Remote Assessment for ALS using Multimodal Dialog Agents: Data Quality, Feasibility and Task Compliance

Vanessa Richter[⋆], Michael Neumann[⋆], Jordan R. Green[*], Brian Richburg[*], Oliver Roesler[⋆], Hardik Kothare[⋆] and Vikram Ramanarayanan[⋆†]

[⋆]Modality.AI, Inc., San Francisco, CA
[*]MGH Institute of Healthcare Professions, Boston, MA
[†]University of California, San Francisco, CA

vikram.ramanarayanan@modality.ai

Abstract

We investigate the feasibility, task compliance and audiovisual data quality of a multimodal dialog-based solution for remote assessment of Amyotrophic Lateral Sclerosis (ALS). 53 people with ALS and 52 healthy controls interacted with Tina, a cloud-based conversational agent, in performing speech tasks designed to probe various aspects of motor speech function while their audio and video was recorded. We rated a total of 250 recordings for audio/video quality and participant task compliance, along with the relative frequency of different issues observed. We observed excellent compliance (98%) and audio (95.2%) and visual quality rates (84.8%), resulting in an overall yield of 80.8% recordings that were both compliant and of high quality. Furthermore, recording quality and compliance were not affected by level of speech severity and did not differ significantly across end devices. These findings support the utility of dialog systems for remote monitoring of speech in ALS.

Index Terms: dialog systems, speech processing, multimodal systems, amyotrophic lateral sclerosis

1. Introduction

The demand for accurate, low-cost, and remotely administered speech assessments is surging from multiple health care sectors including providers, pharmaceutical companies, and educational institutions. The wide appeal of using speech as a diagnostic marker is because (a) changes in speech are associated with a large number of developmental, psychiatric, and neurological conditions [1], and (b) the ease with which speech data can now be collected remotely due to the ubiquity of multimedia enabled personal devices [2]. Remote patient monitoring (RPM) solutions are particularly needed for in-home monitoring of progressive neurological diseases such as Amyotrophic Lateral Sclerosis (ALS) [3, 4, 5]. Despite the ease of capturing speech remotely and advancements in speech and video analytics, there are persistent challenges related to remote administration and recording that, if not addressed, can decrease clinical utility [6, 7]. These include issues that affect signal quality such as environmental noise, cross talk, poor lighting, internet bandwidth, video pixelation, etc. In addition, there are a variety of user-related issues such as correct head position with respect to the image frame, the presence of multiple talkers/faces, optimal distance to the microphone, participants wearing glasses, etc. that can impact the accuracy and reliability of the measured signals. Finally, if the assessment is self-administered, patients may have difficulty complying with the test instructions remotely on their own without the help of a clinician.

Mitigating these challenges is essential for optimizing both user acceptance and accuracy of metrics calculated based on the recorded signal. In this study, we analyze the quality of audio and video recorded by a multimodal dialog-based RPM solution to evaluate the feasibility of such technology for real-world deployments. We define rubrics as part of a human-expert-based evaluation of quality because there is no existing gold-standard to measure quality in *clinical* settings. Manual rubrics allow for tailored rating, which is critical given the unique factors and requirements of audio and video quality in remote assessments. In addition, human raters are able to identify and analyze specific problems, which helps provide actionable guidance for improvements. The outcomes provide data-driven guidance for (1) improving the pre-assessment of each participant's hardware and software, (2) rewording ambiguous standardized operating procedures, participant instructions, and test items, (3) improving algorithms to monitor signal and output quality, and (4) improving the virtual agent scripts and timing of responses to improve interaction and maximize task compliance and completion.

2. Data

We collected audiovisual recordings from people with ALS and from healthy controls as they interacted with Tina, a cloud-based virtual agent for remote patient assessment, in cooperation with EverythingALS and the Peter Cohen Foundation[1]. The multimodal dialog system NEMSI used for data collection is HIPAA compliant and the study was approved by the Advarra Institutional Review Board (IRB). [2]

Subjects were provided with a website link to the secure screening portal and login credentials by their caregiver or study liaison (physician, clinic, a referring website or patient portal). Study participants did not receive help from their caregiver or study liaison in conducting the session with the virtual agent. Each session consists of a structured set of speech tasks, including, among others, a picture description task (where participants are asked to describe an image shown on their screen) and a diadochokinesis (DDK) task (where participants rapidly repeat the syllables /pɑtɑkɑ/ until they run out of breath). For more

The authors gratefully acknowledge support from NIH Grant R42DC019877, and an ongoing partnership with EverythingALS and the Peter Cohen Foundation for participant recruitment, management and advocacy.

[1] https://www.everythingals.org/research

[2] IRBs use a group process to review research protocols and related materials. The aim of IRB review is to ensure the protection of rights and welfare of humans participating in the research. https://www.fda.gov/about-fda/center-drug-evaluation-and-research-cder/institutional-review-boards-irbs-and-protection-human-subjects-clinical-trials

details on the complete protocol and data collection, see [8]. After dialog completion, participants filled out the *ALS Functional Rating Scale-revised (ALSFRS-R)*, a standard instrument for monitoring the progression of ALS [9]. The questionnaire consists of 12 questions about physical functions in activities of daily living. Each question provides five answer options, ranging from *normal function* (score 4) to *severe disability* (score 0). The speech sub score (one question, score 0 to 4) is of particular interest as a measure of level of speech impairment.

When monitoring progressive diseases, such as ALS, it is important that the test is feasible and recordings of sufficient quality for the full range of patient disease severity. Patients in the more severe stages may have increased difficulties completing tasks and adhering to test instructions. Therefore, for this study, we selected two tasks to assess audio/video quality – the DDK task and the picture description task – from participants with ALS who spanned different levels of speech impairment severity and healthy controls. We also ensured adequate coverage across sex and age range (18-80 years). Table 1 shows the distribution of speech severity levels as measured by the first ALSFRS-R question. Note that speech sub scores of 0 and 1 were not represented in the available data, and so could not be chosen for analysis. Also note that healthy controls had a speech score of 4 (and therefore this category is over-represented). The final dataset for this study comprised 250 audio recordings (sampling rate of 44.1 kHz) and 250 corresponding video recordings (resolution of 320 x 240 pixels and a frame rate of 15 frames per second), each resulting in a total duration of 161.2 minutes of data. The set of 250 (129 patients & 121 controls) recordings contained multiple sessions from some participants. Further, some recording sessions included multiple utterances for one task.

Table 1: *Distribution across speech severity levels.*

ALSFRS-R speech score	4	3	2	1	0
Number of recordings	190	44	16	-	-

3. Methods

We rated both audio and video quality on a Likert scale [10] ranging from 1 (very poor) to 5 (very good). We designed the rating rubric to measure the suitability of the audio and video signal for clinically-relevant feature extraction. Scores of 3 and above were deemed *acceptable*, while 2 and 1 ratings were deemed *unacceptable*. We rated the quality of recordings using a two step process. During the first step, we identified distortion categories and assessed preliminary scores for the clear edge cases (5 and 1 ratings). We then refined the rating rubric based on our observations. All recordings were rated by one speech researcher, and audio and video was rated separately. To assess inter-rater agreement, a second speech researcher rated the audio and video corresponding to 125 DDK samples, which comprise half the dataset.

3.1. Audio quality rating

For audio, we identified the following distortion categories:

a) *Echo* (EC): delayed reflection of the participant's speech that is captured in the recording.
b) *Packet loss* (P): audio cuts out for short periods, which is most likely caused by packet loss during network transmission.
c) *Signal noise* (S): this subsumes noises like hiss, hum, and crackling (predominantly caused by the recording equipment).
d) *Environmental noise* (EN): all disturbances that are caused by the surroundings, such as barking dogs, traffic noise, or ticking clocks.

For each of the above defined problem categories, we defined the acceptable degree of distortion severity for each rating score. The lowest scored category determined the overall rating for a recording. The Likert scale audio ratings are based on dominance of the distortions in terms of intensity, frequency, and their position within the recording (occurring during speech or non-speech portions of the signal) and thereby, the potential negative impact to automatic speech analysis. This is reflected in the following evaluation rubric:

1. *Very poor (1)*: Voice barely or not intelligible at all; EC: Participant's speech reflected throughout the entire recording as intensely as the speech signal; P: audio cut off completely or frequently for significant parts; S: Steady signal noise more intense than the speech signal; EN: Intense and frequent noises occurring within the speech portions more dominant than the speech signal.
2. *Poor (2)*: Voice intelligible with artifacts heavily affecting speech portions; EC: Participant's speech reflected throughout major parts of the recording as intensely as the speech signal; P: Frequent, but short audio cut offs; S: Steady signal noise as intense as the speech signal; EN: Intense and frequent noises occurring within the speech portions as dominant as the speech signal.
3. *Acceptable (3)*: Predominant speech signal with few artifacts affecting speech portions directly; EC: Medium intensity echo less dominant than the speech portions; P: Few short cut offs; S: Steady signal noise and few single artifacts at medium intensity; EN: Several occurrences of medium intensity environmental noise in non-speech portions of the signal.
4. *Good (4)*: Mostly clean speech signal with artifacts generally not affecting speech portions directly; EC: Very low intensity echo; P: No cut offs; S: Steady signal noise and few single artifacts at low intensity; EN: Infrequent occurrence of low intensity environmental noise.
5. *Very good (5)*: Clean speech signal; EC: No echo; P: No cut offs; S: Hardly perceivable steady signal noise or single artifact at very low intensity; EN: Single or no occurrence at very low intensity.

3.2. Video quality rating

We rated video quality with a primary focus on the face region, since our goal for this exercise is to ensure adequate quality for the extraction of facial kinematics-based speech measures for ALS. In other words, we relatively down-weighted issues affecting other parts of the video, such as a bright or blurry background, because of their relatively lower contribution to the accuracy of the facial features derived from the video. With that in mind, we identified the following distortion categories:

a) *Pixelation & blocking artifacts* (PB): Pixelation refers to how blurry or fuzzy the signal is, while blocking artifacts are concerned with large blocks of pixels grouping together.
b) *Freezing* (F): a frame being 'stuck' as still picture.
c) *Bad lighting conditions* (L): refers to problems with brightness/darkness and contrast, affecting the visibility of (parts of) the face.
d) *Head pose and movement* (HPM): movements or head posi-

tion (tilted head or instances of the face being out of frame) that can affect the accuracy of facial landmark detection.

e) *Distance to the camera* (D): refers to how far the user is away from the camera.
f) *Glasses* (G): subsumes all problems with glasses impairing the visibility of eyes or eyebrows.
g) *Multiple faces*: problems with automatic processing can occur when multiple faces are at similar distance to the camera.

In line with the audio quality assessments, we defined the acceptable degree of distortion severity for each rating score per problem category, with the lowest scored category determining the overall rating for a video.

1. *Very poor (1)*: Face not or barely visible; PB & F: extreme blocking artifacts and/or freezing and pixelation issues; L: extremely dark or bright; HPM: extremely tilted, excessive movements, head out of frame for major parts of the video or too close so that face is occluded; D: face very small in relation to the frame; G: heavy reflections (eyes/eyebrows not visible at all); more than one person's face clearly visible.
2. *Poor (2)*: Face not well and/or not fully visible; PB & F: major artifacts and/or freezing issues or very coarse pixelation; L: very bright or very dark face; HPM: tilted heavily to the side or backwards, major movements, head out of frame for parts of the video or too close so that face is occluded; D: face small in relation to the frame; G: heavy reflections (eyes/eyebrows not visible for major parts of the video).
3. *Acceptable (3)*: Face mostly well and fully visible with an appropriate distance to the camera; PB: some artifacts/coarser pixelation; F: minor freezing issues; L: dark shadows or bright light on face; HPM: tilted slightly backwards or to the side, few major movements; G: few major reflections making the eyes/eyebrows less visible for short parts of the video.
4. *Good (4)*: Face well and fully visible with an appropriate distance to the camera; PB: minor artifacts; F: no issues; L: rather well lit face that could be evenly somewhat bright or dark with minor shadows; HPM: front view, minor movements; G: minor reflections.
5. *Very good (5)*: Face well and fully visible at all times with an appropriate distance to the camera; PB & F: no issues; L: well lit face without shadows; HPM: front view, minor movements; G: none.

3.3. Compliance

In addition to the audio and video quality ratings, we also assessed if participants were compliant with the given task instructions. This is of particular interest for remote monitoring systems because participants need not have any guidance from a human supervisor. We rated compliance as a binary label (non-compliant/compliant) taking the following factors into consideration:

1. *Completion rate* (Objective: Was the task completed as instructed?)
2. *Comprehension* (Subjective: Did the participant understand the test instructions?)
3. *Commitment/Effort* (Subjective: Did the participant perform the task to the best of their abilities?)

Commitment is the most subjective of these three criteria. Obvious fatigue or any other difficulties due to health condition were not rated as non-compliant. A recording is rated as *compliant* only if all three compliance factors are fulfilled.

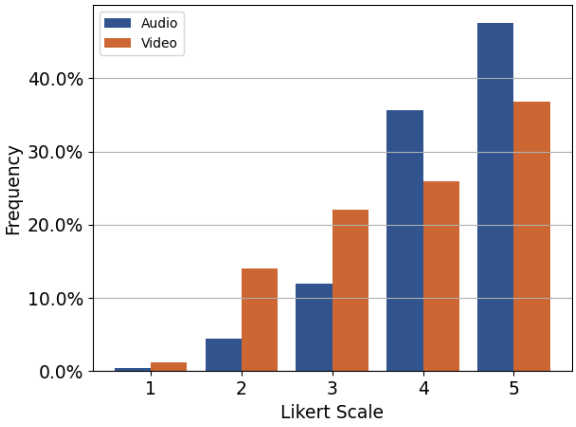

Figure 1: *Distribution of audio and video ratings.*

4. Results

This section presents results based on the judgements of the rater who rated all recordings for ease of visualization. The inter-rater agreement between the two raters, as measured by the quadratic weighted kappa on a subset of 125 samples, was moderate (0.53 for audio and 0.58 for video ratings). However, in looking at the binary categorization into *acceptable* and *unacceptable* samples, the observed percentage agreement was 96.8% for audio and 83.2% for video, suggesting that both raters mostly agreed on which videos were of acceptable quality.

4.1. Audio quality

The evaluation of audio ratings suggests that audio recordings are of high quality (95.2% *acceptable*), as shown in Figure 1. Ratings of *very good* are the most common, with only a small number of recordings that need to be rejected. Looking at the causes for *unacceptable* ratings in Figure 2, only few major problems were encountered overall, *signal noise* being the most common distortion. The second most frequent problem is environmental noise, such as barking dogs.

4.2. Video quality

84.8% of videos were rated within the *acceptable* range. Figure 1 shows the distribution of video ratings. As shown in Figure 2, most *unacceptable* ratings were related to problems with glasses, followed by pixelation and blocking artefacts.

4.3. Compliance

In 98.0% of the investigated recordings, users were compliant with the given instructions and performed the task appropriately. This could be attributed to the fact that Tina, the virtual agent provides clear instructions to participants on how to complete each speech exercise, along with an example demonstration of a successful attempt.
Considering all three aspects – audio and video quality and task compliance – results in a yield of 80.8% recordings that are both compliant and of high audiovisual quality.

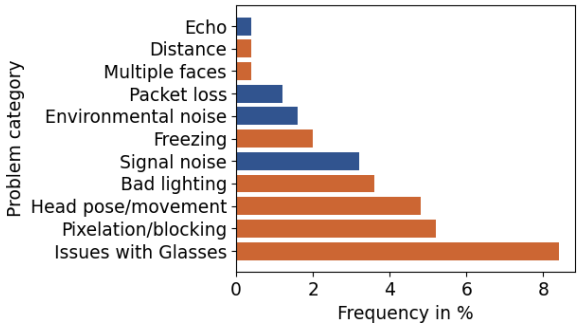

Figure 2: *Distribution of audio and video distortions in unacceptable recordings. Frequency is based on the entire dataset.*

4.4. Relation between severity level and recording quality

To investigate whether the perceptual quality ratings are affected by participants' speech impairment, we computed the Spearman correlation between audio and video ratings and the ALSFRS-R speech sub-score. For this, we sub-sampled the dataset to contain an equal number of samples for each ALSFRS-R speech score (16 samples for each score from 2 to 4; 48 samples total), to mitigate label bias. As can be seen in Figure 3, we found that neither audio ratings ($\rho = 0.23$) nor video ratings ($\rho = 0.10$) were significantly correlated with the ALSFRS-R speech score, suggesting that the system is appropriate for use with participants spanning a wide range of speech disorder severity. In addition, a Kruskal-Wallis test to examine whether rating medians differed significantly between patients and controls revealed no statistically significant effect for video or audio ratings between these groups. These findings are important prerequisites for equitable deployment and accurate analyses of participant speech and video across the spectrum of participants with ALS.

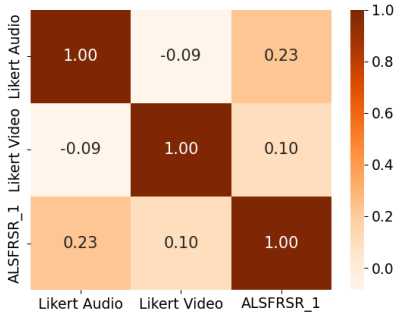

Figure 3: *Spearman correlation between the ALSFRS-R speech scores and the audio as well as video quality ratings for patients.*

4.5. Device-specific analysis

To determine whether the quality of the recordings was affected by the devices used to interact with the system, we compared the median ratings for personal computer versus mobile devices (e.g., smartphones and tablets). Mobile device recordings could be more susceptible to participant-related problems as participants might move the device while recording, accidentally block the microphone, or angle the camera sub-optimally. In the analyzed dataset, 38 out of 250 recordings were captured on mobile devices.

We found that the median audio and video ratings were not significantly different for the different device categories (4.4/3.7 for audio/video on mobile devices and 4.2/3.9 on PCs, compared to overall average ratings of 4.3/3.8; as measured by p-values of greater than 0.05 on Kruskal-Wallis tests). This indicates that remote home assessments can be accurately conducted independent of the users' device. We observed that 18 recordings out of 250 involved the use of headset instead of built-in microphones. We did not find any statistical differences in median quality ratings between these groups (4.1 with headsets, 4.3 with built-in microphones). These results have to be taken with a grain of salt because of the small sample size.

5. Discussion

This paper proposed a rubric to rate audiovisual quality and task compliance of data recorded by a multimodal dialog based RPM solution for ALS. We observed excellent compliance (98%) and high overall audio (95.2%) and visual quality rates (84.8%), which suggests that these data are suitable for automatically computing speech and facial metrics. Importantly, the severity of speech impairment in people with ALS did not affect audio and video quality or compliance which is important for equity, equitability and feasibility of such an RPM solution for ALS. These findings support the feasibility and analytical validity [11, 12] of using such multimodal dialog based solutions for remote speech assessments as our findings suggest that a virtual agent is able to properly explain tasks and engage the user in a sequence of speech exercises without the help of a clinician. In addition, our study indicates that remote speech assessments do not require specialized equipment, as mobile devices and on-device microphones provide adequate recording quality.

Going forward, our work also provides data-driven guidance towards improving our multimodal dialog based RPM solution even further, in terms of: (1) improving the pre-assessment of each participant's hardware and software based on automated or manual checks, and implementing interventions for the different problem categories identified (such as instructions to remove glasses, adjust face position, or move to a quieter environment), (2) rewording ambiguous standardized operating procedures, participant instructions, and test items, to improve the quality and compliance of the recorded signal even further, (3) developing and improving algorithms to monitor signal and output quality, and (4) clever adaptive design of virtual agent responses to improve interaction quality and maximize task compliance and completion. In this manner, the majority of detected problems are potentially addressable by instructing users accordingly and by implementing algorithms that are able to raise warnings if such avoidable and quality-affecting issues are detected.

6. Acknowledgements

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.

7. References

[1] V. Ramanarayanan, A. C. Lammert, H. P. Rowe, T. F. Quatieri, and J. R. Green, "Speech as a biomarker: Opportunities, interpretability, and challenges," *Perspectives of the ASHA Special Interest Groups*, vol. 7, no. 1, pp. 276–283, 2022.

[2] V. Ramanarayananan, M. Neumann, A. Anvar, D. Suendermann-Oeft, O. Roesler, J. Liscombe, H. Kothare, J. Berry, E. Fraenkel, R. Norel *et al.*, "Lessons learned from a large-scale audio-visual remote data collection for amyotrophic lateral sclerosis research (p1-13.002)," in *Proceedings of the Annual Meeting of the American Association of Neurology (AAN) 2022, Seattle, WA*. AAN Enterprises, 2022.

[3] A. Bombaci, G. Abbadessa, F. Trojsi, L. Leocani, S. Bonavita, and L. Lavorgna, "Telemedicine for management of patients with amyotrophic lateral sclerosis through covid-19 tail," *Neurological Sciences*, vol. 42, no. 1, pp. 9–13, 2021.

[4] G. M. Stegmann, S. Hahn, J. Liss, J. Shefner, S. Rutkove, K. Shelton, C. J. Duncan, and V. Berisha, "Early detection and tracking of bulbar changes in ALS via frequent and remote speech analysis," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–5, 2020.

[5] K. P. Connaghan, J. R. Green, S. Paganoni, J. Chan, H. Weber, E. Collins, B. Richburg, M. Eshghi, J.-P. Onnela, and J. D. Berry, "Use of beiwe smartphone app to identify and track speech decline in amyotrophic lateral sclerosis (als)." in *INTERSPEECH*, 2019, pp. 4504–4508.

[6] A. Exner, V. Ramanarayanan, D. Pautler, S. Snyder, H. Kothare, J. Liscombe, S. Sridhar, O. Roesler, W. Burke, M. Neumann, D. Suendermann-Oeft, and J. Huber, "Collecting remote voice and movement data from people with parkinson's disease (pd) using multimodal conversational ai: Lessons learned from a national study." in *Proceedings of the 2022 Motor Speech Conference, Charleston, SC*, 2022.

[7] M. Milling, F. B. Pokorny, K. D. Bartl-Pokorny, and B. W. Schuller, "Is speech the new blood? recent progress in ai-based disease detection from audio in a nutshell," *Frontiers in Digital Health*, vol. 4, 2022.

[8] M. Neumann, O. Roesler, J. Liscombe, H. Kothare, D. Suendermann-Oeft, D. Pautler, I. Navar, A. Anvar, J. Kumm, R. Norel, E. Fraenkel, A. V. Sherman, J. D. Berry, G. L. Pattee, J. Wang, J. R. Green, and V. Ramanarayanan, "Investigating the utility of multimodal conversational technology and audiovisual analytic measures for the assessment and monitoring of amyotrophic lateral sclerosis at scale," in *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 2021*. ISCA, 2021, pp. 4783–4787. [Online]. Available: https://doi.org/10.21437/Interspeech.2021-1801

[9] J. M. Cedarbaum, N. Stambler, E. Malta, C. Fuller, D. Hilt, B. Thurmond, and A. Nakanishi, "The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function," *Journal of the Neurological Sciences*, vol. 169, no. 1-2, pp. 13–21, 1999.

[10] R. Likert, "A technique for the measurement of attitudes," *Archives of Psychology*, vol. 140 22, pp. 5–55, 1932.

[11] J. Goldsack, A. Coravos, J. Bakker, B. Bent, A. Dowling, C. Fitzer-Attas, A. Godfrey, J. Godino, N. Gujar, E. Izmailova, C. Manta, B. Peterson, B. Vandendriessche, W. Wood, K. Wang, and J. Dunn, "Verification, analytical validation, and clinical validation (v3): The foundation of determining fit-for-purpose for biometric monitoring technologies (biomets)," 2020.

[12] J. Robin, J. E. Harrison, L. D. Kaufman, F. Rudzicz, W. Simpson, and M. Yancheva, "Evaluation of speech-based digital biomarkers: Review and recommendations," *Digital Biomarkers*, vol. 4, pp. 99–108, 2020.