Clinical Assessment and Interpretation of Dysarthria in ALS Using Attention-Based Deep Learning AI Models

Supplementary Material

In this supplementary section, we cover the details of the following items:

- Supplementary Note 1: SLPs agreement analysis.
- Supplementary Note 2: Perplexity Calculation: analyzing the complexity of sentences of 11, 13 and 15 words.
- Supplementary Note 3: Attention-based Deep learning AI Models: describing in details the architecture of our proposed models, as well as ablation studies on the effects of different training hyperparemeters on performance.
- Supplementary Note 4: Feature Extraction: regarding the description of the state of the art hand-crafted features plus modeling used as baselines against our attention-based deep learning models, as well as hand-crafted features designed by us on top of the Whisper model.
- Supplementary Note 5: Analysis on Discarded Recordings: looking at our best model's performance on recordings that were discarded according to the filtering described in Section Ground truth calculation in the main text.
- Supplementary Note 6: Longitudinal Analysis.

Supplementary Note 1: SLPs Agreement Analysis

To quantify the agreement between SLPs and assess intra-rater variability, we computed the weighted Cohen's Kappa score (κ) [1]. Only the first grading per audio per SLP is used to calculate the inter-rate variability. We observe high intra-rate agreement, with weighted κ values of 0.87, 0.92, and 0.90, for SLP 1, SLP 2, and SLP 3, respectively. Similar values were obtained when comparing the grading between pairs of SLPs (see the first row of Supplementary Table 1). After removing the files (114) flagged by the SLPs as too noisy or of bad quality, the inter-rate variability between subjects increased, as shown in the second row of Supplementary Table 1. Values increased further, as expected, after removing the recordings with largest discrepancies among the closest rating (see last row of Supplementary Table 1).

Supplementary Table 1 Inter- and intra-rater agreement for speech-language pathologists (SLPs) across different quality-controlled datasets. The first row shows intra-rater agreement with weighted Cohen's Kappa score (κ), reflecting consistency within individual SLPs. The second row shows high inter-rater agreement for comparisons. After removing 114 low-quality files flagged by the SLPs, the third row reflects increased inter-rater variability. Further quality control was applied by excluding 96 recordings where feature extraction failed and 94 recordings with CLES discrepancies greater than 9 points between the two closest SLP ratings, improving inter-rater agreement.

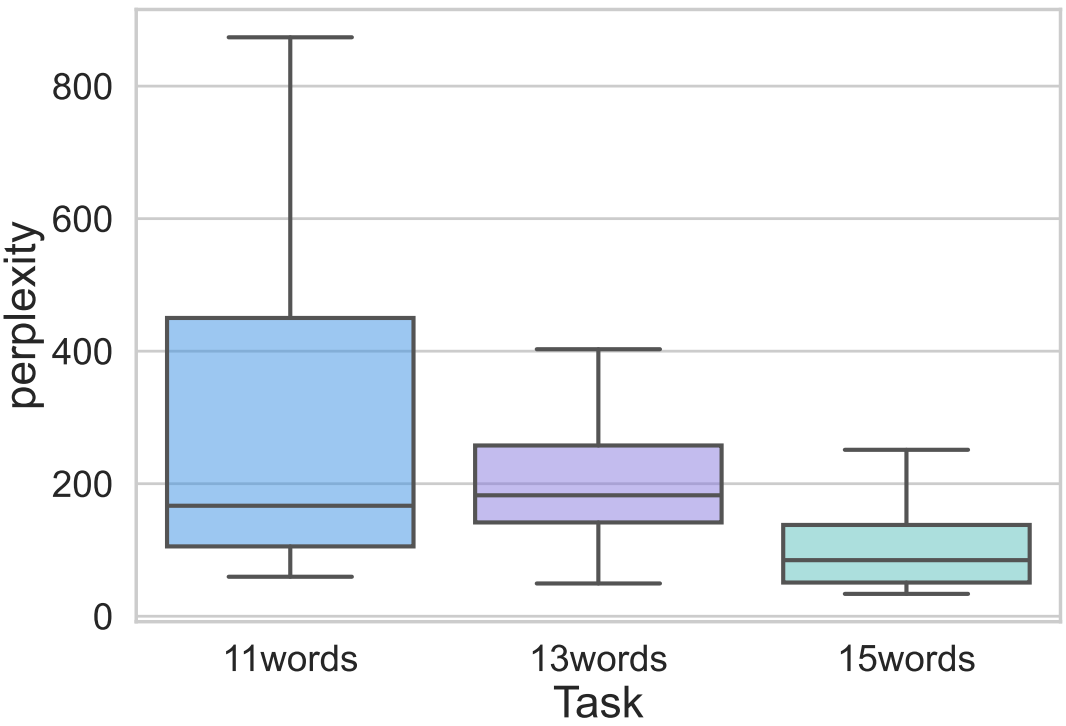
	SLP 1	SLP 2	SLP 3
Intra-rater	0.87	0.92	0.90

	SLP 1 vs. SLP 2	SLP 2 vs. SLP 3	SLP 1 vs. SLP 3
Original	0.90	0.87	0.91
After QC	0.91	0.88	0.92
CLES grading	0.94	0.91	0.94

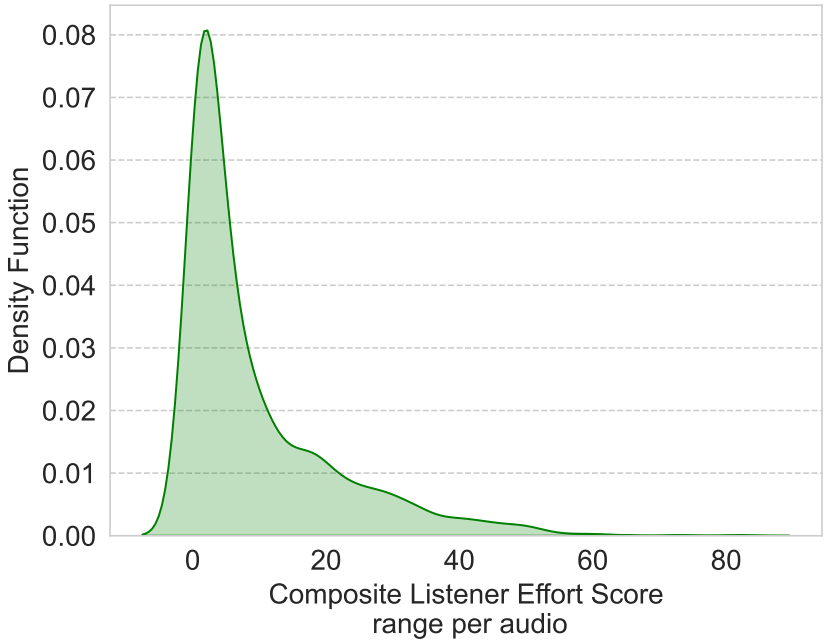
Supplementary Note 2: Perplexity Calculation

We also evaluated the sentences' 'perplexity'. Perplexity, in the context of evaluating sentence complexity within natural language processing, serves as a metric to assess the level of difficulty a probabilistic model encounters in predicting the elements of a sentence. It gauges the intricacy of a sentence by determining how well a language model, which predicts the likelihood of word sequences, can anticipate each subsequent word in a given sentence.

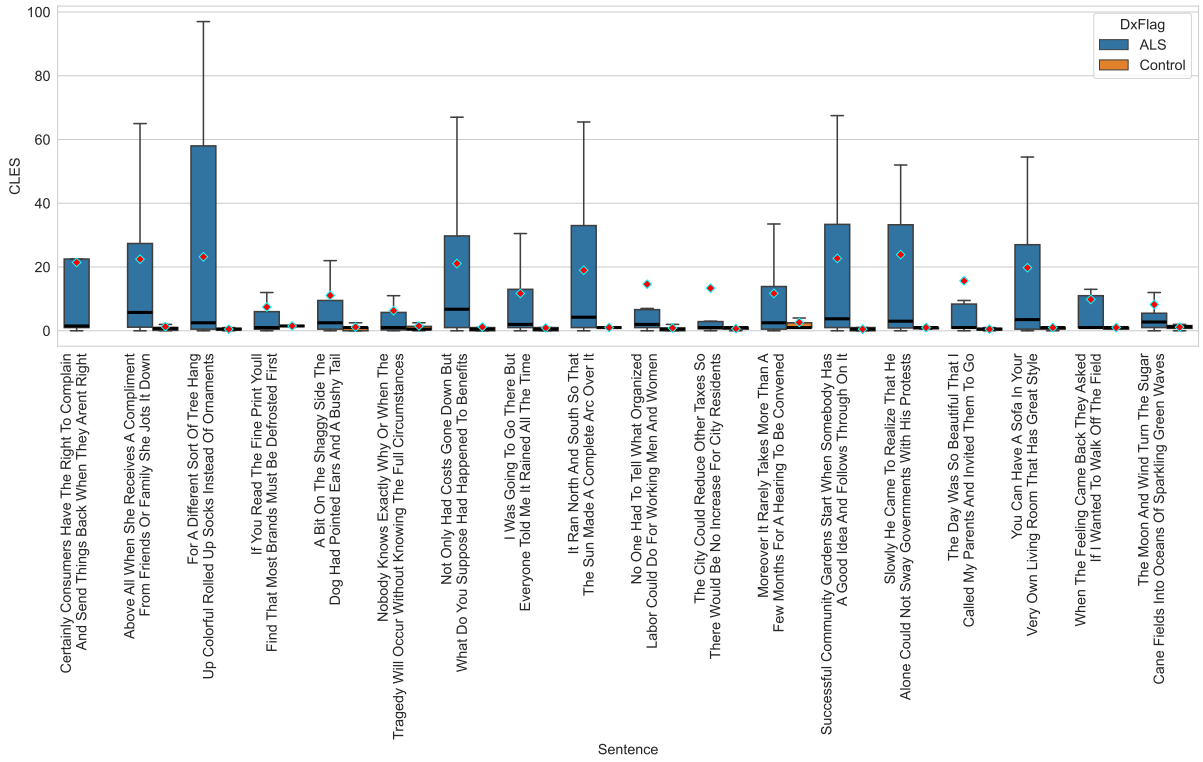
Perplexity for a sentence is obtained by calculating the inverse probability of the sentence according to the model, normalized by the number of words in the sentence. This calculation is often expressed through the exponentiation of the average negative log-likelihood of the words in the sentence. In simpler terms, perplexity measures the weighted average number of choices the model perceives it has for each word it tries to infer. A lower perplexity value signifies that the model finds the sentence less complex, indicating a higher level of predictability or simplicity in the sentence structure according to the model's learned patterns. We analyzed perplexity values and found that 15-word sentences had lower perplexity compared to 13-word sentences (KS = 0.52, p = 0.01), while 11-word sentences showed high variability (Supplementary Figure 1). Additionally, the correlation between CLES and perplexity was not significant (r = 0.02, p = 0.335)



Supplementary Figure 1 Distribution of sentence perplexity scores across the three speech tasks, categorized by sentence length. To investigate the association between sentence length and reading difficulty, we calculated perplexity scores and compared their distributions using the Kolmogorov–Smirnov test. We found a statistically significant difference in the cumulative distribution between 13-word and 15-word sentences ($KS = 0.52, p = 0.01$), with 15-word sentences having lower perplexity, indicating simpler structure and higher predictability. In contrast, 11-word sentences exhibited high variability in perplexity scores. In general, a lower perplexity score suggests that the sentence structure is simpler, as it requires fewer predictive choices according to the model's learned patterns.



Supplementary Figure 2 Distribution of Composite Listener Effort Score range per audio. The high variability in ratings from the three SLPs for a small subset of the data results in a long tail to the right of the distribution.



Supplementary Figure 3 Comparison of Composite Listener Effort Scores (CLES) for different sentences of 15 words. Much variability is noted, particularly in the pALS cohort. Some sentences exhibit low mean and median CLES, while others have high values. Additionally, for certain sentences, individual scores vary significantly

Supplementary Note 3: Attention-based Deep Learning AI Models

Implementation details of the two attention-based deep learning AI models used to analyze listener effort are described below.

AST is a vision transformer model applied to a spectrogram extracted from approximately 10-second audio clips. Following the standard default implementation, a recording is represented by spectrograms with 128 mel bins in the frequency domain and 1024 dimensions in time. The model splits the spectrogram into a set of 16×16 patches with overlap and then linearly projects them to a sequence of 1-D patch embeddings. Each patch embedding is added with a learnable positional embedding. An additional classification token is prepended to the sequence. The output embedding is input to a transformer block, and the output of the classification token is used for classification with a linear layer. We used the base384 model pre-trained on ImageNet and Audioset and fine-tuned it for our regression task by substituting the final linear layer.

Whisper [2] is an encoder-decoder model with a transformer backbone operating on top of spectrograms. Spectrograms are extracted for 30-second temporal windows using 128 log-mel bins in frequency. A positional encoder is followed by the Whisper encoder block, which has two convolutional layers and a transformer with N layers, each comprising self-attention blocks, Gelu activations, and two MLP layers. The decoder block mirrors the encoder one, adding cross-attention layers. We used the encoder part of the Whisper-base model, which has six attention layers and produces a tensor of dimension 512 as output. We added a classification head comprising of a series of lineal layers and Relu activations, like the FT network described in Figure 4. We also tried to use the 512-dimensional output vector of the Whisper encoder directly as features on top of which we trained a simple Lasso regressor. However, we found this approach not to perform as well as fine-tuning the model end-to-end.

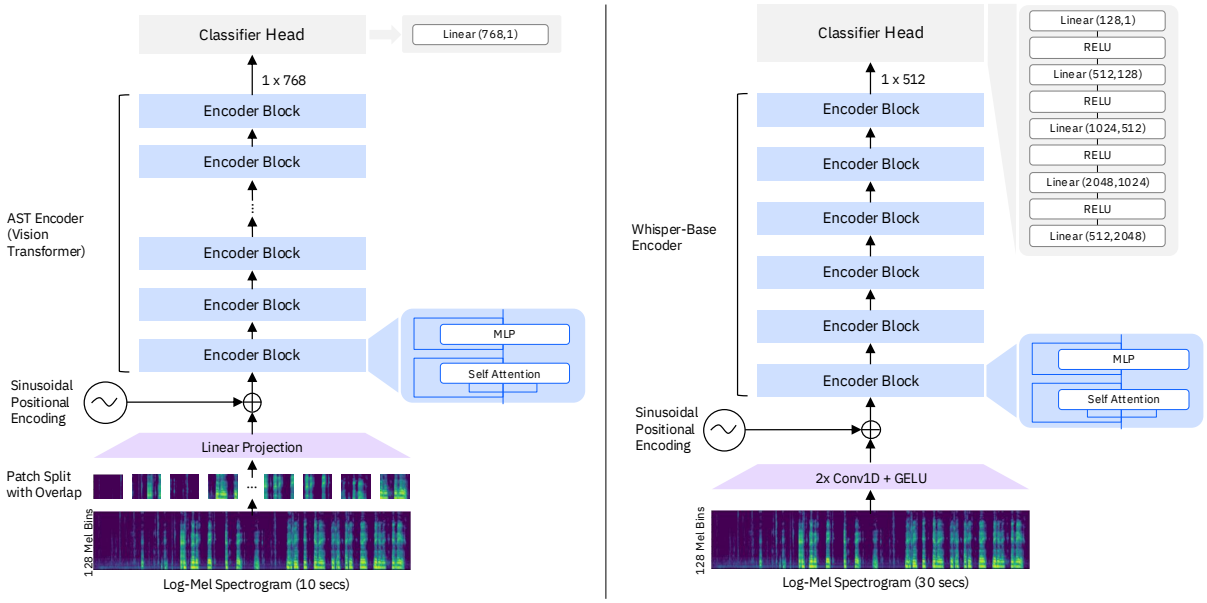
In order to determine the influence of the hyperparameter choices picked in cross-validation on the performance of the Whisper-FT model we conducted a series of ablation studies. Specifically, we studied the effect of learning rate and number of epochs of the fine-tuning process, as well as the size of the whisper encoder model used as base for learning.

The Whisper family of models contains a set of models of different sizes including Base, Small, Medium, Large, Large-v2, Large-v3 and Turbo. For all the experiments reported in Figure 2 (main text) and Supplementary Figure 6 we used the Whisper Base encoder, which processes audio spectrograms into embeddings of dimension 512. We then performed an ablation study to determine the influence of choice of encoder across Whisper model sizes. The list of models that we analyzed and their parameters is reported in Supplementary Table 2. From Supplementary Figure 5 we can see that independently of the choice of model for the encoder of the Whisper-FT, both RMSE and R^2 values are statistically superior to those of the second best model we analyzed (AST). For five folds there is no intersection between standard deviations.

Supplementary Table 2 Variations of Whisper models encoders used to test Whisper-FT performance.

	Number of Parameters	Encoder Embedding Dimension	Pretrained Language
Base	74M	512	English
Small	244M	768	English
Medium	769M	1,024	English
Large	1.55B	1,280	Multilingual
Large-v2	1.55B	1,280	Multilingual
Large-v3	1.55B	1,280	Multilingual
Turbo	809M	1,280	Multilingual

Supplementary Figure 6 reports the effect for five-fold and ten-fold cross validation scores when changing the learning rate in the range $[1e-6, 1e-4]$ and the number of epochs in the range $[5, 50]$ when fine-tuning the Whisper-base model, the same for which results are reported in the tables of Figure 2 (main text). Solid lines represent the mean (RMSE is blue, R^2 in red) and the shaded areas the standard deviation. Note that those are results for the test portion of each fold. For most combinations



(a) (b)

Supplementary Figure 4 Architectures of the two attention-based deep learning AI models used to infer composite listener effort scores: a AST-FT and b Whisper-FT. While implementation details differ, both architectures share the common structure consisting of: input recording spectrogram, projection followed by positional encoding, transformer based encoder, and classifier head.

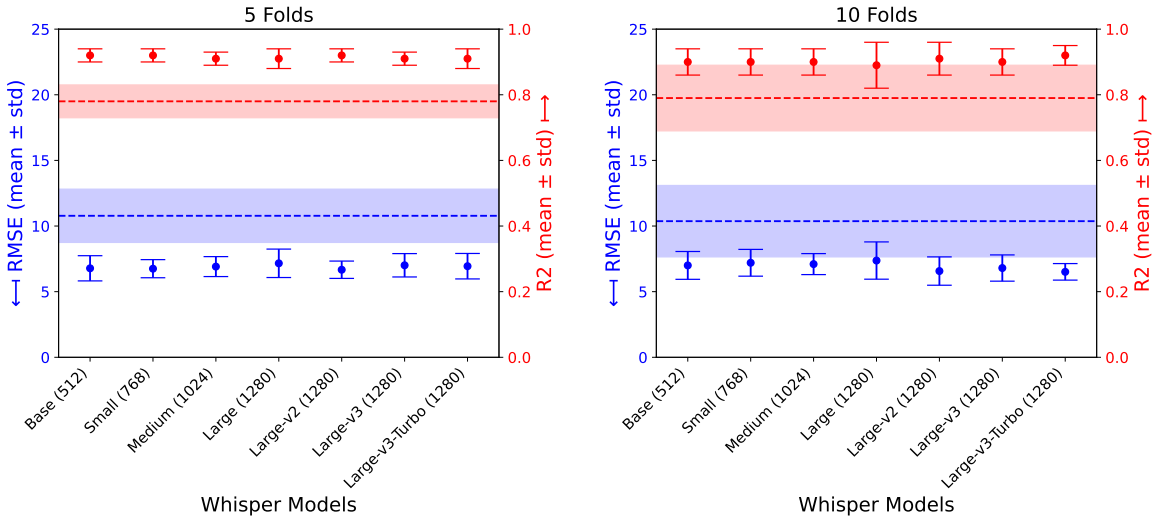
of learning rates and epochs we observe that the performance is substantially homogeneous across parameters values. The only exceptions can be observed for the smallest learning rate 1e-6, where the model requires a larger number of epochs (at least 20) in order to converge to an optimal and stable performance. This is not surprising as lower learning rates require more epochs to properly update the weights of the model to learn the desired signal. Performance with learning rate 1e-4 also fluctuates a bit more across epochs, probably due to the learning rate starting to be too aggressive given the number of training examples. Overall, this ablation study confirms that the performance of the Whisper-FT model can be attributed to the strong representation power of the Whisper encoder for speech patterns coupled with a small deep classification head fine-tuned on top, and not to a particular choice of hyperparameters, as for most combinations of learning rates and number of epochs the test fold validation scores are homogeneous.

Supplementary Note 4: Handcrafted Features

Features from state-of-the-art tools for analyzing speech data were extracted for all the recordings. Specifically, we focus our analysis on these three sets:

- Praat-based: This set is composed of 246 features, which are easily interpretable (i.e. low complexity), and were designed to capture voice properties such as pitch, speech rate, pause duration distribution, voice stability, noise measurements and others. More details on the architecture used to extract these features can be found in [3, 4].
- OpenSMILE-6.3k: These medium complexity features are extracted using a well-known open-source toolbox called OpenSMILE [5]. In particular, we used a feature set called ComParE 2016, which extracts 6,373 features (for full description, please see [6]).
- OpenSMILE-80: Additionally, we used the Geneva Minimalistic Acoustic Parameter Set (GeMAPS v2.0) from OpenSMILE [7], which is a reduced set composed of 80 features.

We used the output of the automatic transcription software (i.e., Whisper [2]) to compute some derived features (31). Specifically, given that the Whisper family comprises models with different levels of complexity (i.e., tiny , medium, large), we compute derived features by comparing the output of



(a) (b)

Supplementary Figure 5 Ablation study: Whisper model size on performance. Whisper model size (encoder final embedding dimension next to name) on performance RMSE (blue) and R^2 (red), mean and standard deviation across five and ten folds, using best learning rate and number of epochs. A dotted straight line is the second-best model (AST).

two models (specifically Tiny and Large) using word error rate metrics [8]. We also added confidence metrics such as *no probability of speech* and *Compression ratio* obtained for each of the following models: Tiny, Base, Small, Medium, Large, Large-v2.

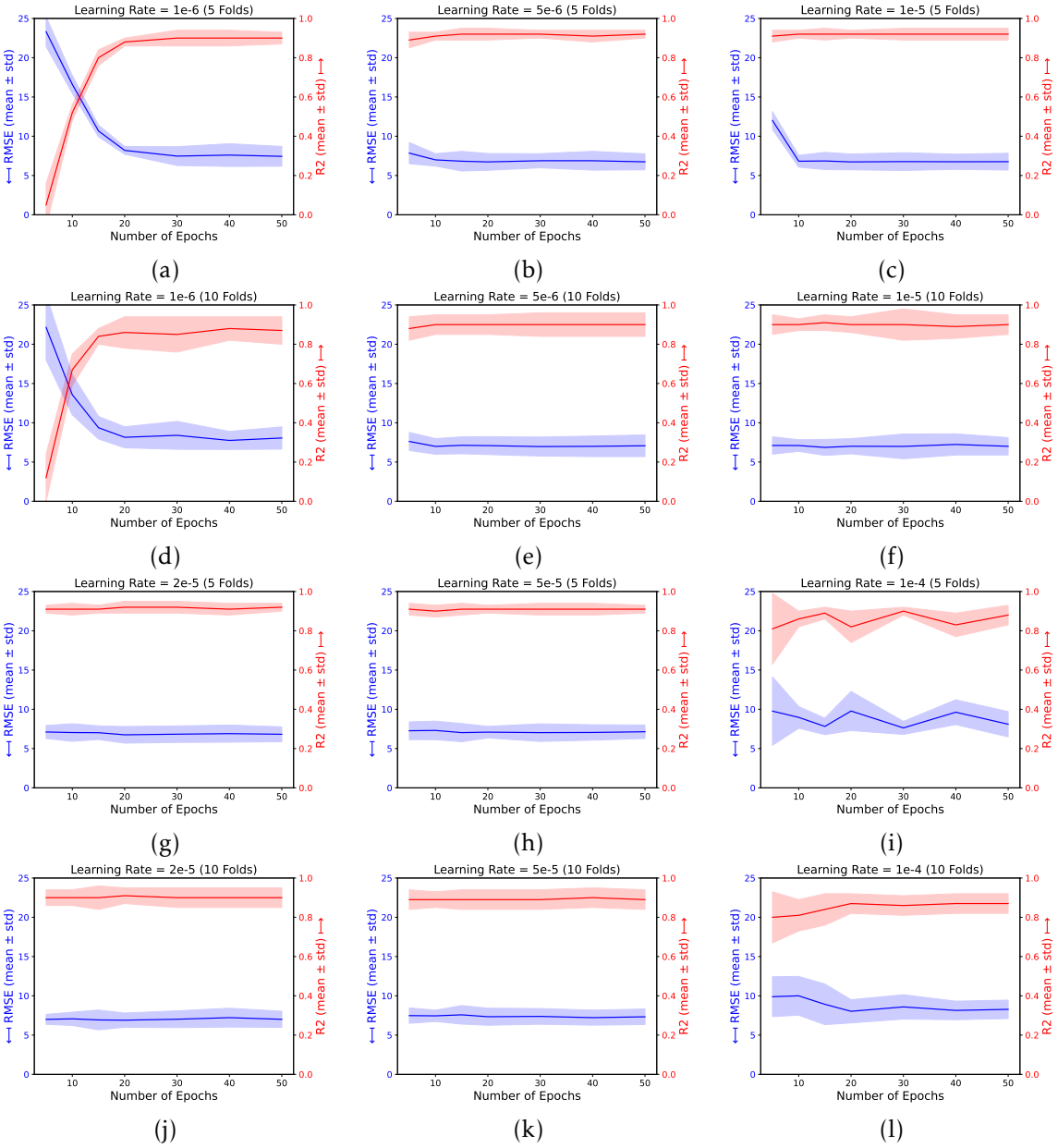
Permutation importance test was performed [9] to determine which features were key for the models developed using handcrafted features. In the case of Praat-derived features, *unvoiced frames*, *audio duration*, *median pitch*, and *speech rate*, among other features, were identified as highly relevant to the model. A similar analysis on the reduced set (80) of OpenSMILE features revealed key features such as *audio duration* and measures of loudness variation. Additional relevant features emerged when using a larger set of OpenSMILE features, including those derived from the audio spectrum. In this expanded feature set, fundamental frequency (F0)-related features also appeared, consistent with the Praat top features. In the case of Whisper-derived features, *no-speech probability* of the Tiny model and *information lost* from the Large to the Tiny model were the most useful. The list of the top 10 features for each modality is in Supplementary Table 4

Supplementary Note 5: Analysis on Discarded Recordings

We inferred CLES in the discarded recordings using the best performance model (Whisper-FT). Then, we calculated the shortest distance between the inferred CLES and the scores given by each of the three SLPs. This approach assumes that, in instances where the raters do not agree, at least one of the scores is likely correct. We compared these distances with a null hypothesis built by shuffling the inferred values. The results are statistically significant (see Supplementary Figure 10) with (KS = 0.24 and p= 0.00002) and medium effect size (Cohen's d = 0.492).

Supplementary Note 6: Longitudinal Analysis

Given the longitudinal nature of our data, we assessed the effectiveness of our approach in capturing the progression of speech deterioration within each participant. From the 124 subjects in our experiments only 6 had a single session due to quality control (e.g., noisy background), while 118 had at least two sessions, with an average and median of 5 sessions per subject. Out of 124 subjects, 96 had sessions over a period longer than six months. In terms of time passed between the first and the last session for each subject, the minimum is 2 months, maximum 27 months, average above 11 months and median



Supplementary Figure 6 Ablation study: effect of (1) number of epochs and (2) learning rate on performance. Performance is computed as RMSE (blue) and R^2 (red), mean and standard deviation across five and ten folds, using Whisper-base encoder for the Whisper-FT model.

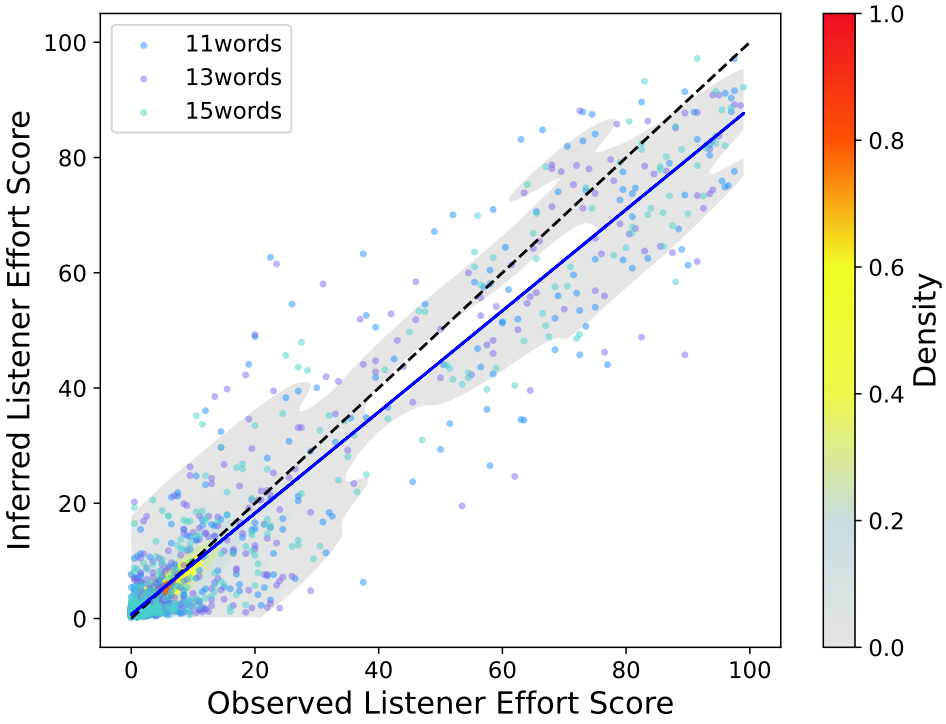
just under 11 months. Fifty-two subjects had at least one year difference between the first and last visit, with an average of almost 8 visits during that timeframe.

To analyze longitudinal trends, we have computed the difference in CLES between the first and last session for each subject, for each task (11 words, 13 words and 15 words) in both ground truth and prediction from our best AI model. We then computed the Pearson correlation of such deltas which are reported in Supplementary Table 5. We can observe from the correlation coefficients and p-values in Supplementary Table 5 that the deltas are indeed statistically correlated. This means that the accuracy of the inference from the model is independent on LE score progression of the individual. The model performs equivalently for early and late stages of the progression and its accuracy does not fluctuate over time. We show in Supplementary Figure 11 four examples where the LE score increases significantly over time, and the predictions from our model mostly follow the ground truth trends.

Supplementary Table 3 Results for five-fold (top) and ten-fold (bottom) experiments for all models generated using handcrafted features and our proposed methodology. RMSE and R^2 scores for different features and models (left). Steiger's Z test for dependent correlation coefficients determines that results obtained with FT are statistically significantly higher ($p < 0.00001$) than the ones obtained with Lasso regres- sion. Among all the results, the models developed with Whisper FT are statistically significantly better than the other methods ($p < 0.00001$).

Model Type + Classifier	Details	Dimension	RMSE ↓ (Mean ± std)	R2 ↑ (Mean ± std)
Handcrafted Features + Lasso	Praat-based	246	13.80 ± 0.84	0.67 ± 0.04
Handcrafted Features + Lasso	OpenSMILE-80	80	14.63 ± 1.76	0.62 ± 0.09
Handcrafted Features + Lasso	OpenSMILE-6.3k	6373	13.35 ± 0.89	0.69 ± 0.03
Handcrafted Features + FT	OpenSMILE-6.3k	6373	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**

Model Type + Classifier	Details	Dimension	RMSE ↓ (Mean ± std)	R2 ↑ (Mean ± std)
Handcrafted Features + Lasso	Praat-based	246	13.98 ± 2.54	0.61 ± 0.17
Handcrafted Features + Lasso	OpenSMILE-80	80	14.37 ± 2.55	0.61 ± 0.13
Handcrafted Features + Lasso	OpenSMILE-6.3k	6373	13.25 ± 1.81	0.67 ± 0.10
Handcrafted Features + FT	OpenSMILE-6.3k	6373	11.62 ± 2.12	0.74 ± 0.10
Attention-based AI model	AST-FT	-	10.37 ± 2.72	0.79 ± 0.10
Attention-based AI model	**Whisper-FT**	-	**7.00 ± 1.06**	**0.90 ± 0.04**



Supplementary Figure 7 CLES Prediction results using ten-fold cross validation. Inferred composite listener effort scores plotted against the observed CLES by SLPs for the best attention-based AI model (Whisper-FT variant), showing high correlation.

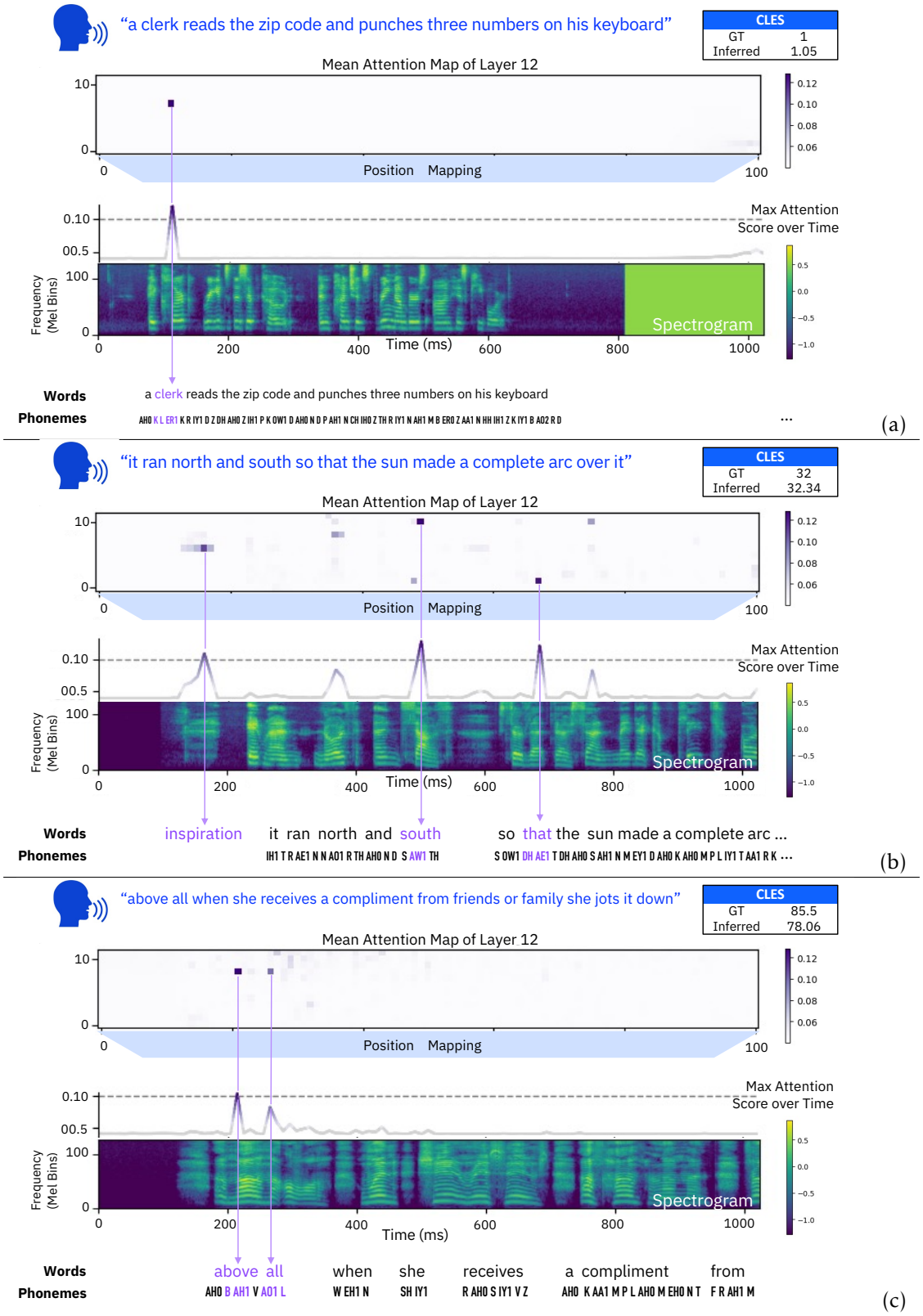
Supplementary Table 4 Top 10 relevant features for models based on lasso regression using five-fold cross validation. Permutation importance was used to calculate coefficients for ranking the features.

Rank	Praat-lasso	Whisper-derived lasso
1	Unvoiced Frames	No speech probability tiny model
2	Audio duration	Information lost from large to tiny model
3	MFCC 1 svd	Match error rate from large to tiny model
4	Formant 3 - 5th percentile	No speech probability largev2 model
5	Norm. total vowel duration	Average log probability base model
6	MFCC 3	Compression ratio largev2 model
7	Formant 2 - 95th percentile	Temperature tiny model
8	Vowel Space overlap	Compression ratio large model
9	Median pitch	Average log probability tiny model
10	Speech rate	Temperature medium model

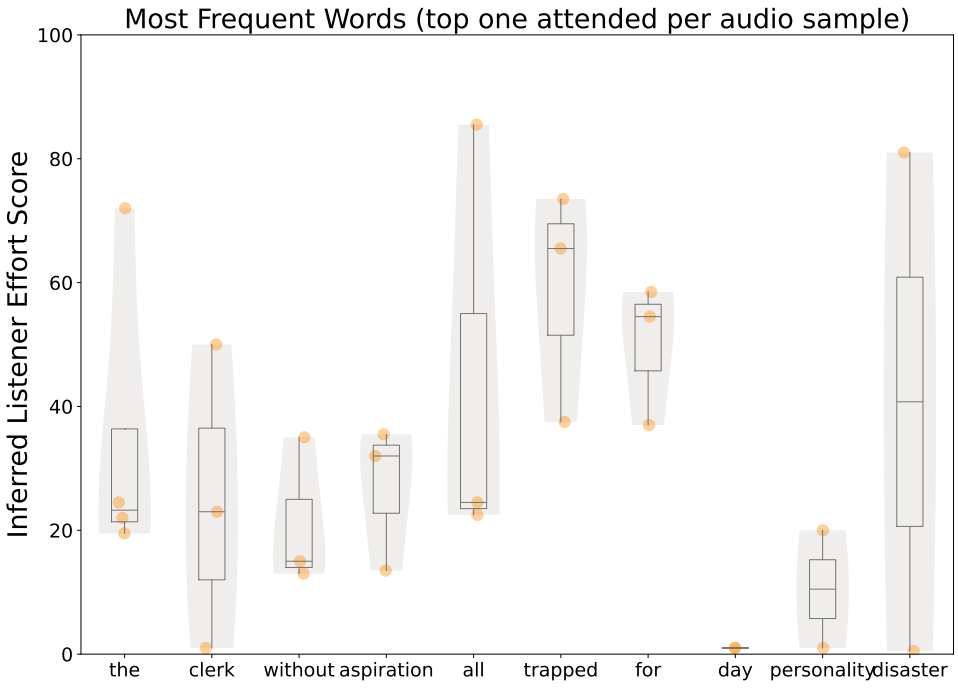
Rank	OpenSMILE-80-lasso	OpenSMILE-6.3K-lasso
1	Audio duration	audspec_lengthL1norm_sma_meanPeakDist
2	F2amplitudeLogRelF0_sma3nz_stddevNorm	pcm_RMSenergy_sma_meanPeakDist
3	loudness_sma3_meanRisingSlope	mfcc_sma[1]_pctlrange0-1
4	mfcc1V_sma3nz_amean	mfcc_sma[1]_rqmean
5	F3amplitudeLogRelF0_sma3nz_stddevNorm	audspec_lengthL1norm_sma_de_lpc0
6	HNRdBACF_sma3nz_amean	F0final_sma_quartile2
7	loudness_sma3_percentile80.0	audspec_lengthL1norm_sma_stddevRisingSlope
8	loudnessPeaksPerSec	audspec_lengthL1norm_sma_iqr2-3
9	loudness_sma3_pctlrange0-2	mfcc_sma[6]_lpc0
10	logRelF0-H1-H2_sma3nz_amean	mfcc_sma[3]_pctlrange0-1

Supplementary Table 5 Pearson correlation coefficients and p-values comparing deltas between ground truth and predicted CLES in first and last session for each patient, for each task.

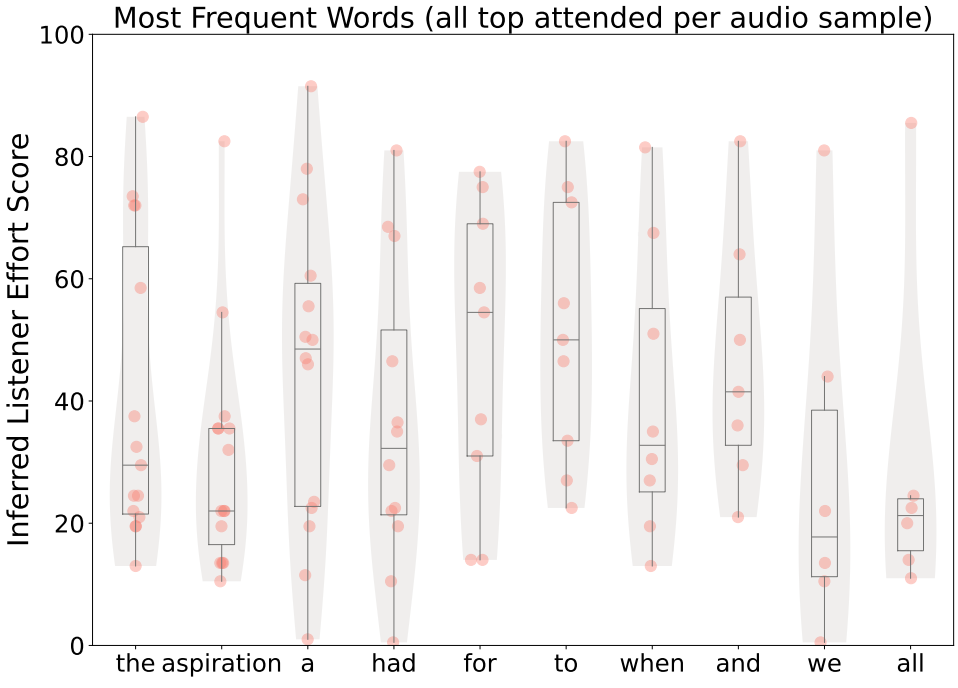
Task	Correlation	p-value
11 Words	0.68	$6.77e^{-18}$
13 Words	0.66	$6.11e^{-17}$
15 Words	0.80	$1.15e^{-28}$



Supplementary Figure 8 Visualization of spectrogram and attention maps for the AST model on different levels of composite listener effor score (GT). **a** 1 - easy to understand, **b** 32 - can be understood with some effort, **c** 85.5 - hard to understand. The model pays attention to the words (and phonemes) highlighted in red to generate the reported inferred composite listener effort scores (CLES).

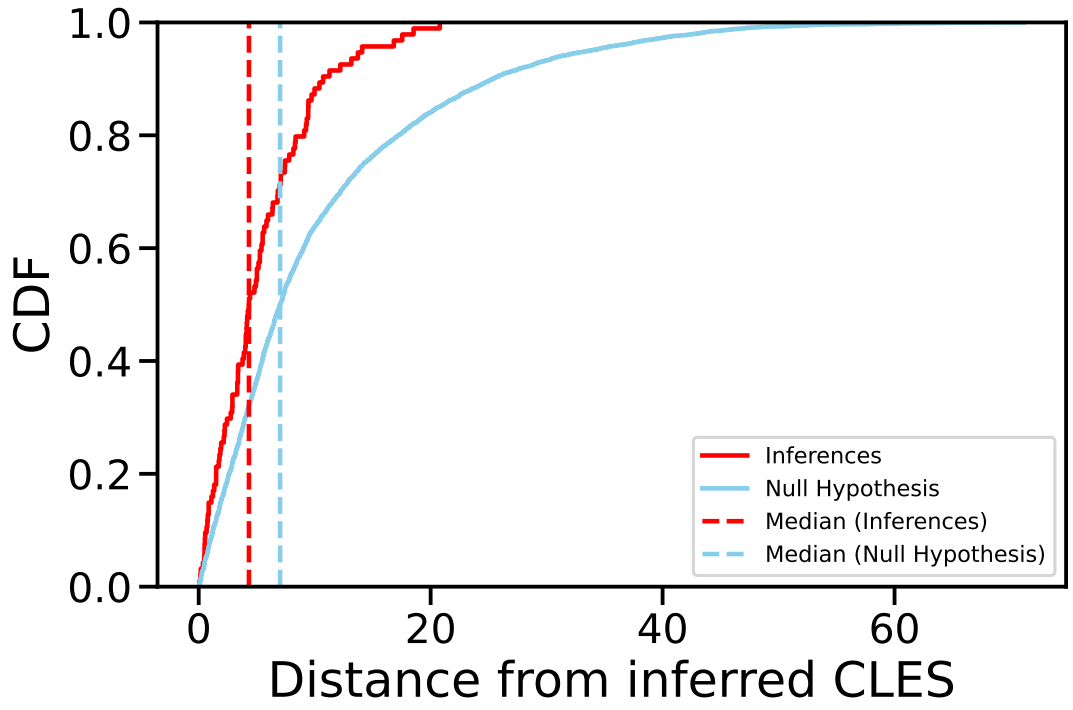


(a)

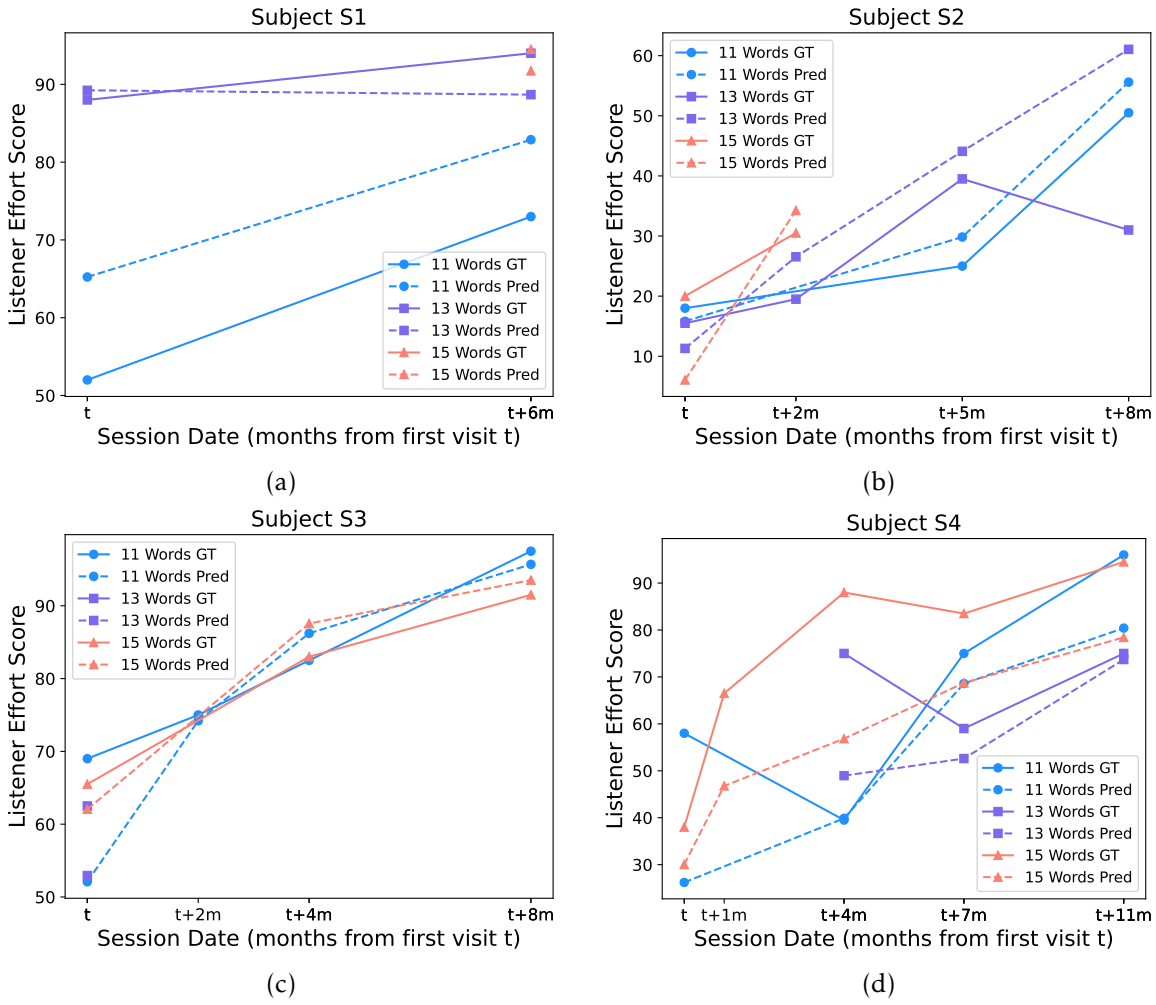


(b)

Supplementary Figure 9 Distribution of the top ten most common words the AST model attends to across composite listener effort score (CLES). Top ten **a** all attended words per audio sample **b** top-1 attended words per audio sample.



Supplementary Figure 10 Cumulative distribution of distances between inferred composite listener effort scores (using Whisper-FT model) and the ratings from each of the three SLPs. We assume that at least one rating is accurate even in cases of disagreement. The red line represents the observed distances, while the blue line represents the null hypothesis generated by shuffling inferred score values. Vertical lines indicate the median of each distribution. The statistically significant difference (KS = 0.24, p = 0.00002, Cohen's d = 0.492) highlights that the inferred CLES values align more closely with at least one rater than random chance.



Supplementary Figure 11 Longitudinal analysis of listener effort score across sessions over a period greater than six months for four subjects. The predictions of our best attention-based AI model (Pred) mostly follow the trends of the ground truth (GT) deterioration progression registered by the SLPs.

Supplementary References

[1] Cohen, J. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin* **70**, 213 (1968).

[2] Radford, A. *et al.* Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, 28492–28518 (PMLR, 2023).

[3] Agurto, C. *et al.* Analyzing progression of motor and speech impairment in als. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 6097–6102 (IEEE, 2019).

[4] Wen, B. *et al.* Accelerating automation of digital health applications via cloud native approach. In *Proceedings of the Eighth International Workshop on Container Technologies and Container Clouds*, 1–6 (2022).

[5] Eyben, F., Wöllmer, M. & Schuller, B. W. Opensmile: the munich versatile and fast open-source audio feature extractor. In Bimbo, A. D., Chang, S.-F. & Smeulders, A. W. M. (eds.) *ACM Multimedia*, 1459–1462 (ACM, 2010). URL http://dblp.uni-trier.de/db/conf/mm/mm2010.html#EybenWS10.

[6] Weninger, F., Eyben, F., Schuller, B. W., Mortillaro, M. & Scherer, K. R. On the acoustics of emotion in audio: what speech, music, and sound have in common. *Frontiers in psychology* **4**, 292 (2013).

[7] Eyben, F. *et al.* The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing* **7**, 190–202 (2016).

[8] Morris, A. C., Maier, V. & Green, P. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *Eighth International Conference on Spoken Language Processing* (2004).

[9] Breiman, L. Random forests. *Machine Learning* **45**, 5–32 (2001).