npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01654-7

Clinical assessment and interpretation of dysarthria in ALS using attention based deep learning AI models

Michele Merler[1,8], Carla Agurto[1,8], Julian Peller[2], Esteban Roitberg[2], Alan Taitz[3], Marcos A. Trevisan[4], Indu Navar[2], James D. Berry[5], Ernest Fraenkel[6], Lyle W. Ostrow[7], Guillermo A. Cecchi[1] & Raquel Norel[1]

Speech dysarthria is a key symptom of neurological conditions like ALS, yet existing AI models designed to analyze it from audio signal rely on handcrafted features with limited inference performance. Deep learning approaches improve accuracy but lack interpretability. We propose an attention-based deep learning AI model to assess dysarthria severity based on listener effort ratings. Using 2,102 recordings from 125 participants, rated by three speech-language pathologists on a 100-point scale, we trained models directly from recordings collected remotely. Our best model achieved R^2 of 0.92 and RMSE of 6.78. Attention-based interpretability identified key phonemes, such as vowel sounds influenced by 'r' (e.g., "car," "more"), and isolated inspiration sounds as markers of speech deterioration. This model enhances precision in dysarthria assessment while maintaining clinical interpretability. By improving sensitivity to subtle speech changes, it offers a valuable tool for research and patient care in ALS and other neurological disorders.

Amyotrophic lateral sclerosis (ALS) is a complex neurodegenerative condition that primarily targets the upper and lower motor neurons responsible for voluntary muscle movement. Initial symptoms include limb weakness and muscle spasms, which progress to severe muscle wasting, paralysis, and respiratory failure, on average within 2–4 years of diagnosis[1–4]. Bulbar-onset ALS is when an individual presents with changes in speech, salivation, and/or swallowing before experiencing limb-related symptoms, as opposed to spinal-onset. Although between 25 and 30% of people with ALS (pALS) initially display bulbar symptoms, the majority will eventually develop speech and swallowing complications[5,6]. Dysarthria is estimated to manifest in over 80% of pALS[7]. Moreover, individuals with bulbar-onset ALS experience earlier decline in speaking rate, speech intelligibility[5,8,9] and lower survival rate[10] compared to those with spinal-onset ALS.

The most common tool for measuring ALS progression is the ALS functional rating scale revised (ALSFRS-R)[11], which evaluates 12 aspects, including speech deterioration, through 12 questions with five alternatives to choose from normal to total loss function. This low-granularity assessment fails to capture timely variations in speech. Similar issues are presented in other health conditions affecting speech, such as Parkinson's disease. To address these limitations, alternative metrics for assessing dysarthria have been explored. Among them, listener effort (LE) has emerged as a promising clinically meaningful measure, offering a more nuanced understanding of the motor aspects of speech production. While its potential is recognized, LE has faced criticism for its variability and bias across different listeners. Carolan et al.[12] highlighted discrepancies between subjective and objective measures of effort, questioning consistency, while Strand et al.[13] noted the difficulty of standardizing LE assessments across individuals. Despite these challenges, research indicates that LE can effectively capture dysarthria severity when evaluated by trained speech-language pathologists (SLPs), thus reducing bias and enhancing reliability. By using expert-driven assessments, the variability associated with untrained listeners is minimized. In a recent study, Stipancic et al.[14] compared intelligibility, speaking rate, their combination, and LE as measures of dysarthria severity. Their findings revealed that LE demonstrated the highest accuracy in distinguishing severity levels, with superior effect sizes. Building on this work, our study described in[15] employed assessments from three expert SLPs, who evaluated LE on a quantitative scale from 0 to 100, allowing us to track the progression of dysarthria in individuals with ALS.

Several approaches have been explored for analyzing speech in ALS to assess dysarthria. Early studies, such as Yunusova et al.[16], focused on characterizing articulatory movements during vowel production in speakers with dysarthria. More complex methods have since been developed to track

[1]IBM Research, Yorktown Heights, NY, USA. [2]EverythingALS, Peter Cohen Foundation, Los Altos, CA, USA. [3]SRI International, Menlo Park, CA, USA. [4]Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales, Departamento de Física - CONICET - Instituto de Física Interdisciplinaria y Aplicada (INFINA), Buenos Aires, Argentina. [5]MGH Institute of Health Professions, Boston, MA, USA. [6]Massachusetts Institute of Technology, Cambridge, MA, USA. [7]Lewis Katz School of Medicine, Temple University, Philadelphia, PA, USA. [8]These authors contributed equally: Michele Merler, Carla Agurto. e-mail: rnorel@us.ibm.com

disease progression, emphasizing the integration of acoustic and articulatory measures for precise quantification of speech deterioration in ALS[8,17,18]. Vowel intelligibility testing has been examined as a reliable indicator of speech impairment severity in ALS[19]. Clustering methods combined with binary classification have been proposed as an alternative for classifying dysarthria severity levels[20]. With recent advancements in machine learning, research has increasingly focused on deep learning approaches to evaluate speech impairments[21–24]. For example, adversarial auto-encoders and ensemble learning have demonstrated potential for classifying speech dysarthria and improving early ALS diagnosis[23,24], and deep convolutional neural networks have been used on top of spectrogram representations of speech audio clips to predict the clinical standard ALSFRS-R[25]. Powerful encoder-decoder architectures such as Whisper[26] have also been used for speech deterioration tasks. In[27], the full encoder-decoder architecture is still tuned for audio transcription on speech data of individuals who suffered from strokes. The frozen embeddings from the tuned model have been concatenated with acoustic, linguistic, and glottal biomarkers as base to train a neural net classifier for post-stroke speech assessment, instead of directly optimizing the whole architecture for speech deterioration prediction. A recent study further contributed to this progress by introducing an interpretable decision tree model to classify dysarthria severity in ALS, potentially offering clinicians guidelines for assessment[28]. For more comprehensive insights, a recent survey on deep learning approaches for pathological speech highlights recent emerging trends in machine learning applications for dysarthria detection can be found in[22].

In parallel with these methodological advancements, research institutions are making significant efforts to collect large-scale datasets for ALS speech analysis, enabling the development of robust models for assessing dysarthria severity and tracking disease progression. Key datasets include the TORGO Database[29], which captures articulatory and acoustic features from individuals with ALS and cerebral palsy, focusing on short words and sentences; Answer ALS[30], an ALS-exclusive dataset containing speech and motor tasks, genetic data, and clinical measures such as forced vital capacity (FVC); Everything ALS[15], which includes ALS patients and matched controls, offering longitudinal speech recordings with ALSFRS-R scores; and the Voice Signals Dataset[31], which provides acoustic features extracted from sustained vowel and diadochokinetic (DDK) tasks, along with genetic and clinical data.

This work focuses on the automatic assessment of LE scores in recordings of sentences read by participants with varying levels of speech deterioration. We recognize that speech dysfunction in ALS results from impairments in multiple components of the speech system, leading to a differential impact on specific words or syllables depending on an individual's condition[32]. As a result, the assigned score is highly dependent on the words uttered during evaluation. Specifically, we propose an automated remote monitoring system based on attention-based deep learning AI models. This system not only achieves high precision (R^2 = 0.92, *RMSE* = 6.78 on a 0-100 scale) on capturing individual subtle speech changes but also identifies critical, clinically meaningful speech elements affected by dysarthria. Through this method, we identified speech patterns essential for more accurate individual assessments, reducing evaluation time and enabling speech therapists to tailor therapeutic exercises to improve patient outcomes.

Results

Composite Listener Effect Score Evaluation

Upon calculating the Composite Listener Effect Score (CLES) from the three SLPs for each analyzed recordings (see Ground truth calculation in Methods), we found that the percentage of gradings used from SLP 1, SLP 2, and SLP 3 were 32%, 34%, and 34%, respectively. No significant difference (Friedman test[33], Q = 4.7, p = 0.1) was found when comparing the distributions of the CLESs for the three groups of sentences, as shown in the top panel of Fig. 1. In the same Figure, we also observed a strong correlation between the CLESs and the ALSFRS-R speech scores (R^2 = 0.5, and $p < 0.00001$). We calculated perplexity values, a measure of how unpredictable or linguistically complex a text is based on a language model (see Supplemental Note 2 for more details) and compared their distribution using Kolmogorov-Smirnov to determine if the number of words was associated with reading difficulty. We found that the cumulative distribution of 13-word sentences is different from 15-word sentences (*KS* = 0.52, p = 0.01), the 15-word sentence being the one with lower perplexity values. In the case of 11-word sentences, a large variability of perplexity values was found (see experimental results Supplementary Fig. 1 for more details). We computed the correlation between CLES and perplexity values and found that it is not significant (r = 0.02, p = 0.335).

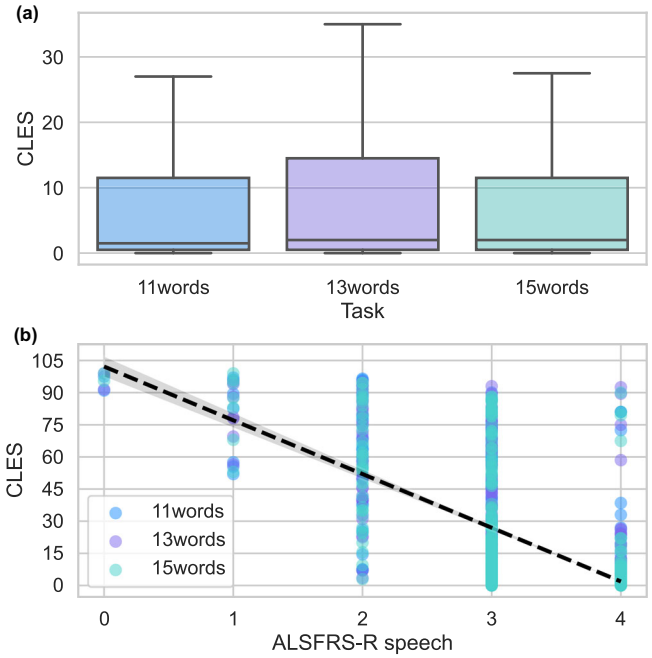

Fig. 1 | Analysis of the Composite Listener Effort Score (CLES). a The top panel illustrates the distribution of CLES across speech tasks categorized by sentence length, showing no significant differences among tasks (p = 0.1, Friedman test). **b** The bottom panel demonstrates a strong correlation between ALSFRS-R speech scores and CLES (R^2 = 0.5, $p < 0.00001$), highlighting the relationship between CLES and established clinical measures. This figure emphasizes the consistency of CLES across tasks and its alignment with clinical evaluations.

Performance results

Table 2, Supplementary Table 2, Fig. 2 and Supplementary Fig. 7 present the performance results of the models using our proposed approach with five- and ten-fold cross-validation. Supplementary Table 3, which also includes results for both five- and ten-fold cross-validation, displays the performance of other models generated for comparison with ours, as described in Section Model Comparison. It can be observed that performance is slightly lower for five-fold compared to ten-fold as expected. For the remainder of the manuscript, we focus on the five-fold results. We used Steiger's Z test for dependent correlation coefficients to determine if one model was better than the other. We found that for handcrafted features results obtained with FT are statistically significantly higher ($p < 0.00001$) than the ones obtained with Lasso regression. Among all the results, the attention based deep learning AI models developed with Whisper-FT are statistically significantly better than the other methods ($p < 0.00001$).

We also developed gender-specific models using Whisper-FT, achieving R^2 of 0.86 ± 0.04 (RMSE of 7.67 ± 1.39) and R^2 of 0.93 ± 0.02 (RMSE 6.90 ± 0.75) for male and female models, respectively. Figure 2 compares the inferred vs. observed CLES for sentences of three different lengths. As shown in the Figure and confirmed by a Friedman test (Q = 1.25, p = 0.53), there is no statistical difference in performance across the three sentence lengths.

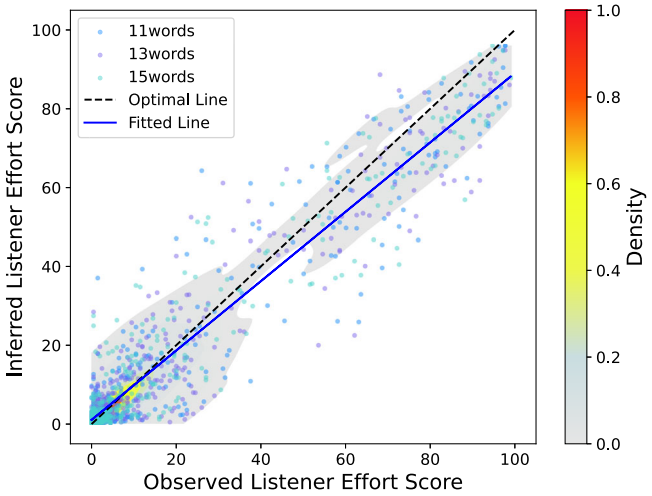

Fig. 2 | CLES Prediction results for five folds cross validation experiments. For the best attention-based AI model (Whisper-FT variant), inferred listener effort scores are plotted against the observed listener effort scores by SLPs, showing high correlation. Same plot for ten-fold can be found in Supplementary Figure 5.

Given the observed power of the Whisper model, we also employed it as a feature extractor, deriving some features on top of its predictions (see Supplemental Note 3 for more details). Training a Lasso regression model on top of those features achieved R^2 of 0.79 ± 0.07 (RMSE of 10.81 ± 1.93). Additionally, we performed an extra analysis in the 94 discarded recording files because of high disagreement in the gradings of the three SLPs. Details can be found in the Supplemental Note 5. Given the longitudinal nature of the data employed in this study, we investigate the behavior of our model over time in Supplemental Note 6, determining that the model performs independently from time of recording nor disease progression of any given subject.

Clinical Interpretation Analysis

Figure 3a shows the most frequent words (top two attended per audio sample) and the findings for phoneme analysis using the Carnegie Mellon University pronouncing dictionary[34]. We can observe that several words, regardless of speech degradation (i.e., CLES), contained key information for the models, the word *a* being the most frequent one. Additionally, we found that the sound made by pALS when inspiring air without uttering a word was informative, especially at lower CLES. When we further broke down words in phonemes (Fig. 3 reports the distributions of top attended Unigrams (b), Bigrams (c) and Trigrams (d), respectively), we found high

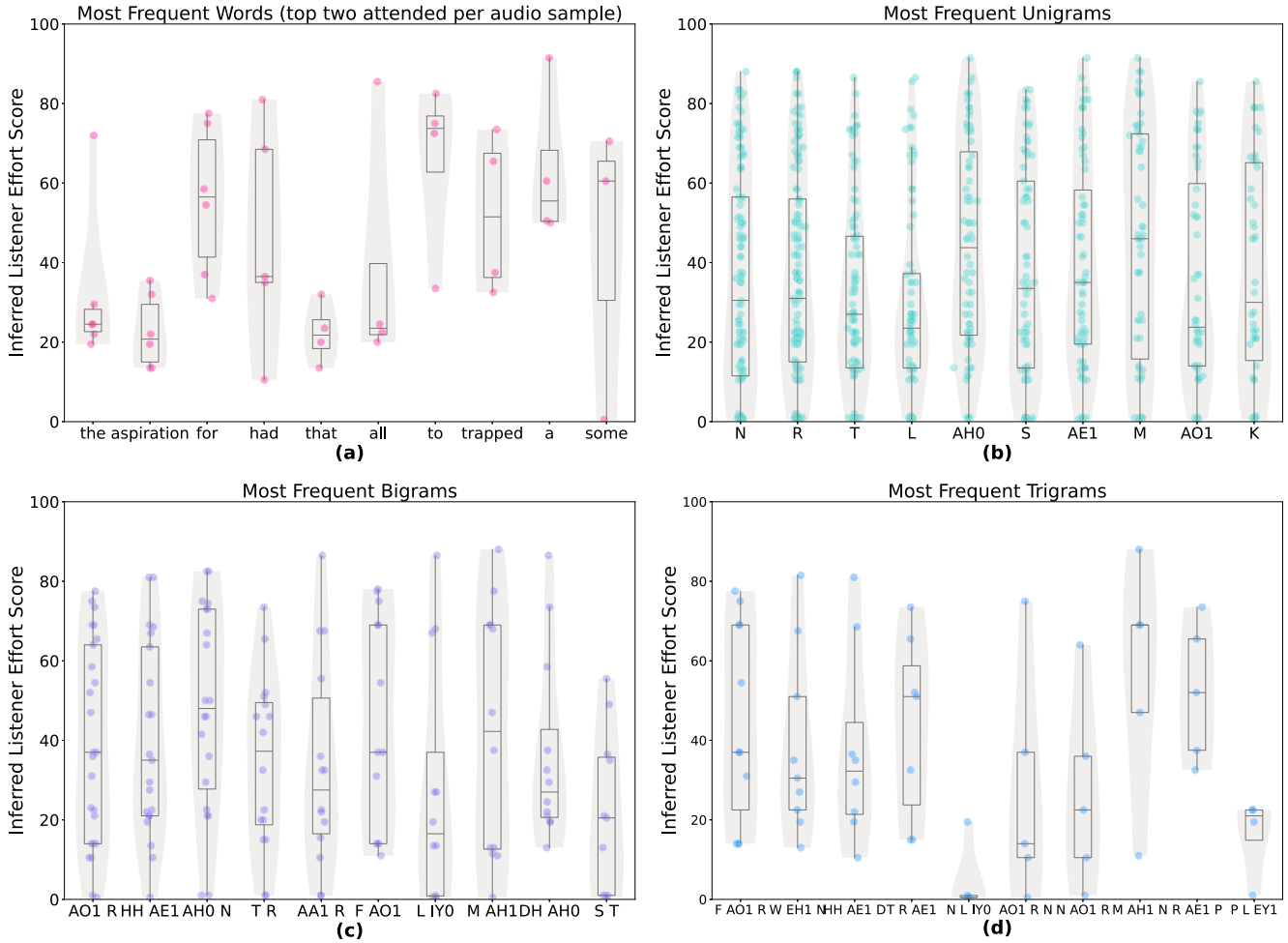

Fig. 3 | CLES distribution among top words and phonemes attended by the model. Distribution of inferred composite listener effort scores (CLES) among the top ten most common words and phonemes identified by the attention-based deep learning AI AST-FT model. Top ten **a** top-2 attended words per audio sample. The word *a* is among the most frequently attended. Additionally, the sound made by pALS when inspiring air without uttering a word was found to be informative, especially at lower CLES scores. Top ten **b** unigrams, **c** bigrams and **d** trigrams. We observe that phonemes such as AO1-R, AA1-R and E-R are highly attended by the model. In individuals with ALS, those phonemes involving complex rhotic vowels are expected to exhibit more pronounced articulatory undershoot, where the intended vowel and rhotic sound may not be fully achieved. Words that contain a voiced dental fricative at the beginning such as the and that are also highly attended by the model.

variability in CLES for phonemes corresponding to vowel sounds as well as phonemes associated to *k, l, m, n, r, s, t*. In our analysis of bigrams (see Fig. 3c), we observed considerable variability in the CLES associated with *AA1-R* (in words like *garden*) and *AO1-R* (in words such as *north*). For trigrams, although the variation related to CLES is reduced, we still identified significant cases, such as *F-AO1-R* (in words such as *fortunately*), that remained relevant across different CLES.

Discussion

Achieving high precision and granularity is essential for quantifying changes in the speech of people experiencing dysarthria, including those with ALS and other conditions. This is useful in clinical trials where quantifying group change in the treated versus placebo group is paramount. It is also highly significant for patients who want to anticipate when they might experience worsening dysarthria or loss of speech to help them plan to use augmentative communication[8]. For this reason, in this work, we proposed an attention-based deep learning AI model to accurately assess listener effort as a proxy for dysarthria severity. Additionally, we identified keywords that specifically highlight speech decline in pALS, enabling precise quantification of subtle changes in this population.

Establishing a robust ground truth was fundamental for achieving high accuracy in our models. Our preliminary analysis of SLPs gradings indicated a large variation (up to 82 points of difference) in scores among them for a small subset of examples (see Supplementary Fig. 2). We suspect that some of the largest discrepancies could be attributed to scoring inconsistencies rather than true differences in rating. However, our approach of excluding the most divergent score among the three SLPs (see Section Ground truth calculation) enhanced the reliability of listener effort severity estimation for each recording.

In Supplementary Fig. 3, we observe that some 15-word sentences are associated with low CLES (especially for pALS). To see if the difficulty of the sentence to be read is a significant factor, we calculated perplexity values for each recording[35,36]. We have found that the relationship between sentence length and perplexity is not evenly distributed (see Supplementary Fig. 1). Specifically, many 11-word sentences exhibit high perplexity values, while many 15-word sentences demonstrate low perplexity. However, there is no significant correlation (r = 0.02, p = 0.335) between perplexity values and CLES.

Demographic factors influencing the performance of the specific model trained in this study were assessed using the attention-based deep learning AI Whisper-FT model with five-fold cross-validation. No significant differences were found regarding gender, comorbidities, or ALS-related dysfunctions such as dyspnea and respiratory insufficiency. However, the RMSE for the 101 recordings from the 6% of non-native English speakers is higher than the overall RMSE, reaching 10.94. A Mann-Whitney test showed significant differences in error distribution being Malayalam and Mandarin first-language speakers exhibiting the highest error rates. Although using Whisper encoders trained on multilingual data reduced these discrepancies (see Supplementary Table 3) while maintaining overall performance (see Supplementary Fig. 5), differences remained, emphasizing the need for further research to enhance the model's generalization across diverse linguistic backgrounds. Furthermore, we evaluated cohort-specific models: the male-specific model achieved a mean R^2 of 0.86, and the female-specific model achieved a mean R^2 of 0.93, possibly due to differences in the LE distribution, which indicated higher LE values for females. Although these models did not significantly surpass the overall model, their promising results with limited data suggest the potential for further refinement of demographic-specific models in future research.

The models' performance results were strong, with R^2 values ranging from 0.62 to 0.92 and RMSE values between 6.78 and 14.63 (in a 0–100 range) for five-fold cross-validation. Similar values were obtained for ten-fold cross-validation. The results showed a statistically significant improvement for the high-dimensional feature set from OpenSMILE-6.3k (Steiger's Z test, $p < 0.00001$) after FT, over Lasso. ElasticNet regression was also explored, but it underperformed across all feature sets. Notably, the most accurate results were obtained using derived features from Whisper and fine-tuning using any Whisper encoder (see ablation study results in Supplementary Fig. 5). This is likely because Whisper's latest models in automatic transcription are trained on vast amounts of data, covering a wide range of accents, dialects, and noisy environments, which enhances their robustness and generalization independent of hyperparameter tuning (see ablation study in Supplementary Fig. 6).

When examining the most relevant handcrafted features, we identified that *audio duration*, *speech rate* and *loudness peak per second*, key features found in our previous work[37] and by others[8,38–40] in discriminating pALS from controls, are also informative for assessing listener effort. Those features may not only reflect the direct effects of dysarthria but also capture indirect effects, such as compensatory strategies often encouraged in speech therapy to improve intelligibility for listeners[41,42]. We also observe that formants 2 (F2) and 3 (F3) features derived from Praat and openSMILE are helpful. It has been documented that one characteristic that occurs in pALS, and perhaps before dysarthria or dysphagia, is decreased tongue strength[43], so it is expected to find changes in F2 since this feature is associated with tongue position[44]. Generally, vowels produced by individuals with dysarthria are typically marked by articulatory undershoot, meaning that vowel sounds do not reach their standard formant frequencies[45].

Interestingly, we found that the inspiration sounds produced by the participants were identified by the model as informative features for low CLES, meaning early speech deterioration. In pALS the presence of laryngeal valving inefficiencies[46] is expected due to vocal fold weakness or spasticity, which can result in reduced subglottal airflow that requires more frequent inspirations[32].

Notably, in this work, we have identified the words and phonemes that were key for the models in inferring the composite listener effort scores with precision. For example, we found that the word *a* and the phonems (unigrams) closely associated with it (*AH0, AE1, AO1*) are relevant to detecting CLES on the whole scale. This confirms previous findings by other groups that distorted vowel production is a primary characteristic of dysarthria, regardless of the neurological condition[47,48]. In fact, previous research has shown that the vowel /a/ was the most sensitive parameter to discriminate ALS patients belonging to the most severe dysarthria categories[28]. Phonemes such as *AO1-R, AA1-R* (see Fig. 3) and *E-R* (see Supplementary Fig. 8) involve complex articulatory tasks generated by the combination of vowels and the *r* sound, forming rhotic vowels. In individuals with dysarthria, particularly those with ALS, those phonemes are expected to exhibit a more pronounced articulatory undershoot, where the intended vowel and rhotic sound may not be fully achieved. This undershoot leads to slurred speech, where the distinction between the vowel and the *r* becomes unclear, further increasing speech dysarthria. This finding is also corroborated by changes in F3 (see Supplementary Table 4), a feature associated with rhotic vowels and lip rounding[49].

Research has indicated that fricative sounds such as *s* in the word *some* are affected in individuals with dysarthria[50]. Furthermore, we observe that words that contains a voiced dental fricative at the beginning such as *the* and *that* (see Supplementary Fig. 8) are also highly attended by the models. We speculate that pronouncing this phoneme is difficult for pALS since it requires fine motor control over the tongue, teeth, and vocal folds. In addition, we also observed the models focusing on words such as *when, we*, and *without* (see Supplementary Fig. 9) that share the *w* phoneme, which is a labial-velar approximant. Producing this phoneme requires the coordination of rounded lips, the back of the tongue toward the velum, and the vibration of the vocal cords-all of which are typically affected in individuals with pALS.

When listening to the word *had* in pALS, as highlighted in Fig. 3, we observe that the *h* sound becomes nasalized. Producing the *h* sound requires continuous airflow from the lungs, while transitioning to the *d* sound demands greater effort as it involves vibrating the vocal cords. When these two actions are combined, it becomes more difficult to control airflow and properly close the soft palate (velum), allowing air to pass through the nasal

cavity. This indicates that combining phonemes is crucial in highlighting speech changes in ALS.

The main contribution of this work is the proposed methodology for modeling listener effort score, which can easily generalize to other populations and conditions given the collection appropriate training data. In this sense, we have identified some areas for future work to improve our approach. Specifically for the model in this study, one of key aspect is language diversity, as our findings demonstrate its impact on model performance. Collecting data from a different specific language of interest (or a combination of multiple languages) could help increase the model's robustness. Another critical aspect is the lack of detailed information about ALS phenotyping and other disease characteristics, which can significantly influence speech impairments. While our cohort reflects the prevalence of bulbar-onset ALS (22%)[51,52], it is predominantly composed of individuals with slow progression rates (0.3). Additionally, neurological conditions exhibit distinct dysarthria subtypes due to varying motor impairments. For instance, amyotrophic lateral sclerosis (ALS) is associated with mixed spastic-flaccid dysarthria, which reduces speech intelligibility, while Parkinson's disease (PD) is linked to hypokinetic dysarthria, characterized by monopitch and articulatory undershoot[48]. Although our approach identified common phonemes relevant across these disorders, such as /a/ and /m/, aligning with prior PD research[53,54], developing a more comprehensive model would further improve performance. Moreover, while our longitudinal analysis in the section Longitudinal Analysis showed that the model's inference accuracy is independent of LE score progression, a larger dataset with more participants is necessary to validate these findings and enable more precise quantification of disease progression and subtle changes over time, both overall and at the phoneme level. One limitation of our study is the exclusion of other neurodegenerative diseases in the control group, which could have provided additional insights into the specificity of the model in assessing LE. Future work should incorporate more detailed clinical information about ALS, consider additional progression rates, as well as exploring other dysarthria subtypes and incorporating controls from other neurological conditions. Long-term speech data and broader range of linguistic backgrounds should also be included to enhance the model's robustness further. These improvements will enhance early detection and contribute to a more comprehensive framework for tracking speech deterioration in neurological disorders.

In summary, our approach offers a strong foundation for detecting subtle dysarthria-related speech impairments, leveraging AI models to systematically select phonemically informative words for tailored disease progression monitoring. By delivering precise, interpretable assessments from remotely collected data, our AI approach empowers patients by reducing barriers to assessment while maintaining clinical rigor. Moreover, this pipeline can be applied directly to other neurodegenerative conditions, such as Parkinson's disease and Huntington's disease, broadening its impact and utility.

Methods

Protocol

Participants for this study were recruited as part of a large longitudinal speech study performed by the Everything ALS organization: the Austen Speech Study[55]. All participants signed informed consent and were asked to perform different speech tasks through Modality.ai Inc.'s secure web-based platform that gathers audio and video data remotely[56].

For this work, we focused on a specific speech task: reading sentences of increasing length. Specifically, we analyzed sentences containing 11, 13, and 15 words. Sentences were randomly selected by Modality.ai from pools of 19, 16, and 18 sentences corresponding to each length, respectively. We think this task effectively captures the complex articulation challenges faced by individuals, particularly those with ALS, as it better reflects real-world communication demands compared to standardized assessments like diadochokinetic tasks.

Recordings chosen for grading followed strict selection criteria: ALS participants were required to have at least two months of recorded data, with all available sessions included while maintaining a minimum gap of two months between sessions. Additionally, twenty non-ALS participants were selected using the same criteria while ensuring demographic matching with the ALS group.

Incorporating recordings from non-ALS participants helped ensure an unbiased evaluation, as graders were blinded to participants' diagnoses. Moreover, this cohort increased the diversity of speech samples, providing a richer dataset for analyzing early speech deterioration.

SLPs Annotations

Three expert SLPs retrospectively evaluated the selected recordings as described in Section Protocol. They were instructed to listen to each recording and assess the listener effort (LE)[57] required to comprehend the sentences, on a scale from 0 to 100, with the request: "How effortful was it for you to understand this person. Remember, we are asking how hard you worked, not how well you did".

To minimize potential variability variability in perception, we have taken several steps to enhance the reliability of the LE ratings. We have carefully controlled the rating conditions by presenting sentences without contextual cues, limiting word complexity, masking speaker condition (i.e., ALS) and dysarthria severity level to prevent bias. Furthermore, recordings were graded in a randomized order to avoid the influence of consecutive samples from the same participant.

In total, 2102 recordings were graded by each SLP. To assess the intra-variability of SLP, each SLP re-evaluated 20% of the recordings (422). Additionally, the SLPs provided feedback on the overall quality of the recordings.

Ground truth calculation

Even though we observed high intra-rate agreement among the SLPs (see Supplemental Section SLPs Agreement Analysis for more details on the method and results obtained), we still observed high variability in a small subset of the data among the gradings from the three SLPs. Therefore, we calculated a CLES for each recording which is more robust to outliers ratings using the following method (see Fig. 4). First, we only considered the first time an SLP scored a recording. Second, we only used the closest two scores among the three graders, discarding the more distant value. Third, if those two values have less than 10 points (10% of the scale) difference, we averaged the grading; otherwise, that audio is not used for modeling. Supplementary Table 1 shows intra-rater agreement levels in each stage of data curation.

Participants

From the 2102 recordings (125 participants) graded by the SLPs, only 1798 were considered for the analysis. Recordings were excluded for the following reasons: flagged as low quality by the SLPs (114), at least one of the feature extraction methods failed (e.g., not detecting any vowel for the formant analysis, 96), and did not meet the criteria for ground truth calculation specified in Section Ground truth calculation (94). Those 1798 recordings correspond to 124 participants (104 from pALS).

Table 1 shows all participants' demographic information with valid data for both cohorts. Additionally, for the pALS cohort, we indicate ALS-specific metrics such as age at onset (median = 59.1 years). Table 2 Specifically for the ALSFRS-R speech score, we found that there are 1069 recordings where the participants assigned a score of 4 (normal function) while the SLPs assigned values between 0 (no effort) and 92.5 (very hard to understand), indicating some level of deterioration. Neither cohort exhibited comorbidities known to cause dysarthria, with muscle disease being the most common in the ALS cohort, accounting for 17.7%.

Experimental design

Our proposed approach is composed of three steps. First, we computed a composite listener effort score (CLES) for each remotely recorded audio using the SLPs annotations (see Section Protocol for more details on the protocol the SLPs followed to rate the recordings and refer to *ground truth calculation* in Fig. 4 for a visual representation of our approach). Second, we

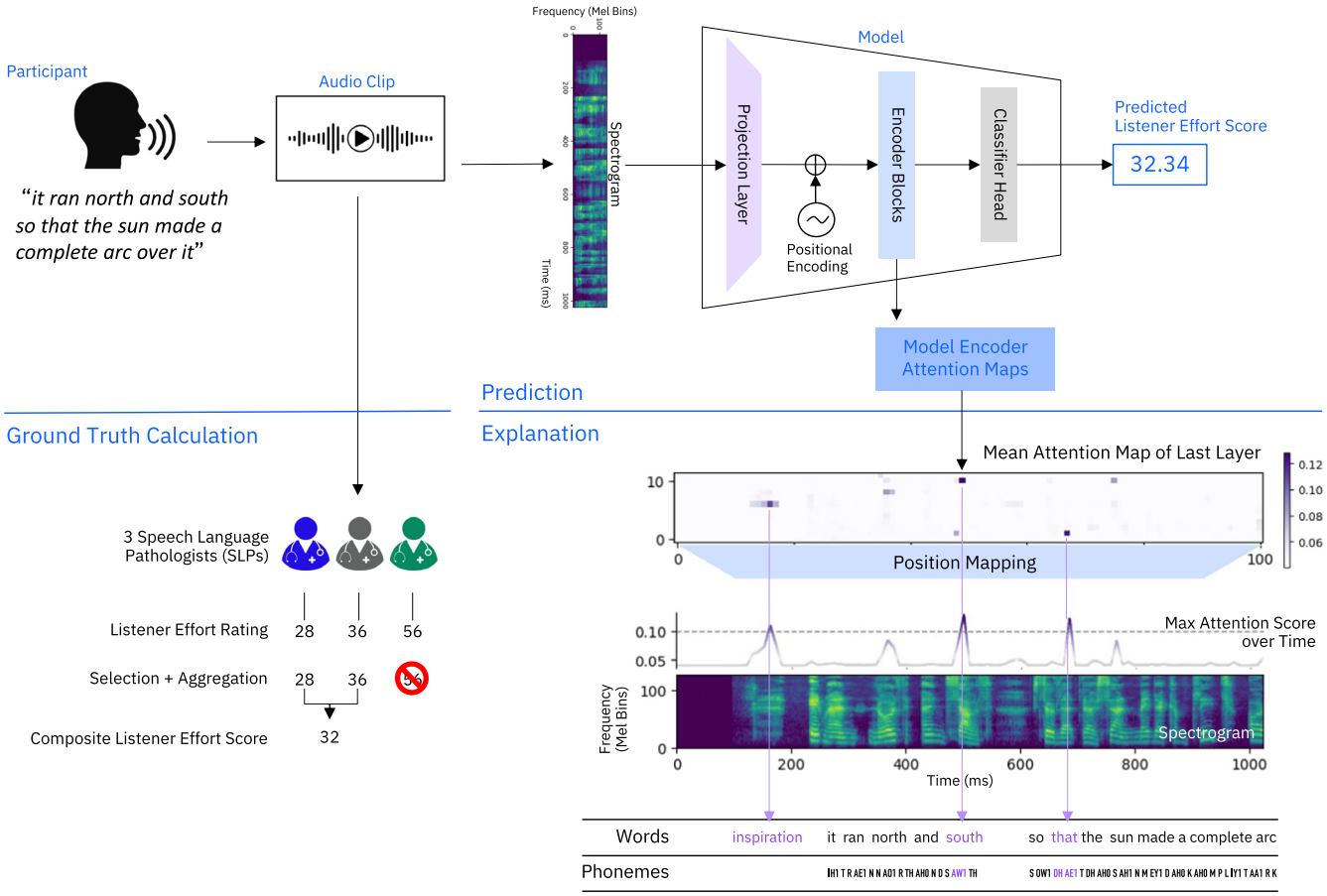

Fig. 4 | System Diagram of our attention-based deep learning AI approach. Composite listener effort score (CLES) for each recording is computed using SLPs annotations, filtered by quality and agreement (Ground Truth Calculation). The attention-based deep learning AI model is trained on top of the spectrogram representation of each audio and consists of a projection layer followed by positional encoding, a transformer encoder plus a classification head to infer a listener effort score. Finally, an explanation of the model's inference is provided using the largest activations in the attention maps and their corresponding regions in the spectrogram to identify the words, and more specifically phonemes, to which the model pays the most attention to when making an score inference (Explanation).

trained attention-based deep learning AI models directly on top of the recordings to learn how to infer CLES. Once trained, the models can be used to infer a listener effort score on a new audio recording. Third, we provided an explanation of the model's inference, using the largest activations in the attention maps and their corresponding regions in the spectrogram to identify the words, and more specifically phonemes, that were relevant to the model (see *explanation* in Fig. 4). In the following sub-sections, we provide more details on each step.

AI modeling

In order to train attention-based deep learning AI models for inferring CLES, we conducted end-to-end fine-tuning of powerful pre-trained encoders on top of which we added a classification head. Our models were trained on the audio signals represented as spectrograms, not on top of handcrafted features. The first encoder we used comes from the Audio Spectrogram Transformer (AST)[58], a vision transformer model applied to spectrograms. This encoder was pre-trained on tens of millions of examples comprising both natural images (ImageNet[59]) and spectrograms (which are treated as images) from audio files (Audioset[60]). The second encoder is from Whisper[26], a sequence-to-sequence (seq2-seq) model pre-trained for automatic speech transcription on 680k hours of labeled audio data. Figure 4 presents the general architecture shared across all the trained attention-based deep learning AI models. Starting from a spectrogram representation of the audio recording which represents the signal in frequency and time, the models have three major components:

- A projection layer followed by positional encoding, which maps different regions of the spectrogram into an embedding space that can be processed by the following modules.
- An encoder based on a series of stacked transformer blocks which provides a powerful representation of the speech signal, utilizing the attention mechanism to learn the relationships among different parts of the speech signal.
- A classifier head tailored for inferring listener effort score, which takes the encoder representation of the speech signal as input, and produces as single listener effort score from his last linear layer.

The specific details of the architectures used for fine-tuning the AST encoder and the Whisper encoder with an added classification head are reported in Supplementary Fig. 4 and further implementation details can be found in the Supplemental Material.

All models were trained using the Mean Squared Error (MSE) loss function and the AdamW optimizer.

Following standard practice in machine learning for datasets the size of ours (1798 examples, 124 subjects)[61,62], performance evaluation was conducted through five-fold and ten-fold cross-validation, with splits performed using a stratified group k-fold, which separated subjects and sessions. Hyper-parameters were optimized within each fold using a train/validation split, and the best hyper-parameters set was then used across all folds. Specifically, we searched over a number of epochs in the range $[5, 50]$ and learning rate $\left[1e^{-6}, 1e^{-5}, 1e^{-4}\right]$. The batch size was set to 16 for all those models. To quantify error, we used RMSE, and for model

Table 1 | Demographic and Clinical Characteristics of the Participants in this Study

Category	Variable	All	pALS	Control
Demo-graphics	Participants	124 (100%)	104 (84%)	20 (16%)
	N sessions (N audios)	655 (1798)	538 (1465)	117 (333)
	Age (years) at baseline	63.6 (56.2, 68.1)	63.6 (56.2, 68.2)	63.9 (56.1, 68.0)
	Education (in years)	17 (16, 18)	17 (16, 18)	17 (16, 18)
	Gender: Female	60 (48%)	50 (48%)	10 (50%)
	Male	64 (51%)	54 (52%)	10 (50%)
	Race: Caucasian	112 (90%)	94 (90%)	18 (90%)
	Asian	5 (4%)	5 (5%)	0 (5%)
	Other/Non Reported	7 (6%)	5 (5%)	2 (10%)
	Ethnicity: Not Hispanic	115 (93%)	97 (93%)	18 (90%)
	Hispanic	6 (5%)	5 (5%)	1 (5%)
	Non Reported	3 (2%)	2 (2%)	1 (5%)
	First Language: English	116 (94%)	98 (94%)	18 (90%)
	Other	8 (6%)	6 (6%)	2 (10%)
	Comorbidities: Muscle disease	22 (17.7%)	22 (21.2%)	0 (0%)
	Depression	20 (%)	15 (%)	5 (%)
	Cancer	14 (%)	13 (%)	1 (%)
	Diabete	8 (%)	7 (%)	1 (%)
	Epilepsy	2 (%)	2 (%)	0 (%)
	Bipolar disorder	2 (%)	1 (%)	1 (%)
ALS-specific metrics	Age at symptom onset		59.1 (51.3, 64.4)	
	Bulbar/Non-bulbar onset		23 (22%)/ 81 (78%)	
	ALSFRS-R total		37.5 ± 8	
	ALSFRS-R speech		3.5 ± 0.7	
	ALSFRS-R bulbar		10.7 ± 1.7	
	ALSFRS-R dyspne		3.5 ± 0.9	
	ALSFRS-R respiratory insufficiency		3.6 ± 0.6	
	Progression rate[65]		0.3 ± 0.7	

Values are reported for overall participants and divided by cohort. Values for continuous variables are expressed as median (Q1, Q3). In the case of discrete variables such as ALSFRS-R subscores are expressed in mean ± standard deviation. Finally, for binary variables like gender are expressed in number (percentage).

Table 2 | CLES prediction results of the AI models for five-fold cross validation experiments

Model Type	Details	RMSE ↓ (Mean ± std)	R2 ↑ (Mean ± std)
Handcrafted Features + Model	OpenSMILE-6.3k	11.27 ± 1.51	0.78 ± 0.05
Attention-based AI model	AST-FT	10.78 ± 2.03	0.78 ± 0.05
Attention-based AI model	Whisper-FT	**6.78 ± 0.96**	**0.92 ± 0.02**

For the proposed models and best model using handcrafted features, we report RMSE and R^2 scores for different features and models (top). Steiger's Z test for dependent correlation coefficients determines that results obtained with our proposed attention-based deep learning AI models are statistically significantly better ($p < 0.00001$) than the ones obtained with handcrafted features. The best model is boldfaced.

interpretability in regression analysis, we employed R^2, as recommended by Chicco et al.[63]. We report mean and standard deviation values across folds, thus robustly reporting the effective performance of the model, not biased by a single potential split containing a set of examples for which the model performs unusually better or worse. Low values of standard deviations reported in the results demonstrate the robustness of the models's performance across splits.

Explainability of Inferences

Understanding how a model achieves accurate score inferences is essential, particularly in uncovering the specific patterns, sounds, or words within the data that the model relies on to infer listener effort scores effectively. To achieve this, we mapped the top activations from the attention maps of the last encoder layer of the model, the one providing the input to the classifier head, to the original input, following the seminal explainability work on transformers[64]. For each recording, we extracted the attention weights from the last layer of the encoder. We then generated a mean attention map by averaging across the attention heads, as shown in Fig. 4. This map is then referenced back to the input spectrogram dimensions by positional mapping and therefore highlights both the spectrogram regions that the model focuses on, as well as their corresponding intensity, reflecting their relative importance. Since our goal is to determine which words and phonemes are most relevant for the model when making predictions, we take the maximum intensity from the map across the frequency domain at each point in time, as represented in the curve over the spectrogram in Fig. 4. The peaks in the intensity plot correspond to the temporal regions most attended to by the model when making the prediction. This procedure was applied to each of the 1798 analyzed recordings. While approximations of such calculations could be theoretically possible using Whisper-FT, we decided to focus on the AST-FT model, for which the encoder vision transformer (ViT) activation maps were extracted using the readily available model's codebase (https://github.com/YuanGongND/ast) and their exact mapping to spectrogram patches can be precisely computed. In contrast, Whisper processes directly the entire spectrogram (see details in Supplementary Fig. 4).

Since the primary aim of this analysis is not only to identify the areas where the model focuses its attention but also to understand how specific words or phonemes contribute to listener effort, including samples with high prediction errors would undermine this goal, as inaccurate predictions do not reliably represent the relationship between speech patterns and perceived effort. Therefore, we focused exclusively on samples with low errors to ensure that our findings accurately reflect the impact of speech deterioration. Specifically, we applied the following constraints: First, to address the skewed nature of the dataset, we divided the recordings into ten intervals and selected 20 samples from each based on prediction error, ensuring balanced representation across score ranges. Then, we excluded recordings with errors exceeding 10 (10% of the LE scale), resulting in a final set of 138 recordings. Finally, relevant speech regions were identified using a peak local maximum function, selecting peaks with a minimum intensity of 0.1 and spaced at least 5 seconds apart. From these 138 recordings, we extracted 714 smaller segments focusing on specific phonemes and sounds, which were analyzed using unigrams, bigrams, and trigrams to uncover key speech patterns indicative of speech deterioration. These patterns could potentially inform the tailoring of speech tasks to optimize the monitoring process.

Model Comparison

To ensure a fair comparison with the current state-of-the-art handcrafted acoustic features (listed in Section Handcrafted Features), we also developed models to assess LE using Lasso regression and fine-tuning the same classifier head as Whisper-FT (Supplementary Fig. 4 (right)).

All models were validated and evaluated using the same approach detailed in section AI modeling. For the fine-tuned model, hyperparameter settings were as previously described, while for the Lasso regression models, parameter α was optimized with grid search in the linear space within the [0.001, 1] range, with step equal to 0.01. Since these models relied on handcrafted features, feature relevance was straightforward to determine and was obtained by analyzing the weight of each feature in the models.

Data availability

The repository of recorded speech and de-identified clinical data from this study is now available to ALS researchers to advance speech research in ALS (https://www.everythingals.org/available-data).

Code availability

The Whisper encoder was taken from the Transformers library 4.23.1 https://huggingface.co/docs/transformers/en/index. Code for the Audio Spectrogram Transformer was based on the following codebase https://github.com/YuanGongND/ast. The code generated and used during the current study is available from the corresponding author upon request.

Received: 5 December 2024; Accepted: 20 April 2025;
Published online: 08 May 2025

References

1. Phukan, J. et al. The syndrome of cognitive impairment in amyotrophic lateral sclerosis: a population-based study. *J. Neurol. Neurosurg. Psychiatry* **83**, 102–108 (2012).
2. Goldstein, L. H. & Abrahams, S. Changes in cognition and behaviour in amyotrophic lateral sclerosis: nature of impairment and implications for assessment. *Lancet Neurol.* **12**, 368–380 (2013).
3. Chiò, A. et al. Cognitive impairment across als clinical stages in a population-based cohort. *Neurology* **93**, e984–e994 (2019).
4. Pender, N., Pinto-Grau, M. & Hardiman, O. Cognitive and behavioural impairment in amyotrophic lateral sclerosis. *Curr. Opin. Neurol.* **33**, 649–654 (2020).
5. Chio, A. et al. Prognostic factors in als: a critical review. *Amyotroph. Lateral Scler.* **10**, 310–323 (2009).
6. Masrori, P. & Van Damme, P. Amyotrophic lateral sclerosis: a clinical review. *Eur. J. Neurol.* **27**, 1918–1929 (2020).
7. Tomik, B. & Guiloff, R. J. Dysarthria in amyotrophic lateral sclerosis: a review. *Amyotroph. Lateral Scler.* **11**, 4–15 (2010).
8. Eshghi, M. et al. Rate of speech decline in individuals with amyotrophic lateral sclerosis. *Sci. Rep.* **12**, 15713 (2022).
9. He, Z. et al. Time of symptoms beyond the bulbar region predicts survival in bulbar onset amyotrophic lateral sclerosis. *Neurological Sci.* **43**, 1817–1822 (2022).
10. Moura, M. C. et al. Prognostic factors in amyotrophic lateral sclerosis: a population-based study. *PLoS ONE* **10**, e0141500 (2015).
11. Cedarbaum, J. M. et al. The alsfrs-r: a revised als functional rating scale that incorporates assessments of respiratory function. *J. neurological Sci.* **169**, 13–21 (1999).
12. Carolan, P. J., Heinrich, A., Munro, K. J. & Millman, R. E. Quantifying the effects of motivation on listening effort: A systematic review and meta-analysis. *Trends Hearing* **26**, 23312165211059982 (2022).
13. Strand, J. F., Brown, V. A., Merchant, M. B., Brown, H. E. & Smith, J. Measuring listening effort: Convergent validity, sensitivity, and links with cognitive and personality measures. *J. Speech Lang. Hearing Res.* **61**, 1463–1486 (2018).
14. Stipancic, K. L. et al. "you say severe, i say mild" Toward an empirical classification of dysarthria severity. *J. Speech, Lang., Hearing Res.* **64**, 4718–4735 (2021).
15. Navar Bingham, I. et al. Listener effort quantifies clinically meaningful progression of dysarthria in people living with amyotrophic lateral sclerosis. *medRxiv* 2024–05 (2024).
16. Yunusova, Y., Weismer, G., Westbury, J. R. & Lindstrom, M. J. Articulatory movements during vowels in speakers with dysarthria and healthy controls. *Journal of Speech, Language, and Hearing Research* 596-611 (2008).
17. Rong, P. et al. Predicting speech intelligibility decline in amyotrophic lateral sclerosis based on the deterioration of individual speech subsystems. *PLoS ONE* **11**, e0154971 (2016).
18. Yunusova, Y., Green, J. R., Lindstrom, M. J., Pattee, G. L. & Zinman, L. Speech in als: Longitudinal changes in lips and jaw movements and vowel acoustics. *J. Med. speech-Lang. Pathol.* **21**, 1 (2013).
19. Krajewski, E., Lee, J., Olmstead, A. J. & Simmons, Z. Comparison of vowel and sentence intelligibility in people with dysarthria secondary to amyotrophic lateral sclerosis. *J. Speech Lang. Hearing Res.* **67**, 1117–1126 (2024).
20. Al-Ali, A. S. et al. Integrating binary classification and clustering for multi-class dysarthria severity level classification: a two-stage approach. *Clust. Comput.* **28**, 136 (2025).
21. Tröger, J. et al. An automatic measure for speech intelligibility in dysarthrias-validation across multiple languages and neurological disorders. *Front. Digital Health* **6**, 1440986 (2024).
22. Sheikh, S. A., Sahidullah, M. & Kodrasi, I. Deep learning for pathological speech: A survey. *arXiv preprint arXiv:2501.03536* (2025).
23. Shabber, S. M. & Sumesh, E. P. Afm signal model for dysarthric speech classification using speech biomarkers. *Front. Hum. Neurosci.* **18**, 1346297 (2024).
24. Devi, V. K., Sreenivas, R., Umamaheshwari, E. & Bacanin, N. Adversarial auto-encoders based model for classification of speech dysarthria. In *2024 15th IEEE International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 1–7 (IEEE, 2024).
25. Ilias, L. & Askounis, D. Recognition of dysarthria in amyotrophic lateral sclerosis patients using hypernetworks. *arXiv preprint arXiv: 10.48550/arXiv.2503.01892* (2025).
26. Radford, A. et al. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, 28492–28518 (PMLR, 2023).
27. Sanguedolce, G., Gruia, D.-C., Naylor, P. & Geranmayeh, F. Latent representation encoding and multimodal biomarkers for post-stroke speech assessment. In *ICLR 2025 Workshop on Foundation Models in the Wild* https://openreview.net/forum?id=b70nL82O5Y (2025).
28. Dubbioso, R. et al. Precision medicine in als: Identification of new acoustic markers for dysarthria severity assessment. *Biomed. Signal Process. Control* **89**, 105706 (2024).
29. Rudzicz, F., Namasivayam, A. K. & Wolff, T. The torgo database of acoustic and articulatory speech from speakers with dysarthria. *Lang. Resour. evaluation* **46**, 523–541 (2012).
30. Baxi, E. G. et al. Answer als, a large-scale resource for sporadic and familial als combining clinical and multi-omics data from induced pluripotent cell lines. *Nat. Neurosci.* **25**, 226–237 (2022).
31. Dubbioso, R. et al. Voice signals database of als patients with different dysarthria severity and healthy controls. *Sci. Data* **11**, 800 (2024).
32. Rong, P., Yunusova, Y., Wang, J. & Green, J. R. Predicting early bulbar decline in amyotrophic lateral sclerosis: A speech subsystem approach. *Behavioural Neurol.* **2015**, 183027 (2015).
33. Friedman, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Am. Stat. Assoc.* **32**, 675–701 (1937).
34. at Carnegie Mellon University, S. G. The CMU Pronouncing Dictionary. http://www.speech.cs.cmu.edu/cgi-bin/cmudict (2014).
35. Tuckute, G. et al. Driving and suppressing the human language network using large language models. *Nature Human Behaviour* 1–18 (2024).

36. Shain, C. Word frequency and predictability dissociate in naturalistic reading. *Open Mind* **8**, 177–201 (2024).
37. Agurto, C. et al. Harnessing remote speech tasks for early als biomarker identification. In *2024 IEEE International Conference on Digital Health (ICDH)*, 161–168 (IEEE, 2024).
38. Yorkston, K. M. Speech deterioration in amyotrophic lateral sclerosis: Implications for the timing of intervention. *Jounal Med. Speech-Lang. Pathol.* **1**, 35–46 (1993).
39. Ball, L. J., Willis, A., Beukelman, D. R. & Pattee, G. L. A protocol for identification of early bulbar signs in amyotrophic lateral sclerosis. *J. Neurol. Sci.* **191**, 43–53 (2001).
40. Wang, J. et al. Automatic prediction of intelligible speaking rate for individuals with als from speech acoustic and articulatory samples. *Int. J. Speech-Lang. Pathol.* **20**, 669–679 (2018).
41. Yunusova, Y. et al. Tongue movements and their acoustic consequences in amyotrophic lateral sclerosis. *Folia Phoniatrica et. Logopaedica* **64**, 94–102 (2012).
42. Green, J. R. et al. Bulbar and speech motor assessment in als: challenges and future directions. *Amyotroph. Lateral Scler. Frontotemporal Degeneration* **14**, 494–500 (2013).
43. Weikamp, J., Schelhaas, H., Hendriks, J., De Swart, B. & Geurts, A. Prognostic value of decreased tongue strength on survival time in patients with amyotrophic lateral sclerosis. *J. Neurol.* **259**, 2360–2365 (2012).
44. Lee, J., Shaiman, S. & Weismer, G. Relationship between tongue positions and formant frequencies in female speakers. *J. Acoustical Soc. Am.* **139**, 426–440 (2016).
45. Kent, R. D. & Kim, Y.-J. Toward an acoustic typology of motor speech disorders. *Clin. Linguist. Phonetics* **17**, 427–445 (2003).
46. Ramig, L. O., Scherer, R. C., Klasner, E. R., Titze, I. R. & Horii, Y. Acoustic analysis of voice in amyotrophic lateral sclerosis: a longitudinal case study. *J. Speech Hearing Disord.* **55**, 2–14 (1990).
47. Darley, F. L., Aronson, A. E. & Brown, J. R. *Motor speech disorders*. (W. B. Saunders, Philadelphia, PA, 1975).
48. Duffy, J. R. et al. *Motor speech disorders: Substrates, differential diagnosis, and management* (Elsevier Health Sciences, 2020), 4th edition edn.
49. Ladefoged, P. & Maddieson, I. *The Sounds of the World's Languages* (Wiley-Blackwell, 1996).
50. Kumar, C. V. T. et al. Spectral analysis of vowels and fricatives at varied levels of dysarthria severity for amyotrophic lateral sclerosis. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 12767–12771 (IEEE, 2024).
51. Pupillo, E. et al. Amyotrophic lateral sclerosis and food intake. *Amyotroph. Lateral Scler. Frontotemporal Degeneration* **19**, 267–274 (2018).
52. Dorst, J. et al. Prognostic factors in als: a comparison between germany and china. *J. Neurol.* **266**, 1516–1525 (2019).
53. Pah, N. D., Motin, M. A. & Kumar, D. K. Phonemes based detection of parkinson's disease for telehealth applications. *Sci. Rep.* **12**, 9687 (2022).
54. Klumpp, P. et al. The phonetic footprint of parkinson's disease. *Computer Speech Lang.* **72**, 101321 (2022).
55. Advancing the diagnosis and prognosis of als from speech. https://www.everythingals.org/available-data. Accessed: 2024-02-14.
56. Neumann, M. et al. Investigating the utility of multimodal conversational technology and audiovisual analytic measures for the assessment and monitoring of amyotrophic lateral sclerosis at scale. *arXiv preprint arXiv:2104.07310* (2021).
57. Nagle, K. F. & Eadie, T. L. Perceived listener effort as an outcome measure for disordered speech. *J. Commun. Disord.* **73**, 34–49 (2018).
58. Gong, Y., Chung, Y.-A. & Glass, J. AST: Audio Spectrogram Transformer. In *Proc. Interspeech 2021*, 571–575 (2021).
59. Deng, J. et al. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255 (2009).
60. Gemmeke, J. F. et al. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 776–780 (2017).
61. Bradshaw, T. J., Huemann, Z., Hu, J. & Rahmim, A. A guide to cross-validation for artificial intelligence in medical imaging. *Radiology Artificial Intelligence* (2023).
62. Wilimitis, D. & Walsh, C. G. Practical considerations and applied examples of cross-validation for model development and evaluation in health care: Tutorial. *JMIR AI* **2**, e49023 (2023).
63. Chicco, D., Warrens, M. J. & Jurman, G. The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. *Peerj computer Sci.* **7**, e623 (2021).
64. Abnar, S. & Zuidema, W. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4190–4197 (Association for Computational Linguistics, 2020).
65. Labra, J., Menon, P., Byth, K., Morrison, S. & Vucic, S. Rate of disease progression: a prognostic biomarker in als. *J. Neurol. Neurosurg. Psychiatry* **87**, 628–632 (2016).

Acknowledgements

We are deeply grateful to the individuals who participated in the study and contributed their data, as well as to the speech-language pathologists who conducted the audio gradings; this work would not have been possible without their support. No funding was granted for this study.

Author contributions

The authors contributed to the following aspects of the work. M.M., C.A., G.A.C. and R.N.: Conceptualization, Investigation, Formal Analysis, Methodology, Validation, Writing—Original Draft; J.P., E.R., A.T., and M.A.T.: Data Curation, Investigation, Writing—Review and Editing; I.N.: Data Acquisition Administration, Writing—Review and Editing; J.D.B., E.F., and L.W.O.: Conceptualization, Formal Analysis, Writing—Review and Editing; All authors have read and approved the manuscript.

Competing interests

Julian Peller, Marcos A. Trevisan and Esteban G. Roitberg are paid consultants to EverythingALS. Alan Taitz was a paid consultant to EverythingALS and his contribution to this work was done while working at EverythingALS. Currently he is an employee of SRI International. James D. Berry has research sup- port from Rapa Therapeutics, MT Pharma of America, ProJenX, Novartis, MDA, ALS Finding a Cure, ALS Association, Association for Frontotemporal Demenia and NINDS. He has been a consultant to Alexion Pharmaceuticals, Amylyx Pharmaceuticals, Biogen, MTPA, Regeneron Pharmaceuticals, Roon, Sanofi and Trace Neuroscience. He is an unpaid scientific advisor to EverythingALS. Indu Navar Bing- ham is the founder of EverythingALS. Lyle W. Ostrow, Ernest Fraenkel, Michele Merler, Carla Agurto, Guillermo Cecchi and Raquel Norel declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-01654-7.

Correspondence and requests for materials should be addressed to Raquel Norel.