Speech Communication 132 (2021) 10–20

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Read speech voice quality and disfluency in individuals with recent suicidal ideation or suicide attempt

Brian Stasak [a,*], Julien Epps [a], Heather T. Schatten [b,c], Ivan W. Miller [b,c], Emily Mower Provost [d], Michael F. Armey [b,c]

[a] School of Elec. Eng. & Telecomm., UNSW, Sydney, Australia
[b] Warren Alpert Medical School, Brown University, Providence, RI, USA
[c] Butler Hospital, Providence, RI, USA
[d] EECS Dept., University of Michigan, Ann Arbor, MI, USA

ARTICLE INFO

Keywords:
Digital phenotyping
Digital medicine
Machine learning
Mental health
Psychogenic voice disorders

ABSTRACT

Individuals that have incurred trauma due to a suicide attempt often acquire residual health complications, such as cognitive, mood, and speech-language disorders. Due to limited access to suicidal speech audio corpora, behavioral differences in patients with a history of suicidal ideation and/or behavior have not been thoroughly examined using subjective voice quality and manual disfluency measures. In this study, we examine the Butler-Brown Read Speech (BBRS) database that includes 20 healthy controls with no history of suicidal ideation or behavior (HC group) and 226 psychiatric inpatients with recent suicidal ideation (SI group) or a recent suicide attempt (SA group). During read aloud sentence tasks, SI and SA groups reveal poorer average subjective voice quality composite ratings when compared with individuals in the HC group. In particular, the SI and SA groups exhibit average 'grade' and 'roughness' voice quality scores four to six times higher than those of the HC group. We demonstrate that manually annotated voice quality measures, converted into a low-dimensional feature vector, help to identify individuals with recent suicidal ideation and behavior from a healthy population, generating an automatic classification accuracy of up to 73%. Furthermore, our novel investigation of manual speech disfluencies (e.g., manually detected hesitations, word/phrase repeats, malapropisms, speech errors, non-self-correction) shows that inpatients in the SI and SA groups produce on average approximately twice as many hesitations and four times as many speech errors when compared with individuals in the HC group. We demonstrate automatic classification of inpatients with a suicide history from individuals with no suicide history with up to 80% accuracy using manually annotated speech disfluency features. Knowledge regarding voice quality and speech disfluency behaviors in individuals with a suicide history presented herein will lead to a better understanding of this complex phenomenon and thus contribute to the future development of new automatic speech-based suicide-risk identification systems.

1. Introduction

The majority of people who make suicide attempts do not die by suicide; in 2017 there were approximately 47,000 suicide deaths and an estimated 1400,000 suicide attempts in the United States of America (CDC, 2017). Due to the potentially harmful nature of suicide methods[†], according to Costache et al. (2004), Wazeer et al. (2015), and Zabel et al. (2005), survivors of attempted suicide often exhibit a debilitating range of irreversible health concerns, such as neuropsychological, neuropsychiatric, physiological, cognitive, and speech-language disorders. Brodnitz et al. (1971) was one of the first studies to report psychological problems found in patients with voice disorders. Their study of over 2000 patients, involving all forms of voice disorders, found that 80% of all voice disorder cases were attributed to vocal abuse and/or psychogenic factors (e.g., anxiety, depression). Further, Marmor et al. (2016) noted that depressive symptoms in patients were accompanied by nearly a two-fold increase in a reported voice problem in the past year when compared to a healthy population.

* Corresponding author.
E-mail addresses: b.stasak@unsw.edu.au (B. Stasak), j.epps@unsw.edu.au (J. Epps), heather_schatten@brown.edu (H.T. Schatten), ivan_miller_iii@brown.edu (I.W. Miller), emilykmp@umich.edu (E.M. Provost), michael_armey@brown.edu (M.F. Armey).

https://doi.org/10.1016/j.specom.2021.05.004
Received 8 May 2020; Received in revised form 9 February 2021; Accepted 10 May 2021
Available online 17 May 2021

Mood disturbances directly impact the speech system, triggering quantifiable divergences in normal healthy speech physiological mechanisms (e.g., respiratory, muscle tension, motor coordination). For individuals with clinical depression and/or those exhibiting suicidal behavior, recorded changes in their voice characteristics are frequently attributed to psychogenic emotional and stress symptoms (Cummins et al., 2015). Examples of psychogenic symptoms include psychomotor retardation and agitation. Disturbances caused by psychomotor retardation include poorer cognitive processing and muscular incoordination, which adversely impact gross/fine motor movement and speech production (Flint et al., 1993; Hoffman et al., 1985; Silverman et al., 1992). Moreover, psychomotor agitation results in abnormal accelerated motor activity and excessive gross/fine motor movements (Day, 1999). In investigations by France et al. (2000), Ellgring and Scherer, (1996), and Yingthawornsuk et al. (2006), careful evaluation of abnormal acoustic vocal manifestations has helped to motivate new ways to automatically identify mood disorders in patients.

Suicidal speech-based literature, such as Cummins et al. (2015), Ozdas et al. (2004), Scherer et al. (2013), and Yingthawornsuk et al. (2006) have indicated that patients with a history of suicidal ideation exhibit lower acoustic energy, unusual glottal control, and breathy voice quality when compared with healthy populations. But, in these aforementioned studies, only a relatively small number of clinically validated patients with suicidal ideation and/or attempts were analyzed (i.e., less than two dozen per study). Furthermore, in many of these studies, patients' voice quality attributes were reliant on spectral acoustic-based features rather than grounded on a standard set of clinical descriptive pathological qualities.

Scherer et al. (2013) examined 'breathiness' and 'tenseness' in a narrow demographic of adolescents with and without suicidal ideation and/or behavior. Their study found that the adolescents' speech exhibited significantly more breathy qualities than adolescents without suicidal behavior based on peak slope and normalized amplitude quotient acoustic feature values. However, Scherer et al. (2013) did not explore other potential common pathological voice quality attributes (i.e., hoarseness, roughness, instability); or verify that these particular acoustic features only captured 'breathiness' quality information (i.e., they could also be capturing information from other voice quality attributes). Studies by Brodnitz et al. (1971), Mamor et al. (2016), and Scherer et al. (2013) hint that abnormal voice quality is associated with suicidal behavior. However, automatic speech-based studies have yet to further investigate several distinct pathological voice qualities associated with suicide present in a more sizable suicidal dataset.

Studies by Esposito et al. (2016), Oxman et al. (1988), Rosenberg et al. (1991), and Rubino et al. (2011) have shown that clinical depression can be identified through patients' spontaneous speech disfluencies. Further, Stasak et al. (2019) found that when compared with healthy controls, patients with clinical depression exhibited significantly greater numbers of speech disfluencies during specific emotionally charged read aloud sentence tasks. For patients with suicidal behavior, it is anticipated that during simple read sentence tasks this population will show an increase in speech disfluencies due to associated depression and cognitive dysfunction (Levens and Gotlib, 2015; Marzuk et al., 2005; Mitterschiffthaler et al., 2008; Roy-Byrne et al., 1986; Rubino et al., 2011; Weingartner et al., 1981).

This research present herein is one of the largest-scale studies of inpatients with suicidal ideation and behavior to-date that investigates voice quality and speech disfluency behaviors found in such samples using text-dependent read aloud elicitation with a range of mood content. As an elicitation protocol, read speech has many advantages over spontaneous speech because it: (1) constrains the phonetic variability; (2) controls the syntactic order of affective word content; (3) isolates a patient's cognitive-processing demands; (4) offers objective clinical repeatability; and (5) reduces potential patient-observer bias caused by interviewer-adaptation, which influences the speaking style of a participant (Bouhuys and Van Den Hoofdakker, 1991).

In this study, we investigate the subjective GRBASI voice pathology quality attributes (e.g., 'grade', 'roughness', 'breathiness', 'asthenia', 'strain', 'instability') to help establish which of these are most associated with the speech of psychiatric inpatients hospitalized for suicidal ideation or suicide attempt. In addition, the speech of healthy controls is compared to the psychiatric inpatients. We hypothesize that voice quality is an important indicator for individuals who are at higher risk for suicide, regardless of whether they exhibit depression. Based on the previous literature (Costache et al., 2004; Wazeer et al., 2015; Zabel et al. 2005), we hypothesize that inpatients with a history of recent suicide attempts will exhibit abnormal voice qualities along with language processing and production difficulties. It is theorized that descriptive voice quality and speech disfluency measures can be applied as discriminative low-dimensional features to help automatically classify individuals with no suicide history and psychiatric inpatients with a recent history of suicidal thoughts or behaviors.

2. Database

The Butler-Brown Read Speech (BBRS) database is a privately collected speech corpus consisting of recordings of participants reading a set of sentences into a microphone. All participants were recorded at a psychiatric hospital in the northeastern United States of America. The BBRS database was developed to investigate verbal behaviors of inpatients hospitalized for recent suicidal ideation (SI group) or suicide attempts (SA group), along with a group of healthy controls recruited from the community with no history of suicide (HC group). Due to the sensitive nature of suicide, inpatient privacy, and obligatory human-listening research safety precautions, both public and private speech corpus collections of this nature are rare in the literature.

The BBRS database has inpatient metadata that is unlike other previously published speech corpora (France et al., 2000; Ozdas et al., 2004; Scherer et al., 2013; Yingthawornsuk et al., 2006) related to suicide. For example, the BBRS database has a relatively large number of inpatients with lifetime and past-month suicidal ideation and suicide attempt history (or lack thereof) validated by a semi-structured clinical interview (i.e., the Columbia-Suicide Severity Rating Scale [C-SSRS]), demographic metadata (e.g., age, gender, suicide history), self-report measures (e.g., Beck Depression Inventory II), and uniform English read aloud sentence recordings. Additionally, all inpatient recordings were conducted in the exact same research laboratory environment under the same protocol instructions with a research team member present.

Presented in Table 1, the BBRS database comprises of participants without a history of suicide ideation or attempt (HC group; $n = 20$) and 226 inpatients with a recent past history of suicide ideation without history of suicide attempt (SI group; $n = 74$) or any past history of attempted suicide with or without suicide ideation (SA group; $n = 152$). Further, within the SA group there were 107 inpatients who had a suicide attempt event occur within the last 30 days of their recordings. An ample portion of the inpatients with a history of suicide ideation or attempt (91%) had a Beck Depression Inventory-II (BDI-II) score of ≥ 13, which is indicative of a '*mild*' to '*severe*' depression (Beck et al., 1996). Note that 10 inpatients (4 SA; 6 SI) did not have their Beck Depression

Table 1
In the BBRS database, of the total 246 participants, approximately 47% were females and 53% were males. The average age per inpatient group: SA (39 years), SI (41 years), and HC (34 years). For the depression group allocation purposes, any participant that had a BDI-II score $\geq$13 was considered 'depressed' and $\leq$12 'non-depressed'.

Group	Non-Depressed	Depressed	Total
Suicide Attempt (SA)	4	142	152
Suicide Ideation (SI)	7	63	74
Healthy Control (HC)	20	–	20

Inventory-II (BDI-II) scores recorded.

Participants were instructed to read aloud unpracticed sentences from a computer screen (refer to Table 2 for text). Once participants completed an entire sentence, they pressed the keyboard spacebar to proceed to the next sentence, which was selected at random until all twenty-one sentences were read aloud. To improve elicitation protocol familiarity and help confirm that the instructions were understood, all participants were given two neutral practice sentences at the beginning of their recording sessions (e.g., '*There are twelve months in a year*' and '*A rose is a type of flower*'). All recordings were conducted using condenser microphone and digital recording software with a 16-bit 44.1 kHz sampling rate frequency. The read sentence audio recordings ranged from approximately 2 to 7 s in length (i.e., ceiling of task).

Fig. 1 shows that the BBRS database participant age ranges were wider for the SI and SA inpatients than the healthy controls (HC). However, the median age for all three participant groups was approximately 39 years old. In Fig. 1, although the SI group had the largest BDI-II score range, the SA group contained inpatients with the most severe BDI-II average scores when compared with the other groups.

In Table 2, the BBRS database recording protocol contained sentence stimuli with five different mood category types. These sentences were chosen based on their referential first-person viewpoint (e.g., *I, me, my*) and different mood content. In studies by Brierley et al. (2007), Cummins et al. (2011), Jiang et al. (2017), Lawson et al. (1999), and Stasak et al. (2019), read aloud sentences containing emotionally charged keywords have been previously used to evaluate abnormal behaviors found in patients exhibiting clinical mood disorders. Using the Flesch-Kincaid Grade Level test (DuBay, 2006), the BBRS twenty-one protocol sentences averaged a readability score of 2.9 (i.e., roughly a 3rd grade reading comprehension); therefore, these sentences had a low-level reading skill demand. BBRS participants were not tested for reading or visual disorders, which might have marginally impacted some inpatients' reading abilities.

3. Methods

3.1. Voice quality assessment measures

Our subjective voice quality scale was based on the GRBASI voice quality evaluation (Yamauchi et al., 2010). The GRBASI perceptual evaluation scale is one of the most common assessments for pathological voice quality. The GRBASI is based on a four-point scale (i.e., 0-normal, 1-mild, 2-moderate, 3-severe), and requires a human listener to subjectively score six voice quality attributes: 'grade' (hoarseness), 'roughness' (vibration irregularity), 'breathiness' (air escaping), 'asthenia' (weakness), 'strain' (hyper-functional phonation), and 'instability' (inconsistent overall quality). Adding the individual voice quality attribute scores generates a voice quality GRBASI composite score. A higher composite score indicates a poorer overall voice quality.

The GRBASI evaluation is subjectively administered using a short text (i.e., read passage of approximately twenty sentences) and sustained vowel. In our study, BBRS database participant voice quality was manually assessed based on the twenty-one read sentence tasks (see: Table 2) by an experienced human annotator with a background in speech-language pathology. During the voice quality evaluation process, the annotator wore headphones and was unaware of participants' BDI-II severities and group allocations to help reduce annotator bias.

Many studies (Bele et al., 2005; Couch et al., 2015; Dejonckere et al., 1996; Lechien et al., 2018; Yamaguchi et al., 2003) have shown that the GRBASI perceptual evaluation has good interrater reliability even with a small number of annotators, especially when assessing sentence-level speech utterances. Furthermore, in the clinical and research field, subjective perceptual voice assessment is the 'gold standard' even when compared to potential automatic objective approaches (Barsties and Bodt, 2015). In Bele et al. (2005), it was demonstrated that experienced annotators (e.g., speech pathologist, voice experts) were more consistent with their GRBASI ratings than unfamiliar annotators. However, it should be noted that this difference between experienced and inexperienced annotators still produced acceptable GRBASI voice quality reliability (Bele, 2005).

3.2. Disfluency feature extraction

As with the voice quality assessment measures, similarly to Stasak et al. (2019), participants' read sentences were individually evaluated using an experienced annotator with a background in speech language pathology. During the speech disfluency annotation process, the annotator wore headphones and was again unaware of participants' BDI-II severities and group allocations. Sentence-level annotations provided detailed information regarding what kinds of speech disfluencies occurred: (1) hesitations; (2) word-repeats; (3) phrase-repeats; (4) speech errors; (5) malapropisms; and (6) uncorrected speech error.

A hesitation was defined as any unnatural abrupt pause, false start, word/phrase repeat, or abnormal prolongation of an utterance. For hesitations, a participant's rate of speech across all twenty-one sentences was taken into consideration, as some participants were slower readers than others. In terms of hesitations, in the form of pauses or word prolongations, the degree of abruptness and speaker inconsistency was a considerable factor. Thus, hesitations were subjectively judged based on a participant's idiosyncratic speech pattern behaviors over the course of the entire set of read sentences.

A speech error was defined as any deviation in a pronunciation from the intended read target word, such as phonological deletions, substitutions, or slips of the tongue. Each word/phrase repeat was also individually recorded per sentence. In addition, any malapropisms (i.e., unrelated substitution for a word; see: Fay and Cutler, 1977) or speech errors that went uncorrected (i.e., failed to make a verbal correction) were recorded per sentence. During the manual disfluency annotations,

Table 2
Categorical list of read aloud elicitation protocol sentences contained in the BBRS database. The twenty-one declarative sentences contained 113 words, 72 of which were unique. There was an average of 5.38 words per sentence with a range of 3 to 8 words. The majority of BBRS sentences {1–17} contained direct self-referential pronouns (i.e. *I, me, my*), which have been shown to evoke greater affective attachment than indirect non-referential text (Frewen and Lundberg, 2011; Salem et al., 2017).

Mood Categories	Sentences	Mood Categories	Sentences
Suicidal	{1} *I wish I were dead*	Positive	{11} *I am a valuable person*
	{2} *I have thought about killing myself*		{12} *People like me*
	{3} *I have a plan to kill myself*		{13} *The things I do are rewarding and enjoyable*
	{4} *I intend to end my life*		{14} *I have fulfilling friendships*
	{5} *I know exactly how I will kill myself*	Negative	{15} *I can't do anything right*
Life Orientation	{6} *I want to live*		{16} *Nothing seems to work out for me*
	{7} *I have reasons for living*		{17} *Nothing good ever happens to me*
	{8} *My life has purpose*	Neutral/Factual	{18} *The grass is green*
	{9} *I value most of life's experiences*		{19} *The sky is blue*
	{10} *I enjoy life*		{20} *There are twelve eggs to a dozen*
			{21} *The sun rises in the east*

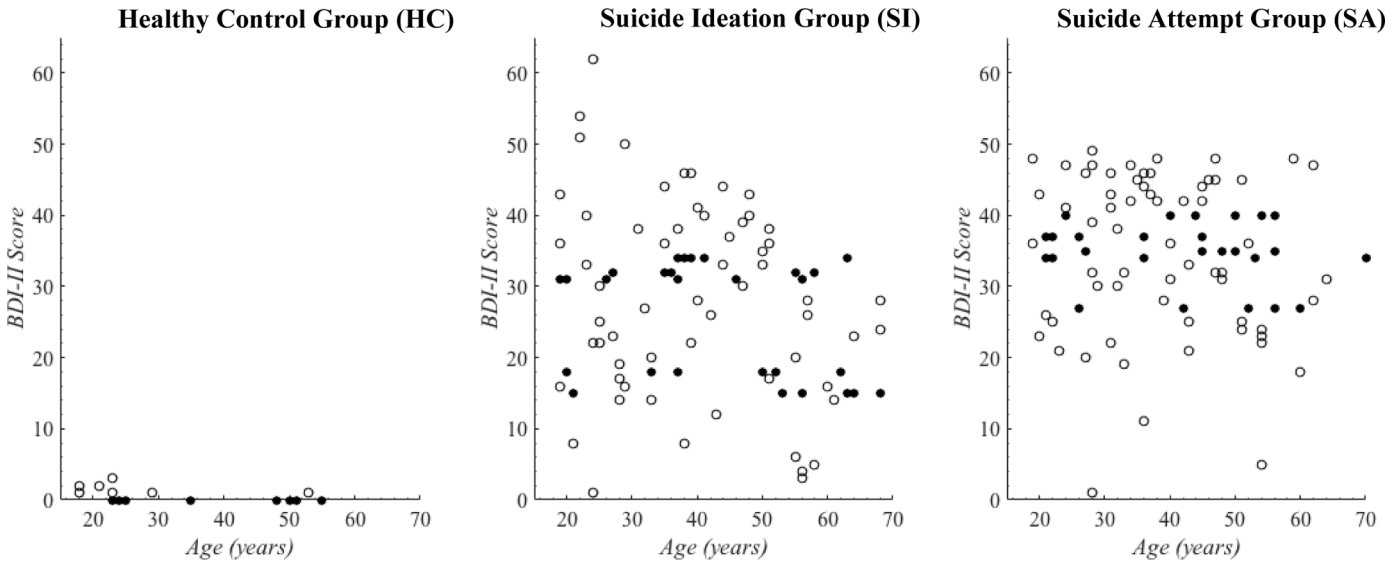

Fig. 1. BBRS database age and BDI-II scores by participant groups. The filled circles indicate ≥5 participants (i.e. higher density), whereas non-filled circles indicate ≤4 participants (e.g. lower density). The average BDI-II score per participant group was: HC (0.68), SI (26.58), and SA (35.90). The BDI-II diagnosis label score ranges are defined as 'minimal' (0–13), 'mild' (14–19), 'moderate' (20–29), and 'severe' (30–63).

regional and foreign accents were taken into consideration. For instance, some of the participants had a distinct regional accent; thus, some phonemes such as */r/* (e.g., *brother* → *bwotha*) and */dz/* (e.g., *jersey* → *djwesy*) were dialectically different from a more generalized American-English accent. In these instances, dialectal norms were considered acceptable so long as the phonemic idiosyncrasies were consistent across *all* sentences.

Surprisingly, many participants used inconsistent word pronunciations, wherein for one sentence a word would be spoken improperly, and then later in another sentence, the same word would be spoken properly. These inconsistencies were recorded as speech errors. Frequent examples of these kinds of pronunciation inconsistencies found were *want to/wanna, there/der*, and *the/da*. Naturally, during the investigation described herein, human-perceptual insights are a major advantage to manual disfluency tracking when compared with automatic tracking methods that would require more sensitive and sophisticated speaker normalization algorithms.

Each of the six disfluency types was individually totaled for all sentences {1–21} to create a set of speech disfluency features per participant. Based on Table 2, speech disfluency types were also selectively summed to create mood-specific category speech disfluency feature sets: suicidal {1–5}, life orientation {6–10}, positive {11–14}, negative {15–17}, and neutral/factual {18–21}. Each of the mood-specific category speech disfluency features were later concatenated into a single fused feature comprised of the five different category type speech disfluency features.

3.3. System configuration

Fig. 2 shows the block diagram for the experiments herein. As previously mentioned in Section 3.2, each of the six speech disfluency features could be calculated using all the sentences {1–21} or by the defined mood category sentence types shown previously in Section 2 Table 2. Furthermore, the speech disfluency features could also include all the sentences {1–21} and also mood-specific categories. While the front-end of our system relied on manual annotations for voice quality and speech disfluencies, the back-end of our system was entirely automated using statistical machine learning methods. To establish the utility of voice quality and speech disfluency features for suicidal classification, human annotations were essential for experimental 'ground truth'.

In Fig. 2, the addition of an automatic speech recognition (ASR) could help to automate the disfluency tracking rather than a manual approach. However, ASR systems are generally not designed to accurately record speech disfluencies, especially false starts, repairs, repetitions, filled pauses, and punctuation (Liu et al., 2006). Therefore, specifically designed speech models (i.e., preferably including individuals with depression) along with human-annotated dysfluency transcripts would be necessary so that an ASR system could be trained to adequately detect disfluencies (i.e., record speech as it was accurately spoken). Furthermore, in Fig. 2, rather than relying on manual voice quality annotations, the automatic feature extraction could include automatic GRBASI voice quality measurements. However, according to Barsties and Bodt (2015), automatic objective-acoustic voice quality

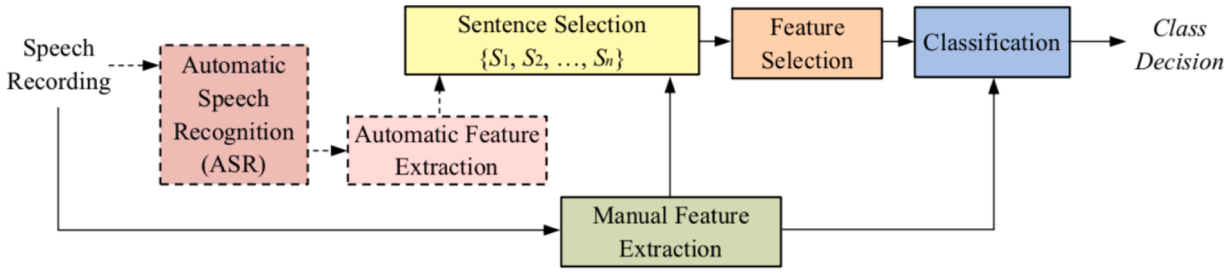

Fig. 2. System configuration, with solid lines indicating exploratory manual methods used herein and the dashed lines showing a suggested automatic approach. 'Manual Feature Extraction' includes the voice quality and speech disfluency annotation process. 'Sentence Selection' refers to calculating the disfluency features based on mood-specific category types (e.g., suicidal, life orientation, positive, negative, neutral/factual).

analysis is currently limited, and multi-parametric automated methods still show large variability in assessing qualities, such as roughness and breathiness qualities. Moreover, although objective-acoustic analysis techniques hold promise, voice quality is a multi-dimensional perceived construct; thus, its performance is typically compared and correlated with subjective acoustic voice quality ratings (Barsties and Bodt, 2015).

3.4. Classification and evaluation metrics

All experiments included the BBRS database described previously in Section 2. Classification experiments utilized a balanced set of class representations per speaker dependent train/test fold; and per fold, a 5-fold cross validation using an 80–20 train/test split was used to further reduce overfitting. Each fold consisted of 20 healthy controls (HC group) and 20 psychiatric inpatients (SI and SA groups). The balanced classes helped to maximize available healthy control data while also curtailing class-weighted bias. With each speaker dependent fold, the HC group remained the same (because there were only 20 HC participants in total), whereas the inpatient groups, SA and SI, were iteratively replaced with another set until all inpatients were evaluated. During a cross-fold evaluations, no participant was ever simultaneously in both train and test partitions. Moreover, each fold contained the same speakers per different classification experiments. In addition, these folds were specifically designed and statistically evaluated using a Kruskal-Wallis test ($p = 0.01$) make sure each had relatively similar age, gender, and depression score representations.

While new computational advances in machine learning, such as deep neural networks (LeCun et al., 2015) and other similar methods (e.g., recurrent neural networks, convolutional neural networks), have become more prevalent in automatic classification tasks, these approaches require very large amounts of data to train on to avoid overfitting. Consequently, until larger depression/suicide-related speech databases become more readily available, the small number of current database resources and deep learning requirements is a limitation (Li et al., 2019). Pampouchidou et al. (2017) conducted a literature review of over sixty automatic depression analysis studies and found that only a few of these publications utilized deep learning methods.

Automatic 2-class classification was conducted using linear support vector machine (SVM) (Cortes and Vapnik, 1995) and non-linear cosine *k*-nearest neighbor (kNN) (Kataria and Singh, 2013; Qamar et al., 2008) methods. In previous studies (Islam et al., 2018; Jiang et al., 2017; Meftah et al., 2012; Schuller & Batliner, 2014; Valstar et al., 2013), these classification methods have been used for automatic speech-based suicide/depression identification. All experiments found in Section 4.3 used MATLAB machine learning software. For classification experiments herein, performance was determined using the average accuracy and individual class F1 scores across all train/test folds (similarly to Valstar et al. (2016)).

4. Results and discussion

4.1. Voice quality

An individual voice quality attribute analysis per group is shown below in Table 3. While some of the automatic speech-based literature (Cummins et al., 2015; Scherer et al., 2013) has mentioned individuals exhibiting depression and/or suicidal behavior having increased 'breathiness', our manual evaluation on the BBRS database analysis found that inpatients with suicide attempt history had considerably higher 'grade' and 'roughness' attributes. For instance, both the SI and SA inpatient groups had vocal quality 'grade' score averages (~0.30) that were six times greater than those of the HC group (0.05). Moreover, for the vocal quality 'roughness' score average, the SI and SA inpatient groups attained average scores four to nearly five times higher than the HC group (0.10).

A further investigation of the GRBASI voice qualities based on gender revealed that the female SI and SA groups had higher 'grade' (≥0.53) and 'roughness' (≥0.51) score averages than the male SI and SA groups. An analysis of inpatients that were non-depressed (ND) versus depressed (D) showed that 'roughness' and 'breathiness' is more commonly found in the ND group (0.63, 0.38) than D group (0.43, 0.12).

As hypothesized in Section 1, the GRBASI composite voice quality score average was significantly greater for the SI (1.25) and SA (1.07) groups when compared with the HC group (0.40). The individual 'grade' and 'roughness' scores contributed the most to the SI and SA inpatients' increased GRBASI composite score. Shown in Fig. 3(a), 75% of the HC group was given a GRBASI composite score of 0, whereas only 44% of the SI and SA groups attained a 0. Furthermore, in Fig. 3(a), SI and SA groups were twice or more likely to have a GRBASI composite score of 1, 2, and 3 or more than the HC group. The comparative group analysis of SI and SA shown in Fig. 3(b) indicated that the SA group had a larger percentage (6% absolute) that had a GRBASI composite score above 0 when compared with the SI group.

Shown in Fig. 4, an examination of GRBASI composite score averages

Table 3
GRBASI individual voice quality attribute score averages per participant group. All group types included both female and male participants; the number of participants per group is shown in parentheses. In addition, independent two sample *t*-tests were conducted by comparing results from the healthy control (HC) group paired with those of the suicidal ideation (SI), suicide attempt (SA), non-depressed inpatient (ND), and depressed inpatient (D) groups; *t*-test differences in statistical significance are indicated by $*p = 0.05$ and $**p = 0.01$. Cohen's *d* analysis of HC-SI and HC-SA paired groups typically demonstrated 'medium' to 'large' effect size differences for grade (0.55, 0.60), roughness (0.55, 0.84), and strain (0.37, 0.56) voice quality averages. While an approximate 'medium' (0.45) Cohen's *d* effect size was recorded for the HC-SI GRBASI voice quality composite, an approximate 'large' (0.78) Cohen's *d* effect size was recorded for HC-SA paired groups. Furthermore, using Hedges' *g* (i.e., more appropriate for samples ≤20), ND-D psychiatric impatient paired group analysis showed 'small' effect size differences for roughness (0.20), breathiness (0.20), asthenia (0.20), strain (0.34), and instability (0.21). Due to audio recording quality issues, there were 25 participants that did not have a GRBASI voice quality composite score.

All Group Types	Grade	Roughness	Breathiness	Asthenia	Strain	Instability
HC (20)	0.05**	0.10**	0.15**	0.05**	0.00**	0.05**
SI (107)	0.32**	0.48**	0.15**	0.04**	0.11**	0.10**
SA (119)	0.30**	0.40**	0.11**	0.06**	0.14**	0.06**
Female						
HC (12)	0.00**	0.08**	0.25**	0.08**	0.00**	0.08**
SI (40)	0.53**	0.69**	0.28**	0.14**	0.14**	0.14**
SA (62)	0.55**	0.51**	0.15**	0.02**	0.11**	0.04**
Male						
HC (8)	0.13**	0.13**	0.00**	0.00**	0.00**	0.00**
SI (67)	0.13**	0.23**	0.08**	0.00**	0.05**	0.05**
SA (57)	0.10**	0.43**	0.06**	0.10**	0.22**	0.12**
Psychiatric Inpatients						
ND (5)	0.25**	0.63**	0.38**	0.00**	0.00**	0.13**
D (211)	0.31**	0.43**	0.12**	0.06**	0.13**	0.08**

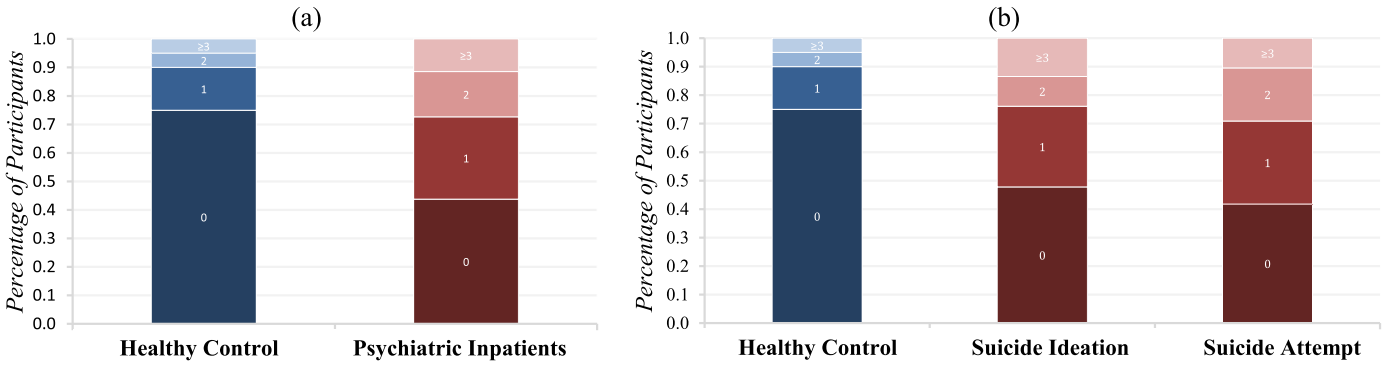

Fig. 3. Comparison of GRBASI composite score distributions (i.e., all voice quality type scores combined) per participant group: (a) healthy control vs. psychiatric inpatients and (b) HC vs. SI vs. SA. The four different color shades represent the GRBASI composite score severity: '0' (dark), '1' (mid-dark), '2' (mid-light), and '≥3' (light).

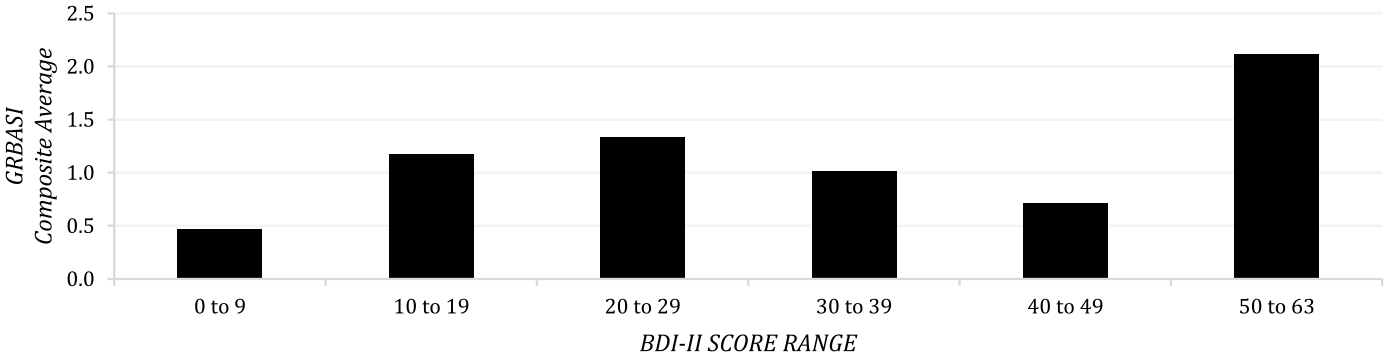

Fig. 4. Comparison of GRBASI composite distributions per participant BDI-II score incremental ranges with number of participants shown in parenthesis: 0–9 (28); 10–19 (23); 20–29 (42); 30–39 (63); 40–49 (49); and 50–63 (9). Due to audio recording quality issues, there were 25 participants that did not have a GRBASI voice quality composite score. Additionally, 8 participants were missing a BDI-II score.

based on low to severe depression scores (i.e., based on BDI-II inpatient metadata) demonstrated that the participants with the highest BDI-II severities also received the highest GRBASI composite score average. Moreover, participants in the 50–63 BDI-II range had a GRBASI composite score average four times higher than that of the participants in the 0–9 BDI-II range. By clustering the participants based on the BDI-II score label scale, results showed that the group with '*moderate*' depression severity attained the highest GRBASI composite score average (1.37). However, participants in the '*mild*' (1.24) and '*severe*' (1.08) BDI-II score ranges were not far behind. Based on results in Fig. 4, a higher GRBASI voice quality score is a strong probable indicator for inpatients with '*mild*' to '*severe*' BDI-II severity ranges.

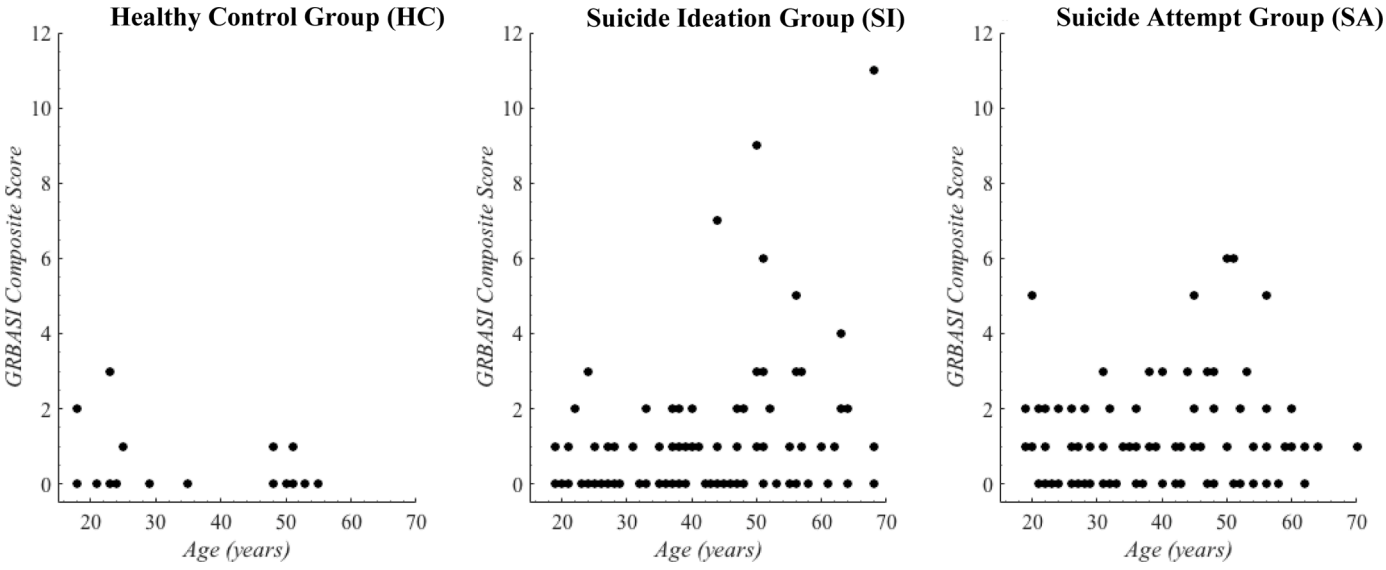

Fig. 5. Comparison of age and GRBASI composite scores distributions per participant groups. The average age per participant group: HC (34 years), SI (41 years), and SA (39 years).

Fig. 5 shows that for SI and SA groups, more than half of patients in these two groups attained ≥0 GRBASI composite scores. Surprisingly, in Fig. 6, for SI and SA groups, there are a considerable number of young inpatients (e.g., 18–30 years old) that attained a GRBASI composite score ≥1 when compared with the HC group. Generally, voice disorder prevalence is higher in elderly adults than younger adults (de Araújo Pernambuco et al., 2014; Roy et al., 2007). The average GRBASI voice quality scores discovered in SI and SA groups herein provide further human-perceptual evidence of increased association with previous vocal apparatus damage, abuse, and/or current psychomotor agitation/retardation vocal effect due symptoms of depression (Costache et al., 2004; Wazeer et al., 2015).

In Tables 4 and 5, 2-class classification accuracy and F1 score results are shown using the GRBASI voice quality composite score feature and automatic SVM/kNN machine learning techniques. Tables 4 and 5 results indicate that, for depression classification, the 1-dimensional (62%) GRBASI voice quality composite score feature and 7-dimensional (64%) GRBASI voice quality type score features attained relatively low depression classification accuracy. However, it should be remembered that our labels for depressed participants included individuals with '*mild*' depression ranges (e.g., 14–19 BDI-II scores), which are generally more challenging to classify than per se '*severe*' ranges (e.g., 30–63 BDI-II scores).

Classification accuracy using the GRBASI features for the SA group was considerably higher than for the SI group. For instance, using the 7-dimensional GRBASI voice quality attribute score features, SA classification accuracy achieved up to 73% with the cosine kNN classifier technique, whereas for the SI classification accuracy attained only 59% using the same classifier. To our knowledge, these results are the first to demonstrate that manually annotated voice quality score ratings can be applied as a novel low-dimensional feature to help automatically distinguish healthy individuals from inpatients with suicide ideation or behavior. Voice quality experiments herein demonstrate that any prior inpatient history of attempted suicide results in increased GRBASI voice quality composite scores.

4.2. Speech disfluency

As shown in Figs. 6 and 7, the SI and SA groups had more hesitations in the form of pauses and syllable word prolongations than the HC group. Similarly to patients with clinical depression evaluated in Stasak et al. (2019), the BBRS inpatients examined herein, a majority with depression, also had a significantly more speech errors while reading aloud when compared with the HC group. In Fig. 6, the SI and SA groups averaged three to four times as many speech errors for the twenty-one sentences than the HC group. Furthermore, the SI and SA groups had more average occurrences of malapropisms (~0.40) along with average uncorrected speech errors (~1.0) when compared with the HC group (0.10, 0.15).

Table 4
SVM and kNN classification accuracy results including the F1 score performance using the 1-dimensional GRBASI voice quality composite score feature.

2-Class Classification	*Linear SVM* Accuracy	F1 Scores		*Cosine kNN* Accuracy	F1 Scores	
D vs. (ND)	62%	*0.70*	*(0.43)*	57%	*0.67*	*(0.40)*
SI vs. (HC)	54%	*0.35*	*(0.65)*	61%	*0.54*	*(0.66)*
SA vs. (HC)	69%*	*0.64*	*(0.72)*	69%*	*0.65*	*(0.72)*

Table 5
SVM and kNN classification accuracy results including the F1 score performance using the 7-dimensional GRBASI voice quality score attribute scores plus total composite score features described previously in Section 3.1.

2-Class Classification	*Linear SVM* Accuracy	F1 Scores		*Cosine kNN* Accuracy	F1 Scores	
D vs. (ND)	64%**	*0.73*	*(0.43)*	61%**	*0.70*	*(0.46)*
SI vs. (HC)	54%	*0.35*	*(0.64)*	59%	*0.45*	*(0.67)*
SA vs. (HC)	69%**	*0.61*	*(0.74)*	73%**	*0.69*	*(0.76)*

Previously, in Stasak et al. (2019), only a relatively small difference (7% absolute) in non-self-corrections was found between depressed and healthy groups; with the depressed inpatient group making greater effort to self-correct speech errors. However, results herein indicated that SA and SI groups were more likely to attempt a self-correction than the HC group. For instance, the HC group failed to self-correct a speech error an average of 67%, whereas the SI and SA groups failed to self-correct at a lower rate of 45%. This average self-correction attempt percentage calculation was derived from the average total number of speech errors uncorrected divided by the average total number of speech errors and calculating the remaining average percentage per participant group.

We observed more distinction between disfluencies produced by the SI group (i.e., history of no suicide attempts) and SA group (i.e., lifetime suicide attempt history). For example, in Fig. 6, the SA group with a lifetime suicide attempt history had a greater average number of hesitations (0.61 absolute gain), word repeats (0.13 absolute gain), phrase repeats (0.12 absolute gain), and speech errors (0.55 absolute gain) than the SI group. These results show that *any* prior inpatient history of attempted suicide results in generally higher rates of speech disfluencies.

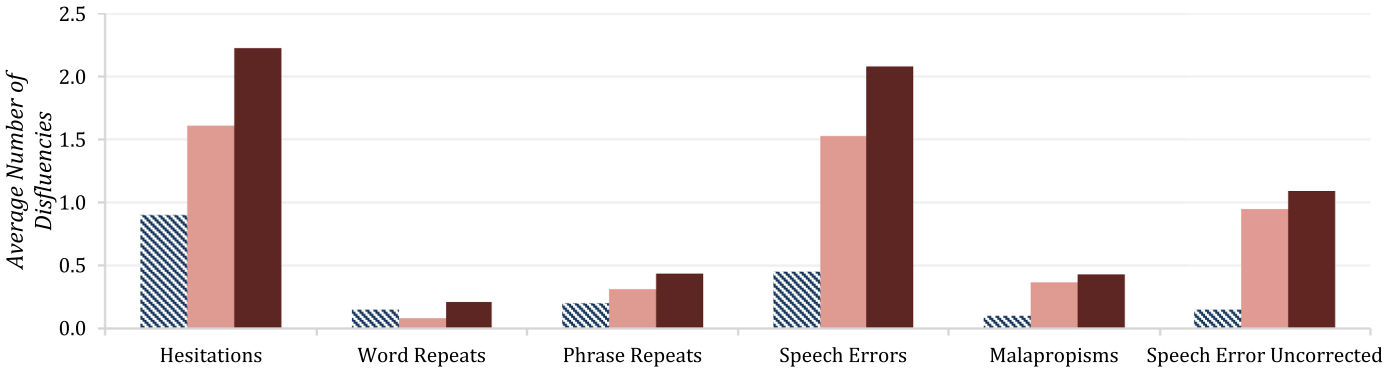

Fig. 6. Comparison of HC (diagonal shade), SI (light shade), and SA (dark shade) participant group average disfluencies for all 21 sentences in the BBRS database. The SA group consists only of inpatients that had at least one suicide attempt during any point during his/her lifetime. Statistical significance differences were recorded using the Kruskal-Wallis test ($p = 0.01$) for the SI and SA groups when compared to the HC group for hesitations, speech errors, and speech error uncorrected disfluencies.

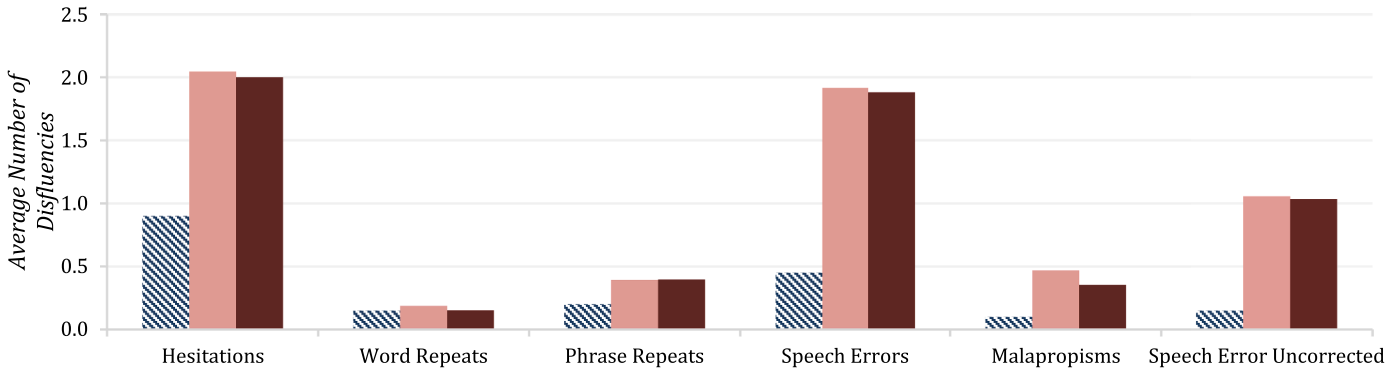

Fig. 7. Comparison of HC (diagonal shade), SI (light shade), and SA (dark shade) participant group average disfluencies for all 21 sentences in the BBRS database. The SA group consists of inpatients that had one or more suicide attempts only within the last 30 days prior to the recordings. Statistical significance differences were recorded using the Kruskal-Wallis test ($p = 0.01$) for the SI and SA groups when compared to the HC group for hesitations, speech errors, and speech error uncorrected disfluencies.

In Fig. 7, only inpatients who had attempted suicide within the last 30-days from the time of the recordings were included in the SA group. Fig. 7 reveals that by grouping inpatients that are currently experiencing suicidal ideation along with other inpatients who had an attempted suicide history outside of the 30-day interval, it generated speech disfluency average results similar to SA group who have made a suicide attempt within the previous 30-days. Figs. 6 and 7 results indicate that inpatients with any attempted suicide history should be collectively analyzed despite the time period in which the suicide attempt occurred. This enables more speech disfluency feature delineation between those with suicide ideation versus suicide attempt (i.e., as shown previously in Fig. 6).

Notably, approximately a quarter of the SA and SI groups produced 3 to 7 speech errors for the twenty-one sentences, whereas the HC group as a whole had less than 1 average speech error. As for the malapropisms, roughly half of the patients in the SA and SI groups produced an average of at least one malapropism (1.0), whereas for individuals in the HC group less than one-sixth (0.15) produced a malapropism. Examples of the malapropisms recorded were verb-tense or preposition substitutions, such as: *was/were, in/to, is/has*, and *would/will*. However, as shown in Table 6, many recorded malapropisms were far less confusable in nature. Malapropism speech errors are rare even during free conversational speech (Cowie, 1985; Fay and Cutler, 1977); therefore, it is unusual to record numerous malapropisms during a small set of text-dependent reading tasks in the BBRS database.

Based on previous clinical depression studies (Hartlage et al., 1993; Goeleven et al., 2006; Gotlib & McCann, 1984; Levens and Gotlib, 2015; Mitterschiffthaler et al., 2008; Silberman et al., 1983; Roy-Byrne et al., 1986; Rubino et al., 2011; Weingartner et al., 1981), it has been hypothesized that individuals exhibiting suicidal behavior also have difficulty with reading tasks due to poorer information retrieval and concentration ability than healthy populations.

In Table 7, an investigation of non-depressed (ND) and depressed (D) inpatient group average disfluencies demonstrated that the D group had a greater frequency of hesitations, speech errors, and malapropisms. For A Shapiro-Wilk normality test was conducted for individual group disfluency types shown in Table 7. The normality test analysis, using a $p = 0.01$, indicated that for each group disfluency type, the data was normal with unspecified mean and variance. Results in Table 7, it appears that ND inpatients have a lower average rate of hesitations when compared with D patients (an absolute difference of 0.74). Thus, the degree of hesitations is a good indicator to help diagnostically differentiate inpatients with suicidal behavior that are non-depressed from those with suicidal behavior that are depressed. However, it should be noted that our comparative sample size of psychiatric inpatients was limited (e.g., ND = 5 and $D = 211$). Therefore, these results should be interpreted with caution until further investigation is conducted.

Further analyses of the speech disfluencies per participant group were evaluated by examining mood-specific categories from the BBRS sentences (see Table 1). Speech disfluency averages, shown in Table 8, summarize how sentence mood-content influenced the elicitation of certain kinds of speech disfluencies within different diagnosed inpatient populations (e.g., ideation/attempt). For the HC group, their hesitations, speech errors, malapropism, and speech error uncorrected averages were much narrower in range (0.01 to 0.05) than those of the SI and SA groups (0.00 to 0.18). This indicates that the HC group's degree disfluencies are more evenly spread across different kinds of mood category types when compared with those of the SI and SA groups.

In Table 8, a greater number of hesitations were recorded during reading of 'positive' mood category sentences for SI (0.17) and SA (0.18) groups than the HC (0.04) group. In addition, the 'positive' mood category sentences produced a greater number of speech errors for SI and SA groups (~0.12) than for the HC group (0.01). Again, according to Table 8, the bulk of malapropisms occurred during 'neutral/factual' mood category sentences rather than other mood types. In addition, Table 9 indicates that for SI and SA groups, generally their efforts to self-correct speech errors were similar (~0.04) for most mood category types.

As described previously in Section 3.3, experiments herein evaluated the classification accuracy using the core speech disfluency features (e.g., hesitations, speech errors, uncorrected speech errors) and also the impact of different mood category types (refer to Section 2, Table 2). Table 9 contains SI and HC classification results using the core speech disfluency features. Using core speech disfluency features derived from *all* sentences (73%) or fusion approaches (72%) attained higher classification accuracy than any single mood sentence type. Among the individual mood sentence types, the 'positive' sentences provided the highest classification accuracy (70%) between HC and SI groups, whereas the 'negative' produced the worst (53%). It should be noted that there were less examples of the 'negative' sentences for analysis than other mood categories.

Table 10 shows SA and HC classification results using the core speech disfluency features. Again, using core speech disfluency features derived

Table 6
Examples of malapropisms recorded in the BBRS dataset; the correct target words are shown in bold and total number of occurrences is shown in parentheses. Many of these malapropism speech errors occurred more than once among different participants.

intend (9) – tend	sun (3) – science, sum	twelve (1) – thousand
valuable (5) – vulnerable	rewarding (2) – regarding	rises (1) – raises
plan (4) – plane, pain	east (1) – dark	seasons (1) – seizing
friendships (3) – relationships	life (1) – long	they (1) – the

Table 7
Individual speech disfluency feature averages per participant group. The SA group consists of inpatients that had one or more suicide attempts within the last 30 days prior to the recordings. In addition, independent two sample *t*-tests were conducted by comparing feature averages from the healthy control (HC) group to each of the suicide ideation (SI), suicide attempt (SA), non-depressed inpatient (ND), and depressed inpatient (D) groups. A *t*-test statistical significance is indicated by $*p = 0.05$ and $**p = 0.01$. Cohen's *d* analysis on HC-SI and HC-SA paired groups typically demonstrated 'medium' to 'large' effect size differences for hesitation all (0.50, 0.86), speech error all (0.81, 0.94), speech error malapropisms (0.52, 0.56), and speech error uncorrected (0.82, 0.87). Furthermore, using Hedges' *g* (i.e., more appropriate for samples ≤20), ND-D psychiatric impatient paired group analysis indicated an approximate 'medium' effect size for hesitation all (0.35) and hesitation word-level (0.44).

	Hesitation	Word Repeat	Phrase Repeat	Speech Error	Malapropism	Speech Error Uncorrected
All Group Types						
HC (20)	0.90**	0.15**	0.20**	0.45**	0.10**	0.15**
SI (107)	2.05**	0.19**	0.39**	1.92**	0.47**	1.06**
SA (119)	2.00**	0.15**	0.39**	1.88**	0.35**	1.03**
Psychiatric Inpatients						
ND (5)	1.36**	0.00**	0.36**	1.82**	0.27**	1.09**
D (211)	2.10**	0.20**	0.40**	1.99**	0.40**	1.05**

Table 8
Average percentage disfluencies for all 21 sentences per participant group: HC, SI, and SA. This average was calculated by averaging the raw count across within group for specific mood category type sentence partitions, as indicated by {}. The word and phrase repeat speech disfluency averages are not shown, as these values had minimal deviation (≤0.03) per group. Values in bold indicate larger deviations from the HC group.

	Hesitations			Speech Errors			Malapropisms			SE Uncorrected		
Mood Category Types	HC	SI	SA	HC	SI	SA	HC	SI	SA	HC	SI	SA
Suicide {1–5}	0.03	0.06	0.10	0.04	0.07	0.11	0.02	0.02	0.03	0.01	0.05	0.07
Life Orientation {6–10}	0.05	0.07	0.10	0.00	0.08	0.10	0.00	0.01	0.01	0.00	0.04	0.05
Positive {11–14}	0.04	0.17	0.18	0.01	0.11	0.12	0.00	0.01	0.02	0.00	0.04	0.04
Negative {15–17}	0.05	0.04	0.06	0.03	0.03	0.07	0.00	0.00	0.00	0.02	0.03	0.05
Neutral/Factual {18–21}	0.01	0.04	0.08	0.03	0.07	0.09	0.00	0.06	0.05	0.01	0.05	0.05

Table 9
Average accuracy and F1 scores for SI versus (HC) binary classification for the BBRS database using core speech disfluency features (e.g., hesitations, speech errors, uncorrected speech errors). Results using only specific mood category type sentences are shown in {}. Both classifier results were derived using the identical sets of participant folds.

	Linear SVM		Cosine kNN	
Mood Category Types	Accuracy	F1 Scores	Accuracy	F1 Scores
All {1–21}	71%**	*0.69 (0.74)*	73%**	*0.71 (0.74)*
Suicidal {1–5}	56%	*0.51 (0.59)*	60%	*0.51 (0.65)*
Life Orientation {6–10}	62%*	*0.58 (0.66)*	65%*	*0.63 (0.68)*
Positive {11–14}	65%	*0.59 (0.70)*	70%**	*0.67 (0.73)*
Negative {15–17}	53%	*0.47 (0.57)*	59%	*0.49 (0.66)*
Neutral/Factual {18–21}	60%**	*0.51 (0.67)*	68%**	*0.65 (0.71)*
Fusion	72%**	*0.69 (0.74)*	68%**	*0.66 (0.70)*
All + Fusion	65%*	*0.59 (0.69)*	66%**	*0.63 (0.69)*

Table 10
Average accuracy and F1 scores for SA versus (HC) binary classification for the BBRS database using core speech disfluency features (e.g., hesitations, speech errors, uncorrected speech errors). Results using only specific mood type sentences are shown in {}. Both classifier results were derived using the identical sets of participant folds.

	Linear SVM		Cosine kNN	
Mood Category Types	Accuracy	F1 Scores	Accuracy	F1 Scores
All {1–21}	78%**	*0.77 (0.79)*	77%**	*0.75 (0.79)*
Suicidal {1–5}	67%	*0.65 (0.69)*	74%**	*0.71 (0.76)*
Life Orientation {6–10}	70%	*0.65 (0.74)*	72%**	*0.68 (0.74)*
Positive {11–14}	73%	*0.69 (0.76)*	75%	*0.72 (0.76)*
Negative {15–17}	65%*	*0.58 (0.69)*	70%**	*0.66 (0.72)*
Neutral/Factual {18–21}	70%*	*0.63 (0.73)*	71%*	*0.66 (0.74)*
Fusion	74%**	*0.70 (0.76)*	78%**	*0.76 (0.79)*
All + Fusion	71%*	*0.67 (0.74)*	80%**	*0.77 (0.81)*

all sentences (78%) or fusion approaches (80%) attained higher classification accuracy than any single mood type. Among the individual mood types, the 'positive' sentences provided the highest HC and SA classification accuracy (75%), whereas the 'negative' produced the worst (65%).

In comparing results found previously in Table 9 and those found in Table 10, inpatients with a recent history of suicide attempt were easier to classify than inpatients exhibiting only suicidal ideation. Moreover, the fusion method worked better for the SA group (78%−80%) than SI group (66%−68%). As indicated previously in Section 2 in Fig. 1, the SA group had a higher BDI-II score average ('*severe*') than the SI group ('*moderate*'). These results agree with Schotte et al. (1997), where it was shown that melancholic symptoms (e.g., psychomotor disturbances) emerge as depression severity increases.

Table 11 contains SI and HC classification results using feature fusion comprised of the GRBASI voice quality composite and six speech disfluency features. Again, using core speech disfluency features derived from *all* sentences (69%) or fusion approaches (70%) attained higher classification accuracy than any single mood type. Among the individual mood types, the 'positive' sentences provided the highest HC and SI classification accuracy (67%), whereas the 'negative' produced the worst (54%).

Table 12 shows SA and HC classification results using the fusion features derived from the GRBASI voice quality composite and six

Table 11
Average accuracy and F1 scores for SI versus (HC) binary classification for the BBRS database using GRBASI voice quality composite (1) and speech disfluency (6) features. Results using specific mood types are shown in {}. Both classifier results were derived using the identical sets of participant folds.

	Linear SVM		Cosine kNN	
Mood Category Types	Accuracy	F1 Scores	Accuracy	F1 Scores
All {1–21}	69%**	*0.67 (0.70)*	68%**	*0.70 (0.65)*
Suicidal {1–5}	59%	*0.56 (0.63)*	61%**	*0.53 (0.66)*
Life Orientation {6–10}	66%**	*0.61 (0.70)*	68%**	*0.64 (0.72)*
Positive {11–14}	67%**	*0.60 (0.72)*	65%**	*0.62 (0.68)*
Negative {15–17}	54%	*0.46 (0.60)*	60%**	*0.52 (0.66)*
Neutral/Factual {18–21}	64%**	*0.56 (0.69)*	67%**	*0.63 (0.71)*
Fusion	66%**	*0.64 (0.68)*	69%**	*0.69 (0.68)*
All + Fusion	69%**	*0.66 (0.72)*	70%**	*0.70 (0.69)*

Table 12
Average accuracy and F1 scores for SA versus (HC) binary classification for the BBRS database using GRBASI voice quality composite (1) and speech disfluency (6) features. Results using only specific mood types are shown in {}. Both classifier results were derived using the identical sets of participant folds.

	Linear SVM		Cosine kNN	
Mood Category Types	Accuracy	F1 Scores	Accuracy	F1 Scores
All {1–21}	72%**	*0.70 (0.73)*	77%**	*0.75 (0.78)*
Suicidal {1–5}	72%**	*0.70 (0.74)*	74%**	*0.71 (0.76)*
Life Orientation {6–10}	70%**	*0.65 (0.72)*	67%**	*0.61 (0.71)*
Positive {11–14}	74%**	*0.71 (0.76)*	74%**	*0.70 (0.76)*
Negative {15–17}	66%**	*0.61 (0.70)*	71%**	*0.68 (0.72)*
Neutral/Factual {18–21}	70%**	*0.64 (0.74)*	72%**	*0.67 (0.74)*
Fusion	73%**	*0.68 (0.75)*	77%**	*0.74 (0.78)*
All + Fusion	72%**	*0.67 (0.75)*	77%**	*0.74 (0.79)*

speech disfluency features. Again, using core speech disfluency features derived *all* sentences (77%) or fusion approaches (77%) attained higher classification accuracy than any single mood type. Among the individual mood types, the 'positive' sentences provided the highest HC and SA classification accuracy (74%), whereas the 'negative' produced the worst (66%).

4.3. Limitations

We acknowledge that the participants included in the BBRS database were not pre-screened for potential medical illness comorbidity, medication (e.g., anti-depressants, anti-psychotics) and/or drug (e.g., alcohol, illicit drugs) use during recordings. It is recognized that some disorders/diseases and medications/drugs can cause abnormal changes in cognitive processing and speech behaviors. However, we highlight that the BBRS database was an authentic representation of health controls and also inpatients seeking support in a clinical environment; and a situation wherein inpatient comorbidity and medication/drug use is often widespread (Brook et al., 2002; Henriksson et al., 1993; Tondo et al., 1999).

We also recognize that our manual annotation voice quality and speech disfluencies assessments did not implement multiple evaluators. Due to costs and specialized training we were unable to procure multiple annotators. In future manual studies, it is recommended to have multiple annotators to reduce possible bias and human-error factors. Another limitation was the training size of the models, as 20 participants per class was the maximum due to the BBRS total available number of speakers in the healthy control group. It is believed that having a larger training size per class would benefit classification performance. We also note differences from the BBRS database herein to the dataset used in Stasak et al. (2019), which examined clinically depressed speakers during affectively charged read sentences. In the current study, the BBRS database read declarative sentences were not linguistically constructed using near-identical pairs (i.e. one word difference, similar lengths), mood label types were generalized (i.e., not derived from the Affective Norms for English Words ratings), and many read sentence stimuli contained both positive and negative words, such as {1} *wish/dead*, {4} *life/end*, and {4} *nothing/good* (i.e., possibly introducing a degree of mood uncertainty).

5. Conclusion

Our study examined voice quality and speech disfluency behaviors found in psychiatric inpatients with a suicide history and healthy controls with no suicide history in the BBRS database. When compared with a healthy control, an analysis of voice qualities in inpatients exhibiting suicidal ideation or behavior yielded new valuable insights, revealing that they have increased 'grade' and 'roughness' voice qualities (i.e., not only 'breathiness' as indicated in previous literature). Further, we discovered that even for non-depressed inpatient groups (i.e., BDI-II score $\leq$13), their voice quality scores were poorer than healthy controls.

In several instances, the inpatient group without depression had poorer voice quality averages than the inpatient group with depression (see Table 2). This discovery indicates that unlike psychomotor retardation/agitation symptomology often associated with clinical depression, inpatients with non-depressed suicidal behavior must have another reason for exhibiting poorer voice quality. Based on the increased average GRBASI score in the non-depressed SA group is believed that prior inpatient suicide attempts may have contributed to speech mechanism dysfunction; thus, resulting in an increase in poorer voice quality.

An evaluation of BBRS by gender also indicated that the females had higher average GRBASI composite. This result is consistent with literature (Martins et al., 2015; Roy et al., 2005; Piccinelli and Wilkinson, 2000), which notes that females are more likely to suffer from depression and/or voice disorders than males. Using common machine learning techniques, we demonstrated that the subjective GRBASI voice quality measures could be converted into a compact composite feature or a set of low-dimensional features – attaining HC and SA two-class classification accuracy up to 73% with F1 scores of 0.69 (0.76).

To our knowledge, this is the first large-scale study to use read aloud sentences to evaluate speech disfluencies exhibited by inpatients with suicidal ideation and behavior. We found that the increase in speech disfluencies was indicative of a more severe BDI-II score and a higher association with recent suicide attempt. Our speech disfluency analysis showed that SI and SA groups had approximately twice as many hesitations and four times as many speech errors than the HC group. Moreover, in examining the SA group (i.e., who made any suicide attempt during their lifetime) we discovered that this particular population had the greatest frequency of hesitations, speech errors, and malapropisms when compared to SI group (i.e., suicide ideation but no history of attempted suicide) and HC group.

While we recorded only a few minor gains using only disfluency features from specific-mood sentence categories, we noted that the 'negative' and 'suicidal' sentences categories consistently performed the worst for HC/SI and HC/SA classification, whereas the 'positive' sentences performed the best. In general, the HC and SI/SA group classification results presented herein demonstrates that a read speech protocol containing a wide variety of mood types (i.e., all 21 sentences or mood-type sentence fusion) provides enhanced speech disfluency feature classification performance over narrower single mood sentence types.

Using machine learning techniques and a set of three core disfluency features (i.e., voice quality composite, hesitation, speech error) derived from fusion of different mood-specific sentence types, HC versus SI group classification attained up to 73%; and HC versus SA group classification attained up to 80% accuracy. It is believed that the baseline features, which included all twenty-one sentences, resulted in strong HC/SI and HC/SA classification performance because these contained the maximum number of sentences encompassing all different mood types. Although the depression and suicidal type classification results presented herein rely on manual annotation, we note that human listening has important practical considerations, especially since careful listening is a vital part of any clinical interview/assessment protocol. The manually extracted voice quality and speech disfluency features provide an upper bound on the likely classification accuracy from a fully automated system. In future, we intend to investigate a system that utilizes automatic methods to extract voice quality and speech disfluency features.

Larsen et al. (2015) advocates that further efforts involving cross-discipline research will lead to new ways to automatically identify and monitor suicidality. Results gleaned from this study will help provide a new direction for future automatic speech-based suicide detection development. Further exploration into speech behaviors associated with suicide is needed to help uncover potential biomarkers, which can be automatically extracted and leveraged to help identify at-risk populations.

Declaration of Competing Interest

None.

Acknowledgements

The authors would like to thank Butler Hospital and the Warren Alpert Medical School of Brown University for providing the BBRS dataset. This research was partly made possible by funding from the National Institute of Health (Grants: R01MH108610; R01MH095786; R01MH097741). We also thank Lifeline Harbour to Hawkesbury in Sydney, Australia for their safety-wellness 'Accidental Counselor' training certification. Additionally, we thank Nancy Briggs with the UNSW Mark Wainwright Analytic Centre.

References

Barsties, B., De Bodt, M., 2015. Assessment of voice quality: current state-of-the-art, 2015 Auris Nasus Larynx 42, 183–188.
Beck, A.T., Steer, R.A., Brown, G.K., 1996. Beck Depression Inventory-II. The Psychological Corporation.
Bele, I.V., 2005. Reliability in perceptual analysis of voice quality. J. Voice 19 (4), 555–573.
Bouhuys, A.L., Van Den Hoofdakker, R.H., 1991. The interrelatedness of observed behavior of depressed patients and of a psychiatrist: an ethological study on mutual influence. J. Affect. Disord. 23, 63–74.
Brook, D.W., Brook, J.S., Zhang, C., M.A., Cohen, Whiteman, M., 2002. Drug use and the risk of major depressive disorder, alcohol dependence, and substance use disorders. Arch. Gen. Psychiatry 59 (11), 1039–1044.
Centers for Disease Control (CDC), 2017. Centers for disease control and prevention data & statistics fatal injury report for 2017.
Cortes, C., Vapnik, V.N., 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297.
Costache, V.S., Renaud, C., Brouchet, L., Toma, T., Le Balle, F., Berjaud, J., Dahan, M., 2004. Complete tracheal rupture after a failed suicide attempt. Ann. Thorac. Surg. 77 (4), 1422–1423.
Couch, S., Zieba, D., Van der Linde, J., Van der Merwe, A., 2015. Vocal effectiveness of speech-language pathology students: before and after voice use during service delivery. S. Afr. J. Comm. Disord. 62 (1), 1–7.
Cummins, N., Scherer, S., Krawjewski, J., Schnieder, S., Epps, J., Quatieri, T.F., 2015. A review of depression and suicide risk assessment using speech analysis. Speech Commun. 71, 10–49.
Day, R.K., 1999. Psychomotor agitation: poorly defined and badly measured. J. Affect. Disord. 55 (2–3), 89–98.
de Araújo Pernambuco, L., Espelt, A., Balata, P.M-M., Costa de Lima, K., 2014. Prevalence of voice disorders in the elderly: a systematic review of population-based studies. Eur. Arch. Oto-Rhino-Laryngol. 272 (10), 2601–2609.
Dejonckere, P.H., Remacle, M., Fresnel-Ebaz, E., Woisard, V., Crevier-Buchman, L., Millet, B., 1996. Differential perceptual evaluation of pathological voice quality: reliability and correlations with acoustic measurements. Rev. Laryngol. Otol. Rhinol. (Bord.) 117 (3), 219–224.
Ellgring, H., Scherer, K.R., 1996. Vocal indicators of mood change in depression. J. Nonverbal. Behav. 20 (2), 83–110.
Esposito, A., Esposito, A.M., Likforman-Sulem, L., Maldonato, M.N., Vinciarelli, A., 2016. On the significance of speech pauses in depressive disorders: results on read and spontaneous narratives. In: Recent Advances in Nonlinear Speech Processing, 48. SIST, pp. 73–82.
Fay, D., Cutler, A., 1977. Malapropisms and the structure of the mental lexicon. Linguist. Inq. 8 (3), 505–520.
Flint, A.J., Black, S.E., Campbell-Taylor, I., Gailey, G.F., Levinton, C., 1993. Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression. J. Psych. 27 (3), 309–319.
Frewen, P.A., Lundberg, E., 2011. Visual-verbal self/other-referential processing task: direct vs. indirect assessment, valence, and experimental correlates. Pers. Individ. Dif. 52 (2012), 509–514.
Henriksson, N.M., Aro, H.M., Marttunen, M.J., Isometsa, E.T., Kuoppasalmi, K.I., Lonngvist, J.K., 1993. Am. J. Psychiatry 150 (6), 935–940.
Hoffman, G.M.A., Gonze, J.C., Mendlewicz, J., 1985. Speech pause time as a method for the evaluation of psychomotor retardation in depressive illness. British J. Psych. 146 (5), 535–538.
Islam, M.-.R., Kamal, A.-.R., Sultana, N., Islam, R., Moni, M.-.A., Ulhaq, A., 2018. Detecting depression using k-nearest neighbors (KNN) classification technique. In: Proc. of 2018 Intern. Conf. on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), pp. 1–4.
H Jiang, H., Hu, B., Liu, Z., Yan, L., Wang, Liu, F., Kang, Li, X., 2017. Investigation of different speech types and emotions for detecting depression using different classifiers. Speech Comm. 90, 39–46.
Kataria, A., Singh, M.D., 2013. A review of data classification using k-nearest neighbor algorithm. Intern. J. Emerg. Technol. Adv. Eng. 3 (6), 354–360.
Larsen, M.E., Cummins, N., Boonstra, T.W., O'Dea, B., Tighe, J., Nicholas, J., Shand, F., Epps, J., Christensen, H., 2015. The use of technology in suicide prevention. In: Proc. of IEEE, pp. 7316–7319.
Lechien, J.R., Morsomme, D., Finck, C., Huet, K., Delvaux, V., Piccaluga, M., Harmegnies, B., Saussez, S., 2018. The effect of speech task characteristics on perceptual judgement of mild to moderate dysphonia: a methodological study. Folia Phoniatr. Logop. 70, 156–164.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep Learn. 521, 436–444.
Levens, S.M., Gotlib, I.H., 2015. Updating emotional content in recovering depressed individuals: evaluating deficits in emotion processing following a depressive episode. J. Behav. Ther. Exp. Psych. 48, 156–163.
Li, Y., Lin, Y., Ding, H., Li, C., 2019. Speech databases for mental disorders: a systematic review. General Psych. 32, e100022.
Liu, Y., Shriberg, E., Stolcke, A., Hillard, D., Ostendorf, M., Harper, M., 2006. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. IEEE Trans. Audio, Speech Lang. Process. 14, 1526–1540.
Marmor, S., Horvath, K., Lim, K.O., Misono, S., 2016. Voice problems and depression among adults in the United States. Laryngoscope 126 (8), 1859–1864.
Martins, R.H.G., Abrantes do Amaral, H., Mendes Tavares, E.L., Martins, M.G., Gonçalves, T.M., Hernandes Dias, N., 2015. Voice disorders: etiology and diagnosis. J. Voice 30 (6), 761.e1–761.e9.
Marzuk, P.M., Hartwell, N., Leon, A.C., Portera, L., 2005. Executive functioning in depressed patients with suicidal ideation. Acta Psych. Scand. 112 (4), 294–301.
Meftah, I.T., Thanh, N.L., Amar, C.B., 2012. Detecting depression using multimodal approach of emotion recognition. In: Proc. 2012 IEEE International Conf. on Complex Systems (ICCS), pp. 1–6.
Mitterschiffthaler, M.T., Williams, S.C.R., Walsh, N.D., Cleare, A.J., Donaldson, C., Scott, J., Fu, C.H.Y., 2008. Neural basis of the emotional Stroop interference effect in major depression. Psych. Med. 38, 247–256.
Oxman, T.E., Rosenberg, S.D., Schnurr, P.P., Tucker, G.J., 1988. Diagnostic classification through content analysis of patients' speech. Am. J. Psychiatry 145 (4), 464–468.
Ozdas, A., Shiavi, R.G., Silverman, S.E., Silverman, M.K., Wilkes, D.M., 2004. Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk. IEEE Trans. Biomed. Eng. 51 (9), 1530–1540.
Piccinelli, M., Wilkinson, G., 2000. Gender differences in depression: critical review. Br. J. Psychiatry 177 (6), 486–492.
Qamar, A.M., Gaussier, E., Chevallet, J.-P., Lim, J.-.H., 2008. Similarity learning for nearest neighbor classification. In: Proc. of the IEEE Intern. Conf. on Data Mining (ICDM), pp. 983–988.
Roy, N., Merrill, R.M., Gray, S.D., Smith, E.M., 2005. Voice disorders in the general population: prevalence, risk factors, and occupational impact. Laryngoscope 115, 1988–1995.
Roy-Byrne, P.P., Weingartner, H., Bierer, L.M., Thompson, K., Post, R.M., 1986. Effortful and automatic cognitive processes in depression. Arch. Gen. Psych. 43, 265–267.
Roy, N., Stemple, J., Merrill, R.M., Thomas, L., 2007. Epidemiology of voice disorders in the elderly: preliminary findings. Laryngoscope 117, 1–6.
Rubino, I.A., D'Agostino, L., Sarchiola, L., Romeo, D., Siracusano, A., Docherty, N.M., 2011. Referential failures and affect reactivity of language in schizophrenia and unipolar depression. Schizophr. Bull. 37 (3), 554–560.
Salem, S., Weskott, T., Holler, A., 2017. Does narrative perspective influence readers' perspective-taking? An empirical study on free indirect discourse, psycho-narration, and first-person narrative. Glossa 2 (1), 1–18.
Scherer, S., Pestian, J., Morency, L.-P., 2013. Investigating the speech characteristics of suicidal adolescents. In: Proc. Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Vancouver, BC, Canada, pp. 709–713.
Schotte, c.K.W., Maes, M., Cluydts, R., Cosyns, P., 1997. Cluster analytic validation of the DSM melancholic depression. the threshold model: integration of quantitative and qualitative distinctions between unipolar depressive subtypes. Psych. Res. 71 (3), 181–195.
Stasak, B., Epps, J., Goecke, R., 2019. Automatic depression classification based on affective read sentences: opportunities for text-dependent analysis. Speech Comm. 115, 1–14.
Tondo, L., Baldessarini, R.J., Hennen, J., Minnai, G.P., Salis, P., Scamonatti, L., Masia, M., Ghiani, C., Mannu, P., 1999. Suicide attempts in major affective disorder patients with comorbid substance use disorders. J. Clin. Psychiatry 60 (2), 63–69.
Wazeer, M.M., John, S., Rajashekhar, B., 2015. Neurogenic speech sequelae following suicide attempt by hanging: a case report. Intern. J. Adolesc. Med. Health 29 (2), 20150039.
Weingartner, H., Cohen, R.M., Martello, J.D.I., Gerdt, C., 1981. Cognitive processes in depression. Arch. Gen. Psych. 38, 42–47.
Yamaguchi, H., Shrivastav, R., Andrews, M.L., Niimi, S., 2003. A comparison of voice quality ratings made by Japanese and American listeners using the GRBAS scale. Folia Phoniatr. Logop. 55, 147–157.
Yingthawornsuk, T., Keskinpala, K., France, D., Wilkes, D.M., Shiavi, R.G., Salomon, R. M., 2006. Objective estimation of suicidal risk using vocal output characteristics. In: Proc. of INTERSPEECH 2006, Pittsburgh, USA, pp. 649–652.
Zabel, T.A., 2005. Neuropsychological profile following suicide attempt by hanging: two adolescent case reports. Child. Neuropsychol. 11 (4), 372–388.