 2023).

Using these technologies in non-clinical settings like mobile apps or telehealth platforms could improve their accessibility and clinical efficacy. Home-based voice monitoring would allow patients greater autonomy in tracking symptoms and assist prompt interventions, especially in underserved or rural locations (Adams et al., 2020; De Marchi et al., 2021). As computational models improve and datasets grow in scale and diversity, the vision of using speech as a non-invasive, cost-effective biomarker for ALS and related disorders has become increasingly attainable.

A. **Random Forest model's predictive power of features.**



References

Adams, Jamie, Myers, Taylor, Waddell, Emma, Spear, Kelsey, and Schneider, Ruth. Telemedicine: a valuable tool in neurodegenerative diseases. *Current Geriatrics Reports*, 9, 06 2020. doi: 10.1007/s13670-020-00311-z.

Andrade, Bárbara Moreira, Ferreira, Ana Paula, de Oliveira Lemos, Izabella, Moreira, Luana, and Almeida, Gustavo. Acoustic parameters for the evaluation of voice quality in patients with voice disorders. *ResearchGate Preprint*, 2020. URL https://www.researchgate.net/figure/Fundamental-frequency-F-0-standard-deviation-of-the-fundamental-frequency-F-0-SD_tbl1_348456895.

Baevski, Alexei, Zhou, Henry, Mohamed, Abdelrahman, and Auli, Michael. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2006.11477.

Bar, David, Yochai, Michael, Kantor, Ido, Tal, Shira, Mizrahi, Amir, Fuchs, Yael, Adi, Yossi, and Keshet, Joseph. Als speech detection: A novel dataset and baseline system. In *IEEE Spoken Language Technology Workshop (SLT)*, 2022. URL https://ieeexplore.ieee.org/document/10041907.

Breiman, Leo. Random forests. *Machine Learning*, 2001. URL https://doi.org/10.1023/A:1010933404324.

Brown, Robert H. and Al-Chalabi, Ammar. Amyotrophic lateral sclerosis. *The New England Journal of Medicine*, 2017. URL https://doi.org/10.1056/NEJMra1603471.

Bäckström, Tom, Räsänen, Okko, Zewoudie, Abraham, Zarazaga, Pablo Pérez, Koivusalo, Liisa, Das, Sneha, Mellado, Esteban Gómez, Mansali, Marieum Bouafif, Ramos, Daniel, Kadiri, Sudarsana, Alku, Paavo, and Vali, Mohammad Hassan. *Introduction to Speech Processing*. 2022. URL https://speechprocessingbook.aalto.fi.

Cedarbaum, Jesse M., Stambler, Norman, Malta, E., Fuller, C., Hilt, D., Thurmond, B., and Nakanishi, A. The alsfrs-r: A revised als functional rating scale that incorporates assessments of respiratory function. *Journal of the Neurological Sciences*, 1999. URL https://doi.org/10.1016/S0022-510X(99)00210-5.

Chen, Tianqi and Guestrin, Carlos. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016. URL https://doi.org/10.1145/2939672.2939785.

Chiò, Adriano, Logroscino, Giancarlo, Hardiman, Orla, Swingler, Robert, Mitchell, Jeremy, Beghi, Ettore, and Traynor, Bryan J. Prognostic factors in als: A critical review. *Amyotrophic Lateral Sclerosis*, 2009. URL https://www.tandfonline.com/doi/full/10.3109/17482960802566824.

Darley, Frederick L., Aronson, Arnold E., and Brown, Joe R. Differential diagnostic patterns of dysarthria. *Journal of Speech and Hearing Research*, 1969. URL https://pubs.asha.org/doi/10.1044/jshr.1202.246.

Davis, Steven and Mermelstein, Paul. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1980. URL https://doi.org/10.1109/TASSP.1980.1163420.

De Marchi, Fabiola, Contaldi, Elena, Magistrelli, Luca, Cantello, Roberto, Comi, Cristoforo, and Mazzini, Letizia. Telehealth in neurodegenerative diseases: Opportunities and challenges for patients and physicians. *Brain Sciences*, 11:237, 02 2021. doi: 10.3390/brainsci11020237.

Dubbioso, Raffaele, Spisto, Myriam, Verde, Laura, Iuzzolino, Valentina Virginia, Senerchia, Gianmaria, Salvatore, Elena, Pietro, Giuseppe De, Falco, Ivanoe De, and Sannino, Giovanna. Voice signals database of als patients with different dysarthria severity and healthy controls. *Scientific Data*, 2024. URL https://doi.org/10.1038/s41597-024-03597-2.

et al., González. Voc-als: A voice database for als patients with acoustic and clinical features. *Synapse Data Repository*, 2024. URL https://www.synapse.org/Synapse:syn53009474/wiki/624731.

Eyben, Florian, Scherer, Klaus R., Schuller, Björn W., Sundberg, Johan, Andre, Elisabeth, Busso, Carlos, Devillers, Laurence, Epps, Julien, Laukka, Petri, Narayanan, Shrikanth S., et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 2016. URL https://doi.org/10.1109/TAFFC.2015.2457417.

Facebook AI Research. Wav2vec: Self-supervised learning for speech recognition (fairseq example). https://github.com/facebookresearch/fairseq/blob/main/examples/wav2vec/README.md, 2021.

Fernandes, Joana, Teixeira, Felipe, Guedes, Vitor, Junior, Arnaldo, and Teixeira, João Paulo. Harmonic to noise ratio measurement - selection of window and length. In *Procedia Computer Science*, 2018. URL https://doi.org/10.1016/j.procs.2018.10.040.

Fu, Yao, Zevallos, Juan, Gorman, Kyle, Jurgens, David, Zelikman, Rachel, Chang, Shiyue, Beigi, Homa, and

Prabhakaran, Vinodkumar. Towards inclusive automatic speech recognition: Investigating demographic bias in speech datasets and recognition systems. *Speech Communication*, 2023. doi: 10.1016/j.specom.2023.07.003. URL https://www.sciencedirect.com/science/article/pii/S0885230823000864.

Green, Jordan R., Yunusova, Yana, and Kuruvilla, Mili S. Bulbar and speech motor assessment in als: Challenges and future directions. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, 2013. URL https://pubmed.ncbi.nlm.nih.gov/23898888/.

Hardiman, Orla, Al-Chalabi, Ammar, Chio, Adriano, Corr, Enda M., Logroscino, Giancarlo, Robberecht, Wim, Shaw, Pamela J., Simmons, Zachary, and van den Berg, Leonard H. Amyotrophic lateral sclerosis. *Nature Reviews Disease Primers*, 2017. URL https://www.nature.com/articles/nrdp201771.

Hastie, Trevor, Tibshirani, Robert, and Friedman, Jerome. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009. URL https://hastie.su.domains/ElemStatLearn/.

Helleman, J., van Eenennaam, R.M., van Reenen, E.T. Kruitwagen, Faber, C.G., van den Berg, L.H., Visser-Meily, J.M.A., van der Kooi, A.J., Hendriks, J.C.M., de Visser, M., and Schröder, C.D. Telehealth as part of specialized als care: Feasibility and user experiences with 'als home-monitoring and coaching'. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, 2020. URL https://doi.org/10.1080/21678421.2020.1718712.

Hoerl, Arthur E. and Kennard, Robert W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 1970. URL https://doi.org/10.1080/00401706.1970.10488634.

Ioffe, Sergey and Szegedy, Christian. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, 2015. URL http://proceedings.mlr.press/v37/ioffe15.html.

Kimura, Fumihito, Fujimura, Chie, Ishida, Shinichi, Nakajima, Hiroyuki, Furutama, Daisuke, Uehara, Hiroko, and Shibasaki, Hiroshi. Progression rate of alsfrs-r at time of diagnosis predicts survival time in als. *Neurology*, 2006. URL https://doi.org/10.1212/01.wnl.0000194316.91908.8a.

Luz, Saturnino. Longitudinal monitoring and detection of alzheimer's type dementia from spontaneous speech data. In *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*, 2017. doi: 10.1109/CBMS.2017.41.

Riviere, Marc'Aurelio, Joulin, Armand, and Dupoux, Emmanuel. Unsupervised pretraining transfers well across languages. In *ICASSP 2020 - IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020. URL https://arxiv.org/abs/2002.02848.

Rudzicz, Frank, Namasivayam, Arun Kumar, and Wolff, Tommie. The torgo database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation*, 2012. URL https://www.cs.toronto.edu/~complingweb/data/TORGO/torgo.html.

Sannino, Giovanna. Process_all_files.py – data processing script for voc-als dataset. https://github.com/giovannasannino/VOC-ALS/blob/main/Process_all_files.py, 2024.

Schindler, A. and Gullì, R. An italian version of the yorkston dysarthria self-assessment questionnaire. Available in VOC-ALS dataset documentation, 2012. URL https://www.researchgate.net/publication/323577226_Italian_Validation_of_a_Test_to_Assess_Dysarthria_in_Neurologic_Patients_A_Cross-Sectional_Pilot_Study.

Schneider, Steffen, Baevski, Alexei, Collobert, Ronan, and Auli, Michael. wav2vec: Unsupervised pre-training for speech recognition. In *Interspeech*, 2019. URL https://doi.org/10.48550/arXiv.1904.05862.

Seber, George A.F. and Lee, Alan J. Linear regression analysis. *Wiley Series in Probability and Statistics*, 2012. URL https://www.wiley.com/en-us/Linear+Regression+Analysis%2C+2nd+Edition-p-9780471415404.

Smith, Leslie N. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2017. URL https://doi.org/10.1109/WACV.2017.58.

Talbott, Evelyn O., Malek, Amy M., and Lacomis, David. The epidemiology of amyotrophic lateral sclerosis. *Handbook of Clinical Neurology*, 2016. URL https://pubmed.ncbi.nlm.nih.gov/27637961/.

Teixeira, João Paulo, Oliveira, Carla, and Lopes, Carla. Vocal acoustic analysis – jitter, shimmer and hnr parameters. In *Procedia Technology*, 2013. URL https://doi.org/10.1016/j.protcy.2013.12.124.

Tibshirani, Robert. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1996. URL https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Tripathi, Ayush, Bhosale, Swapnil, and Kopparapu, Sunil Kumar. Improved speaker independent dysarthria intelligibility classification using deepspeech posteriors. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6114–6118. IEEE, 2020.

wen Yang, Shu, Chi, Po-Han, Chuang, Chia-Hao, et al. Superb: Speech processing universal performance benchmark. In *Interspeech*, 2021. URL https://arxiv.org/abs/2105.01051.

Yunusova, Yana, Green, Jordan R., Wang, Jinghua, and Zinman, Lorne. Speech in als: Longitudinal changes and subgroups of motor decline. *Neurology*, 2016. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4955603.