Linguistic Features of Automatically Transcribed Speech from ALS Patients

Anirudh Arbatti, Michael Neumann, David Suendermann-Oeft, and Vikram Ramanarayanan
Modality.AI, Inc., San Francisco, CA, USA

v@modality.ai

How accurate is Whisper ASR for impaired speech?

Which linguistic features show differences between pALS and HC?

What is the relative utility of linguistic features within a multimodal assessment of ALS?

Motivation and Research Questions

- **People with ALS (pALS) may present with deficits in verbal fluency** due to neuronal loss. Therefore **analysis of linguistic abilities** can provide insights into disease progression (*Boschi et al., Frontiers in Psychology 2017*)
- **Objective: analysis of linguistic characteristics** in pALS (as additional source of information in a multimodal assessment approach)
- Analytical validation: **Can we extract such features reliably from automatic speech recognition (ASR)**, even when speech is impaired?
- Clinical validation: In addition to previously validated speech and facial measures *(Neumann et al., Interspeech 2021; Kothare et al., Interspeech 2023; Neumann et al., Computers in Biology and Medicine 2024)*, **which linguistic features are clinically meaningful** for remote assessment of ALS?

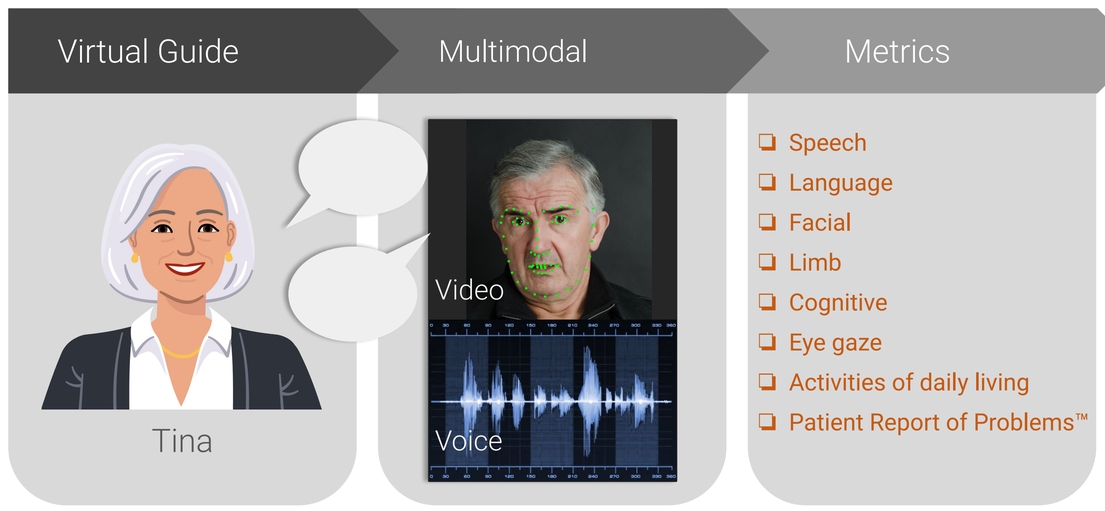

Figure 1. Schematic diagram of the Modality.AI dialogue platform.

Methods

- **Audiovisual recordings from 72 pALS and 60 age and sex matched healthy controls (HC)** were recorded using a web-based multimodal dialog platform (Fig. 1); 397 sessions in total
- **Acoustic, orofacial, and linguistic features were automatically extracted** from established tasks, including DDK, reading passage (RP), sentence intelligibility test (SIT), and picture description (PD), see Table 1
- Linguistic features were extracted based on ASR transcriptions using *spaCy* (**compared two ASR models: Whisper and AWS Transcribe**)
- For analytical validation, a subset of 68 samples from 18 pALS with varying degree of speech impairment was transcribed manually
- **Analytical validation: word error rate** (WER) between transcriptions and ASR output, and **mean absolute error (MAE) for linguistic features** extracted from ASR output (as compared to features extracted from transcriptions)
- **Clinical validation: Kruskal-Wallis tests for all features** to identify statistical differences between pALS and HC (only pALS with symptom onset within three years prior to study enrollment were included, n=60)

	Domain	Features
Audio	Energy	shimmer (%), intensity (dB), signal-to-noise ratio (dB)
	Timing	speaking duration (sec.), speaking rate (WPM), percent pause time (PPT, %), canonical timing alignment (CTA, %), cycle-to-cycle temporal variability (cTV, sec.), syllable rate (syl./sec.), number of syllables
	Voice Quality	cepstral peak prominence (CPP, dB), harmonics-to-noise ratio (HNR, dB)
	Frequency	mean, max., min. fundamental frequency F0 (Hz), first three formants F1, F2, F3 (Hz), slope of 2nd formant (Hz/sec.), jitter (%)
Text	Lexico-semantic	Noun Rate, Verb Rate, Demonstrative Rate, Pronoun Rate, Adjective Rate, Adverb Rate, Conjunction Rate, Possessive Rate, Noun-to-pronoun ratio, Closed-Class Word Rate, Open-Class Word Rate, Content Density, Honore's Statistic, Brunet's Index, Type-Token Ratio
	Morphosyntactic	Inflected Verb Rate, Auxiliary Verb Rate, Gerund Rate
	Discourse-Pragmatic	Word Count, Number of Subjects, Number of Objects, Number of Places, Number of Actions
	Syntactic	Average Dependency Tree Height
Video	Mouth (distances)	lip aperture/opening, lip width, mouth surface area, mean symmetry ratio between left and right half of the mouth
	Lip/Jaw Movement	velocity, acceleration, and jerk of lower lip and jaw center
	Eyes	number of eye blinks per sec., eye opening, vertical displacement of eyebrows

Table 1. Overview of multimodal features.

Analytical Validation

- Acoustic and visual features have been validated in previous work
  - Liscombe et al. evaluated speech event detection, crucial for extracting acoustic features *(Liscombe et al., Motor Speech Conference 2022)*
  - Zhang et al. assessed the accuracy of automatic facial landmark detection *(Zhang et al., Motor Speech Conference 2024)*
- Accuracy of ASR is key for text based features: We found the **WER between Whisper ASR and manual transcriptions (13.3%)** was significantly smaller than for AWS Transcribe (26.8%)
- **Normalized mean absolute error** for linguistic features based on ASR vs. manual transcriptions **ranged from 0.95% to 8.96%**

Clinical Validation

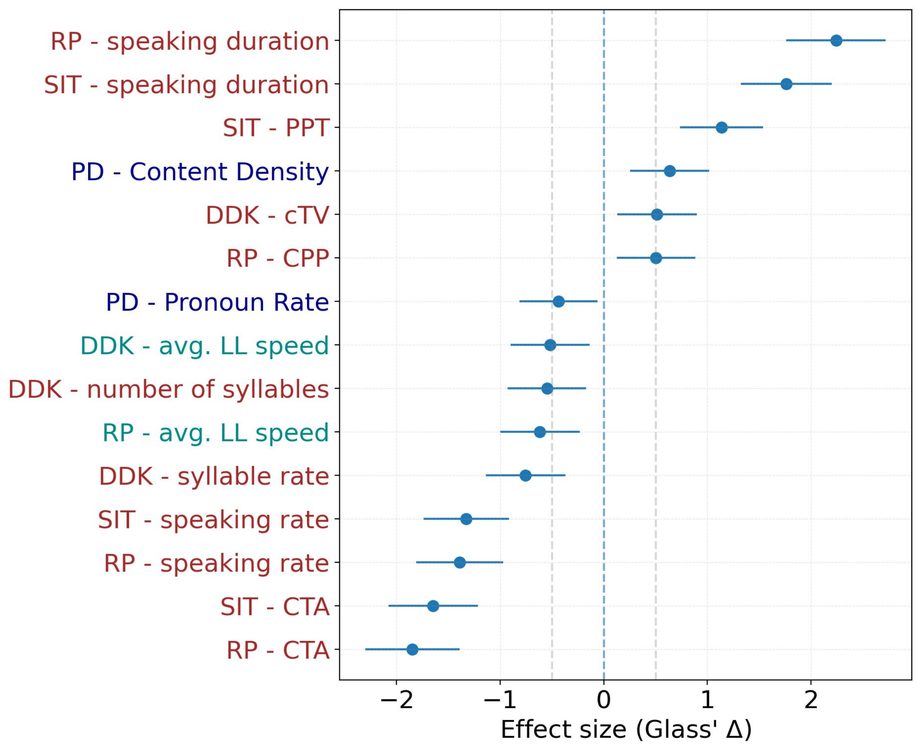

Figure 2. Effect sizes as Glass'Δ of **speech**, **orofacial**, and **linguistic** features that show statistically significant differences between pALS and HC at α = 0.05.
(Positive effect sizes indicate that feature values for pALS are on average greater than for HC)

- **Large effects for timing related speech measures**, such as speaking duration, PPT, CTA, speaking rate
- Among linguistic features, **content density (CD) and pronoun rate (PR)** show statistically significant differences between cohorts
- Possible interpretation: Increased CD and decreased PR in pALS suggest a **preference for content-rich words to maximize clarity and efficiency in effortful speech**

Conclusion: Beyond well established acoustic features, linguistic characteristics such as *content density* and *pronoun rate* reveal not just *how* patients speak, but can provide more insight into *what they say*

This work was funded by the National Institutes of Health grant R42DC019877. We thank all study participants for their time and we gratefully acknowledge the contribution of the Peter Cohen Foundation and EverythingALS towards participant recruitment and data collection.