The Reproducibility of Bio-Acoustic Features is Associated With Sample Duration, Speech Task, and Gender

Shaykhah A. Almaghrabi, Dominic Thewlis, Simon Thwaites, Nigel C. Rogasch, Stephan Lau, Scott R. Clark, and Mathias Baumert, *Senior Member, IEEE*

***Abstract*— Bio-acoustic properties of speech show evolving value in analyzing psychiatric illnesses. Obtaining a sufficient speech sample length to quantify these properties is essential, but the impact of sample duration on the stability of bio-acoustic features has not been systematically explored. We aimed to evaluate bio-acoustic features' reproducibility against changes in speech durations and tasks. We extracted source, spectral, formant, and prosodic features in 185 English-speaking adults (98 w, 87 m) for reading-a-story and counting tasks. We compared features at 25% of the total sample duration of the reading task to those obtained from non-overlapping randomly selected sub-samples shortened to 75%, 50%, and 25% of total duration using intraclass correlation coefficients. We also compared the features extracted from entire recordings to those measured at 25% of the duration and features obtained from 50% of the duration. Further, we compared features extracted from reading-a-story to counting tasks. Our results show that the number of reproducible features (out of 125) decreased stepwise with duration reduction. Spectral shape, pitch, and formants reached excellent reproducibility. Mel-frequency cepstral coefficients (MFCCs), loudness, and zero-crossing rate achieved excellent reproducibility only at a longer duration. Reproducibility of source, MFCC derivatives, and voicing probability (VP) was poor. Significant gender differences existed in jitter, MFCC first-derivative, spectral skewness, pitch, VP, and formants. Around 97% of features in both genders were not reproducible across speech tasks, in part due to the short counting task duration. In conclusion, bio-acoustic features are less reproducible in shorter samples and are affected by gender.**

***Index Terms*— Bio-acoustic features, features' reproducibility, speech signal processing, speech task.**

Manuscript received May 27, 2021; revised October 27, 2021 and December 12, 2021; accepted December 28, 2021. Date of publication January 17, 2022; date of current version January 31, 2022. *(Corresponding author: Mathias Baumert.)*

This work involved human subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by the University of Adelaide's Human Research Ethics Committee and performed in line with the Declaration of the University of Helsinki.

Shaykhah A. Almaghrabi is with the School of Electrical and Electronic Engineering, The University of Adelaide, Adelaide, SA 5005, Australia, and also with the Biomedical Engineering Department, College of Engineering, Imam Abdulrahman Bin Faisal University, Dammam 31441, Saudi Arabia.

Dominic Thewlis and Simon Thwaites are with the Centre for Orthopaedic and Trauma Research, The University of Adelaide, Adelaide, SA 5000, Australia.

Nigel C. Rogasch is with the Discipline of Psychiatry, Adelaide Medical School, The University of Adelaide, Adelaide, SA 5000, Australia, also with the Hopwood Centre for Neurobiology, Lifelong Health Theme, South Australian Health and Medical Research Institute, Adelaide, SA 5000, Australia, and also with the Turner Institute for Brain and Mental Health, School of Psychological Sciences, Monash University, Melbourne, VIC 3800, Australia.

Stephan Lau is with the Australian Institute for Machine Learning, School of Computer Science, The University of Adelaide, Adelaide, SA 5005, Australia, and also with the South Australian Health and Medical Research Institute, Adelaide, SA 5000, Australia.

Scott R. Clark is with the Discipline of Psychiatry, The University of Adelaide, Adelaide, SA 5000, Australia.

Mathias Baumert is with the School of Electrical and Electronic Engineering, The University of Adelaide, Adelaide, SA 5005, Australia (e-mail: mathias.baumert@adelaide.edu.au).

Digital Object Identifier 10.1109/TNSRE.2022.3143117

I. Introduction

HUMAN speech produces acoustic waves that carry information about the speaker's gender, physiological condition, and pathophysiological state [1]. These waves are generated when the mechanical vibration of vocal folds, affected by aerodynamic factors, are converted into acoustic energy (acoustic source signal). This signal is then filtered and modulated based on the vocal tract configuration shaped by speech articulators [2]–[4]. The ability to control articulatory and phonatory speech processes is affected by neuro-physiological changes in the brain associated with the speaker's mental state. Such changes are encoded into acoustic speech signals and quantified through bio-acoustic qualities such as source, spectral, prosodic, and formants properties [5]–[7].

The control of the vocal fold, which generates the source signal, can be measured through source features such as jitter and shimmer [8]. Spectral features can be obtained from the speech waveform's frequency distribution at a specific time [9] and effectively represented by Mel-frequency cepstral coefficients (MFCCs), for example, to distinguish mood states [10], [11]. Prosodic features, on the other hand, reflect the differences in individuals' speaking styles. They include fundamental frequency (F0) and intensity, expressed through pitch and loudness, respectively [9], [12], [13]. Formants are spectral peaks representing the vocal tract's resonance frequencies and capture essential spectral characteristics for speech analysis [14].

Bio-acoustic properties correlate with mental health disorders [9] and are easily quantifiable using speech analysis techniques [15]; they represent objective biomarkers of mental

health [9]. The feasibility and validity of analyzing speech features through machine learning algorithms to assess mental health conditions such as major depressive disorder were examined previously [16]–[18].

Experimental protocols and methodologies across studies on the association of the speech with clinical outcomes vary significantly [19]–[21], limiting the comparability of results. Studies on speech signal processing use speech samples that differ in speech task type and duration. Some of the studies on depressed individuals, for example, were conducted on three speaking tasks, including an interview, reading-a-story, and picture description, with the overall recording lengths differing 14.5 h, 5.9 h, and 4.5 h, respectively, and average duration of speech recording was 18.3 s [17], [18]. Other researchers used only interview samples with a duration range between 7 to 30 minutes [22], [23]. Therefore, it is critical to determine whether differences in speaking tasks and task duration impact the stability of bio-acoustic feature measurements.

Kiss and Vicsi reported that the measurement of speech features, mainly those calculated over sustained vowels or voiced parts of reading-a-story, is affected by the type of speech task [16]. It was also found that quantifying spectral and cepstral acoustic features, whether from vowel or continuous speech, is dependent on speech content [24]. A study of healthy speakers revealed that different speech types, such as counting, reading passages, and spontaneous speech, impacted the vibration frequency of the vocal folds in connected speech (speaking fundamental frequency) [25].

Vogel and Morgan documented that the length of obtained speech data impacted the measurement accuracy of bio-acoustic features [26]. Although several efforts have been made to explore the accuracy of short-duration speech samples for detecting a disease or estimating a physical parameter [27]–[30], only a few studies have explored the impact of voice sample length on speech characteristics [31]–[33]. Scherer *et al.* have shown that, in sustained vowel tasks, the stability of perturbation measurements, jitter and shimmer, is affected by the task duration. At least 3 s of speech is required to provide accurate measurement [31]. Another study also found that reducing the speech duration from 60 s to 30 s affects the pitch measurements [34]. Additionally, there is high variability in optimal sample duration across a type of predictive task, reflecting the complexity of the outcome measure. For example, complex neurological phenotypes, such as dementia, may take up to 12 minutes of interview speech [28], and one-minute picture descriptions to distinguish individuals with dementia from healthy controls using only acoustic features [35].

Variation in sampling accuracy may also be influenced by gender. Several differences between men's and women's speech have been found related to the vocal folds' mass and vocal tract length, leading to significant differences in phonetics and the quality of voice [36], [37]. Simpson reported that both vocal fold vibration rate and the formants frequencies are higher in women than in men [36].

The first aim of this paper was to examine the effect of speech duration on the reproducibility of women and men adults' bio-acoustic features by determining whether there is a difference between the features extracted from a full-duration task and those measured over shorter durations of the same task. The second aim was to investigate the difference in these parameters between different speech tasks, reading a predefined story versus counting, on the measurements of bio-acoustics qualities.

TABLE I
CHARACTERISTICS OF THE PARTICIPANTS ENROLLED IN THE STUDY

Gender	Men (n=87)	Women (n=98)	*p*-value
Age (years)	26.16 (±6.66)	27.77 (±7.11)	0.1192
Mood score (session 1)	2.32 (±2.04)	2.76 (±2.28)	0.2006
Mood score (session 2)	2.30 (±2.02)	2.69 (±2.25)	0.2470

II. METHODS

A. Dataset

The database contained 796 audio recordings of 199 English speakers aged between 18 and 45 years that were collected at the University of Adelaide as part of a larger study. From every participant, four voice recordings were collected over two separate assessment sessions, with sessions spaced at least three days and at most two weeks apart, at a sampling rate of 44.1 kHz and 16-bit sampling depth in uncompressed WAV format. By using a headset type microphone, the distance between the speaker's mouth and the microphone was kept constant. Each of the four recordings contained a different speaking task: reading a pre-selected story, re-telling the story in the participants' own words, counting from 1 to 20, and telling the capital cities of Australia loudly. Speech was recorded in a 10×14m isolated room within a research facility. Only the investigators and participants were present and the door remained closed at all times. Participants also completed the Mood and Feeling Questionnaire (containing 13 items) before speech recording.

Fourteen subjects were removed to match mood score and age between men and women; nine of them had a mood score suggesting depression (>10 points), while others were relatively older ($=45$ years). Two types of structured, controlled speech tasks were analyzed: reading-a-story, with a mean duration of about 124.4 s (standard deviation $=$ 25.0 s), and counting, with a mean duration of about 26.0 s (standard deviation $=$ 7.7 s). Table I presents the basic characteristics of the participants.

All procedures were approved by the University of Adelaide's Human Research Ethics Committee. All participants provided written informed consent in compliance with the Declaration of the University of Helsinki.

B. Speech Signal Processing

Speech analysis included preprocessing, bio-acoustic feature extraction and statistical analysis of the extracted features to examine reproducibility (Fig. 1).

1) Pre-Processing: Several preprocessing steps were applied to the speech signals to improve the performance of the feature extraction algorithm [38]. Linear down-mixing was

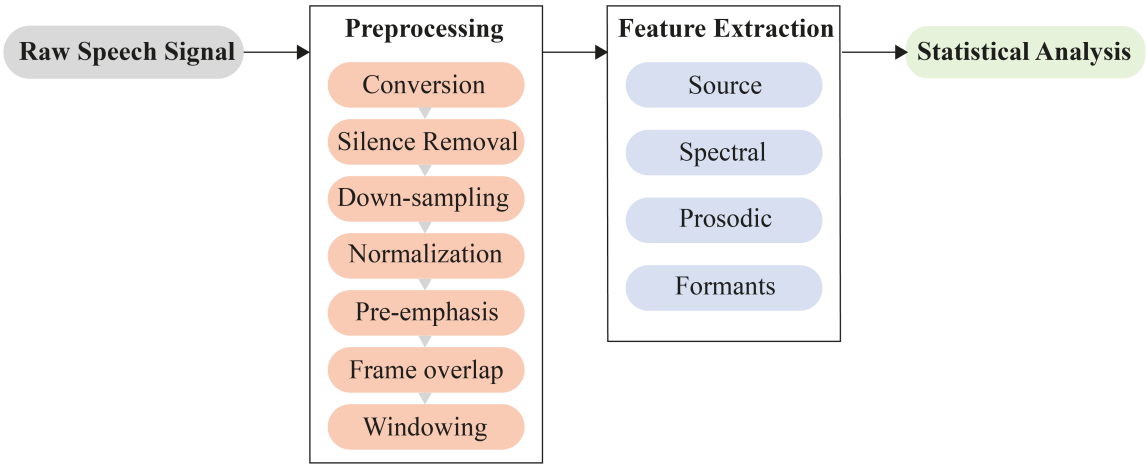

Fig. 1. A block diagram illustrating the steps to examine bio-acoustic features' reproducibility. These steps mainly including preprocessing steps, features extraction, and statistical analysis. Preprocessing steps comprise down-mixing signal, removing silent pauses, resampling speech signal (16 kHz), z-score normalization, and signal pre-emphasis. Moreover, the features extraction step focuses on quantifying acoustic features. Statistical analysis using Intraclass Correlation Coefficient tests was applied to the quantified features.

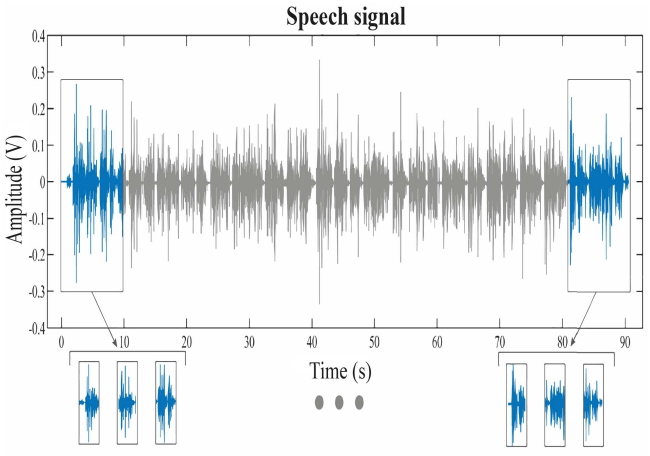

Fig. 2. Segmentation of speech signals into non-overlapping frames of 20 ms duration.

used to convert each recording from two channels (stereo) into a single channel (mono). Silent pauses were eliminated from the input signal to avoid extracting acoustic features from the background acoustical noise [39], [40], by detecting the speech boundaries using the MATLAB® detectSpeech function (The MathWorks, USA). The signals were then down-sampled to 16 kHz, commonly used for speech processing, to reduce the computational load [38], [41]. Samples were then normalized to eliminate differences from the recording environment using the z-score method that centres data to have a zero mean and unit variance [42]. Finally, a pre-emphasis filter was implemented with a coefficient value equal to 0.97, commonly used for speech applications, to enhance the signal-to-noise ratio [41], [43]:

$$H(z) = 1 - 0.97z^{-1} \tag{1}$$

Since the speech signal is non-stationary and considered stable only in short time intervals [38], short-time analysis (framing) is required for analysis, as shown in Fig. 2. We segmented the speech signal into frames of 20 ms duration, as recommended [38], [44]. The frames were overlapped by 50% to avoid introducing any spurious frequency components while processing each frame [38], [43]. Afterwards, the Hamming window, commonly used for speech processing, was applied to all frames to reduce spectral leakage [38], [43].

2) Feature Extraction: Speech feature extraction is at the core of the ability of speech processing systems to derive descriptive attributes of the signal [43]. Speech features can be categorized into two branches: acoustic and linguistic [45]. In this study, we considered only acoustic features, which can be divided into source, spectral, prosodic, and formants features [9]. A summary of the extracted features is provided in Table II. The features were measured with the help of MATLAB®2021a (The MathWorks, USA) [46].

The source features calculated over voiced regions included jitter, which quantifies the cycle-to-cycle variation in the glottal pulse timing period, and shimmer, which quantifies the cycle-to-cycle variation in the amplitude of the glottal pulse [8], [47]. They are defined by the following equations,

$$Jitter(\mu s) = \frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i-1}| \tag{2}$$

$$Shimmer(dB) = \frac{1}{N-1}\sum_{i=1}^{N-1}|20\log_{10}\frac{A_{i+1}}{A_i}| \tag{3}$$

where T_i denotes the time period of the glottal pulse, N denotes the number of periods, and A_i represents the peak-to-peak amplitude [47], [48].

Both jitter and shimmer were determined by utilizing the glottal closure instants within each glottal cycle, which were detected automatically, over 60 ms frame duration, using the Dynamic Programming Projected Phase-Slope Algorithm [49], [50].

MFCCs are one of the most common spectral features employed in speech processing [18], and are analyzed using a bank of band-pass filters. These filters are equally spaced triangular filters in the logarithmic Mel-scale to map the

TABLE II
SUMMARY OF THE EXTRACTED BIO-ACOUSTIC FEATURES

Features Category [a]	Features	Statistical measurements
Source	Jitter	Mean, SD, percentile range (90-10%).
	Shimmer	Mean, SD, percentile range (90-10%).
Spectral	MFCC(1-13)	Mean, SD, percentile range (90-10%), skewness, kurtosis.
	MFCC delta	Mean.
	MFCC delta-delta	Mean.
	SR	Mean, skewness, kurtosis.
	SC	Mean, percentile range (90-10%).
	SE	Mean, SD.
	SF	Mean.
	SS	Mean.
	SK	Mean.
Prosodic	Pitch	Mean, SD, percentile range (90-10%), skewness, kurtosis.
	Loudness	Mean, SD.
	VP	Mean.
	ZCR	Mean, SD, skewness, kurtosis.
Formants	F1	Mean, SD, percentile range (90-10%).
	F2	Mean, SD, percentile range (90-10%).

[a] MFCC: Mel-frequency cepstral coefficients; SR: Spectral roll-off; SC: Spectral centroid; SE: Spectral entropy; SF: Spectral flatness; SS: Spectral skewness; SK: Spectral kurtosis; VP: Voicing probability; ZCR: Zero-crossing rate; F1: First formant; F2: Second formant; SD: Standard deviation.

frequency range of human hearing [38], [51]. We obtained the first 13 MFCCs, MFCC delta, and MFCC delta-delta. Spectral shape descriptors including spectral roll-off points (SR), spectral centroid (SC), spectral entropy (SE), spectral flatness (SF), spectral skewness (SS), and spectral kurtosis (SK) were also computed by converting a time-domain signal into a frequency-domain using Short-time Fourier transform. Most of these descriptors are related to the timbre characteristics of the speech signal.

Pitch is a subjective psychoacoustical attribute of sound and is closely correlated to the physical quantity F0. We estimated the pitch in the short-time domain via the normalized correlation function method [52], with a range set to 75-300 Hz for men and 100-500 Hz for women, as recommended [53]. Loudness, referring to the human ear perception of a sound wave strength [47], was estimated in dB(A) using an A-weighted sound pressure level as a proxy for the measurement of perceived loudness [54].

A voicing probability (VP) determines the speech-silence pattern in the participants' speech. It was obtained by applying a voice activity detector algorithm, introduced by Sohn *et al.* [55], over a segmented speech signal, in the frequency domain to detect speech-present. The probability threshold of transition from voiced to unvoiced frames was set to 0.2, while the transition from unvoiced to voiced frames was set to 0.1.

The zero-crossing rate (ZCR), the number of times the speech signal passing the zero [38], was also extracted on a frame level. Additionally, we tracked the first two formant frequencies (F1 and F2) frame-by-frame throughout the speech signal using the automated formant tracking tool, which employs a recurrent neural network to consider temporal information of signal's frames [56].

Once these features are extracted across each speech sample on a frame basis, several statistical functions were applied over these frames to reduce the influence of tracking errors (e.g., pitch-halving or pitch-doubling errors). These functions include the mean, standard deviation (SD), third- and fourth-order statistical moments (skewness and kurtosis), and percentile range (90-10%) value. This lead to 125 bio-acoustic features per speech sample.

3) Statistics: Statistical analysis was carried out to determine how speech task length and speech task type affected bio-acoustic features' reproducibility in men and women. Each speech sample was segmented three times at different percentages of speech recording length (25%, 50%, and 75%) with a 25% sliding window, resulting in nine speech sub-samples. In seconds, these percentages are approximately equivalent to 31 s ($\pm$6 s), 62 s ($\pm$12 s), and 93 s ($\pm$19 s), respectively. Features calculated over 25% sub-samples were correlated with those obtained from 25%, 50%, and 75% non-overlapping randomly selected sub-samples. The correlation between features calculated over 25% randomly selected sub-samples and the full-length recording were also tested. Features extracted from a similar duration of 50% from the beginning and end of each recording were correlated. Therefore, correlation of five-pair sub-samples were examined: 25% vs. 25%, 25% vs. 50%, 25% vs. 75%, 25% vs. 100%, and 50% vs. 50%. Additionally, the speech characteristics calculated over the first ten seconds of the counting task were correlated to those extracted from the same duration of the reading-a-story task.

To assess the agreement level of the extracted features, intraclass correlation coefficients (ICC) [57] were calculated for individual features. ICC values of $0.00-0.39$, $0.40-0.59$, $0.60-0.74$, and $0.75-1.00$ were used to indicate poor, fair, good, and excellent agreement, respectively. Features with an excellent agreement level (ICC$\geq$0.75) were considered reproducible in this study. A two-way analysis of variance (ANOVA) test was performed over the ICC values to determine significant differences within the five correlation measurements and gender.

III. RESULT

A. Effect of Speech Task Duration

Fig. 3 summarizes the number of reproducible features (ICC $\geq$ 0.75) at different lengths of speech data for the reading-a-story task. The number of reproducible features increased with the duration of the speech task. Comparing features obtained at 25% with 100% speech duration, 82 and 81 acoustic features were deemed reproducible in men and women, respectively. The number of reproducible features decreased to 53 in men and 57 in women when the same duration (25% sub-samples) were correlated. There was no statistical difference between

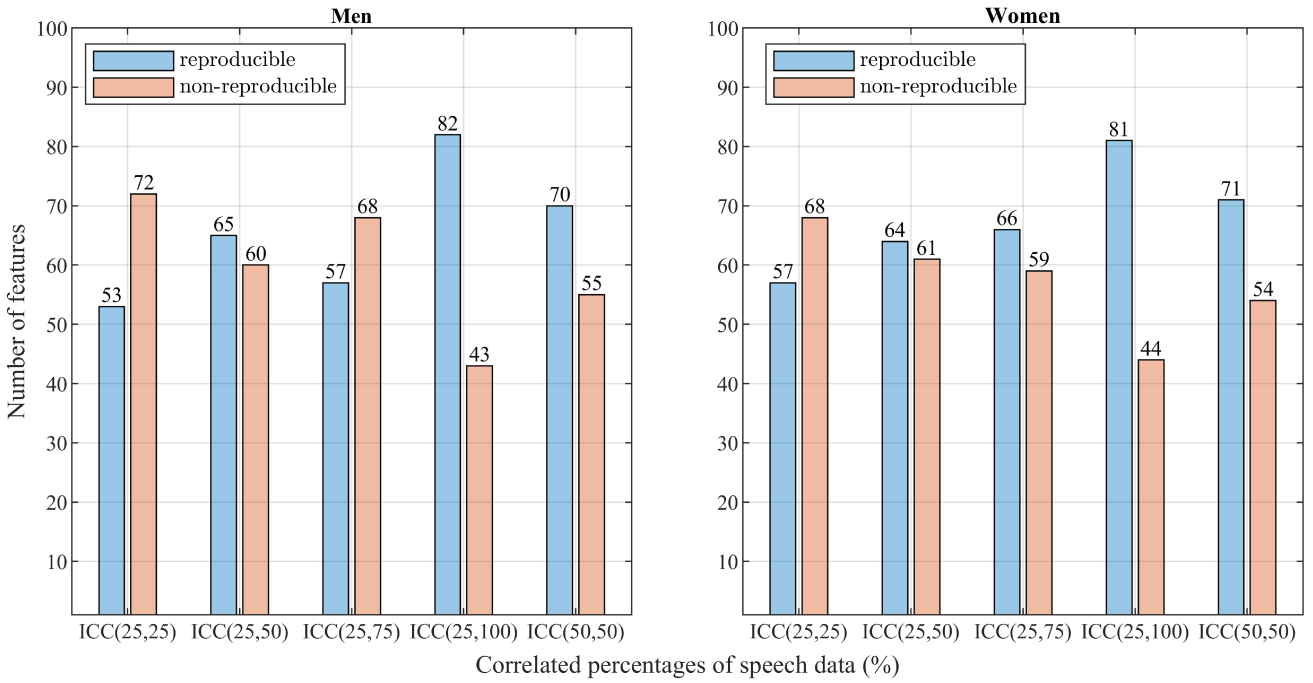

Fig. 3. Comparison of the number of reproducible bio-acoustic features as a function of correlated percentages speech data for men (left) and women (right). Data were extracted for different durations of the reading-a-story task.

men and women (P = 0.52) in ICC values (out of 625) across five paired measurements.

The ICC values for feature categories are shown in Fig. 4. The duration had a considerable impact on source features' reproducibility. Jitter parameters (out of 3) achieved poor to fair reproducibility (ICC < 0.59) across different speech durations (P = 0.30; Fig. 4a); gender difference is significant (P = 0.05). Shimmer parameters' agreement level (ICC < 0.71) was similar (P = 0.38) when considering different speech lengths; there was no significant difference between men and women (P = 0.33; Fig. 4b).

MFCC coefficients, MFCC delta, and MFCC delta-delta contributed about 73% to the total of measured features (91 out of 125). MFCC parameters were affected by speech duration (P < 0.05). Gender has no effect on ICC values of these parameters (P = 0.73). MFCC features also had fair to excellent reproducibility, with a mean ICC value around 0.75 in each measurement (Fig. 4c). Both MFCC delta and MFCC delta-delta attributes were influenced by reducing speech task length, resulting in a poor agreement level (ICC < 0.32; Fig. 4d, Fig. 4e). Gender has a significant impact on ICCs of MFCC delta (P = 0.02), but it has no effect on ICCs of MFCC delta-delta (P = 0.60).

Spectral shape characteristics showed high stability across reduction in speech task lengths. SR parameters (out of 3) achieved excellent reproducibility (ICC > 0.75) when speech duration decreased from full recording to 25%, with no significant gender difference was found (P = 0.23; Fig. 4f). SC parameters (out of 2) had excellent and good-to-excellent agreement level in men and women, respectively, when speech duration is shortened (Fig. 4g). No significant difference was observed in SC ICCs between men and women (P = 0.14). SE and SF were reproducible across different speech durations; no gender effect was found (P > 0.05; Fig. 4h, Fig. 4i). SS and SK showed excellent reproducibility across duration reduction in men and women, as shown in Fig. 4j and Fig. 4k; only a statistical difference was found in SS between genders (P < 0.05).

In terms of prosodic features, gender and speech duration reduction had a significant impact on ICCs of pitch parameters (P<0.05); however, pitch achieved an excellent agreement level (ICC > 0.75) across all comparisons in both genders (Fig. 4l). A wide variation in loudness parameters was found, ranged from fair to excellent agreement; no gender effect was observed (P=0.75; Fig.4m). ICC values of ZCR parameters showed fair to excellent agreement (> 0.40) in both men and women, with no statistical difference was found between genders (P = 0.20; Fig. 4n). VP attributes were varied and considered non-reproducible in men, while women maintained good to excellent ICC values (Fig. 4o).

ICC values of formants features' in men and women were statistically different at P<0.05. Although F1 and F2 parameters were considered reproducible across all duration comparisons (ICC > 0.75), duration reduction impacted ICC values (P<0.05). At full sample duration, the ICCs of F1 parameters was around 0.95 for men and 0.93 for women. These values decreased gradually to nearly 0.90 and 0.83 for men and women, respectively, at the shortest speech length (Fig. 4p). Higher stability was observed in men than in women in ICCs of F2 parameters (Fig.4 q).

B. Effect of Speech Task Type

This experiment examined the reproducibility of bio-acoustic qualities calculated over the first ten seconds of

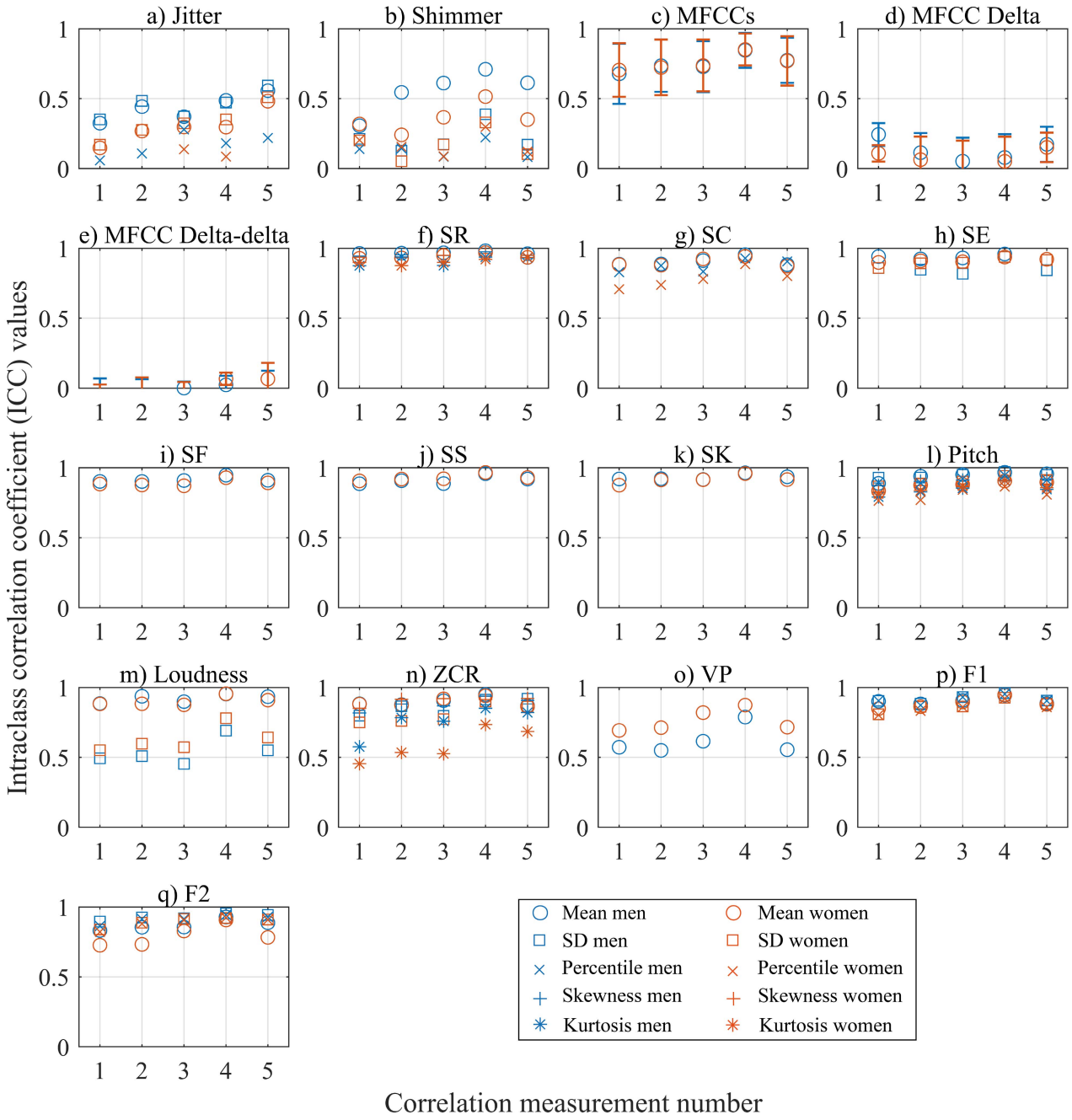

Fig. 4. The ICC values of bio-acoustic features for both men and women is shown in the scatter plot. The numbers on the x-axis can be interpreted as follows; 1: ICC(25% vs. 25%), 2: ICC(25% vs. 50%), 3: ICC(25% vs. 75%), 4: ICC(25% vs. 100%), 5: ICC(50% vs. 50%).

two different tasks; reading-a-story and counting. Table III summarizes the ICC values obtained by comparing these tasks for men and women. Most features showed high variability. Source features lost their reproducibility by changing speech tasks. The mean ICC values of jitter parameters were -0.05 in men and 0.03 in women. Shimmer ICC values were ranged between 0.002 and 0.07 and between -0.03 and -0.22 for men and women, respectively. For both genders, a poor agreement level was found in MFCC, MFCC delta, and MFCC delta-delta parameters. SR showed good to excellent stability in men (ICC $>$ 0.60) and fair to excellent stability in women (ICC $>$ 0.47). For both genders, poor reproducibility was found in SC and SE. Good and fair agreement levels were observed in SS and SK for men and women, respectively. The ICC of SF was around 0.5 for both genders. Speech task type impacted pitch reproducibility in men (ICCs: 0.36-0.67) and women (ICCs: 0.27-0.58). Variation in loudness and ZCR attributes was observed when comparing counting and reading tasks. VP was not a reproducible feature in both genders when speech task type is changed. In men and women, F1 parameters presented fair to good reproducibility, and F2 parameters showed fair reproducibility.

TABLE III
ICC VALUES OF MEASURED BIO-ACOUSTIC FEATURES COMPARING TWO SPEECH TASKS: COUNTING AND READING-A-STORY

Feature	Statistical measurements	ICC Value	
		Men	Women
Jitter	Mean	−0.12	−0.06
	SD	−0.13	−0.03
	Percentile range	0.09	0.18
Shimmer	Mean	0.07	−0.22
	SD	0.05	−0.07
	Percentile range	0.002	−0.03
MFCCs	Mean, SD, Percentile range, Skewness, Kurtosis	0.36 (±0.21)	0.35 (±0.20)
MFCC Delta	Mean	0.13 (±0.16)	0.09 (±0.16)
MFCC Delta-delta	Mean	0.03 (±0.08)	−0.03 (±0.08)
SR	Mean	0.79	0.78
	Skewness	0.80	0.72
	Kurtosis	0.60	0.47
SC	Mean	0.68	0.54
	Percentile range	0.48	0.37
SE	Mean	0.57	0.53
	SD	0.35	0.47
SF	Mean	0.55	0.54
SS	Mean	0.74	0.52
SK	Mean	0.62	0.45
Pitch	Mean	0.67	0.27
	SD	0.68	0.58
	Percentile range	0.60	0.41
	Skewness	0.44	0.33
	Kurtosis	0.36	0.33
Loudness	Mean	−0.18	−0.25
	SD	0.31	0.40
VP	Mean	−0.32	−0.30
ZCR	Mean	0.66	0.54
	SD	0.40	0.45
	Skewness	0.60	0.59
	Kurtosis	0.23	0.27
F1	Mean	0.46	0.56
	SD	0.67	0.65
	Percentile range	0.47	0.55
F2	Mean	0.51	0.47
	SD	0.44	0.50
	Percentile range	0.43	0.56

IV. DISCUSSION

We studied the effect of speech task duration, speech task type, and gender on the reproducibility of bio-acoustic features in normal adults. The main findings of our study are as follows: (i) the reproducibility of acoustic features steadily reduces proportional to speech duration down to about 30 s across gender; (ii) acoustic speech properties are less reproducible in less complex counting versus reading-a-story tasks; and, (iii) Some spectral (spectral shape descriptors), prosodic (pitch), and formants (F1, F2) features reached excellent reproducibility in both genders at different speech duration. Some spectral features (MFCC) and prosodic features (Loudness, ZCR) achieved excellent reproducibility at a longer duration. The reproducibility of source (Jitter, Shimmer), and other spectral (MFCC delta, and MFCC delta-delta) features were lost when speech duration was changed. There were significant gender differences in jitter, MFCC delta, SS, pitch, VP, and formants (F1, F2).

Interview based diagnostic and prognostic assessments for common psychiatric illnesses, such as major depression, have limited reliability and predictive accuracy [58]. Examining reproducibility of acoustic features at different speech durations has become of clinical interest to improve the accuracy of the assessment and provide valuable insights that can drive the assessment. Few studies have explored the impact of voice sample length on speech characteristics [31]–[33]. Previous work has largely focused on evaluating only one type of acoustic property against time. Scherer *et al.* suggested that at least 3 s of recording are required for accurate reading of speech perturbations [31]. To the best of our knowledge, no study has systematically investigated the influence of decreasing the length of a speech signal on the reproducibility of bio-acoustic features in healthy individuals.

In our study, the source features' measurements were not reproducible when fewer voice samples were considered. Perturbation measurement stability is dependent on the components of speech in the location of the selected segments, for instance, there is high variability between different vowels [31]. Selecting a more stable speech segment, periodic (repetitive) or nearly periodic (nearly-repetitive) waves, leads to a more consistent result [59]. Based on the measurements of logMel and MFCC features of cropped signals from about 8 s to about 1 s, Neumann and Vu reported that a system for emotion detection performed sufficiently, despite a slight loss in accuracy compared to the use of full samples [32]. Our study found that the reproducibility of MFCC features reduces as duration shortens, which might cause a loss of prediction accuracy in such a system [32].

We showed that the pitch parameters were reproducible with reduced sample duration in both women and men. A study on German speakers showed a substantial effect of utterance length on the variability of F0 measurements [33]. Zraick *et al.* also showed that the estimated pitch value of White women varied for different speech durations [34]. Several factors may have contributed to the differences between our study and prior investigations, including speech task type, differences of speakers' language, and method used to compute pitch. In this study, we limited our analysis to English speakers who read a story, and the normalized correlation function method was used to extract pitch. Nishinuma *et al.* report an effect of shorter sample duration on the loudness measurement [60]. Similarly, we showed a wide variation in loudness parameters across all duration comparisons. We demonstrated

the considerable impact of duration on VP feature reproducibility in men and women [61], [62].

A study of French and German speakers has shown that decreasing the speech duration influences both F1 and F2 as a function of vowel duration [63]. Although we observed a similar pattern, where short speech data impacted formants' qualities, the ICC values remained high (>0.75). For men and women, formants measures were reproducible across durations. We found higher stability in men's formants than women's.

Several studies have investigated the impact of the speech task on acoustic parameters [24], [64], [65]. Our study tested the reproducibility of a wide range of voice parameters during counting and reading-a-story tasks. Our results demonstrate that changing speech tasks impacted at least 96% and 98% of the measured acoustic qualities for men and women, respectively, even if the duration was identical (10 s).

Several studies suggest that vowel type impacts shimmer parameters [66], [67]. Similarly, we found that the shimmer feature was not reproducible between tasks. Our results indicate that although the task type had a significant effect on the measurements of pitch features in women, it achieved good reproducibility in men. This finding that is in line with Sandage *et al.* [68] and Zraick *et al.* [25]. Hence, some spectral shape descriptors (i.e., SR and SC) are relatively stable across speaking tasks.

Gender differences in speech arising from difference in vocal cord anatomy lead to dissimilarities in some acoustic features such as F0 and jitter [36], [48]. In our study, when men and women were analyzed separately, we found that significant differences in the correlation analysis of some speech properties, including jitter, MFCC delta, SS, pitch, VP, and formants, suggesting that the pattern of reliable markers may be different across gender.

Our study has several limitations. First, we assessed the reproducibility of bio-acoustic features derived from native English speakers only and in a dataset with an identical recording setup; we did not validate our findings on voice samples across different datasets, languages or environments. Hence we may have overestimated real-world generalized reproducibility. Second, we only examined acoustic parameter reproducibility in individuals with healthy voices. Furthermore, we conducted our experiments in a restricted sample of participants aged 18 to 44 years (mean = 27 years), limiting generalizability to older or younger individuals. Additionally, scripted speech tasks, while allowing standardized comparisons, do not elicit natural speech [69]. Our results need to be validated with future studies on natural speech samples across more diverse age groups, languages and environments, compared between clinical and healthy control samples. Finally, given that the content of the speech data impact the acoustic features, thus when 25% of full speech duration was correlated with the full recording duration, speech content of 25% is included within the full recording, which may affect the correlation results.

V. Conclusion

This study has examined the effect of speech duration and speech task on the reproducibility of bio-acoustic qualities. Shortening the speech duration from full duration to 25% of total speech duration reduces the speech features' reproducibility from 82 and 81 to 53 and 57 in men and women, respectively. Thus, clinicians may have to collect a minimum of speech data to achieve a high number of reproducible bio-acoustic features (at least one minute and a half in the case of the reading-a-story task). In addition, changing the speech task has a significant effect on the measurements of features; around 97% of features in both genders lost reproducibility. Therefore, researchers may have to build and train speech-task specific models (classifier/regressor). Gender factor has a significant impact on the reproducibility of jitter, MFCC delta, SS, pitch, VP, and formants qualities.

References

[1] R. K. Sharma and A. K. Gupta, "Estimation and statistical analysis of physical task stress on human speech signal," *Int. J. Image, Graph. Signal Process.*, vol. 8, no. 10, pp. 29–34, Oct. 2016.

[2] M. H. Farouk, "Speech production and perception," in *Application of Wavelets in Speech Processin*. Cham, Switzerland: Springer, 2018, pp. 5–10.

[3] J. Kreiman and D. Sidtis, "Producing a voice and controlling its sound," in *Foundations of Voice Studies*. Oxford, U.K.: Wiley, 2011, pp. 25–71.

[4] P. Song, "Assessment of vocal cord function and voice disorders," in *Principles and Practice of Interventional Pulmonology*. New York, NY, USA: Springer, 2012, pp. 137–149.

[5] N. Cummins, A. Baird, and B. W. Schuller, "Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning," *Methods*, vol. 151, pp. 41–54, Dec. 2018.

[6] T. Akkaralaertsest and T. Yingthawornsuk, "Classification of depressed speech samples with spectral energy ratios as depression indicator," in *Proc. 14th Int. Joint Symp. Artif. Intell. Natural Lang. Process. (iSAI-NLP)*, Oct. 2019, pp. 1–6.

[7] Y. Tahir *et al.*, "Non-verbal speech cues as objective measures for negative symptoms in patients with schizophrenia," *PLoS ONE*, vol. 14, no. 4, Apr. 2019, Art. no. e0214314.

[8] T. F. Quatieri and N. Malyska, "Vocal-source biomarkers for depression: A link to psychomotor activity," in *Proc. 13th Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, Sep. 2012, pp. 1059–1062.

[9] D. M. Low, K. H. Bentley, and S. S. Ghosh, "Automated assessment of psychiatric disorders using speech: A systematic review," *Laryngoscope Invest. Otolaryngol.*, vol. 5, no. 1, pp. 96–116, Jan. 2020.

[10] V. Sethu, E. Ambikairajah, and J. Epps, "Speaker dependency of spectral features and speech production cues for automatic emotion classification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Apr. 2009, pp. 4693–4696.

[11] N. Cummins, J. Epps, M. Breakspear, and R. Goecke, "An investigation of depressed speech detection: Features and normalization," in *Proc. 12th Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, Aug. 2011, pp. 1–4.

[12] N. Cummins, "Automatic assessment of depression from speech: Paralinguistic analysis, modelling and machine learning," Ph.D. dissertation, School Elect. Telecommun. Eng., UNSW, Sydney, NSW, Australia, 2016.

[13] V. Mitra and E. Shriberg, "Effects of feature type, learning algorithm and speaking style for depression detection from speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 4774–4778.

[14] M. Lech, "Detection of adolescent depression from speech using optimised spectral roll-off parameters," *Biomed. J. Sci. Tech. Res.*, vol. 5, no. 1, p. 10, Jun. 2018.

[15] T. Boonla and T. Yingthawornsuk, "Assessment of vocal correlates of clinical depression in female subjects with probabilistic mixture modeling of speech cepstrum," in *Proc. 11th Int. Conf. Contr., Autom., Syst.*, Oct. 2011, pp. 387–391.

[16] G. Kiss and K. Vicsi, "Mono- and multi-lingual depression prediction based on speech processing," *Int. J. Speech Technol.*, vol. 20, no. 4, pp. 919–935, Sep. 2017.

[17] H. Jiang *et al.*, "Investigation of different speech types and emotions for detecting depression using different classifiers," *Speech Commun.*, vol. 90, pp. 39–46, Jun. 2017.

[18] H. Jiang *et al.*, "Detecting depression using an ensemble logistic regression model based on multiple speech features," *Comput. Math. Methods Med.*, vol. 2018, Sep. 2018, Art. no. 6508319.

[19] C. Perez, Y. C. Roca, L. Naranjo, and J. Martin, "Diagnosis and tracking of Parkinson's disease by using automatically extracted acoustic features," *J. Alzheimer's Disease Parkinsonism*, vol. 6, no. 5, p. 45, 2016.

[20] Z. Du, W. Li, D. Huang, and Y. Wang, "Bipolar disorder recognition via multi-scale discriminative audio temporal representation," in *Proc. Audio/Visual Emotion Challenge Workshop*, Oct. 2018, pp. 23–30.

[21] P. Lopez-Otero, L. Dacia-Fernandez, and C. Garcia-Mateo, "A study of acoustic features for depression detection," in *Proc. 2nd Int. Workshop Biometrics Forensics*, Mar. 2014, pp. 1–6.

[22] E. A. Stepanov *et al.*, "Depression severity estimation from multiple modalities," in *Proc. 20th IEEE Int. Conf. E-Health Netw., Appl. Serv. (Healthcom)*, Sep. 2018, pp. 1–6.

[23] S. Scherer, G. Stratou, J. Gratch, and L.-P. Morency, "Investigating voice quality as a speaker-independent indicator of depression and PTSD," in *Proc. 14th Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, Aug. 2013, pp. 847–851.

[24] S. H. Choi and C.-H. Choi, "The effect of gender and speech task on cepstral and spectral-measures of Korean normal speakers," *Audiol. Speech Res.*, vol. 12, no. 3, pp. 157–163, Jul. 2016.

[25] R. I. Zraick, S. D. Skaggs, and J. C. Montague, "The effect of task on determination of habitual pitch," *J. Voice*, vol. 14, no. 4, pp. 484–489, Dec. 2000.

[26] A. Vogel and A. Morgan, "Factors affecting the quality of sound recording for speech and voice analysis," *Int. J. Speech-Lang. Pathol.*, vol. 11, no. 6, pp. 431–437, Dec. 2009.

[27] S. B. Kalluri, D. Vijayasenan, and S. Ganapathy, "Automatic speaker profiling from short duration speech data," *Speech Commu.*, vol. 121, pp. 16–28, May 2020.

[28] J. Weiner, M. Angrick, S. Umesh, and T. Schultz, "Investigating the effect of audio duration on dementia detection using acoustic features," in *Proc. 19th Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, Sep. 2018, pp. 2324–2328.

[29] S. B. Kalluri, D. Vijayasenan, and S. Ganapathy, "A deep neural network based end to end model for joint height and age estimation from short duration speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6580–6584.

[30] M. Sigmund, "Gender distinction using short segments of speech signal," *Int. J. Comput. Sci. Netw. Secur.*, vol. 8, no. 10, pp. 159–162, Oct. 2008.

[31] R. C. Scherer, V. J. Vail, and C. G. Guo, "Required number of tokens to determine representative voice perturbation values," *J. Speech, Lang., Hearing Res.*, vol. 38, no. 6, pp. 1260–1269, Dec. 1995.

[32] M. Neumann and N. Thang Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," 2017, *arXiv:1706.00612*.

[33] C. Draxler, F. Schiel, and T. Ellbogen, "F0 of adolescent speakers-first results for the German Ph@ ttSessionz database," in *Proc. 6th Int. Conf. Lang. Resour. Eval.*, May 2008, pp. 2275–2278.

[34] R. I. Zraick, K. Y. Birdwell, and L. Smith-Olinde, "The effect of speaking sample duration on determination of habitual pitch," *J. Voice*, vol. 19, no. 2, pp. 197–201, Jun. 2005.

[35] A. Satt, R. Hoory, A. König, P. Aalten, and P. H. Robert, "Speech based automatic and robust detection of very early dementia," in *Proc. 15th Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, Sep. 2014, pp. 1–5.

[36] A. P. Simpson, "Phonetic differences between male and female speech," *Lang. Linguistics Compass*, vol. 3, no. 2, pp. 621–640, Mar. 2009.

[37] D. R. R. Smith and R. D. Patterson, "The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age," *J. Acoust. Soc. Amer.*, vol. 118, no. 5, pp. 3177–3186, Nov. 2005.

[38] T. Özseven and M. Däenci, "SPeech ACoustic (SPAC): A novel tool for speech feature extraction and classification," *Appl. Acoust.*, vol. 136, pp. 1–8, Jul. 2018.

[39] C. Dong and K. Jingming, "A robust voice activity detector applied for AMR," in *Proc. 5th Int. Conf. Signal Process.*, 2000, pp. 687–692.

[40] J. A. Morales-Cordovilla, N. Ma, V. Sánchez, J. L. Carmona, A. M. Peinado, and J. Barker, "A pitch based noise estimation technique for robust speech recognition with missing data," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 2011, pp. 4808–4811.

[41] F. Eyben, *Real-Time Speech Music Classification by Large Audio Feature Space Extraction*. Cham, Switzerland: Springer, 2015.

[42] B. Schuller *et al.*, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Trans. Affect. Comput.*, vol. 1, no. 2, pp. 119–131, Jul. 2010.

[43] I. McLoughlin, *Application Speech Audio Processing: With MATLAB Examples*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[44] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech signals*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1980.

[45] T. Polzehl, A. Schmitt, F. Metze, and M. Wagner, "Anger recognition in speech using acoustic and linguistic cues," *Speech Commun.*, vol. 53, nos. 9–10, pp. 1198–1209, Nov.-Dec. 2011.

[46] The MathWorks. (2021). *Audio Toolbox*. [Online]. Available: https://au.mathworks.com/help/audio/index.html

[47] R. Singh, *Profiling Humans From Their Voice*. Singapore: Springer, 2019.

[48] J. P. Teixeira and P. O. Fernandes, "Jitter, shimmer and HNR classification within gender, tones and vowels in healthy voices," *Proc. Technol.*, vol. 16, pp. 1228–1237, Mar. 2014.

[49] M. Brookes. (1997). *VoiceBox: Speech Processing Toolbox for MATLAB*. [Online]. Available: https//www.ee.ic.ac.U.K./hp/staff/dmb/voicebox/voicebox.html

[50] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 1, pp. 34–43, Jan. 2007.

[51] V. Sethu, "Automatic emotion recognition: An investigation of acoustic and prosodic parameters," Ph.D. dissertation, School of Elect. Telecommun. Eng., UNSW, Sydney, NSW, Australia, 2009.

[52] B. S. Atal, "Automatic speaker recognition based on pitch contours," *J. Acoust. Soc. Amer.*, vol. 52, no. 6B, pp. 1687–1697, Dec. 1972.

[53] P. Boersma, and D Weenink. (2019). *Praat: Doing Phonetics by Computer*. [Online]. Available: https://www.fon.hum.uva.nl/praat/

[54] S. Namba and S. Kuwano, "Psychological study on Leq as a measure of loudness of various kinds of noises," *J. Acoust. Soc. Jpn.*, vol. 5, no. 3, pp. 135–148, 1984.

[55] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[56] Y. Dissen, J. Goldberger, and J. Keshet, "Formant estimation and tracking: A deep learning approach," *J. Acoust. Soc. Amer.*, vol. 145, no. 2, pp. 642–653, Feb. 2019.

[57] S. Arash, *Intraclass Correlation Coefficient*. Portola Valley, CA, USA: Mathworks, 2018. [Online]. Available: https://au.mathworks.com/

[58] M. Cearns *et al.*, "Predicting rehospitalization within 2 years of initial patient admission for a major depressive episode: A multimodal machine learning approach," *Transl. Psychiatry*, vol. 9, no. 1, pp. 1–9, Nov. 2019.

[59] A. E. Olszewski, L. Shen, and J. J. Jiang, "Objective methods of sample selection in acoustic analysis of voice," *Ann. Otol., Rhinol. Laryngol.*, vol. 120, no. 3, pp. 155–161, Mar. 2011.

[60] Y. Nishinuma, A. Di Cristo, and R. Espesser, "How does vowel duration affect loudness in a CV syllable?" *Speech Commun.*, vol. 3, no. 1, pp. 39–47, Apr. 1984.

[61] C. Shih, B. Möbius, and B. Narasimhan, "Contextual effects on consonant voicing profiles: A cross-linguistic study," in *Proc. 14th Int. Congr. Phonetic Sci.*, Aug. 1999, vol. 2, pp. 989–992.

[62] B. Möbius, "Corpus-based investigations on the phonetics of consonant voicing," *Societas Linguistica Europaea*, vol. 38, nos. 1–2, pp. 5–26, 2004.

[63] C. Gendrot and M. Adda-Decker, "Impact of duration on F1/F2 formant values of oral vowels: An automatic analysis of large broadcast news corpora in French and German," in *Proc. 9th Eur. Conf. Speech Commun. Technol. (Interspeech-Eurospeech)*, Sep. 2005, pp. 2453–2456.

[64] S. Y. Lowell and J. A. Hylkema, "The effect of speaking context on spectral and cepstral-based acoustic features of normal voice," *Clin. Linguistics Phonetics*, vol. 30, no. 1, pp. 1–11, Jan. 2016.

[65] S. N. Awan, A. Giovinco, and J. Owens, "Effects of vocal intensity and vowel type on cepstral analysis of voice," *J. Voice*, vol. 26, no. 5, p. 670, Sep. 2012.

[66] S. H. Choi and C.-H. Choi, "The stability and variability based on vowels in voice quality analysis," *Phonetics Speech Sci.*, vol. 7, no. 1, pp. 79–86, Mar. 2015.

[67] M. C. Franca, "Acoustic comparison of vowel sounds among adult females," *J. Voice*, vol. 26, no. 5, p. 671, Sep. 2012.

[68] M. J. Sandage, L. W. Plexico, and A. Schiwitz, "Clinical utility of CAPE-V sentences for determination of speaking fundamental frequency," *J. Voice*, vol. 29, no. 4, pp. 441–445, Jul. 2015.

[69] MA. Romana, "Automatically detecting errors and disfluencies in read speech to predict cognitive impairment in people with parkinson disease," in *Proc. 21th Annu. Conf. Int. Speech Commun. Assoc.*, 2021, pp. 1907–1911.