On Timing and Pronunciation Metrics for Intelligibility Assessment in Pathological ALS Speech

Jackson Liscombe[1], Michael Neumann[1], Hardik Kothare[1], Oliver Roesler[1], David Suendermann-Oeft[1], Vikram Ramanarayanan[1,2]
[1]Modality.AI,Inc., [2]University of California, San Francisco

Introduction

We investigate three speech metrics -- goodness of pronunciation (GoP), percent pause time (PPT), and a new measure of canonical timing alignment (CTA) -- with respect to how well they characterize the temporal and spectro-acoustic aspects of pathological speech intelligibility in people with amyotrophic lateral sclerosis (pALS).

Methods and Materials

- We selected 2174 read speech (SIT) utterances of varying lengths spoken by 40 distinct participants (comprising both pALS and healthy controls) from a large corpus of speech and video data collected remotely via an interactive dialog agent.
- The data was divided into four cohorts by diagnosis, responses to the ALSFRS-R surveys, and age. BULBAR were those patients with a diagnosis who scored < 12 on the the 12 on the speaking, swallowing, or salivating questions; otherwise they were considered to by pre-symptomatic for speech dysarthria (PRESYMP).
    - BULBAR: 11 users, 109 sessions, 568 utterances
    - PRESYMP: 13 users, 153 sessions, 845 utterances
    - CONTROL45+: 10 users, 78 sessions, 454 utterances
    - CONTROL45-: 10 users, 53 sessions, 311 utterances
- The following metrics were computed:
    - **Percent Pause Time (PPT)**
      Used Praat to estimate speaking and articulation time then computed (speaking time - articulation time) / speaking time * 100
    - **Goodness of Pronunciation (GoP)**
      Modified Kaldi gop_speechocean762 package to obtain phonene GOP score. An utterance-level GOP score was arrived at by calculating the mean GOP value over all phonemes in the utterance. A GOP score of 0 indicates perfect pronunciation according to the acoustic model. Negative values indicate pronunciations that differ from the pronunciations in the model: the more negative, the more different.
    - **Canonical Timing Agreement (CTA)**
      Used Montreal Forced Aligner (based on Kaldi) to get predicted word-level timing information based on the expected text of the prompt, as read first by automated agent. A number between 0% (no alignment) and 100% (perfect alignment), as measured by the normalized inverse Levenstein edit distance between the word and silence boundaries.
- Human annotated **listener effort** was used to evaluate the informativeness of these metrics. Three annotators listened to a subset of utterances and were asked, "How effortful was it for you to understand?" on a scale of 0% (no listener effort at all in parsing the speech) to 100% (total listener effort; could not understand speech). (Stipancic et al., 2021). Listener effort was calculated as the mean of the two closest ratings. One novel SIT utterance from each participant was annotated and this listener effort score was applied to all utterances for that user.

Results

	GOP	CTA
CTA	0.4722	-
PPT	-0.2923	-0.4395

Table 1. Intra-metric Correlations. Pearson's r statistic. All correlations significant at p ≤ 0.00001. Low correlations suggest that the metrics are not redundant. CTA correlates most strongly with both GOP and PPT.

	BULBAR	PRESYM	CONTROL✳
GoP	-0.31	-0.17✦	-0.15✦
CTA	66.71%	77.31%	80.72%
PPT	4.15%	2.72%	1.14%

Table 2. Average metric value per cohort.
✳The CONTROL45+ and CONTROL45- cohorts showed no difference and were conflated into CONTROL. ✦Mann Whitney pairwise tests indicated means were significantly different at p ≤ 0.00001 for all combinations except these ones where p = 0.01.

feature set	F-measure	relative improvement
GoP+CTA+PPT	0.496	49.90%
GoP+CTA	0.477	43.67%
CTA+PPT	0.455	37.05%
GoP+PPT	0.431	29.82%
CTA	0.417	25.60%
GoP	0.412	24.10%
PPT	0.332	NA

Table 3. Cohort Classification and Metric Informativeness. Results of 3-way cohort classification (BULBAR v PRESYM v CONTROL) with 10-fold cross validation using a Random Forest classifier. Given three classes, PPT performance is no better than randomly guessing a cohort, which is 0.340 +/- 0.010.

feature set	correlation	mean absolute error	relative improvement
GoP+CTA+PPT	0.800	15.15	81.66%
GoP+CTA	0.792	15.46	79.75%
GoP+PPT	0.726	16.71	64.80%
CTA+PPT	0.697	17.95	58.17%
GoP	0.681	17.97	54.47%
CTA	0.679	18.50	54.02%
PPT	0.441	23.03	NA

Table 4. Mean absolute error and relative improvement in correlations between different feature combinations and human-scored listener effort based on linear regression of listener effort using different feature set combinations.

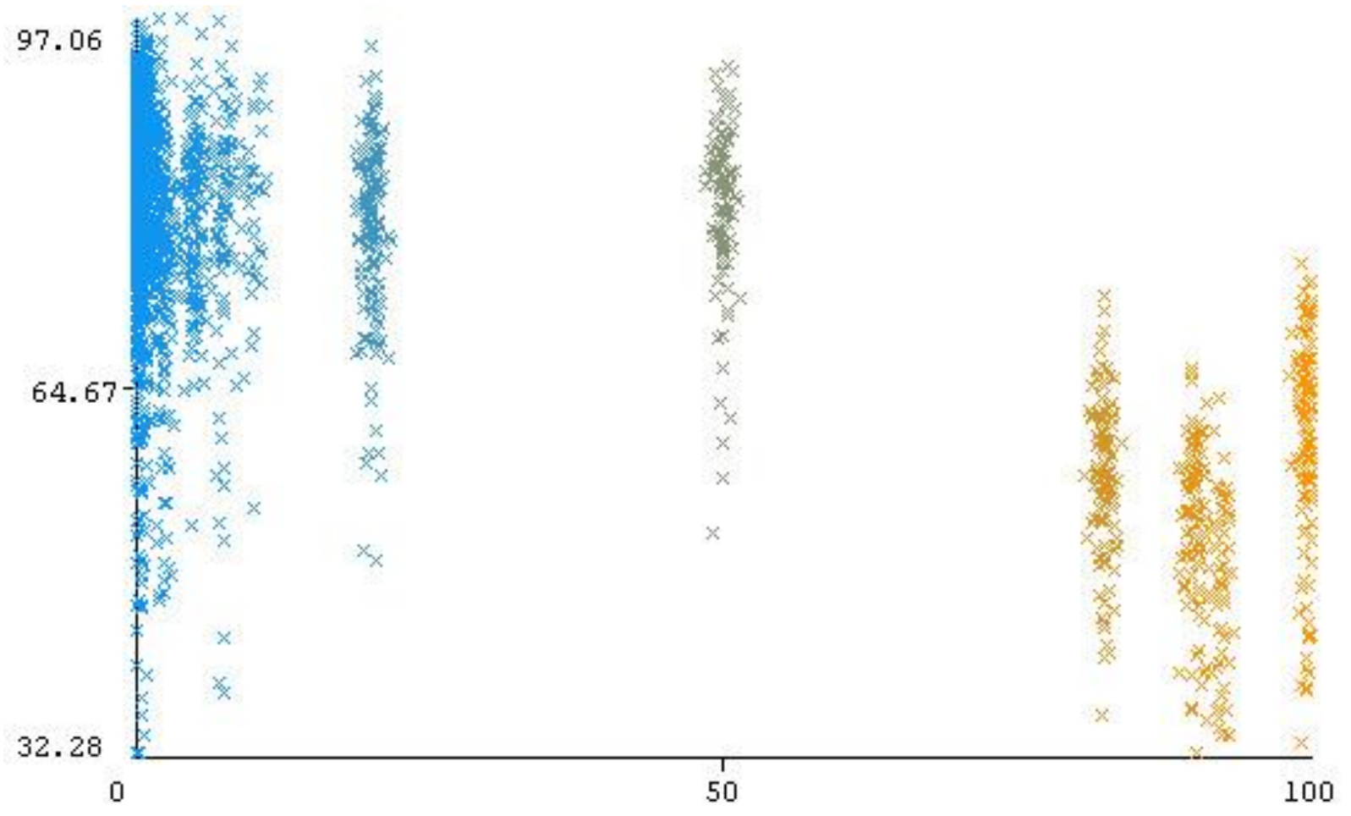

Figure 1. CTA v Listener Effort.

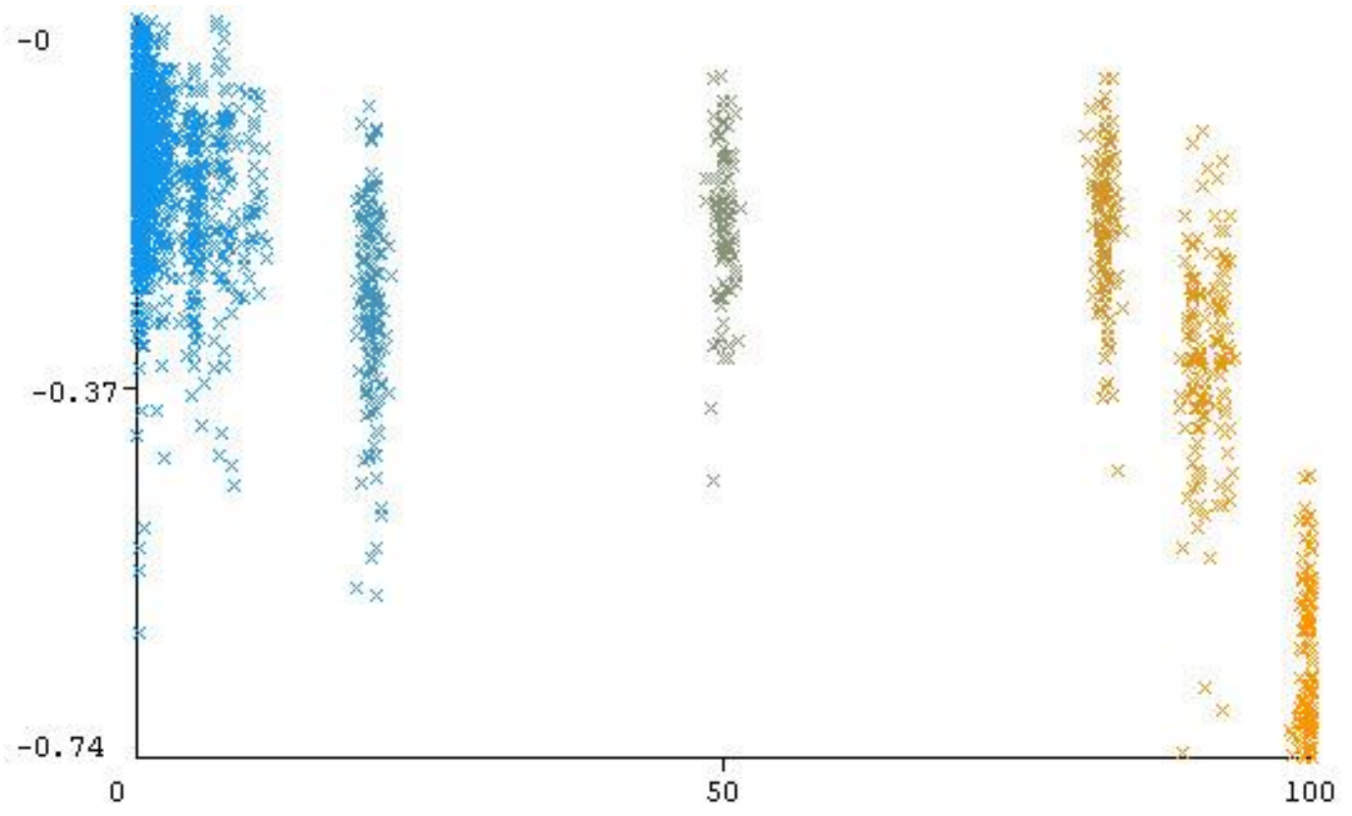

Figure 2. GoP v Listener Effort.

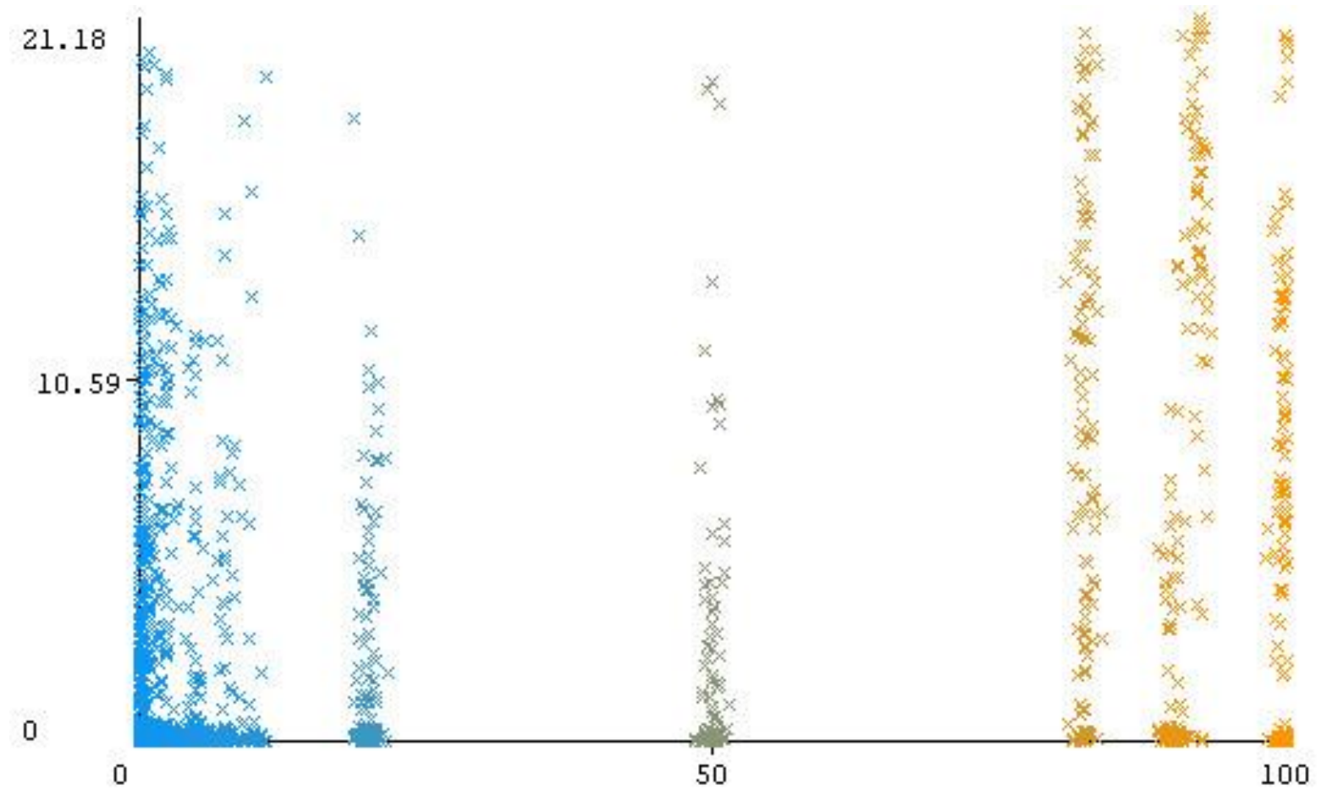

Figure 3. PPT v Listener Effort.

Cohort	mean	sdev
CONTROL	3.92%	10.87
PRESYMP	10.58%	24.67
BULBAR	41.36%	41.41

Table 5. Average listener effort scores per cohort along with the standard deviation. Mean scores indicate difference per cohort.

The figures above show little relationship between PPT and listener effort, though there does seem to be a relationship with respect to listener effort and GoP and CTA. With respect to GoP, though, this relationship only seems to hold and the very extreme end of the scale when the speech is unintelligible (listener effort = 100%). CTA seems a bit more robust in that listener effort scores of 75% and above correspond to lower CTA scores.

Conclusions

- Highest correlation to listener effort was achieved when combining all three metrics.
- CTA is as, if not more, informative than GoP and PPT in distinguishing controls from bulbar pre-symptomatic and bulbar symptomatic ALS patients in our cohort.
- Both CTA and GoP displayed moderate to high correlations with human listener effort and low correlation with each other. This potentially highlights the relative importance of timing over spectral information in characterizing ALS pathological speech.

References

- K. L. Stipancic, K. M. Palmer, H. P. Rowe, Y. Yunusova, J. D. Berry, and J. R. Green, "'You say severe, I say mild': Toward an empirical classification of dysarthria severity," *Journal of Speech, Language, and Hearing Research*, vol. 64, no. 12, pp. 4718–4735, 2021.
- W. Hu, Y. Qian, F. K. Soong, and Y. Wang, "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers," *Speech Commun.*, vol. 67, pp. 154–166, 2015.
- M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: trainable text-speech alignment using Kaldi," in *Proceedings of the 18th Conference of the International Speech Communication Assoc.*, 2017.