Digital
Biomarkers

Digit Biomark 2020;4:109–122

DOI: 10.1159/000511671
Received: July 22, 2020
Accepted: September 16, 2020
Published online: December 2, 2020

Validation – Research Article

Repeatability of Commonly Used Speech and Language Features for Clinical Applications

Gabriela M. Stegmann[a, b] Shira Hahn[a, b] Julie Liss[a, b] Jeremy Shefner[c] Seward B. Rutkove[d] Kan Kawabata[a] Samarth Bhandari[a] Kerisa Shelton[c] Cayla Jessica Duncan[c] Visar Berisha[a, b]

[a]Arizona State University, Phoenix, AZ, USA; [b]Aural Analytics, Scottsdale, AZ, USA; [c]Barrow Neurological Institute, Phoenix, AZ, USA; [d]Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA

Keywords

Automatic speech analysis · Repeatability · Digital biomarkers · Mobile technology · Speech

Abstract

Introduction: Changes in speech have the potential to provide important information on the diagnosis and progression of various neurological diseases. Many researchers have relied on open-source speech features to develop algorithms for measuring speech changes in clinical populations as they are convenient and easy to use. However, the repeatability of open-source features in the context of neurological diseases has not been studied. ***Methods:*** We used a longitudinal sample of healthy controls, individuals with amyotrophic lateral sclerosis, and individuals with suspected frontotemporal dementia, and we evaluated the repeatability of acoustic and language features separately on these 3 data sets. ***Results:*** Repeatability was evaluated using intraclass correlation (ICC) and the within-subjects coefficient of variation (WSCV). In 3 sets of tasks, the median ICC were between 0.02 and 0.55, and the median WSCV were between 29 and 79%. ***Conclusion:*** Our results demonstrate that the repeatability of speech features extracted using open-source tool kits is low. Researchers should exercise caution when developing digital health models with open-source speech features. We provide a detailed summary of feature-by-feature repeatability results (ICC, WSCV, SE of measurement, limits of agreement for WSCV, and minimal detectable change) in the online supplementary material so that researchers may incorporate repeatability information into the models they develop.

Gabriela M. Stegmann
Aural Analytics
1355 N Scottsdale Rd. Unit 110
Scottsdale, AZ 85257 (USA)
gabriela.stegmann@auralanalytics.com

Karger

Introduction

Speech is known to change in a host of neurological conditions. There has been recent interest in using machine learning models applied to speech recordings to assess changes in clinical conditions automatically [1, 2]. For example, it has recently been used to identify individuals with major depressive disorder [3], amyotrophic lateral sclerosis (ALS) [4], and Alzheimer disease [5]; predict speech severity in participants with Parkinson's disease [6]; and detect changes in affective states in individuals with bipolar disorder [7]. Speech is a high-dimensional acoustic signal, which is sampled at tens of thousands of times per second for acoustic analysis. To make problem-solving tractable under these conditions, open-source feature extraction software packages have been developed to extract a smaller number of low-level features that serve to reduce the dimensionality of the signal. Wide availability of open-source tool kits for acoustic and language feature extraction has democratized the development of algorithms that aim to detect and track diseases.

However, despite the popularity of using speech as a biomarker, the features commonly used in models have not undergone the rigorous validation work which is typically done in the medical field when evaluating new assessment tools. To develop reliable clinical models using speech, it is important to distinguish natural variation in speech production from disease-related change in the measures of interest. An individual undergoing repeated measurements is likely to show natural variation such that the scores vary across different days (and even within the same day). When the natural variation is large, it becomes difficult to detect when there is an important disease-related change. For example, in Parkinson disease, weakness of muscles used in speaking causes a decrease in loudness; however, many factors unrelated to Parkinson disease can also impact the day-to-day speaking loudness. The more loudness variation is exhibited by healthy individuals, the harder it becomes to detect a true loudness decline due to Parkinson disease progression. Even in cross-sectional studies, high-variance features combined with small sample sizes increase the risk of overfitting to a dataset. Therefore, it is important to understand the typical natural variation of speech features if speech is to be used as a biomarker.

In this paper, we define *repeatability* as the average variation over a short period of time in which disease-related decline is unlikely to be evident or manifested through the features. Although publicly available speech features have been used in many clinical research studies, no study has systematically evaluated their repeatability. Therefore, in this study we evaluated the repeatability of a commonly used set of acoustic and language features extracted through open-source tool kits. As the results show, we found that the majority of the features had repeatability below typical acceptable limits for clinical practice regardless of the measure of repeatability and regardless of the population used.

Methods

We evaluated 2 classes of repeatability measures, i.e., those that considered the variability in measurements within a single person, and within-person variability relative to the variability in the full sample. Furthermore, we evaluated repeatability using a sample of healthy individuals and 2 samples of individuals with ALS. The healthy sample allowed for evaluating repeatability in a sample with no speech impairment. ALS, on the other hand, is a neurodegenerative disease which impacts speech, resulting in a highly heterogeneous sample, allowing evaluation of repeatability in participants with a wide dynamic range of values. The 2 complementary repeatability measures and the 2 contrasting samples allowed a thorough evaluation of the repeatability of the speech features.

Repeatability Measures

We measured repeatability using 2 unit-independent measures [8], i.e., the within-subjects coefficient of variation (WSCV) and intra-class correlation (ICC). These 2 measures offer different information about a feature's repeatability, which we describe briefly in the following paragraphs.

The WSCV indicates the expected variation in an individual's score expressed as a percentage from the mean [8]. It is the ratio of the within-person SD to the mean, and it assumes that the variability within individuals is proportional to the mean. It provides an estimate of how close adjacent measurements are expected to be and, because it is a percentage, it can be compared across features directly. Low WSCV indicate a low variability (i.e., high repeatability in the feature).

The ICC [8] is a measure of within-person variability relative to the variability of the full sample; rather than indicating how much an individual is expected to vary (i.e., WSCV), it additionally indicates whether individuals generally maintain the same rank in the sample across repeated measures. This measure of repeatability is useful beyond within-person variability alone because it provides context for whether the amount of variability is small or large relative to the possible range of the data. For example, consider a feature where individuals are expected to change by 10 units on average across occasions; this change may be small or large depending on how different participants are from each other and the dynamic range of the feature. High ICC values indicate that the measure is repeatable. However, it needs to be interpreted within the context of the sample. A measurement with little within-person variability may result in low ICC if the sample is highly homogeneous. Therefore, the ALS sample provided an excellent test case, as it was heterogeneous with a wide dynamic range on the speech features due to the participants' speech impairment.

Although we do not discuss the following in the main body of the paper, we additionally computed: (1) the standard error of measurement (SEM), which is the expected variation in a given individual's score expressed in the measure's scale; (2) the minimal detectable change (MDC), which is the smallest change in scores that indicates a change in the individual's true or average score and is calculated using the SEM; and (3) the limits of agreement for the WSCV (LOACV), which is conceptually similar to the MDC but expressed as a percentage and therefore useful when the within-subject variability is proportional to the subject's mean. The SEM is similar to the WSCV in that both provide an indication of how much each individual is expected to change across different measurements; however, the SEM is useful when the participant's variability is independent from the participant's mean score while the WSCV is useful when the participant's variability is proportional to the participant's mean score. The SEM, the MDC, and the LOACV are not discussed in the paper since they depend on the scale of the features (SEM and MDC require understanding the scale and units of the measures) or are redundant (LOACV is calculated based on the WSCV). However, we provide these values for each feature in our online supplementary material (for all online suppl. material, see www.karger.com/doi/10.1159/000511671).

In Data Analysis, we provide more detail about the equations used for the measures of repeatability. The results focus on summarizing the findings in the unit-independent measures (ICC and WSCV), which are the most appropriate for discussion as they do not require the reader to know the scale of each feature in order to interpret them.

Samples

In order to evaluate repeatability for a range of values that might be expected in both healthy individuals and individuals with impaired speech, we collected speech from 3 separate samples.

Samples 1 and 2

Sample 1 consisted of participants with ALS and sample 2 consisted of healthy individuals. These 2 samples were obtained from "ALS at Home," an observational, longitudinal study [9]. Healthy participants had no history of generalized neurological conditions. ALS is a neurodegenerative disease characterized by an eventual loss of muscle function, including those used to breathe and speak. It causes impairment of speech motor control and weakness of muscles required for vocalization. The resulting speech disturbance is referred to as dysarthria, which is evident perceptually and acoustically.

Participants provided speech samples on a daily basis by reading a set of sentences and holding out an "ahh" sound (sustained phonation). The sentences and phonations were used for measuring acoustic features. In an attempt to reduce the impact of nuisance variables on the acoustic features, the participants were requested to provide recordings under the same conditions and using the same devices each time. Acoustic features were extracted from each task from all participants using openSMILE [10] and Praat [11], 2 open-source tool kits.

To avoid capturing variability due to disease progression, we limited the sample to each participant's first 7 days of data. We did this for both the ALS and the healthy participants. A mobile application was used for data collection. It led participants through a series of speech tasks, including sentence reading and sustained phonation.

Sample 3

The third sample was obtained from a separate observational, longitudinal study that is currently in progress. It consisted of ALS participants with cognitive symptoms secondary to suspected frontotemporal dementia. Therefore, these participants changes experienced not only in speech motor control but also in cognition. Previous studies have utilized language features from connected speech for assessing cognition, such as identifying participants with Alzheimer disease [5]. Therefore, the participants completed the same speech tasks as the participants in "ALS at Home" (the same set of sentences and phonations), with an additional task for eliciting connected speech. In the connected speech task, participants were asked to describe in detail a displayed picture. All speech samples for this task were manually transcribed. In each session, this picture was rotated among a set of 8 to mitigate familiarization effects. Pictures were normed to elicit responses of a similar length. Language features were extracted from the transcripts using Talk2me, an open-source tool kit [5].

Unlike "ALS at Home" samples 1 and 2, the speech samples in this study were provided on a weekly basis; we used the participants' first 2 sessions, which were typically 1 week apart, unless a participant missed a session. If the first 2 sessions were more than 13 days apart, the participant was excluded.

This sample allowed us to replicate the repeatability results from "ALS at Home" and to use the connected speech task to evaluate the repeatability of open-source language measures extracted from the picture descriptions in a sample exhibiting a range of cognitive impairment. Thus, acoustic features were extracted from samples 1, 2, and 3, and language features were extracted from sample 3.

Feature Extraction

A large number of speech and language analysis tools are available for analyzing speech. For example, Amazon's Comprehend, Google's Cloud Natural Language API, and IBM's Natural Language Understanding provide algorithm designers with NLP tools for analyzing text data. Similarly, the Python libraries librosa and scipy provide practitioners with signal processing routines for analyzing speech acoustics. These tools do not calculate speech features directly; rather, they provide code that can be used to engineer features. In addition to these, there are also several tools for extracting speech and language features directly from audio or tran-

Table 1. Description of features

Task	Category (features, *n*)	Description
Sentences and phonations	Energy (135)	These are features related to the energy generated in the acoustic signal, such as loudness and the harmonics-to-noise ratio
	Frequency (31)	These are features related to the frequencies in the acoustic signal, such as the frequencies of formants 1, 2, and 3
	MFCC (1,521)	These are features related to mel-frequency cepstral coefficients. MFCC are a parametric representation of the slow-changing part of the spectrum. They are often used to represent how the vocal tract shape manifests itself in the envelope of the spectrum
	Pitch (356)	These are features related to the pitch (F0), such as the mean pitch or the variability in pitch
	Spectral (4,493)	These are features that describe the energy distribution across different center frequencies in the spectrum. Examples include spectral slope and vowel formant energy
	Temporal (130)	These are features related to the rhythm of speech, such as the rate of loudness peaks and voiced/unvoiced regions
Connected speech	Lexical (92)	These are features that are related to word use, such as the number of filler words and the proportion of words that are unique
	Pragmatic (119)	These are features related to contextually appropriate language use, such as identifying topics within the text
	Semantic (45)	These are features that compute semantic similarity between words
	Syntactic (149)	These are grammatical features, such as part-of-speech counts and propositional density

scripts. These features have become increasingly popular in clinical applications as they allow digital health algorithm designers to build sophisticated speech-based models without requiring significant expertise in speech and language processing.

In this paper, we focused on the repeatability of open-source speech and language features and not on the tool kits that can be programmed to do speech and language analysis. As a result, for the evaluation we identified 3 open-source packages that have been used in clinical applications and that measure both acoustic and language aspects of speech. These include openSMILE [10], Talk2me [5], and Praat [11]. The output of these feature extraction programs were used without modifications. We did not derive new features and we used the default settings for extracting each set of features.

Acoustic features were extracted from the sentences and phonations from participants. Two tools used for this purpose were openSMILE [10] and Praat [11], 2 open-source feature extraction tool kits which extract acoustic features from audio files. We extracted features using the openSMILE emo_large configuration, which extracts over 6,000 acoustic features, the openSMILE eGeMAPS configuration, a minimalistic set which extracts 88 features [12], and Praat Voice Report, which extracts 26 features. *Language* features were extracted from text transcripts obtained from the connected speech in the picture description task using Talk2Me [5], an open-source library which extracts 405 language features from transcripts.

Given the large number of features, we grouped them into 6 acoustic categories [12], i.e., energy, frequency, MFCC, pitch, spectral, and temporal, and 4 language categories [5], i.e., lexical, pragmatic, semantic, and syntactic. Table 1 shows the number of features within each category and a brief description of each category. We included examples of what we considered to be easily understood or meaningful features within each category.

Note that there were over 6,000 features computed in this study. Table 1 provides a summary for each category of features. The description of each category provides an explanation of representative features for that category but it is not an exhaustive list of all of the features associated with that category. Readers interested in learning more can refer to openSMILE [10], eGeMAPS [12], Praat [11], and Talk2Me [5].

Data Analysis

In this article we provide results relating to the ICC and WSCV, and in the online supplementary material we include the SEM, the MDC, and the LOACV. Therefore, we describe here how each was calculated.

For computing the ICC, we used a two-way mixed effects model allowing random intercepts for the participants and the testing dates [13]. Therefore,

$$ICC = \sigma_s^2/(\sigma_s^2 + \sigma_T^2 + \sigma_e^2)$$

such that σ_s^2 was the between-subject variance, σ_T^2 was the testing date variance, and σ_e^2 was the error variance for each subject within each window. The SEM was obtained from σ_e^2 in the same mixed-effects model [8]. The MDC was:

$$MCD = 1.96 \times \sqrt{2} \times SEM$$

For obtaining the WSCV, a log transformation of the data was obtained. WSCV only make sense for measures such that all values are positive. Therefore, we excluded from this analysis any features that had values of 0 or below. A mixed-effects model was fit to the log-transformed data, and the residual variance $\sigma_{ln,e}^2$ was extracted, such that [8]:

$$WSCV = (e^{\sigma_{ln,e}} - 1) \times 100\%$$

For computing the Bland-Altman LOACV, we used the Rousson [14] adaptation using $\sigma_{ln,e}$ and transformed it back into percentages using the same transformation as the one for WSCV:

$$LOA_{WSCV} = (e^{\pm 1.96 \times \sqrt{2} \times \sigma_{ln,e}} - 1) \times 100\%$$

All analyses were done in R [15]. The R Package lme4 [16] was used for fitting the mixed-effects models.

Results

Samples 1 and 2

The data consisted of 72 ALS participants (sample 1: a total of 430 sessions, mean age = 59.8 years, SD = 10.4; 24 females) and 22 healthy controls (sample 2: a total of 132 sessions, mean age = 50.1, SD = 14.7; 16 females). The ALS participants were characterized by a wide range of dysarthria severity. The clinical standard for measuring motor speech impairment in ALS is the ALS Functional Rating Scale – Revised (ALSFRS-R), which ranges from 0 (total loss of speech) to 4 (normal speech). Participants in the sample had mean speech ALSFRS-R = 3.22 with scores between 0 and 4.

Sample 3

The data consisted of 24 ALS participants (a total of 49 sessions, mean age = 67.4 years, SD = 11.3; 7 females). Participants in the sample had a mean speech ALSFRS-R of 3.35, with scores between 2 and 4. The participants' cognitive function was measured with the Montreal Cognitive Assessment (MoCA), with scores that ranged from 0 (most impaired) to 30 (no impairment). The participants in the sample had a mean MoCA of 22.5, with scores between 9 and 27.

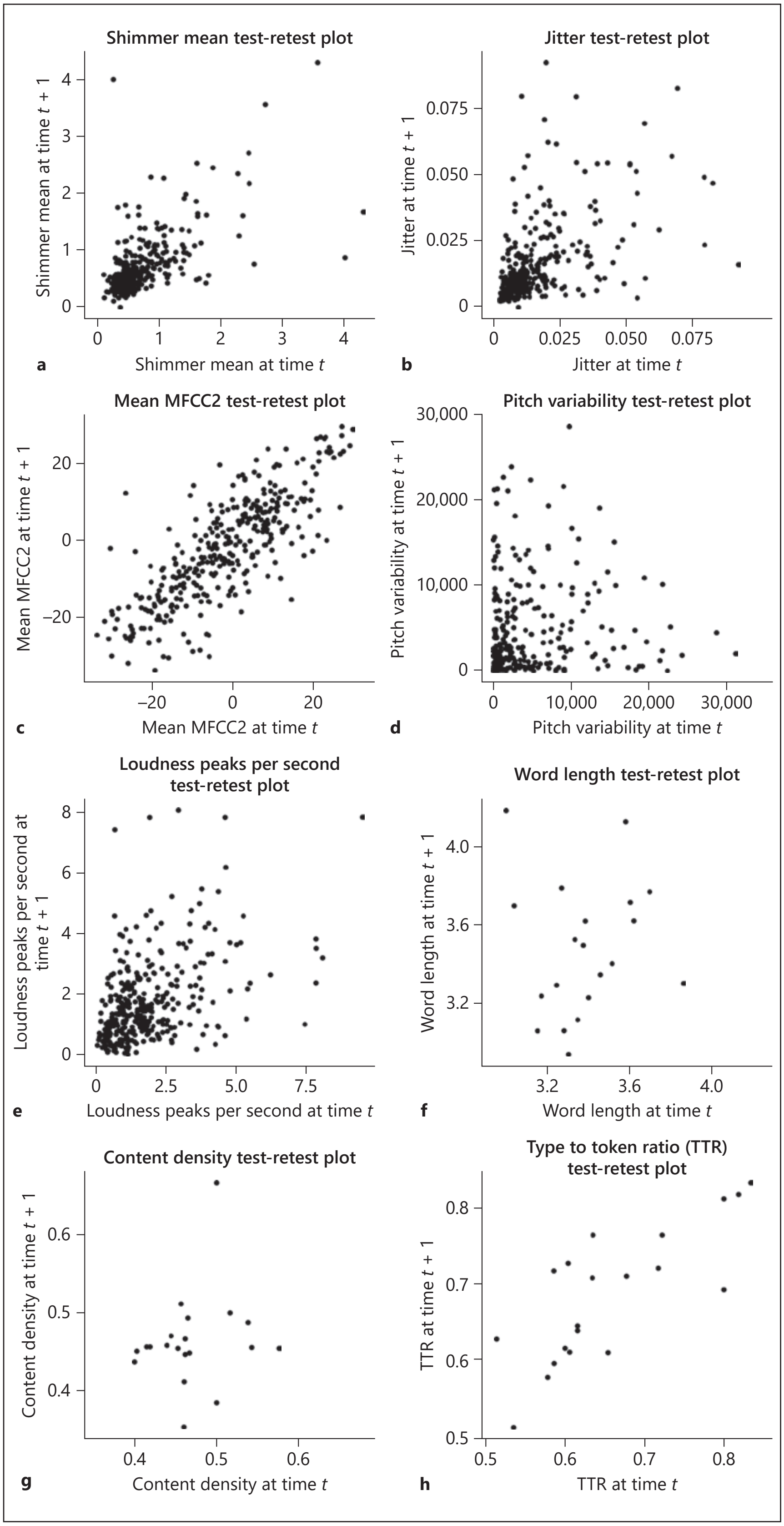

Fig. 1. Sample test-retest plots for a set of open-source features. Test-retest plots: shimmer mean (**a**), jitter (**b**), mean MFCC2 (**c**), pitch variability (**d**), loudness peaks per second (**e**), and word length (**f**). TTR, type-to-token ratio.

Table 2. Percentiles, medians, and means by task

Task	5th percentile	25th percentile	Median	Mean	75th percentile	95th percentile
a ICC for the features separated by task						
Connected speech	0.00	0.00	0.02	0.15	0.32	0.49
Phonation	0.00	0.14	0.31	0.33	0.52	0.71
Sentences	0.14	0.42	0.56	0.55	0.70	0.86
b WSCV for the features separated by task (%)						
Connected	6.61	23.37	52.45	93.49	61.15	365.29
Phonation	14.48	40.05	79.19	109.04	126.69	315.46
Sentences	5.11	12.32	28.87	39.99	56.35	130.35

Summary of Results

The 3 samples were analyzed separately, such that there was a set of repeatability measures for ALS speech (sample 1), healthy speech (sample 2), and ALS with cognitive decline speech (sample 3). Using openSMILE and Praat, acoustic features were extracted from sustained phonations and sentences separately from healthy controls and ALS participants. Repeatability measures were calculated for the phonations and sentences separately.

The median ICC for the phonations, sentences, and language features were: 0.31, 0.55, and 0.02, and the median WSCV were: 79, 29, and 52%, respectively.

To better visualize the spread of the data, we chose a sample of features and plotted the test-retest plots in Figure 1a–h using the phonations from the ALS sample (sample 1) for the acoustic features and the language features from the ALS with cognitive decline sample (sample 3). The disease groups (as opposed to the healthy sample) were chosen for visualizing the plots because they exhibited a wider dynamic range in the data. All observations from all of the participants are included such that the *x*-axis shows the participants' observations on a given assessment date and the *y*-axis shows the participants' observations in the following assessment date. Given that there was a large number of features, we selected features that have been used in papers related to clinical speech analytics and features that readers would be familiar with. For example, jitter and shimmer aim to measure instability in the voice, which has been found to become pronounced in ALS and Parkinson disease patients [17–19]. The second MFCC coefficient has been used to classify between patients with Parkinson disease and healthy controls [20]. Content density is the number of content words (nouns, verbs, adjectives, and adverbs) normalized by the total number of words produced. The type-to-token ratio is the proportion of words that are unique. Changes in noun use, the time-to-token ratio, and the density of information content are linked to cognitive changes seen in Alzheimer disease [21]. Pitch, loudness, and length of words are expected to be familiar to most readers, and we therefore included them.

Table 2a and b shows the 5th percentile, the 25th percentile, the median, the 75th percentile, the 95th percentile, and the mean ICC and WSCV (as percentages) for the 3 tasks separately (sentences, phonation, and connected speech) across the 3 samples (ALS, ALS with cognitive decline, and healthy controls). Table 3a and b shows the ICC and WSCV separated by categories (energy, frequency, etc.) and samples. As Tables 2 and 3 show, the sentences had a higher repeatability (higher median ICC and lower median WSCV) than the phonations, and the connected speech features had the lowest repeatability scores (lowest ICC and highest WSCV).

DOI: 10.1159/000511671

© 2020 The Author(s). Published by S. Karger AG, Basel
www.karger.com/dib

Table 3. ICC and WSCV by category and sample

Diagnosis	Category	5th percentile	25the percentile	Median	Mean	75the percentile	95th percentile
a ICC separated by category and diagnosis							
ALS (sample 1)	Energy	0.14	0.40	0.67	0.60	0.80	0.91
	Frequency	0.01	0.41	0.58	0.51	0.70	0.87
	MFCC	0.09	0.37	0.60	0.55	0.75	0.88
	Pitch	0.11	0.31	0.52	0.49	0.68	0.86
	Spectral	0.09	0.38	0.56	0.53	0.71	0.84
	Temporal	0.01	0.37	0.62	0.57	0.80	0.90
ALS cognitive decline (sample 3)	Energy	0.00	0.12	0.46	0.42	0.65	0.87
	Frequency	0.00	0.07	0.54	0.44	0.71	0.92
	MFCC	0.00	0.08	0.36	0.37	0.62	0.82
	Pitch	0.00	0.13	0.37	0.37	0.59	0.76
	Spectral	0.00	0.18	0.40	0.39	0.58	0.80
	Temporal	0.00	0.12	0.33	0.39	0.67	0.85
	Lexical	0.00	0.00	0.01	0.14	0.25	0.57
	Pragmatic	0.00	0.00	0.32	0.23	0.32	0.39
	Semantic	0.00	0.00	0.19	0.25	0.44	0.66
	Syntactic	0.00	0.00	0.00	0.05	0.01	0.35
Healthy subjects (sample 2)	Energy	0.08	0.27	0.48	0.46	0.65	0.74
	Frequency	0.01	0.14	0.35	0.35	0.55	0.81
	MFCC	0.02	0.22	0.37	0.37	0.53	0.66
	Pitch	0.03	0.23	0.47	0.46	0.65	0.87
	Spectral	0.01	0.25	0.44	0.40	0.55	0.69
	Temporal	0.08	0.30	0.46	0.44	0.60	0.72
b WSCV separated by category and diagnosis							
ALS (sample 1)	Energy, %	7.40	11.63	21.56	46.11	45.93	165.45
	Frequency, %	2.75	7.07	18.32	29.34	44.10	85.68
	MFCC, %	5.07	9.90	16.73	34.52	33.79	74.42
	Pitch, %	7.20	14.87	29.35	39.88	48.12	98.98
	Spectral, %	9.63	33.74	60.95	81.32	95.40	229.41
	Temporal, %	9.64	16.28	21.78	47.21	47.03	165.70
ALS cognitive decline (sample 3)	Energy, %	7.93	13.60	29.43	52.36	56.90	153.44
	Frequency, %	2.93	6.08	16.11	34.98	57.74	95.64
	MFCC, %	4.72	10.43	22.11	40.59	42.30	97.38
	Pitch, %	9.62	16.72	30.57	47.04	45.59	106.56
	Spectral, %	9.85	38.76	63.49	95.37	116.80	285.77
	Temporal, %	7.99	12.62	29.66	62.63	67.39	188.73
	Lexical, %	4.22	7.32	16.61	37.48	27.49	150.01
	Pragmatic, %	52.45	52.45	52.45	138.88	147.77	584.25
	Semantic, %	7.95	10.50	15.21	22.40	30.42	49.90
	Syntactic, %	16.55	30.82	44.69	42.64	54.50	64.74
Healthy subjects (sample 2)	Energy, %	5.12	10.75	23.10	42.26	50.84	124.29
	Frequency, %	2.20	8.21	14.16	24.38	32.01	92.52
	MFCC, %	4.81	8.55	17.69	36.49	34.22	91.74
	Pitch, %	5.67	10.44	22.32	34.48	37.18	94.13
	Spectral, %	5.97	31.42	66.39	84.85	110.06	229.70
	Temporal, %	6.57	21.71	26.57	55.46	49.89	204.46

Table 4. Features with the highest and lowest ICC per category

Category	Highest ICC	Lowest ICC
Energy	pcm_LOGenergy_sma_de_de_linregerrA (ICC = 0.93) pcm_LOGenergy_sma_de_de_nzabsmean (ICC = 0.93) pcm_LOGenergy_sma_meanPeakDist (ICC = 0.94)	pcm_LOGenergy_sma_de_centroid (ICC = 0.39) pcm_LOGenergy_sma_de_de_qregerrA (ICC = 0.39) pcm_LOGenergy_sma_numPeaks (ICC = 0.23)
Frequency	F0semitoneFrom27.5Hz_sma3nz_amean (ICC = 0.89) F0semitoneFrom27.5Hz_sma3nz_percentile50.0 (ICC = 0.92) F0semitoneFrom27.5Hz_sma3nz_percentile80.0 (ICC = 0.88)	F0semitoneFrom27.5Hz_sma3nz_meanFallingSlope (ICC = 0.28) F0semitoneFrom27.5Hz_sma3nz_stddevFallingSlope (ICC = 0.21) F0semitoneFrom27.5Hz_sma3nz_stddevRisingSlope (ICC = 0.24)
MFCC	mfcc_sma_de_de.0._percentile95.0 (ICC = 0.95) mfcc_sma.12._meanPeakDist (ICC = 0.98) mfcc_sma.12._qregerrA (ICC = 0.94)	mfcc_sma_de_de.11._centroid (ICC = 0.20) mfcc_sma_de_de.11._linregc1 (ICC = 0.18) mfcc_sma_de_de.8._quartile2 (ICC = 0.06)
Pitch	voiceProb_sma_linregerrQ (ICC = 0.89) voiceProb_sma_qregerrA (ICC = 0.90) voiceProb_sma_qregerrQ (ICC = 0.91)	F0_sma_quartile1 (ICC = 0.09) F0env_sma_de_de_quartile2 (ICC = 0.02) F0env_sma_de_quartile2 (ICC = 0.00)
Spectral	pcm_fftMag_melspec_sma.1._meanPeakDist (ICC = 0.92) pcm_fftMag_melspec_sma.2._meanPeakDist (ICC = 0.92) pcm_fftMag_melspec_sma.2._qregerrA (ICC = 0.93)	mfcc2_sma3_stddevNorm (ICC = 0.00) mfcc4_sma3_stddevNorm (ICC = 0.00) slopeV0.500_sma3nz_stddevNorm (ICC = 0.00)
Temporal	loudnessPeaksPerSec (ICC = 0.94) pcm_zcr_sma_de_de_numPeaks (ICC = 0.92) pcm_zcr_sma_de_numPeaks (ICC = 0.92)	pcm_zcr_sma_de_de_amean (ICC = 0.27) pcm_zcr_sma_de_de_qregc1 (ICC = 0.21) pcm_zcr_sma_de_de_qregc2 (ICC = 0.25)
Lexical	brunet (ICC = 0.71) MATTR_50 (ICC = 0.59) TTR (ICC = 0.75)	adverbs (ICC = 0.00) coordinate (ICC = 0.00) familiarity (ICC = 0.00)
Pragmatic	rst_num_attribution (ICC = 0.39) rst_num_background (ICC = 0.39) rst_num_cause (ICC = 0.39)	topic18 (ICC = 0.00) topic20 (ICC = 0.00) topic22 (ICC = 0.00)
Semantic	avg_wn_ambig (ICC = 0.69) kurt_wn_ambig_vb (ICC = 0.72) sd_wn_ambig (ICC = 0.64)	avg_max_wn_depth_vb (ICC = 0.00) avg_min_wn_depth_vb (ICC = 0.00) avg_wn_sim_LC (ICC = 0.00)
Syntactic	S (ICC = 0.73) VP_..._VB_VP (ICC = 0.50) W (ICC = 0.67)	ADJP_..._JJ (ICC = 0.00) ADJP_..._RB_JJ (ICC = 0.00) ADVP_..._RB (ICC = 0.00)

For information about the features, readers can refer to openSMILE and Praat for the acoustic features and Talk2me for the language features.

Table 4 shows the 3 features with the highest and lowest ICC for each category, using the sentences from sample 1 (ALS sample) for the acoustic features and transcripts of the picture description task from sample 3 (ALS with cognitive decline sample) for the language features.

The category of features that had the highest repeatability scores was the energy category (i.e., features related to loudness and the energy produced in the acoustic signal). However, there was considerable variability, and there was not a clear explanation for what differentiated the most repeatable from the least repeatable features. For example, among the energy features, pcm_LOGenergy_sma_meanPeakDist and pcm_LOGenergy_sma_numPeaks were 2 features related to the peaks of energy in the acoustic signal, yet they had the highest and lowest ICC (0.94 and 0.23) in the energy features of sample 1 in the sentences. Because the features are not directly interpretable, it is difficult to identify the reasons why a small subset of features has a high reliability, while most of the features are unreliable.

Discussion

Our study evaluated the repeatability of speech features from several complementary perspectives. In this paper, we summarized the findings of the repeatability measures that were not unit dependent (ICC and WSCV); however, in the online supplementary material we provide all measures of reliability for every acoustic feature from healthy individuals and individuals with ALS (samples 1 and 2) separately, as well as every language feature from the ALS individuals with cognitive decline (sample 3).

Overall, we found that the average repeatability scores were well below acceptable limits for clinical decision making. In medical applications, it is recommended that reliabilities exceed ICC = 0.75 [22], with some authors arguing that they should be above ICC = 0.90 [23]. However, only 11% of the features were above ICC = 0.75. Low ICC are indicators that participants are not well separated according to their features. The WSCV were generally large, with half of the metrics having a WSCV above 48%, indicating that half of the features had within-subject SD of 48% or larger than the mean. High WSCV indicate a large natural variability within subjects, which makes it difficult to detect disease-related changes in individuals. In the online supplementary material, we include repeatability scores from the acoustic features from samples 1 and 2 (ALS and healthy) and the language features from sample 3 (ALS with cognitive decline).

Sources of Speech Variation and Mitigation Strategies

The results of this study beg the question: why are speech and language features so variable, even when collected under approximately the same conditions? Every person's speech characteristics vary from utterance to utterance, and day to day, for a variety of reasons unrelated to neurological health. Speech and language characteristics can change as a function of the physiological state of the speaker (e.g., fatigue, hydration, hormonal state, mood, and engagement level) and the degrees of freedom in the speaking task (e.g., repeating a word vs. describing a picture). Additional sources of variability are introduced by the recording setup and environment (e.g., background noise, reverberation, and mic-to-mouth distance), as well as the algorithms themselves. Quality research paradigms can help to attenuate the variability in the study design. In our study, we attempted to control for this by having participants use the same app on the same device and perform the same tasks in the same location session after session. Nevertheless, there was still considerable variation in the measured features. We posit that this is because most of the features in our consideration set were developed for other applications and have only been adopted in the clinical literature out of convenience. For example, OpenSMILE was developed for speech-based emotion recognition and music information retrieval [10, 12, 24]; many of the linguistic features in Talk2Me were developed to assess second-language proficiency [25]. Developing features that are useful in clinical applications and robust to the nuisance variables that drive variability requires targeted work.

Another way of reducing variability in the features is to collect data from a larger number of tasks and average features across them. It is appealing to expect that, from a single task (e.g., picture description), we can extract measures that can be reliably extracted from session to session; however, it is well-known that performance on any single task can vary. Many neuropsychological tests accommodate this variation by measuring the same construct in multiple ways and ensuring that the *internal validity* – the correlation between related items – is high [26]. For this reason, it is critical to oversample speech collection from multiple, related tasks such that clinical variables of interest can be estimated more reliably.

Implications for Digital Health Algorithm Design and Research

Broadly speaking, open-source features have been used in 2 different ways in the clinical-speech literature, i.e., longitudinal tracking of speech features and combining features to build complex machine learning models for detecting disease. Our results have implications for both use cases, as described below.

Longitudinal Tracking of Speech

Some studies have tracked speech features (or combinations of speech features) longitudinally to assess disease progression or the effects of an intervention. For example, D'Alatri et al. [27] used changes in features such as jitter and shimmer to evaluate the effects of medications on Parkinson disease progression. In studies such as this, the repeatability analysis can be used for study design. That is, the SEM and the MDC results can be used for power analysis and to determine the minimum effect size required to detect a change in that feature.

Machine Learning Model Development

A large number of studies in the literature use combinations of features to develop machine learning models that predict a clinical variable of interest. For example, the various challenges in speech-based prediction of Parkinson disease severity [28] and depression [29, 30] have led to a slew of papers utilizing this approach. Studies have shown that machine learning models, and deep learning models especially, are sensitive to noise in the input features [31, 32]; this is especially true in cases where the sample sizes are small, such as in clinical applications [33]. In this case, our results highlight the need for integration of feature repeatability into the design of speech-based machine learning algorithms. The results in the online supplementary material provide another criterion that algorithm designers can use for feature selection and model selection.

Limitations and Future Work

One limitation of our work is that we only considered the repeatability of the features for healthy speech and ALS speech. Our focus on ALS allowed us to explore the repeatability of the features across a range of dysarthria severity and cognitive impairment levels. The broad nature of the impairment in ALS provides a useful test case for assessing the reliability in a heterogeneous population with a wide variety of speech acoustic and natural language features. Therefore, our expectation is that the findings of this study will likely be generalizable to other patient populations as well; however, this needs to be empirically confirmed in other repeatability studies using speech data from other diseases.

The current study explored a very large number of features for a broad overview of the repeatability of open-source speech measures. However, we did not conduct an in-depth exploration of each individual feature to identify the reasons for the low reliability. That effort is challenged by the lack of interpretability in the feature set, which makes it difficult to diagnose the sources of variation in the features. Future work should focus on developing a robust clinically interpretable set of speech features customized to targeted applications with improved repeatability.

Statement of Ethics

Both studies in this paper were approved by the Institutional Review Board at the Barrow Neurological Institute. All of the participants provided written informed consent to participate in this study.

DOI: 10.1159/000511671

Conflict of Interest Statement

Dr. Visar Berisha is an associate professor at Arizona State University. He is a co-founder of Aural Analytics. Dr. Julie Liss is a professor and associate dean at Arizona State University. She is a co-founder of Aural Analytics. Dr. Jeremy Shefner is the Kemper and Ethel Marley Professor and Chair of Neurology at the Barrow Neurological Institute. He is a scientific advisor to Aural Analytics.

Funding Sources

This work was supported by NIH SBIR (1R43DC017625-01), NSF SBIR (1853247), and NIH R01 (5R01DC006859-13), as well as by an ALS Finding a Cure Grant.

Author Contributions

G.M.S. conducted the statistical analyses and led the writing of this paper. S.H. provided expertise on which speech features to measure, helped with writing and editing, and provided input in statistical analyses. J.L. and V.B. are responsible for the speech study design, helped with writing and editing, and provided input in statistical analyses. J.S. and S.B.R. are responsible for the ALS studies' conception and for supervision of the studies. K.K. and S.B. analyzed speech and provided expertise on the speech features. K.S. and C.J.D. are responsible for study execution, data management, and the preliminary analysis.

References

1 Ranjan Y, Rashid Z, Stewart C, Conde P, Begale M, Verbeeck D, et al.; RADAR-CNS Consortium. The Hyve, Dobson R, Folarin A, The RADAR-CNS Consortium. RADAR-Base: open source mobile health platform for collecting, monitoring, and analyzing data using sensors, wearables, and mobile devices. JMIR Mhealth Uhealth. 2019 Aug;7(8):e11734.
2 Narendra NP, Alku P. Dysarthric speech classification from coded telephone speech using glottal features. Speech Commun. 2019;110:47–55.
3 Taguchi T, Tachikawa H, Nemoto K, Suzuki M, Nagano T, Tachibana R, et al. Major depressive disorder discrimination using vocal acoustic features. J Affect Disord. 2018 Jan;225:214–20.
4 Norel R, Pietrowicz M, Agurto C, Rishoni S, Cecchi G. Detection of amyotrophic lateral sclerosis (ALS) via acoustic analysis. Proc Interspeech. 2018:377–81.
5 Komeili M, Pou-Prom C, Liaqat D, Fraser KC, Yancheva M, Rudzicz F. Talk2Me: automated linguistic data collection for personal assessment. PLoS One. 2019 Mar;14(3):e0212342.
6 Bayestehtashk A, Asgari M, Shafran I, McNames J. Fully automated assessment of the severity of Parkinson's disease from speech. Comput Speech Lang. 2015 Jan;29(1):172–85.
7 Faurholt-Jepsen M, Busk J, Frost M, Vinberg M, Christensen EM, Winther O, et al. Voice analysis as an objective state marker in bipolar disorder. Transl Psychiatry. 2016 Jul;6(7):e856.
8 Bland M. An Introduction to Medical Statistics. United Kingdom: Oxford University Press; 2015.
9 Rutkove SB, Qi K, Shelton K, Liss J, Berisha V, Shefner JM. ALS longitudinal studies with frequent data collection at home: study design and baseline data. Amyotroph Lateral Scler Frontotemporal Degener. 2019 Feb;20(1-2):61–7.
10 Eyben F, Wöllmer M, Schuller B. OpenSMILE: the Munich versatile and fast open-source audio feature extractor. Proceedings of the International Conference on Multimedia – MM '10; 2010 Oct 25–29. Florence, Italy. New York: ACM Press; 2010. p. 1459–62.
11 Boersma P. Praat, a system for doing phonetics by computer. Glot Int. 2001;5:341–5.
12 Eyben F, Scherer KR, Schuller BW, Sundberg J, Andre E, Busso C. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Trans Affect Comput. 2016;7(2):190–202.
13 Li L, Zeng L, Lin ZJ, Cazzell M, Liu H. Tutorial on use of intraclass correlation coefficients for assessing intertest reliability and its application in functional near-infrared spectroscopy-based brain imaging. J Biomed Opt. 2015 May;20(5):50801.

14 Rousson V, Gasser T, Seifert B. Assessing intrarater, interrater and test-retest reliability of continuous measurements. Stat Med. 2002 Nov;21(22):3431–46.
15 R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2019. Available from: https://www.R-project.org/.
16 Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Softw. 2015; 67(1):1–48.
17 Chiaramonte R, Bonfiglio M. Acoustic analysis of voice in bulbar amyotrophic lateral sclerosis: a systematic review and meta-analysis of studies. Logoped Phoniatr Vocol. 2019 Nov;1–13.
18 Vieira H, Costa N, Sousa T, Reis S, Coelho L. Voice-based classification of amyotrophic lateral sclerosis: where are we and where are we going? A systematic review. Neurodegener Dis. 2019;19(5-6):163–70.
19 Bang YI, Min K, Sohn YH, Cho SR. Acoustic characteristics of vowel sounds in patients with Parkinson disease. NeuroRehabilitation. 2013;32(3):649–54.
20 Lipsmeier F, Taylor KI, Kilchenmann T, Wolf D, Scotland A, Schjodt-Eriksen J, et al. Evaluation of smartphone-based testing to generate exploratory outcome measures in a phase 1 Parkinson's disease clinical trial. Mov Disord. 2018 Aug;33(8):1287–97.
21 Slegers A, Filiou RP, Montembeault M, Brambati SM. Connected speech features from picture description in Alzheimer's disease: A systematic review. J Alzheimers Dis. 2018;65(2):519–42.
22 Fleiss JL. The Design and Analysis of Clinical Experiments. Wiley; 1999.
23 Portney L, Watkins M. Foundations of Clinical Research: Applications to Practice. Philadelphia: Davis Company; 2015.
24 Eyben F, Huber B, Marchi E, Schuller D, Schuller B. Real-time robust recognition of speakers' emotions and characteristics on mobile platforms. In: 2015 International Conference on Affective Computing and Intelligent Interaction (ACII). IEEE; 2015:778–80.
25 Lu X. The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. Mod Lang J. 2012; 96(2):190–208.
26 Sherman EM, Brooks BL, Iverson GL, Slick DJ, Strauss E. Reliability and validity in neuropsychology. The little black book of neuropsychology. Boston (MA): Springer; 2011. pp. 873–92.
27 D'Alatri L, Paludetti G, Contarino MF, Galla S, Marchese MR, Bentivoglio AR. Effects of bilateral subthalamic nucleus stimulation and medication on parkinsonian speech impairment. J Voice. 2008 May;22(3):365–72.
28 Schuller B, Steidl S, Batliner A, Hantke S, Honig F, Orozco-Arroyave JR, et al. The INTERSPEECH 2015 computational paralinguistics challenge: Nativeness, Parkinson's & eating condition. In Proceedings of INTERSPEECH 2015. 2015; 478–482.
29 Ringeval F, Schuller B, Valstar M, Cummins N, Cowie R, Tavabi L, et al. AVEC 2019 workshop and challenge: State-of-mind, detecting depression with AI, and cross-cultural affect recognition. In Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop. 2019; 3–12.
30 Valstar M, Pantic M, Gratch J, Schuller B, Ringeval F, Lalanne D, et al. AVEC 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge. Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge – AVEC '16; 2016 Oct; Amsterdam, The Netherlands. New York; Amsterdam;2016. p. 3–10.
31 Ng AY. Preventing overfitting of cross-validation data. In Proceedings of the Fourteenth International Conference on Machine Learning. 1997.
32 Akhtar N, Mian A. Threat of adversarial attacks on deep learning in computer vision: A survey. IEEE Access. 2018;6:14410–30.
33 Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. PLoS One. 2019 Nov 7;14(11):e0224365.