Supplementary Material for "Automatic Assessment of Speech Intelligibility for Individuals with Aphasia"

Duc Le, Keli Licata, Carol Persad, and Emily Mower Provost

I. Annotator Instructions

The instructions for the scoring process were given orally to individual annotators and followed a predefined format without a fixed script. We informally described each scoring category and stepped through all possible answer choices, along with their associated prototypical examples.

Clarity is defined as the degree to which a sentence can be clearly understood. It is intended to capture the overall pronunciation quality of a sentence. The elicitation question for this category is: *How clear is the pronunciation?* The possible answer choices, from low to high quality, are: *Very Unclear*, *Mostly Unclear*, *Mostly Clear*, and *Very Clear*.

Fluidity is defined as the degree to which a sentence can be uttered at an appropriate speed and without pauses or hesitation. The elicitation question for this category is: *How fluid is the speech?* The possible answer choices, from low to high quality, are: *Very Broken*, *Mostly Broken*, *Mostly Fluid*, and *Very Fluid*.

Prosody is arguably the most difficult category to define. In this work, we define it broadly as the correctness of intonation. Utterances that are overly monotonous or have widely varying pitch are both considered incorrect. We found that this definition of *Prosody* resulted in higher human agreement compared to directly quantifying the degree of monotonicity. The elicitation question for this category is: *Is the intonation correct?* The possible answer choices, from low to high quality, are: *Very Incorrect*, *Mostly Incorrect*, *Mostly Correct*, and *Very Correct*.

Note that all scoring categories have one additional answer choice, *Not Enough Data*, which is reserved for utterances that the annotators deem to have insufficient data for analysis.

II. GeMAPS Baseline

The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) is a collection of acoustic features commonly used for affect recognition [1]. GeMAPS may contain useful features for our classification tasks since they share certain similarities with speech emotion recognition.

TABLE I: Classification UAR (%) using GeMAPS features.

	Clarity	**Fluidity**	**Prosody**
3-class	32.8 (SVM)	50.1 (LR)	44.6 (NB)
2-class	58.0 (LR)	64.2 (LR)	55.8 (LR)

SVM: Support Vector Machine | NB: Naïve Bayes | LR: Logistic Regression

TABLE II: A reference alignment extracted for the target sentence "The people clapped." The search must descend into the syllable and phone level for the OOV word "people."

Target	**Level**	**Reference**	**Context**	**Instances**
the	WORD	the	L	6,436
people	SYL.	p iy	L	65
-	PHONE	p	L + R	12
-	PHONE	ah	L + R	22
-	PHONE	l	L + R	139
clapped	WORD	clapped	R	20

Table I shows the UARs achieved on this feature set using the same classification pipeline presented in the paper. All results are statistically significantly worse than those achieved using the proposed features (paired t-test, $p < 0.001$). Further, we found that adding GeMAPS features to our existing feature set does not improve performance; they actually worsened the results in certain cases. As discussed in the paper, this may be caused by the non-ideal recording conditions of the dataset. Further, our proposed high-level acoustic features may already capture relevant information encoded by GeMAPS features, thus making them redundant. These observations help emphasize the importance of feature engineering in this work.

III. Reference Alignment

Reference alignment is an algorithm used for matching a PWA's target sentence with a reference database of healthy speech [2]. As a motivating example, suppose the PWA says "The people clapped." We would want to find the same sentence spoken by a healthy speaker and analyze how the two differ in terms of pronunciations, durations, or pitch contours. The challenge is that some target words may not be present in the reference database, possibly because of the PWA's speech-language impairment causing deviation from the prompt.

The proposed algorithm is able to find a reference alignment of any target utterance by breaking the search down at three different levels, word, syllable, and phone. For example, when encountering the out-of-vocabulary (OOV) word "people," the algorithm breaks it down to two syllables "p iy" and "p ah l", and resumes the search at the syllable level. "p iy" is present in the reference database and can be matched. However, "p ah l" is absent from the database. The algorithm then breaks this syllable down to three individual phones and resumes the search at the phone level. The algorithm is context-sensitive in the sense that it prefers matches with similar context as the target token. For instance, the reference phone "ah" with

TABLE III: Selected features sorted in decreasing information gain (IG) with respect to the *2-class* ground-truth labels.

		Selected Features	Mean IG
Oracle	*Clarity*	wordGOP (mean), contentWordGOP (mean), phoneGOP (median), contentWordGOP (min), PVE (M=1), PVE (M=4), phoneGOP (stdev), contentWordGOP (median), weightedIWR (3-best), IWR (2-best), syllablesPerMin, contentSyllablesPerMin, weightedIWR (1-best), contentWordGOP (stdev), wordGOP (max), clearSpeechRate, intensityDTW (mean), contentWordsProduced	0.18 ± 0.07
	Fluidity	totalDuration, PVE (M=4), PVE (M=3), PVE (M=2), wordsPerMin, syllablesPerMin, PVE (M=1), contentWordsPerMin, nonSpeechDuration, contentSyllablesPerMin, voicedDuration, intensityDTW (min), wordGOP (median), intensityDTW (mean), wordGOP (max), pitchDTW (min), intensityDTW (median)	0.27 ± 0.14
	Prosody	totalDuration, PVE (M=2), PVE (M=3), wordsPerMin, PVE (M=1), voicedDuration, syllablesPerMin, clearSpeechDuration, phoneGOP (median), wordGOP (min), intensityDTW (mean), intensityDTW (min), wordGOP (max), fillerDuration, contentWordsProduced, intensityDTW (median), pitchDTW (min), weightedIWR (1-best)	0.09 ± 0.06
Merged	*Clarity*	phoneGOP$_E$ (mean), contentWordGOP$_E$ (mean), phoneGOP$_E$ (median), phoneGOP$_S$ (median), IWR$_E$ (3-best), weightedIWR$_E$ (2-best), IWR$_E$ (1-best), phoneGOP$_E$ (stdev), wordGOP$_S$ (median), contentWordGOP$_E$ (min), wordGOP$_E$ (median), contentWordGOP$_S$ (median), totalDuration$_E$, PVE$_E$ (M=2), IWR$_S$ (2-best), contentWordsPerMin$_S$, PVE$_S$ (M=4), PVE$_S$ (M=2), contentWordGOP$_S$ (mean), wordGOP$_S$ (max), weightedIWR$_S$ (3-best), syllablesPerMin$_E$, wordsPerMin$_E$, wordGOP$_E$ (max), contentWordGOP$_S$ (max), phoneGOP$_E$ (min), intensityDTW$_S$ (mean), phoneGOP$_S$ (max)	0.16 ± 0.06
	Fluidity	totalDuration$_E$, totalDuration$_S$, wordsPerMin$_S$, syllablesPerMin$_S$, contentSyllablesPerMin$_S$, wordsPerMin$_E$, syllablesPerMin$_E$, PVE$_E$ (M=4), PVE$_S$ (M=4), PVE$_S$ (M=1), PVE$_E$ (M=2), contentWordsPerMin$_E$, clearSpeechDuration$_E$, phoneGOP$_E$ (mean), phoneGOP$_E$ (median), wordGOP$_S$ (max), wordGOP$_E$ (max), pitchDTW$_S$ (min), intensityDTW$_S$ (min), intensityDTW$_S$ (mean), pitchDTW$_S$ (max)	0.22 ± 0.11
	Prosody	totalDuration$_E$, totalDuration$_S$, syllablesPerMin$_S$, clearSpeechDuration$_E$, wordsPerMin$_E$, PVE$_S$ (M=1), PVE$_S$ (M=4), voicedDuration$_S$, wordGOP$_S$ (mean), phoneGOP$_S$ (median), IWR$_E$ (2-best), wordGOP$_S$ (max), pitchDTW$_S$ (min), intensityDTW$_S$ (mean), pitchDTW$_S$ (median)	0.09 ± 0.04

GOP: Goodness of Pronunciation | IWR: Isolated Word Recognition Accuracy | PVE: Pairwise Variability Error | DTW: Dynamic Time Warping Distance
$_S$: features extracted with *Simple* forced alignment | $_E$: features extracted with *Extended* forced alignment

left context "p" and right context "l" will be given more weight than the same phone with different context. This is based on the hypothesis that the characteristics of an acoustic unit (word, syllable, or phone) are influenced by its immediate neighbors. Table II shows a sample alignment of this sentence.

IV. Selected Features

Table III lists the mRMR feature selection result with respect to the 2-class ground-truth labels using features extracted with *Oracle* and *Merged* transcripts. We also include the mean and standard deviation of Information Gain (IG) for each feature set. The overall IG statistics mirror the task difficulty. *Fluidity* is the easiest to classify and corresponds to the feature set with the highest IG, followed by *Clarity* and *Fluidity*, respectively. We also observe a decrease in average IG when moving from *Oracle* to *Merged* transcripts, indicating the impact of automatic transcription.

References

[1] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. Andr, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Transactions on Affective Computing*, pp. 14–27, 2016.

[2] D. Le and E. M. Provost, "Modeling Pronunciation, Rhythm, and Intonation for Automatic Assessment of Speech Quality in Aphasia Rehabilitation," in *Proc. of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Singapore, 2014.