Speech Communication 115 (2019) 1–14

Contents lists available at ScienceDirect

Speech Communication

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

Automatic depression classification based on affective read sentences: Opportunities for text-dependent analysis

Brian Stasak[a,b,*], Julien Epps[a,b], Roland Goecke[c]

[a] School of Elec. Eng. & Telecom., UNSW, Sydney, Australia
[b] Data61-CSIRO, Sydney, Australia
[c] Human-Centred Technology, University of Canberra, Canberra, Australia

ARTICLE INFO

Keywords:
Digital phenotyping
Digital medicine
Paralinguistics
Machine learning
Speech elicitation
Valence

ABSTRACT

In the future, automatic speech-based analysis of mental health could become widely available to help augment conventional healthcare evaluation methods. For speech-based patient evaluations of this kind, protocol design is a key consideration. Read speech provides an advantage over other verbal modes (e.g. automatic, spontaneous) by providing a clinically stable and repeatable protocol. Further, text-dependent speech helps to reduce phonetic variability and delivers controllable linguistic/affective stimuli, therefore allowing more precise analysis of recorded stimuli deviations. The purpose of this study is to investigate speech disfluency behaviors in non-depressed/depressed speakers using read aloud text containing constrained affective-linguistic criteria. Herein, using the Black Dog Institute Affective Sentences (BDAS) corpus, analysis demonstrates statistically significant feature differences in speech disfluencies, whereby when compared to non-depressed speakers, depressed speakers show relatively higher recorded frequencies of hesitations (55% increase) and speech errors (71% increase). Our study examines both manually and automatically labeled speech disfluency features, demonstrating that detailed disfluency analysis leads to considerable gains, of up to 100% in absolute depression classification accuracy, especially with affective considerations, when compared with the affect-agnostic acoustic baseline (65%).

1. Introduction

During mental health evaluations, it is standard practice for a clinician to evaluate a patient's spoken language behavior. Individuals suffering from depression disorders often exhibit psychogenic voice disturbances that adversely change their autonomic system and personality (Perepa, 2017). Recorded exemplars of speech-language disruptions in clinically depressed patients include disfluent speech patterns, abandonment of phrases, and unusually long response latencies (Breznitz & Sherman, 1987; Greden & Carroll, 1980; Hoffman et al., 1985). Patients with clinical depression also exhibit a greater number of speech hesitations (i.e. pauses, repeats, false-starts) than non-depressed populations during communication due to psychomotor agitation/retardation and cognitive processing delays (Alpert et al., 2001; Cannizzaro et al., 2004; Darby et al., 1984; Duffy, 2008; Ellgring & Scherer, 1996; Fossati et al., 2003; Hartlage et al., 1993; Nilsonne, 1987; Nilsonne et al., 1988; Szabadi et al., 1976).

Speech-based depression studies (Alghowinem et al., 2012; Alpert et al., 2001; Esposito et al., 2016; Mundt, 2012; Nilsonne et al., 1988; Stassen et al., 1998; Szabadi et al., 1976) have evaluated pause durations and frequency ratios (e.g. filled-pause rate, empty pause rate) with varied success. For example, Alghowinem et al. (2012) and Esposito et al. (2016) observed that average spontaneous speech pause durations were significantly longer in depressed speakers than non-depressed speakers. Due to unrestrained spontaneous speech variables, pause and rate ratios (i.e. the total number of pauses divided by the total number of words; the total pause time divided by the total recording time) have often been used to help compare utterances of different lengths (Liu et al., 2017).

Concerning speech errors, Rubino et al. (2011) discovered that depressed speakers exhibited significantly greater numbers of referential failures (i.e. including word replacement errors, such as malapropisms) than non-depressed speakers during spontaneous tasks. A malapropism is an incorrect substitution word for an intended word. By definition, a malapropism is unrelated in meaning and has a similar pronunciation, grammatical category, word stress, and syllable length (Fay & Cutler, 1977). Since Rubino et al. (2011), no automatic speech-based depression studies have pursued using speech errors as prospective discriminative depression features – and surprisingly, not even for predetermined read tasks.

Diagnostic applications based on automatic speech-based depression classification are often reliant on the uniformity of the patient elicitation

* Corresponding author.
E-mail addresses: b.stasak@unsw.edu.au (B. Stasak), j.epps@unsw.edu.au (J. Epps), roland.goecke@ieee.org (R. Goecke).

https://doi.org/10.1016/j.specom.2019.10.003
Received 10 December 2018; Received in revised form 27 September 2019; Accepted 11 October 2019
Available online 14 October 2019

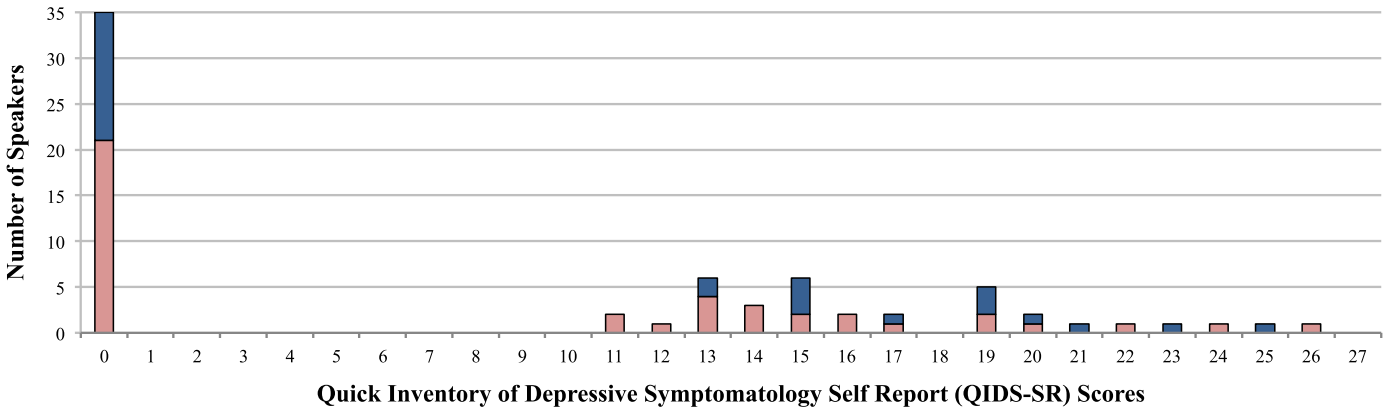

Fig. 1. Depression severity histogram of Black Dog Affective Sentences corpus; 35 non-depressed speakers (0 QIDS-SR) and 35 depressed speakers (11–27 QIDS-SR). The number of females and males are shown in red and blue, respectively. The average QIDS-SR score for the depressed speaker group was 16.7, which is clinically labeled as 'severe' depression. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

procedure. Each elicitation method can include a wide range of linguistic structure, affect information, and task requirements (Howe et al., 2014; Stasak et al., 2017, 2018b). To date, there is no clinically approved speech-depression automatic diagnosis protocol for widespread use, and there is not even consensus on elicitation methods among researchers. Many exploratory speech-based depression techniques have been systematically investigated using a combination of speech-related features (i.e. acoustic, text-based), machine learning techniques, and speech elicitation protocols (see e.g.: Cummins et al., 2015; Jiang et al., 2017; Liu et al., 2017; Stasak et al., 2017; Valstar et al., 2016). However, still to date, no specific system to aid depression diagnosis holds dominance over all others.

The aim of experiments herein is to investigate speech disfluency behaviors in non-depressed/depressed speakers using read text containing specific affective-linguistic criteria. We hypothesize that unlike spontaneous speech, sentences containing target words within specific valence ranges will provide more accurate ground-truth for disfluency analysis as a result of their phonetic and affective constraints; further affording directly comparable speech data between different speakers. While natural speech disfluencies are common in spontaneous speech (Johnson et al., 2004; Gósy, 2003), due to the abnormal cognitive-motor effects of depression on cognitive skills (Breznitz & Sherman, 1987; Greden & Carroll, 1980; Hoffman, et al., 1985), we hypothesize that during simple sentence reading tasks depressed speakers will demonstrate greater numbers of abnormal pauses and speech errors than non-depressed speakers.

According to structural affect theory (Brewer & Lichtenstein, 1982) and information structure studies (Arnold et al., 2013; Dahan, 2015), the emotions of a reader can be systematically manipulated by the sensitive order in which information is presented in a text. Based the aforementioned studies, it is hypothesized that positioning the affective target word at the beginning of sentence rather than middle or end of the sentence will allow a speaker earlier cognitive cues for mood processing and appropriate prosodic phrasing. We suspect that healthy speakers will utilize initial sentence position affective keyword information differently than depressed speakers (i.e. less prosodic range or appropriateness (Cummins et al., 2015)).

In regards to the reading of texts, Salem et al. (2017) found that direct discourse (e.g. first-person) narratives elicited a stronger feeling of taking over the perspective of the character than indirect discourse (e.g. third-person) narratives. Therefore, it is anticipated that features extracted from first-person read narrative sentences will generate better classification results than third-person narrative sentences due to greater emotional attachment. We hypothesize that depressed speakers will exhibit less dynamic emotional vocal range behaviors than healthy speakers on account of their negative fixation (Goeleven et al., 2006; Gotlib & McCann, 1984) and/or less vocal control due to psychomotor agitation/retardation (Flint et al., 1993; Hoffman et al., 1985). Due to passive avoidance strategies exhibited by people with depression (Holahan & Moos, 1987; Holahan et al., 2005), it is also anticipated that depressed speakers will attempt fewer self-corrections after speech errors when compared with non-depressed speakers.

Due to the constraints of text-dependent stimuli, both manual and automatic speech recognition disfluency annotations disfluency attributes are evaluated. It is anticipated that text-dependent constraints will improve the precision of the ASR output since the acoustic phonetic variability in the elicited speech is smaller than that of, for example, spontaneous speech.

2. Database

For all experiments herein, the Black Dog Institute Affective Sentences (BDAS) corpus, a collected data extension of that found in Alghowinem et al. (2012, 2013a, 2013b, 2015), Cummins et al. (2011), and Joshi et al. (2013a, 2013b), was used on account of its clinically validated depression diagnosis. Furthermore, the BDAS corpus had a controlled speech elicitation mode, which comprised read sentences with deliberately designed affective target words (see Section 3.6, Table 1). The speakers were chosen according to the following criteria: relatively good recording quality, completion of all instructed affective read tasks, and equal cohort balance with regards to gender, age, and depression severity. The majority of the speakers had Australian-English as their primary language, while a few non-Australian-English accents were also present (e.g. Indian-English, Irish-English, American-English) in each of the non-depressed and depressed groups.

All audio recordings were conducted in a clinical setting at the Black Dog Institute in Sydney, Australia. Informed consent was obtained from all participants and the study proceeded with approval from the local institutional Human Research Ethics committee in line with the guidelines for human research from the National Health and Medical Research Council. All speakers were recorded during a single session using QuickTime Pro, at a 44.1 kHz sampling frequency subsequently downsampled to 16 kHz. As shown in Fig. 1, in total, there were 21 female and 14 male speakers per non-depressed/depressed group. Speaker ages ranged from 21 to 75 years old with a median age of 40 in the healthy and depressed groups.

Depressed/Non-depressed speakers were recruited to the Black Dog Institute via a tertiary first-time referral and patient consultation diagnosis from the Black Dog Depression Clinic. The depressed speakers were then verified as currently exhibiting a major depression episode

Table 1
BDAS corpus read phrases with 'positive', 'neutral', and 'negative' affective target words emboldened in light green, dark green, and red, respectively. Additionally, sentence affect values for arousal, valence, dominance, and fear-disgust are shown based on the SEANCE text-processing analysis toolkit (Crossley et al., 2017). Arousal, dominance, and fear-disgust were not experimentally evaluated in our study due to range and number of example limitations. Bold scores indicate the most extreme valence sentences {5,16}. (For interpretation of the references to color in this table caption, the reader is referred to the web version of this article.)

#	Sentences	Valence	Arousal	Dominance	Fear-Disgust
{1}	He would **abuse** the children at every party	5.58	6.35	4.87	0.72
{2}	There was a crowd gathering around the **entrance**	0.00	0.00	0.00	0.00
{3}	The teacher made the class **afraid**	3.84	5.36	4.55	0.21
{4}	There had been a lot of **improvement** to the city	6.03	5.24	5.74	0.00
{5}	The **devil** flew into my bedroom	**2.21**	6.07	5.35	1.26
{6}	The chef's hands were covered in **filth**	4.21	4.76	4.58	0.69
{7}	My next door neighbor is a **tailor**	5.13	3.80	4.69	0.00
{8}	The **pain** came as he answered the door	4.63	5.27	4.75	0.59
{9}	She gave her daughter a **slap**	2.95	6.46	4.21	0.57
{10}	There was a **spider** in the shower	3.33	5.71	4.75	0.57
{11}	There was a **fire** sweeping through the forest	3.22	7.17	4.49	0.08
{12}	The **swift** flew into my bedroom	6.46	5.39	6.29	0.00
{13}	There had been a lot of **destruction** to the city	4.59	5.53	4.83	0.37
{14}	The teacher made the class **listen**	5.68	4.05	5.11	0.00
{15}	There was a crowd gathering around the **accident**	2.05	6.26	3.76	0.59
{16}	He would **amuse** the children at every party	**7.47**	6.12	5.47	0.00
{17}	My uncle is a **madman**	3.91	5.56	4.79	0.74
{18}	The **post** came as he answered the door	5.88	4.60	5.27	0.00
{19}	She gave her daughter a **doll**	6.09	4.24	4.61	0.00
{20}	There was a **puddle** in the shower	0.00	0.00	0.00	0.00

based on the Mini International Neuropsychiatric Interview (MINI) (Sheehan et al., 1998). All speakers were also vetted and excluded from the recordings if they had current/prior drug dependencies, prescription medication use, neurological disorders, additional illnesses, or a history of traumatic brain injury.

Utilizing a well-established reading level complexity metric called the Flesch-Kincaid Grade Level test (DuBay, 2006; Kincaid et al., 1975), the 20 BDAS sentences yielded an average readability score of 2.7 (i.e. requiring approximately a 3rd grade reading level). In addition, due to an elicitation mode that required read sentences; all speakers were given the Wechsler Test of Adult Reading (WTAR) (Wechsler, 2001), wherein any speakers that score below an 80 were excluded. Before each recording session began, the Quick Inventory of Depressive Symptomatology Self Report (QIDS-SR) (Rush et al., 2003) was used to re-evaluate the speaker's current severity. As shown previously in Fig. 1, note that only non-depressed speakers with 'none' and depressed speakers with depression greater than 'moderate' severity levels were included in the experiments presented. Similarly to many previous studies (Hashim et al., 2017; Stasak et al., 2017; Liu et al., 2017; Long et al., 2017), our analysis concentrated on speakers with higher severities of depression.

Our BDAS corpus research investigation focused specifically on the task of reading 20 unpracticed short sentences found in Section 3.6, Table 1. This task was self-administered by individual speakers via a computer screen without physician interaction. Speakers viewed and read each entire sentence on a computer monitor one-by-one, pressing the spacebar to proceed to the next sentence. There were no time limits per sentence or intermediary neutralizing stimuli between each sentence. The affective keyword speech elicitation read sentence design herein was loosely based on Brierley et al. (2007) and Lawson et al. (1999), which demonstrated that depressive speakers display negative bias in the interpretation of short sentences. Similar to these studies, our sentences were constructed using opposite affective keyword sentence pairings containing the same approximate lexical frequency, pronunciation likeness, grammar, and syllable length.

3. Methods

3.1. Acoustic feature extraction

For experiments herein, the openSMILE speech toolkit was used to extract 88 eGeMAPS (Eyben, et al., 2015) acoustic speech features (i.e. derived from fundamental frequency, loudness, formants, mel-cepstral coefficients) from all 20 sentences in the BDAS corpus. The eGeMAPS features were calculated by extracting features from 20 ms frames with 50% frame-overlap, wherein an aggregated mean functional was computed per eGeMAPS feature. The eGeMAPS feature set was chosen because it has been used previously as a baseline for speech-based depression research (Stasak, 2018a; Valstar et al., 2016). An analysis of individual eGeMAPS features was omitted in BDAS experiments herein because a prior acoustic feature study of similar data subset already exists (Alghowinem et al., 2013a; Cummins et al., 2011). In general, however, depressed speakers in the BDAS corpus exhibited less prosodic variability, in terms of pitch and loudness, than non-depressed speakers.

3.2. Speech voicing extraction

For speech voicing extraction, the COVAREP toolkit (Degottex et al., 2014) voice activity detection (VAD) algorithm was used. This particular VAD was chosen because it uses frame-level probabilistic decisions based on more than one VAD algorithm (e.g. MFCC-based, summation of the residual harmonics, multi-voicing measures). Also, the COVAREP VAD has been applied in previous speech-based depression studies (Scherer et al., 2014; Valstar et al., 2016). For more details on the COVAREP VAD, the reader is referred to Degottex et al. (2014), Drugman et al. (2016), Eyben et al. (2013), and Sadjadi et al. (2013).

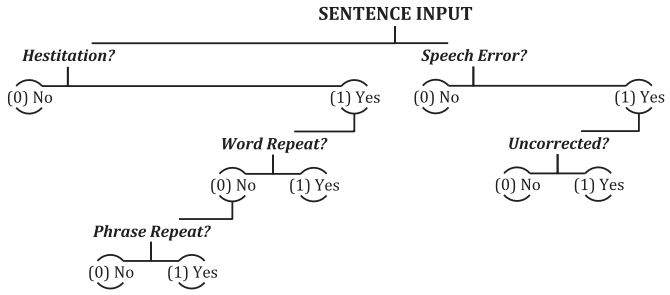

Fig. 2. Speech disfluency feature binary tree showing how each disfluency type relates to each other. The text in italics indicates the five verbal disfluency decision nodes (i.e. disfluency features). A hesitation was defined by the annotator as any unnatural abrupt pause, false start, word/phrase repeat, or abnormal prolongation of an utterance. A speech error was defined by the annotator as any deviation in pronunciation from the intended read target word, such as phonological deletions, substitutions, or slips of the tongue.

Based on the VAD, for every speaker's individual sentence, counts of 'unvoiced' and 'voiced' frames were gathered. Silence frames and/or any fundamental frequency (F_0) values less than 65 Hz and greater than 400 Hz were omitted from unvoiced/voiced counts. For both adult females and males, this F_0 range is within the relative range of normal speech production, excluding extreme vocal modes (e.g. singing, shouting). Per sentence and F_0 threshold, 'voiced' and 'unvoiced' frames were summed into their corresponding groups based on their VAD label, in addition to a combined 'speech' group that included a summation of both 'voiced' and 'unvoiced' frames.

3.3. Manual speech disfluency extraction

All speakers' read sentences were manually analyzed by an experienced annotator with a background in speech-language pathology. From BDAS speaker recordings, the annotator identified any considerable audible hesitations (e.g. abrupt pauses, word repeat, phrase repeat) and speech errors (e.g. malapropisms, omissions, spoonerisms, substitutions). Additionally, the manual disfluency evaluation process was conducted using blind evaluation (i.e. the annotator did not know whether a speaker was depressed/non-depressed), which aided in reducing any possible annotator bias. Differences in disfluency methodology and labeling terminology exist across previous studies, making it more difficult to subjectively or automatically label each with undisputed precision (Garman, 1990; Shriberg, 1994). Hence, as a novel investigation, for each read sentence, three different methods for calculating speech disfluencies were explored.

Firstly, a broad binary single value per sentence was recorded simply if *any* disfluency occurred during the read sentence, even if a self-correction was attempted. Secondly, as shown by Fig. 2, a binary decision tree method was used to tally specific disfluency types, to allow for more insight into hesitation/error type and series-based decision consequences, such as self-correction. Thirdly, a total raw count of disfluencies per sentence was recorded across all of the speech disfluency decision criteria nodes in Fig. 2.

3.4. Automatic speech recognition disfluency extraction

Each recorded sentence was processed using forced alignment via automatic speech recognition (ASR). The Bavarian Archive for Speech Signals (BAS) EMU Magic ASR system (Kisler et al., 2017) was used to obtain the estimated transcript timings for words and pauses (see Fig. 3).

The EMU Magic is a free, cloud-based ASR system that has many language models available, including Australian-English. According to Kisler et al. (2017), BAS generates a word alignment accuracy of ~97% across six example languages using open large vocabulary ASR models (for more details, refer to Kisler et al. (2017)). The automatic monitoring of hesitations was investigated using the ASR transcripts. All sentence files had any start/end labeled pauses removed (e.g. silence) and the total number of ASR entries per speaker was calculated on a per-sentence basis. This calculation included false-starts, word repeats, word additions, and/or pauses. A higher number of observed segmented transcript token word entries per sentence were considered indicative of a speaker's speech disfluency errors.

3.5. Linguistic context

For automatic speech-based processing depression classification, there are no studies that have evaluated the impact of affective keyword locations and narrative point-of-view on read speech. Therefore, linguistic text-based measures centered on the location of the affective keyword and read narrative point-of-views were used to group each of the 20 sentences shown in Section 3.6, Table 1. The affective keyword positions included the following sentence groups: beginning {1,5,8,12}, middle {4,10,11,13,20}, and end {2,3,6,7,9,14,15,17}. For the linguistic read narrative point-of-view measures, each sentence was placed into three different groups: direct discourse first-person (i.e. *I, my, me*) {5,7,12,17}, indirect discourse third-person (i.e. *he, she*) {1,8,9,16,18,19}, and ambiguous {2,3,4,6,10,11,13,14,15,20} narratives.

3.6. Affective measures

Affective text-based measures were evaluated per sentence using Sentiment Analysis and Cognition Engine (SEANCE), which is a free affective text processing toolkit (Crossley et al., 2017). The SEANCE text-processing toolkit's Affective Norms for English Words (ANEW) was used to define valence group ranges for the 'negative' (less than 4), 'neutral' (4 to 9), and 'positive' (6 to 9) sentences. Sentences {2, 20} did not generate valence scores; however, these sentences were placed in the 'neutral' group based on their estimated subjective affective keyword neutrality. The relatively moderate affective target word rating reference size (i.e. approximately 12k English words) is a limitation concerning the use of automatic affective text-processing applications. The use of ANEW valence measures to determine the affective valence negative/neutral groups are a method previously used in Brierley et al. (2007) and Lawson et al. (1999).

Although arousal, dominance, and fear-disgust were also evaluated per sentence, experiments herein focused on valence measures. An affect-text analysis of the arousal scores for sentences in Table 1 produced a narrow score range (i.e. most sentences had an affect score between 4 and 6). In particular, only one sentence {7} had an arousal score less than 4. Therefore, the sentence groupings for arousal were not optimally balanced for affective multi-range analysis. Likewise, the dominance affect score analysis showed that the read sentences only contained ranges from 4 to 6. Consequently, it would be difficult to assess the extreme dominance value ranges (i.e. low, high) since these sentences do not contain this information. The fear-disgust affect score range was the narrowest of the four affective types shown in Table 1. The fear-disgust affect score range was only from 0 to 1. However, it is worth mentioning that analysis shown in Table 1 indicates that any fear-disgust score greater than '0' assured that a 'negative' mood emotion was present. Based solely on the fear-disgust affective sentence scores, it is proposed that this affective measure is useful in automatically labeling sentences 'negative' and 'neutral'.

3.7. System configuration

The speech voicing and disfluency features were average per sentence and concatenated into a 20-dimensional feature vector, which retained the valence sentence-specific feature information. For all 'negative' and 'neutral' sentences shown previously in Table 1, and the additional extended valence sentence group 'positive' group, the summation

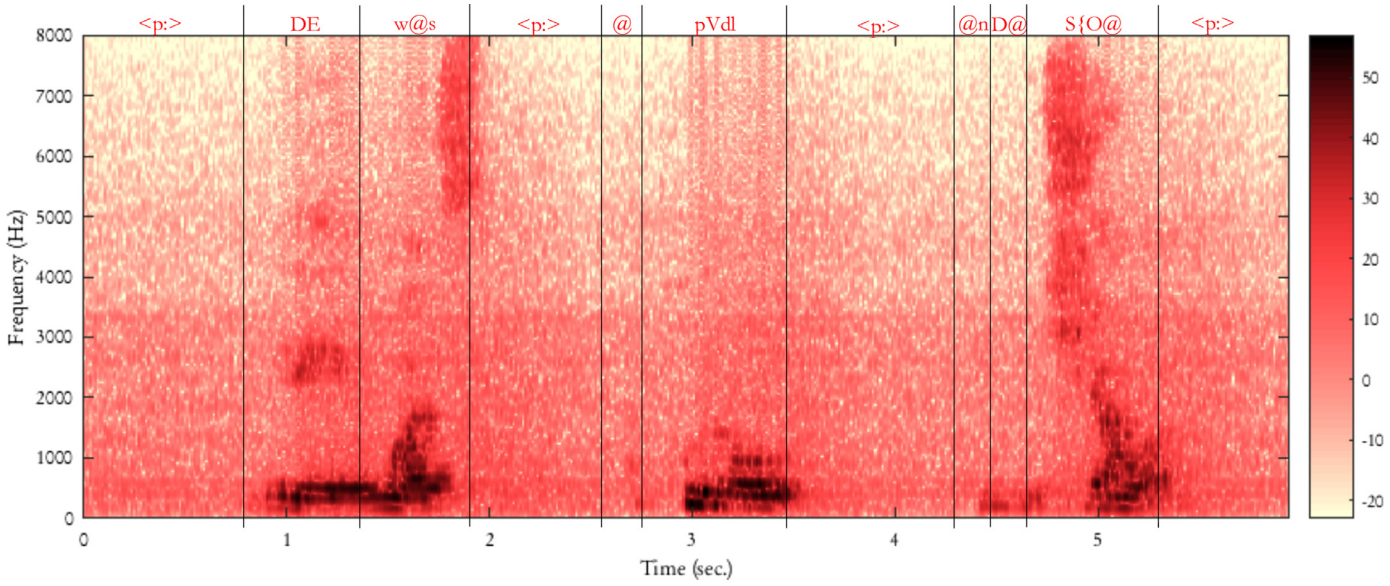

Fig. 3. An example of a spectrogram generated from the BDAS corpus of a severely depressed (i.e. QIDS-SR score of 23) female reading the 'neutral' sentence "*There was a puddle in the shower*". The text-transcript pseudo-phoneme convention is shown along the top of the spectrogram. The vertical lines indicate segmented word tokens (i.e. corresponding to a phoneme-level ASR transcript). The speech hesitations (e.g. pauses) are denoted by <p:>. The loudness (dB-SPL) is indicated by the color bar on the right. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

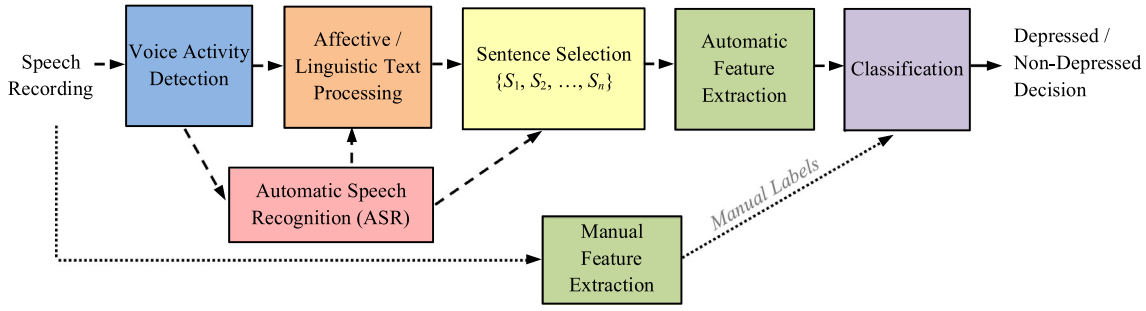

Fig. 4. System configuration, with dashed lines indicating experimental configurations employing data selection based on valence affective text-processing parameters (e.g. 'negative', 'neutral', 'positive'). The thin dotted lines indicate manually analyzed disfluencies.

of manual speech disfluency, speech voicing, ASR token word count, and ASR pause duration were calculated using valence-based feature groupings:

$$S_{all} = \sum_{i=1}^{20} \quad (1)$$

$$S_{neg} = \sum_{i\in\{1,3,5,6,8,9,10,11,13,15,17\}} \quad (2)$$

$$S_{neut} = \sum_{i\in\{2,4,7,12,14,16,18,19,20\}} \quad (3)$$

$$S_{pos} = \sum_{i\in\{4,12,16,19\}} \quad (4)$$

S_{all} contains summed feature values for all 20 sentences, whereas S_{neg}, S_{neut}, and S_{pos} are specific to the respective valence affect groups. Fig. 4 shows the block diagram for the manual and automatic methods used to extract the features evaluated herein. Note that either manual labeling or Automatic Speech Recognition can be used to extract speech disfluency features.

3.8. Classification and evaluation metrics

All experiments utilized the BDAS corpus described previously in Section 2, with speaker independent 10-fold cross validation in a 90/10 training/test split to mimic an initial consultation screening, help maximize data available for training and minimize overfitting. For the acoustic, speech voicing, and fusion experiments presented in Sections 4.1, 4.2 and 4.4, LDA was applied because these features generally had higher dimensions than the majority of disfluency features. Depression classification for the speech disfluency features, both manual and automatic, was conducted using decision trees (similarly to Mitra et al. (2014)) because they allow for the evaluation of all possible consequences of a decision and make no assumptions regarding linearity in the data. All speech disfluency experiments found in Section 4.3 used the simple decision tree classifier from the MATLAB toolkit using a coarse distinction setting with a maximum parameter of 4 splits. For all classification experiments, performance was determined using overall accuracy and individual class F1 scores (similarly to Valstar et al. (2016)).

4. Results and discussion

4.1. Acoustic analysis

The accuracy of the baseline eGeMAPS features depression classification was 65% using all sentences, with F1 scores of 0.68 (0.63) for depressed and non-depressed classes respectively. The accuracies and F1 scores for individual sentences are shown in Table 2.

Table 2
Depressed (D) and Non-Depressed (ND) classification accuracy and F1 score performance using eGeMAPS features. Based on Brierley et al. (2007) and Lawson et al. (1999), affective sentences are listed with 'negative' (red) and 'neutral' (dark green) marked words; whereas, the 'positive' (light green) marked words based on valence scores above 6.0 derived previously from Table 1. {} indicates the sentence number. (For interpretation of the references to color in this table caption, the reader is referred to the web version of this article.)

#	Sentence	Accuracy	F1 D	F1 ND
{1}	He would **abuse** the children at every party	60%	0.66	0.52
{2}	There was a crowd gathering around the **entrance**	59%	0.63	0.52
{3}	The teacher made the class **afraid**	57%	0.64	0.46
{4}	There had been a lot of **improvement** to the city	59%	0.60	0.57
{5}	The **devil** flew into my bedroom	64%	0.67	0.62
{6}	The chef's hands were covered in **filth**	66%	0.69	0.61
{7}	My next door neighbor is a **tailor**	53%	0.56	0.49
{8}	The **pain** came as he answered the door	47%	0.52	0.41
{9}	She gave her daughter a **slap**	51%	0.53	0.50
{10}	There was a **spider** in the shower	57%	0.61	0.53
{11}	There was a **fire** sweeping through the forest	56%	0.58	0.54
{12}	The **swift** flew into my bedroom	56%	0.61	0.49
{13}	There had been a lot of **destruction** to the city	56%	0.56	0.55
{14}	The teacher made the class **listen**	56%	0.58	0.54
{15}	There was a crowd gathering around the **accident**	53%	0.55	0.51
{16}	He would **amuse** the children at every party	**70%**	**0.72**	**0.68**
{17}	My uncle is a **madman**	57%	0.60	0.55
{18}	The **post** came as he answered the door	56%	0.55	0.56
{19}	She gave her daughter a **doll**	61%	0.63	0.60
{20}	There was a **puddle** in the shower	59%	0.58	0.59

Table 3
Depressed (D) and Non-Depressed (ND) classification accuracy and F1 score performance using concatenated sets of eGeMAPS features. Specific sentences that were grouped together are shown in {}.

Affective groups	Accuracy	F1 D	F1 ND
Positive {4,12,16,19}	**67%**	**0.69**	**0.66**
Neutral {2,4,7,12,14,16,18,19,20}	63%	0.66	0.59
Negative {1,3,5,6,8,9,10,11,13,15,17}	56%	0.59	0.52

Table 4
Depressed (D) and Non-Depressed (ND) classification accuracy and F1 score performance using concatenated sets of eGeMAPS features. Specific sentences that are grouped together are shown in {}.

Affective keyword position	Accuracy	F1 D	F1 ND
Beginning {1,5,8,12,16,18}	**70%**	**0.70**	**0.70**
Middle {4,10,11,13,20}	60%	0.62	0.58
End {2,3,6,7,9,14,15,17,19}	61%	0.65	0.57

Generally, single sentence depression classification performance for eGeMAPS features was relatively low when compared with the all-sentence eGeMAPS baseline (i.e. this accuracy decrease was attributed to the reduction in training material). In comparing the acoustic feature depression classification performance produced by specific sentences, the most positive valence score sentence {16} achieved the best accuracy and F1 scores. Furthermore, based on the sentence valence scores previously shown in Table 1, the two of three the best individual sentence depression classification results {5,6,16} occurred for the most extreme valence scores. Sentence {1} (containing highly negative 'abuse' and positive 'party' valence affective keywords) did not perform as well on depression classification in comparison with its opposite-affect paired sentence {16} (containing only positive 'amuse/party' valence affective keywords), or other sentences without opposing valence words. This indicates that read sentences with a single valence keyword or a constrained valence score range are best for acoustic speech-based depression analysis.

Experimentation using training/testing based on concatenated acoustic eGeMAPS features from positive-negative valence sentence opposite pair combinations (i.e. sentences that differed by a single keyword), such as {1,16} and {2,15}, resulted in depression classification results of 49%–67%. As shown in Table 3, training and testing of valence range-specific sentence groups (e.g. 'negative', 'neutral', 'positive') using their concatenated eGeMAPS features demonstrated that the 'neutral' eGeMAPS group did not surpass the depression classification performance of the all-sentence eGeMAPS baseline. Notably, the designated 'negative' valence eGeMAPS group performed the worst when compared with the all-sentence eGeMAPS baseline (9% absolute difference).

Upon extending the sentences into an additional 'positive' valence group, based on extracted valence scores described previously in Section 3.6 in Table 1, the higher valence score sentence group produced a relatively small accuracy improvement (2% absolute gain) over the all-sentence eGeMAPS depression classification baseline. Most importantly, the 'negative' valence sentence group performed the poorest – indicating that for broad acoustic-based feature sets, 'negative' valence sentences are less effective than 'neutral' or 'positive' valence sentences for automatic depression classification. These results concur with Stasak et al. (2017), wherein for spontaneous speech it was demonstrated that acoustic-based speech features derived from phrases with higher valence resulted in better depression classification performance than phrases with lower valence.

With respect to training and testing read sentences based on linguistic affective keyword positions and their concatenated eGeMAPS features as shown in Table 4, sentences with the affective target words towards the beginning of the sentence performed better than those with

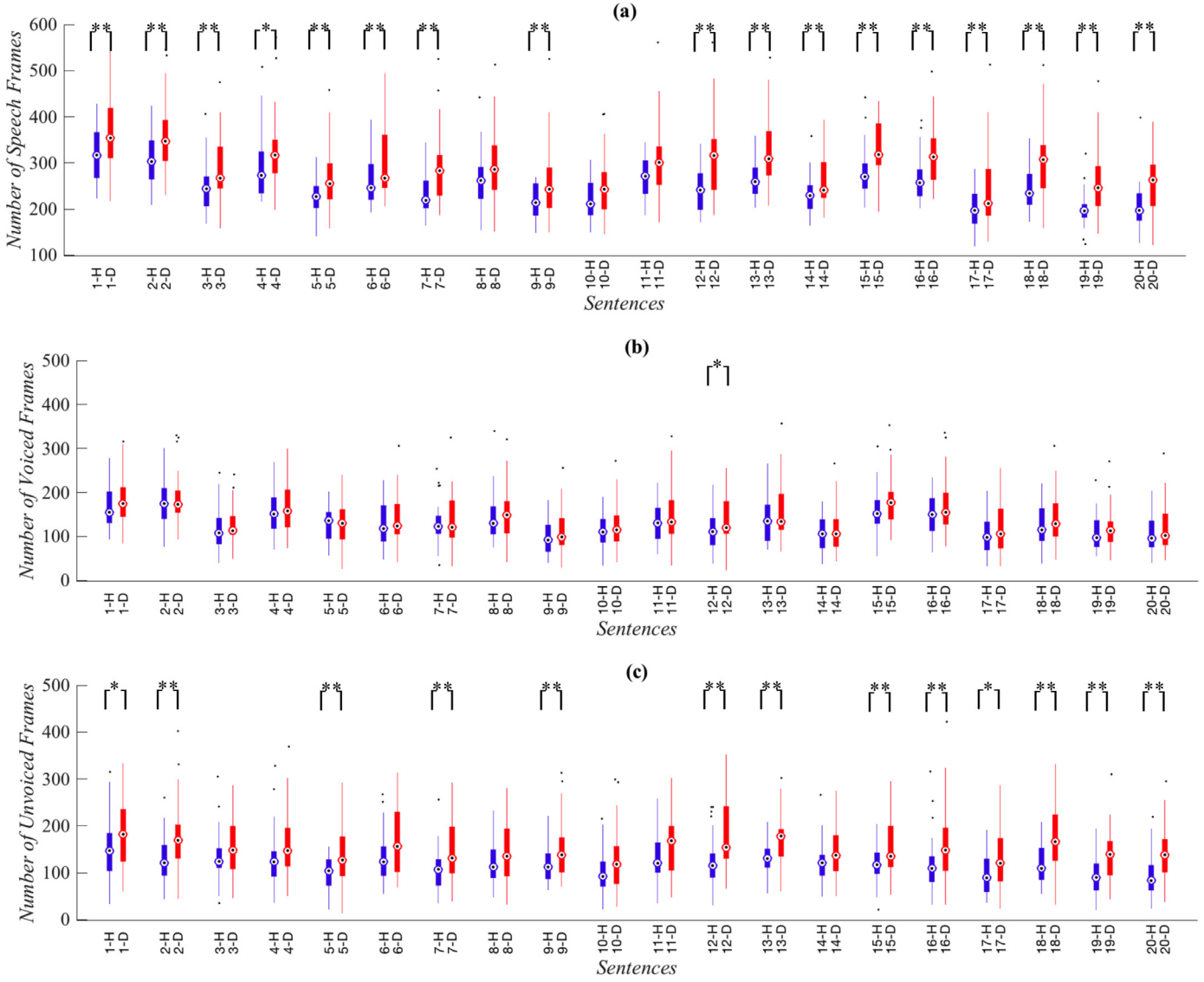

Fig. 5. (a) Number of 'Speech' (voiced + unvoiced); (b) 'Voiced'; and (c) 'Unvoiced' frames per sentence for non-depressed speakers (blue); depressed speakers (red). The mean is indicated using a circle, with the 25th to 75th percentile range shown as a thick bar. The narrower line indicates the outer percentile ranges, while outliers are indicated by small individual dots. Starred and double-starred bracket indicates pairs of results that were statistical different based on a paired *t*-test with $p = 0.05$ and $p = 0.01$ settings, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

affective target words in the middle or end of sentences. The sentence group with affective keywords at the beginning of the sentence using eGeMAPS features achieved a 5% absolute improvement in depression classification accuracy over the all-sentence eGeMAPS baseline. Results in Table 4 support the hypothesis in Section 1 that read sentences with affective target words located at the beginning are more effective for depression classification for acoustic-based features. These results provide syntactic evidence that for acoustic features, an early affective target keyword position induces more distinctive affective paralinguistic cues throughout the remainder of the sentence, whereas for other affective target word positions this is not the case.

4.2. Speech voicing analysis

An initial analysis comparing depressed/non-depressed speakers demonstrated remarkable differences in the overall average of the number of frames of speech per 'voiced' and 'unvoiced' groups, as shown in Fig. 5. In Fig. 5(a), for all sentences, the mean number of 'speech' (i.e. 'voiced' and 'unvoiced') frames for non-depressed speakers was statistically significantly lower on a sentence-by-sentence basis. Interestingly, Fig. 5(b) indicates that the mean number of 'voiced' frames for both non-depressed and depressed speakers is relatively similar; only one sentence {12} showed a statically significant increase in the number of 'voiced' frames for depressed speakers.

The increased speech duration for depressed speakers shown in Fig. 5(a) is thus a consequence of depressed speakers having a greater number of 'unvoiced' frames than non-depressed speakers. Fig. 5(c) shows several statistically significant differences in the 'unvoiced' frame counts between non-depressed and depressed speakers. For non-depressed and depressed speakers, it is understood that sentences {1} and {2} have a greater number of 'speech' frames than other sentences because speakers were adjusting to the reading task expectations.

According to Barrett et al. (2002), for depressed and non-depressed speakers, read sentences with negative (e.g. sad) affect were significantly longer in duration than positive (e.g. happy) affect sentences. The affect group results summarized in Table 5 agree with Barrett's findings,

Table 5
Comparison of average non-depressed (ND) and depressed (D) 'voiced', 'unvoiced', and 'speech' frame counts per sentence group. Those pairs with statistically significant differences are shown by * ($p = 0.05$) and ** ($p = 0.01$).

Affective groups	ND voiced	D voiced	ND unvoiced	D Unvoiced	ND speech	D speech
Positive {4,12,16,19}	133**	148**	116**	162**	249**	310**
Neutral {2,4,7,12,14,16,18,19,20}	131**	144**	115**	158**	246**	303**
Negative {1,3,5,6,8,9,10,11,13,15,17}	130**	142**	123**	155**	252**	296**

Table 6
Depressed/Non-Depressed classification performance using 'speech' and 'unvoiced' frame count features.

'Speech' feature	Classification accuracy	F1 depressed	F1 non-depressed
All {1–20}	70%	0.67	0.73
Positive {4,12,16,19}	**77%**	**0.75**	**0.79**
Neutral {2,4,7,12,14,16,18,19,20}	**76%**	**0.74**	**0.77**
Negative {1,3,5,6,8,9,10,11,13,15,17}	67%	0.64	0.70
'Unvoiced' feature	**Classification accuracy**	**F1 depressed**	**F1 non-depressed**
All {1-20}	70%	0.67	0.73
Positive {4,12,16,19}	**73%**	**0.71**	**0.75**
Neutral {2,4,7,12,14,16,18,19,20}	71%	0.69	0.74
Negative {1,3,5,6,8,9,10,11,13,15,17}	66%	0.63	0.68

Table 7
Comparative depressed/non-depressed classification performance using 'speech' and 'unvoiced' frame count features with linguistic affective target word location groupings.

'Speech' feature	Classification accuracy	F1 depressed	F1 non-depressed
Beginning {1,5,8,12,16,18}	**77%**	**0.77**	**0.78**
Middle {4,10,11,13,20}	71%	0.70	0.73
End {2,3,6,7,9,14,15,17,19}	63%	0.58	0.67
'Unvoiced' feature	**Classification accuracy**	**F1 depressed**	**F1 non-depressed**
Beginning {1,5,8,12,16,18}	70%	0.69	0.71
Middle {4,10,11,13,20}	69%	0.67	0.70
End {2,3,6,7,9,14,15,17,19}	69%	0.72	0.65

Table 8
Depressed/Non-depressed classification performance using 'speech' and 'unvoiced' frame count features with linguistic point-of-view measure groupings.

'Speech' feature	Classification accuracy	F1 depressed	F1 non-depressed
Speech First-Person {5,7,12,17}	**74%**	**0.72**	**0.76**
Speech Third-Person {1,8,9,16,18,19}	70%	0.66	0.73
Speech Ambiguous {2,3,4,6,10,11,13,14,15,20}	67%	0.65	0.69
'Unvoiced' feature	**Classification accuracy**	**F1 depressed**	**F1 non-depressed**
Unvoiced First-Person {5,7,12,17}	71%	0.68	0.74
Unvoiced Third-Person {1,8,9,16,18,19}	69%	0.66	0.71
Unvoiced Ambiguous {2,3,4,6,10,11,13,14,15,20}	67%	0.65	0.69

as the overall average speech duration is longer for the 'negative' valence sentences than for the 'neutral' ones. However, for the depressed speakers, the 'neutral' valence sentence durations were found to be slightly longer than their 'negative' sentences.

This text-dependent analysis of 'voiced' and 'unvoiced' frames helps to substantiate previous observations in studies (Flint et al., 1993; France et al., 2000; Hasham et al., 2012; Kiss et al., 2015; Sahu & Espy-Wilson, 2016) that depressed speakers exhibit lower intensity and breathy speech characteristics. Depression classification results using the speech voicing features are shown in Tables 6–8. In comparison with the eGeMAPS features including baseline, the 'speech' voicing (e.g. 'voiced' and 'unvoiced'), and 'unvoiced' features performed considerably higher (up to 12% absolute gain). The 'voiced' feature results are not shown in Tables 6–8 because they performed no better than chance level.

In Table 6, similarly to the eGeMAPS features, the 'speech' voicing features performed best for the sentence group with the affective target word occurring at the beginning position. Further, as shown in Table 7, for the 'speech' voicing features, the sentence group with the affective target word towards the end of the sentence performed the poorest. These results again support our claim that for acoustic-based features especially, elicitation of the affective target word at the beginning of a sentence provides better depression classification performance than other locations.

For the linguistic narrative point-of-view groups, results in Table 8 show that for the 'speech' voicing (i.e. 'voiced' and 'unvoiced') and 'unvoiced' features, the first-person point-of-view group produced better depression classification results than the other groupings. Note that the first-person point-of-view had sizably less data (e.g. fewer sentences) than the third-person and ambiguous point-of-view groups.

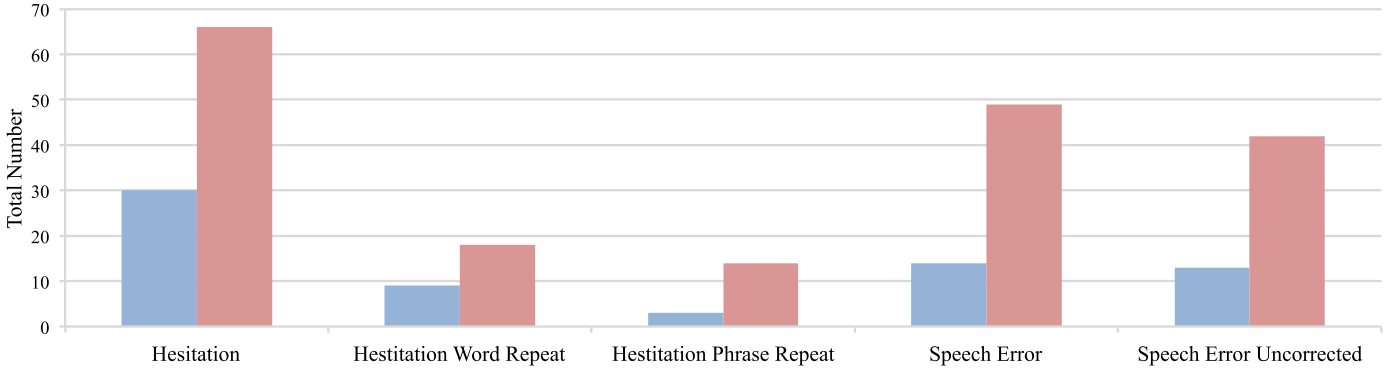

Fig. 6. Total number of manually annotated disfluencies recorded for non-depressed (blue) and depressed (red) speakers for all sentences shown previously in Section 2 in Table 1. Note that the hesitation category includes speech pauses and word/phrase repeats. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.3. Verbal disfluency analysis

A comparison between non-depressed and depressed speakers of the BDAS corpus indicated a statistically significant (*t*-test, $p = 0.01$) higher prevalence of hesitations and spoken word errors for depressed speakers. The average sentence hesitation prevalence across all read sentences was ~9% for depressed speakers, whereas only ~4% for non-depressed speakers (see Fig. 6). The non-depressed speaker average is close to the average percentage of hesitations/repetition occurrences found in Gósy et al. (2003), wherein a speech error of approximately 5% was recorded for interview-driven spontaneous conversational speech.

In general, for conversational speech, the Goldman-Eisler (1968) and Gósy et al. (2003) experiments suggest that hesitations occur naturally roughly every 7–9 words. However, these studies (Gósy et al., 2003; Goldman-Eisler, 1968) evaluated speech generated from spontaneous conversational interview speech, which is expected to be more cognitively demanding than read speech than read speech. It is also known that unlike read speech, conversational speech contains a considerable amount of natural interruptions and interjections (Garman, 1990; Shriberg, 2005). Moreover, in a comparative study on read versus spontaneous speech, Howell & Kadi-Hanifi (1991) found that read speech had fewer hesitations.

As shown in Fig. 6, the depressed speaker group had nearly four times as many speech errors (~8%) as the non-depressed speaker group (~2%). Of particular interest in the manual speech error analysis conducted herein was the unusually high number of recorded malapropisms generated by depressed speakers when compared to non-depressed speakers. Examples of malapropisms produced - many more than once - by depressed speakers in the BDAS corpus were: *abuse*/amuse, *accident*/incident, *chef*/chief, *destruction*/disruption, *she*/he, *tailor*/traitor, *puddle*/poodle and *madman*/madam. As further evidence of the rarity of read speech errors, in a three-month study by Cowie (1985), entire word-level 'slips of the eye' (e.g. malapropisms) were exceedingly rare among adult participants.

It is known that depression disorders limit the degree of cognitive planning and strategies during multitasking (Hartlage et al., 1993). Based on our read speech error analysis, and previous studies on depression disorder speaker cognitive declines (Levens & Gotlib, 2015; Mitterschiffthaler et al., 2008; Silberman et al., 1983; Roy-Byrne et al., 1986; Rubino et al., 2011; Weingartner et al., 1981), the high rate of word errors produced by depressed speakers further substantiates that they are more prone to mental-lexicon word retrieval deficits, referential failures, or concentration restrictions than non-depressed speakers.

Unexpectedly, the analysis of speech error corrections was contrary to our initial hypothesis that depressive speakers would omit verbal self-corrections due to their reliance on avoidance coping strategies (Holahan & Moos, 1987; Holahan et al., 2005). During the sentence reading tasks, as shown previously in Fig. 6, the depressed speakers made more effort to identify and verbally correct their speech errors than the non-depressed speakers. The depressed speakers made attempts to self-correct 14% of their speech errors, whereas non-depressed speakers' attempts at self-corrections were lower at 7%. This percentage calculation was derived from the total number of speech errors uncorrected divided by the total number of speech errors and calculating the remaining percentage.

Although the overall occurrence of hesitations differed between non-depressed and depressed speaker groups, sentences {6,12,15} had the highest number of hesitations for both groups. The similarity in hesitations and speech errors in specific sentences could indicate that these sentences are generally more difficult to read than others due to linguistic syntactical or lexical content. For both non-depressed and depressed speakers, sentences {2,15} exhibited the most recorded speech errors. It is believed that this was mainly due to the word "*around*"; its pronunciation by speakers was often phonetically reduced to "*round*" (i.e. this error may have been influenced by a speaker's dialect). This kind of articulatory deletion indicates that careful consideration should be taken in designing read sentence tasks for speech elicitation and speech error analysis; e.g. avoiding sentence structure with potential colloquialisms, contractions, and acronyms to help maintain consistency of the text-dependent material.

For all combined sentences, as shown by the automatic analysis in Fig. 7, the depressed speaker group showed longer pause durations and a larger number of ASR token word entries than most of the non-depressed speaker group. These findings indicate that the greater number of token word entries for depressed speakers is due to their increase in overall speech disfluencies.

Prior studies (Alghowinem et al., 2012; Esposito et al., 2016; Szabadi et al., 1976) that evaluated speech pauses have generally focused a great deal on rate-of-speech type features. Ultimately, the rate of speech is calculated using the number of phonemes produced over a designated time period. It is known that the rate-of-speech in spontaneous conversation is highly idiosyncratic (Goldman-Eisler, 1961, 1968). Studies have shown that spontaneous speech is faster and has significantly greater interval variability than read speech (Cichocki, 2015; Trouvain et al., 2001). Additionally, spontaneous speech uses different hierarchical acoustic-prosodic cues when compared with read speech, such as intonation, phoneme duration, and spectral features (Haynes et al., 2015; Laan, 1992). Therefore, speech rate ratio type features may be less effective for read speech unless examined at incremental sentence/phrase levels. During our experimentation herein, we experimented briefly with speech phoneme rate ratios. However, our preliminary results on the BDAS read speech performed poorly for depression classification, which included both individual sentence evaluation and concatenated sentence speech rate ratio features.

Table 9
Depressed/Non-depressed classification performance using manually annotated hesitation and speech error features; summed raw counts for all sentences; simple tree classifier.

Speech disfluency feature types	Classification accuracy	F1 depressed	F1 non-depressed
Hesitation (H)	63%	0.57	0.68
Hesitation Word Repeat (HWR)	55%	0.24	0.69
Hesitation Phrase Repeat (HPR)	60%	0.42	0.70
Speech Error (SE)	**71%**	**0.71**	**0.72**
Speech Error Uncorrected (SEU)	**70%**	**0.69**	**0.71**

Speech disfluency types & valence groups	Classification accuracy	F1 depressed	F1 non-depressed
H Positive {4,12,16,19}	63%	0.54	0.69
H Neutral {2,4,7,12,14,16,18,19,20}	**86%**	**0.88**	**0.83**
H Negative {1,3,5,6,8,9,10,11,13,15,17}	61%	0.60	0.63
SE Positive {4,12,16,19}	37%	0.37	0.37
SE Neutral {2,4,7,12,14,16,18,19,20}	64%	0.59	0.68
SE Negative {1,3,5,6,8,9,10,11,13,15,17}	**70%**	**0.64**	**0.74**
Neutral {H} + Negative {SE} + Positive {SE}	**83%**	**0.85**	**0.81**
Neutral {SE} + Negative {H} + Positive {H}	63%	0.57	0.68
Neutral {H,SE} + Negative {H,SE} + Positive {H,SE}	**83%**	**0.84**	**0.81**

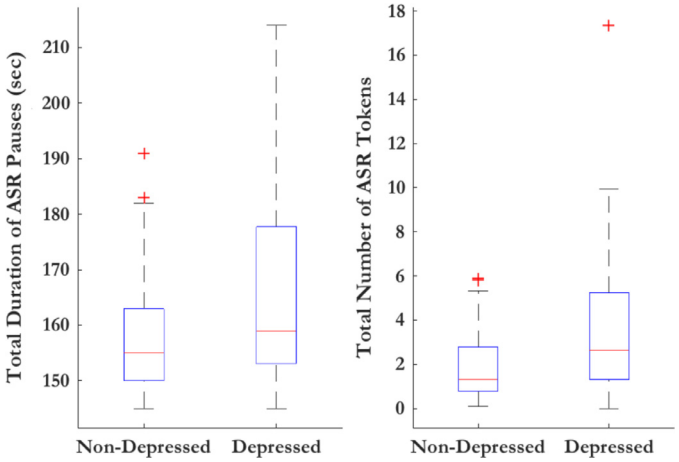

Fig. 7. (left) the total duration of ASR 'pauses' based on transcripts; (right) the total ASR word token entries (i.e. includes words, pauses, disfluencies). These boxplots are based on combined totals for all sentences in Table 1. The mean is indicated using a horizontal line, with the 25th to 75th percentile range shown as a thick bar. The narrower dashed line indicates the outer percentile ranges, while outliers are indicated by '+' symbols.

By using a manually annotated broad binary disfluency single feature approach per sentence, a depression classification accuracy of 67% was obtained with F1 scores of 0.61 (0.68). Statistical analysis using a McNemar test (Adedokun & Burgess, 2012) on non-depressed/depressed speaker binary speech disfluency feature pairs indicated that these type feature values were significantly statistically ($p = 0.05$) different from each other. Furthermore, evaluation of the five binary speech disfluency features using the decision tree classifier, a depression classification accuracy of 67% and F1 scores of 0.58 (0.73) was achieved, indicating that the monitoring and generic identification of any kind of speech disfluency type (e.g. hesitation, repeat, speech error) is useful for identifying depressive characteristics.

By summing the manual raw disfluency values per disfluency feature for all sentences with a decision tree approach, a depression classification average of 69% and F1 of 0.69 (0.69) was attained. Further, Table 9 shows the manual speech disfluency feature results for all individual disfluency raw counts and valence range groups. Combinations of many of the speech disfluency feature types and valence ranges were explored; however, the 'neutral' hesitations contributed the most to higher classification performance. The 86% classification accuracy produced by manual hesitation features on the 'neutral' sentence group was unsurpassed even using many other speech disfluency and valence range combinations. Interestingly, for speech errors, the 'negative' valence sentence group generated the best depression classification results (70%), whereas the 'positive' performed the worst (37%). The 'negative' valence sentences generated more disfluencies because it has been shown that individuals with depression have an affinity to fixate on negative stimuli, which adversely interrupts their task concentration (Goeleven et al., 2006; Gotlib & McCann, 1984). For speech disfluency features, the linguistic narrative point-of-view was examined, as well. However, it appeared to have little impact on speech disfluency features performance.

As shown in Table 10, the ASR approaches to automatically identifying speech disfluencies requires more investigation and refinement, as the general depression classification performance of these features was poorer than for manual methods. The best-automated speech disfluency feature result was observed for the 'positive' valence group, which attained 69% depression classification accuracy. The ASR token word entry results differ from the previously shown manual results in Table 9, wherein 'negative' valence affect sentences are more discriminative of depression; however, the 'negative' valence affect sentences were not far behind (64%). It is believed that an ASR parameter threshold on the number of token word entries allowed per sentence is a major factor (i.e. pause insertions may have been too aggressive). Also, the automatic pause durations excluded abnormal syllable/word prolongations, whereas the manual hesitations included these types of speech disfluency instances.

4.4. *Affect-based feature fusion*

The approaches using the fusion of *n*-best features included the summation of manual speech disfluencies, voicing, and ASR-based derived features. Our final fusion experiments focused on these particular features because for experimental analysis/results previously mentioned in Section 4, they resulted in the best performance. For our fusion approach, valence group features were concatenated into a single feature vector, which is unique from other features proposed previously in that it contains information about each separate valence ranges.

Results shown in Table 11 indicate that used individually, some valence sentence groups and feature types performed better than others, or better than using all sentences. For instance, for manually annotated speech errors, the 'negative' valence sentence group performed the best (70%), whereas for 'speech' voicing features the 'positive' valence sentence group performed the best (79%). The fusion of separate valence groups generated the best result for a single feature type based on manual hesitation features (91%). Moreover, the two automatically derived speech features ('speech' voicing, ASR word token entries, ASR pause

Table 10
Depressed/Non-depressed classification performance using ASR token word entry (TWE) and ASR pause duration (PD) features; summed raw counts for sentences; simple tree classifier.

Automatic disfluency feature	Classification accuracy	F1 depressed	F1 non-depressed
Token word entries all	53%	0.42	0.60
Pause duration all	60%	0.61	0.59
Automatic disfluency feature & valence groups	**Classification accuracy**	**F1 depressed**	**F1 non-depressed**
TWE Positive {4,12,16,19}	**69%**	**0.62**	**0.73**
TWE Neutral {2,4,7,12,14,16,18,19,20}	50%	0.34	0.60
TWE Negative {1,3,5,6,8,9,10,11,13,15,17}	64%	0.53	0.71
PD Positive {4,12,16,19}	60%	0.63	0.56
PD Neutral {2,4,7,12,14,16,18,19,20}	60%	0.64	0.55
PD Negative {1,3,5,6,8,9,10,11,13,15,17}	57%	0.58	0.56

Table 11
Depressed/Non-depressed 2-class classification performance based on raw summed hesitation, speech error, 'speech' voicing, ASR token word entry, and ASR pause duration features. Individual feature results are shown for 'all', 'negative', 'neutral', and 'positive' valence sentence groups. Also, multi-valence fusion results shown include these four types of features (e.g. 'all' + 'negative' + 'neutral' + 'positive'). Manual (blue) and automatic (orange) feature methods have their results in bold to indicate the best results per individual feature type. (For interpretation of the references to color in this table caption, the reader is referred to the web version of this article.)

Feature Types	ALL	NEGATIVE	NEUTURAL	POSITIVE	FUSION
Hesitations	70% 0.64 (0.74)	61% 0.60 (0.63)	70% 0.63 (0.75)	61% 0.60 (0.63)	**91%** **0.92 (0.91)**
Speech Errors	66% 0.61 (0.69)	**70%** **0.64 (0.74)**	64% 0.59 (0.68)	40% 0.36 (0.43)	**70%** **0.63 (0.75)**
Hesitations + Speech Errors	70% 0.67 (0.73)	70% 0.64 (0.74)	69% 0.62 (0.73)	60% 0.56 (0.63)	**87%** **0.88 (0.87)**
Speech Voicing	70% 0.67 (0.73)	70% 0.68 (0.72)	**79%** **0.78 (0.80)**	77% 0.75 (0.79)	76% 0.75 (0.77)
ASR Token Word Entries	59% 0.52 (0.63)	57% 0.50 (0.63)	60% 0.53 (0.65)	**67%** **0.64 (0.70)**	**67%** **0.63(0.70)**
ASR Pause Durations	60% 0.53 (0.65)	59% 0.51 (0.64)	64% 0.59 (0.68)	64% 0.59 (0.68)	**70%** **0.68 (0.72)**
Speech Voicing + ASR Token Word Entries + ASR Pause Durations	73% 0.70 (0.75)	66% 0.61 (0.69)	70% 0.67 (0.73)	**77%** **0.76 (0.78)**	69% 0.68 (0.70)

durations) using the multi-valence sentence group fusion approach were able to surpass (4% absolute gain) the eGeMAPS baseline performance (65%).

By fusing 'speech' voicing and ASR token word entry features using only the 'positive' valence sentences, a depression classification accuracy of 84% and F1 score of 0.83 (0.85) was achieved. Table 12 shows that the majority of disfluency features generated statistically significantly different depression classification accuracy folds results from each other, especially the manual disfluency features (i.e. this compared all possible pairs of speech disfluency types based on their valence group criteria).

Based on the best multi-valence fusion accuracy results shown in Tables 11 and 12, a combination of manual hesitation and automatically derived ASR pause duration feature fusion resulted in an accuracy of 100% for 2-class non-depressed/depressed classification. The combination of sentence groups based on valence and analysis of different feature types indicates that there is discriminative feature-specific information sensitive to exclusive valence ranges. When employed collectively, this multi-affect fusion information was more powerful for depression classification than any individual valence and/or valence-agnostic grouping.

Based on decision tree and LDA classifier depression classification results presented in Tables 9–11, a performance comparison showed the following for both classifiers, respectively: (1) for manual hesitation features the 'neutral' sentence group was the best (86% | 70%); (2) for manual speech error features the 'negative' sentence group performed the best (70% | 70%), whereas the 'positive' sentence group performed the worst (37% | 40%); (3) for the automatic token word entries feature the 'positive' sentence group performed the best (69% | 67%); and (4) for the automatic pause duration feature, all of the affect sentence results were close in performance. A more generalized, hesitation and speech error feature comparison of the decision tree and LDA classifiers based on an their entire depression classification accuracy result set averages (i.e. consisting of all, positive, neutral, and negative sentence groups) resulted in less than a 2% absolute difference in accuracy between these two classifiers.

4.5. Limitations

Like other studies of this kind, a relatively small quantity of speech data was available. Similarly to the minimized number of assessment tasks evaluated in experiments herein, many clinical depression evaluations (e.g. BDI-II, MINI, PHQ-9, QIDS) routinely utilized by physicians consist of only of less than two dozen patient-rated enquiry tasks (Sheehan et al., 1998). Although the best of these results depend on manual annotation from a single annotator, we note that even manual assessment approaches has important practical considerations, as these would be very quick and simple objective tests to administer in a clinical setting or via a web-based service, including by a non-clinician/non-experts. In the future, by using multiple annotators for this type of

Table 12
Statistical significance using a two-sided *t*-test for depression classification results shown previously in Table 11. These results were compiled using the depression classification results per all possible paired sets of 10-fold speech disfluency features. Starred and double-starred bracket indicates pairs of accuracy results that were statistical different based on a paired *t*-test with $p = 0.05$ and $p = 0.01$ settings, respectively.

Disfluency Feature Accuracy *t*-Test Pairings	ALL	NEGATIVE	NEUTRAL	POSITIVE	FUSION
Hesitations ↔ Speech Errors	**	**	**	**	**
Hesitations ↔ Speech Voicing	**	**	**	**	**
Hesitations ↔ ASR Token Word Entries	**	**	**	**	**
Hesitations ↔ ASR Pause Durations	**	**	**	**	**
Speech Errors ↔ Speech Voicing	-	-	-	**	**
Speech Errors ↔ ASR Token Word Entries	**	**	-	**	*
Speech Errors ↔ ASR Pause Durations	**	**	**	**	-
Speech Voicing ↔ ASR Token Word Entries	**	**	**	**	-
Speech Voicing ↔ ASR Pause Durations	**	**	**	**	-
ASR Token Word Entries ↔ ASR Pause Durations	*	-	**	**	-

speech-disfluency research it could provide more insight as to the degree of agreement among annotators.

During the collection of the BDAS corpus, none of the depressed/non-depressed cohorts were receiving anti-depressive or other prescription medication. Some medications can impact speech-language behaviors. Approximately less than half of the depressed/suicidal speech databases published on over the last decade have provided metadata whether or not its participants were prescribed mental illness medications during the recordings (Cummins et al., 2015). It is suggested that future depression speech databases include medication information along with indications of patient comorbidity.

5. Conclusion

In this study, a read sentence protocol was explored as an evaluation method for automatic speech-based depression classification. In comparison to spontaneous speech, text-dependent speech has advantages because the read linguistic and affective content can be explicitly designed to observe behaviors in a repeatable, controlled manner. Further, in a clinical context, text-dependent speech does not rely on the individual interviewer's expertise, bias, and skill level, which has been previously been shown to considerably impact medical diagnostic effectiveness (Chevrie-Muller et al., 1985; Segrin & Flora, 1998).

Experimental results based on acoustic eGeMAPS features show that for text-dependent material, sentences containing affective keywords with the highest valence scores outperformed (4% to 11% absolute gain) sentences with lower valence keywords. An analysis of speech voicing frame-based features based on read sentence experiments showed that depressed speakers have a statistically significantly higher number of 'unvoiced' frames (23% increase) when compared to the non-depressed speakers. Moreover, for 'unvoiced' and 'speech' voicing frames, there was no instance wherein the depressed speakers had a shorter sentence frame average than the non-depressed speakers. For all sentences combined, the speech voicing frame-based features (e.g. 'speech', 'unvoiced') achieved 70% depression classification accuracy. However, by further grouping the 'speech' voicing feature into affective 'negative', 'neutral', and 'positive' groups based on keyword valence scores, greater improvements were recorded, especially for the 'neutral' (76%) and 'positive' (77%) sentence groups.

Our proposed monitoring of specific word-level speech errors for read speech is a new area of exploration for automatic speech depression classification. Manual disfluency analysis presented in our study shows that depressed speakers have a considerable increase in hesitations (55% increase) and speech errors (71% increase) when compared with a non-depressed population. Our investigation of manually annotated speech errors revealed that depressed speakers although given a simple reading task, have an unusually high propensity to produce malapropisms. With regards to automatically extracted disfluencies, the analysis of depressed speakers showed longer pause durations and a greater number of token words than non-depressed speakers based on ASR transcripts (see Fig. 7).

Manual disfluency experiments conducted herein examined broad (e.g. error, no error) versus detailed (e.g. hesitation, word repeat, phrase repeat, self-correction) speech disfluency label feature sets. Using broad manual speech disfluency labels per sentence, results showed that 67% depression classification accuracy was achieved by simply recording whether or not a sentence was read aloud properly. While the examination of the detailed manual disfluency labels per sentence only generated a relatively small depression classification improvement (69%), this method allowed for the examination of individual hesitation and error types per sentence type. For example, it was shown that hesitation features performed the best on neutral sentences (86%), whereas the speech error features performed the best on 'negative' sentences (70%). These results provide evidence that affect contained within a sentence can influence the elicitation of specific speech disfluency types, and furthermore, that detailed disfluency analysis contains rich depression discriminative information.

Experimental results presented in this study demonstrate, in a collective sense, that *all* types of affective valence speech samples (e.g. negative, neutral, positive) are important in the speech-based analysis of depression. The newly proposed approach of fusing feature sets extracted from multiple valence ranges ('all', 'negative', 'neutral', 'positive') during elicited sentences aids the classification of depression, a mood disorder. Remarkably, the majority of the speech disfluency features presented in our depression classification experiments had a relatively low feature dimension (less than 20), yet were highly effective. By fusing the individual affective sentence groups based on valence and combining the most discriminative feature sets explored, significant improvements were recorded when compared to the all-sentence baseline. Our automatically derived disfluency feature investigation showed that specific 'positive' valence sentence fusion of token word entry and pause duration features led to 84% depression classification, which surpassed the affect-agnostic 69% depression classification performance. Furthermore, the crucial technique of fusing two multi-valence fusion systems – based on manual hesitation and automatic pause duration – lead to 100% depression classification.

In terms of implications for elicitation design, our results show that carefully planned linguistic-affective stimuli (e.g. affective keyword position, point-of-view) can help to boost depression classification performance. For example, for acoustic-based features, by evaluating only sen-

tences with affective keyword positions located at the beginning, 5% absolute gains in depression classification accuracy over the all-sentence baseline were recorded. Furthermore, sentences with affective keyword positions located at the beginning generated up to 9% absolute depression classification accuracy gains in comparison to sentences with affective keywords located towards at the end. For the 'speech' voicing feature (e.g. 'voiced' and 'unvoiced'), first-person linguistic point-of-view sentences generated depression classification accuracy gains of 9% and 7% absolute when compared with the all-sentence baseline and third-person sentences, respectively.

The different ranges recorded in depression classification performance based on linguistic (i.e. syntax, grammar) and affective valence facets indicate that read speech protocols (i.e. speech protocols, in general) necessitate many design considerations for automatic depression classification applications to further help maximize effectiveness. To our knowledge, this study is the first examination that explores the manner in which clinically depressed and non-depressed individuals produce disfluencies during spoken affective sentence reading tasks. The newly proposed speech error features have been shown to provide very strong depression classification performance.

Declaration of Competing Interest

None.

CRediT authorship contribution statement

Brian Stasak: Formal analysis. **Julien Epps:** Formal analysis. **Roland Goecke:** Formal analysis.

Acknowledgements

The work of Brian Stasak and Julien Epps was partly supported by ARC Discovery Project DP130101094 led by Roland Goecke and partly supported by ARC Linkage Project LP160101360, Data61-CSIRO. The Black Dog Institute (Sydney, Australia) provided the clinical depression speaker database.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.specom.2019.10.003.

References

Adedokun, O.A., Burgess, W.D., 2012. Analysis of paired dichotomous data: a gentle introduction to the McNemar test in SPSS. J. MultiDiscip. Eval. 8 (17), 125–131.
Alghowinem, S., Goecke, R., Wagner, M., Epps, J., Breakspear, M., Parker, G., 2013a. Detecting depression: a comparison between spontaneous and read speech. In: Acoustics, Speech and Signal Processing (ICASSP) 2013, Vancouver, B.C., Canada, pp. 7547–7551.
Algohowinem, S., Goecke, R., Wagner, M., Epps, J., Breakspear, M., Parker, G., 2012. From joyous to clinically depressed: mood detection using spontaneous speech. In: Proc. of the 25th Intern. Florida Artificial Intell. Research Society Conf., Florida, pp. 141–146.
Alghowinem, S., Goecke, R., Wagner, M., Epps, J., Parker, G., Breakspear, M., 2013b. Characterising depressed speech for classification. In: INTERSPEECH 2013, Lyon, France, pp. 2534–2538.
Alghowinem, S., 2015. Multimodal Analysis of Verbal And Nonverbal Behavior on the Example of Clinical Depression Doctor of Philosophy Thesis. The Australian National University.
Alpert, M., Pouget, E.R., Silva, 2001. Reflections of depression in acoustic measures of the patient's speech. J. Affect. Disord. 66 (1), 59–69.
Arnold, J.E., Kaiser, E., Kahn, J.M., Kim, L.K., 2013. Information structure: linguistic, cognitive, and processing approaches. Wiley Interdiscip. Rev. Cogn. Sci. 4 (4), 403–413.
Barrett, J., Paus, T., 2002. Affect-induced changes in speech production. Exp. Brain Res. 156, 531–537.
Brewer, W.F., Lichtenstein, E.H., 1982. Stories are to entertain: a structural-affect theory of stories. J. Pragmat. 6, 473–483.
Breznitz, Z., Sherman, T., 1987. Speech patterning of natural discourse of well and depressed mothers and their young children. Child Dev. 58, 395–400.
Brierley, B., Medford, N., Shaw, P., David, A., 2007. Emotional memory for words: separating content and context. Cognit. Emot. 21 (3), 495–521.
Cannizzaro, M., Harel, B., Reilly, N., Chappell, P., Snyder, P.J., 2004. Voice acoustical measurement of severity of major depression. Brain Cogn. 56 (1), 30–35.
Chevrie-Muller, C., Sevestre, P., Seguier, N., 1985. Speech and psychopathology. Lang. Speech 28 (1), 57–79.
Cichocki, W., 2015. The timing of accentual phrases in read and spontaneous speech: data from Acadian French. J. Can. Acoust. Assoc. 43 (3).
Cowie, R., 1985. Reading errors as clues to the nature of reading. In: Ellis, A. (Ed.), Progress in the Psychology of Language I & II. London, England, pp. 23–107.
Crossley, S.A., Kyle, K., McNamara, D.S., 2017. Sentiment analysis and social cognition engine (SEANCE): an automatic tool for sentiment, social cognition, and social order analysis. Behav. Res. Meth. 49 (3), 803–821.
Cummins, Nicholas, Epps, J., Breakspear, M., Goecke, R., 2011. An investigation of depressed speech detection: features and normalization. In: INTERSPEECH 2011, Florence, Italy, pp. 2997–3000.
Cummins, N., Scherer, S., Krajewski, 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.
Dahan, D., 2015. Prosody and language comprehension. WIREs Cogn. Sci. 6 (5), 441–452.
Darby, J.K., Simmons, N., Berger, P.A., 1984. Speech and voice parameters of depression: a pilot study. J. Commun. Disord. 17, 75–85.
Degottex, G., Kane, J., Drugman, T., Raitio, T., Scherer, S., 2014. COVAREP – A collaborative voice analysis repository for speech technologies. In: Proc. of ICASSP, Florence, Italy, pp. 960–964.
Drugman, T., Stylianou, Y., Kida, Y, Akamine, M., 2016. Voice activity detection: merging source and filter-based information. IEEE Signal Process. Lett. 23 (2), 252–256.
DuBay, W.H., 2006. Smart Language: Readers, Readability, and the Grading of Text. Impact Information, Costa Mesa.
Duffy, J., 2008. Psychogenic speech disorders in people with suspected neurologic disease: diagnosis and management. In: Proc. of ASHA, presentation, Rochester, USA, pp. 1–48.
Ellgring, H., Scherer, K., 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 (SIST), 48, pp. 73–82.
Eyben, F., Scherer, K., Schuller, B., Sundberg, J., André, E., Busso, C., Devillers, L., Epps, J., Laukka, P., Narayanan, S., Truong, K., 2015. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Trans. Affect. Comp. 7, 190–202.
Eyben, F., Weninger, F., Gross, F., Schuller, B., 2013. "Recent developments in opensmile, the Munich open-source multimedia feature extractor. In: Proc. of the 21st ACM international Conf. on Multimedia, pp. 835–838.
Fay, D., Culter, 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.
Fossati, P., Guilllaume, L., Ergis, A., Allilaire, J., 2003. Qualitative analysis of verbal fluency in depression. Psych. Res. 117 (1), 17–24.
France, D.J., Shiavi, R.G., Silverman, S., Silverman, M., Wilkes, M., 2000. Acoustical properties of speech as indicators of depression and suicidal risk. IEEE Trans. Biomed. Eng. 47 (7), 829–837.
Garman, M., 1990. Psycholinguistics. Cambridge University Press, Cambridge, Great Britain.
Goeleven, E., De Raedt, R., Baert, S., Koster, E., 2006. Deficient inhibition of emotion information in depression. J. Affective Disorders 93 (1-3), 149–157.
Goldman-Eisler, F., 1961. The significance of changes in the rate of articulation. Lang. Speech 4, 171–174.
Goldman-Eisler, F., 1968. Psycholinguistics: Experiments in Spontaneous Speech. Academic Press, London, England.
Gósy, M., 2003. The frequency and interrelations of disfluencies in spontaneous speech. Magyar Nyelvőr 127, 257–277.
Gotlib, I.H., McCann, C.D., 1984. Construct accessibility and depression: an examination of cognitive and affective factors. J. Pers. Soc. Psychol. 47 (2), 427–439.
Greden, J.F., Carroll, B.J., 1980. Decrease in speech pause times with treatment of endogenous depression. Biol. Psych. 15 (4), 575–587.
Hartlage, S., Alloy, L.B., Vazquez, C., Dykman, B., 1993. Automatic and effortful processing in depression. Psychol. Bull. 113 (2), 247–278.
Hasham, N.W., Wilkes, M., Salomon, R., Meggs, J., 2012. Analysis of timing pattern of speech as possible indicator for near-term suicidal risk and depression in male patients. In: Proc. of 2012 Intern. Conf. on Signal Processing Systems, (ICSPS 2012), 58, pp. 6–13.
Hashim, N.W., Wilkes, M., Salomon, R., Meggs, J., France, D.J., 2017. Evaluation of voice acoustics as predictors of clinical depression scores. J. Voice 31 (2) 256.e1–256.e6.
Haynes, R.M., White, L., & Mattys, S.L., 2015. What do we expect spontaneous speech to sound like?, ICPhS.
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.
Holahan, C.J., Moos, R.H., 1987. Personal and contextual determinants of coping strategies. J. Pers. Soc. Psychol. 52 (5), 946–955.
Holahan, C.J., Moos, R.H., Holahan, C.K., Brennan, P.L., Schutte, K.K., 2005. Stress generation, avoidance coping, and depressive symptoms: a 10-year model. J. Consult. Clin. Psychol. 73 (4), 658–666.

Howe, C., Purver, M., McCabe, R., 2014. Linguistic indicators of severity and progress online text-based therapy for depression. In: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, Baltimore, USA, pp. 7–16.
Howell, P., Kadi-Hanifi, K., 1991. Comparison of prosodic properties between read and spontaneous speech material. Speech Commun. 10, 163–169.
Jiang, H., Hu, B., Liu, Z., Yan, L., Wang, T., Liu, F., Kang, H., Li, X., 2017. Investigation of different speech types and emotions for detecting depression using different classifiers. Speech Commun. 90, 39–46.
Johnson, K., 2004. Massive reduction in conversational American English. In: Yoneyama, K., Maekawa, K. (Eds.), Spontaneous Speech: Data and Analysis. The National International Institute for Japanese Language, pp. 29–54.
Joshi, J., Dhall, A., Goecke, R., Cohn, J., 2013a. Relative body parts movement for automatic depression analysis. In: Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII), pp. 492–497.
Joshi, J., Goecke, R., Alghowinem, S., Dhall, A., Wagner, M., Epps, J., Parker, G., Breakspear, M., 2013b. Multimodal assistive technologies for depression diagnosis and monitoring. J. Multimodal User Interfaces 7 (3), 217–228.
Kincaid, J.P., Fishburne, R.P., Rogers, R.L., Chissom, B.S., 1975. Derivation of new readability formulas (automated readability index, Fog count, and Flesch reading ease formula) for Navy enlisted personnel. Instit. Simulation and Training, pp. 8–75 (CNTECHRA Research Branch Report).
Kisler, T., Reichel, U.D., Schiel, F., 2017. Multilingual processing of speech via web services. Comput. Speech Lang. 45, 326–347.
Kiss, G., Vicsi, K., 2015. Seasonal affective disorder speech detection on the base of acoustic-phonetic speech parameters. Acta Universitatis Sapientiae Elect. Mech. Eng. 7, 62–79.
Laan, G.P.M., 1992. Perceptual differences between spontaneous and read aloud speech. In: Proc. of the Institute of Phonetic Sciences Amsterdam, 16, pp. 65–79.
Lawson, C., MacLeod, C., 1999. Depression and the interpretation of ambiguity. Behav. Res. Ther. 37, 463–474.
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.
Liu, Z., Hu, B., Li, X., Liu, F., Wang, G., Yang, J., et al., 2017. In: Zeng, Y., et al. (Eds.), Detecting Depression in Speech Under Different Speaking Styles and Emotional Valences. Springer International Publishing, pp. 261–271.
Long, H., Guo, Z., Wu, X., Hu, B., Liu, Z., Cai, H., 2017. Detecting depression in speech: comparison and combination between different speech types. IEEE Inter. Conf. Bioinfo. and Biomed.
Mitra, V., Shriberg, E., McLaren, M., Kathol, A., Richey, C., Vergyri, D., Graciarena, M., 2014. The SRI AVEC-2014 evaluation system. In: Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, Orlando, FL, pp. 93–101.
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. Psychol. Med. 38, 247–256.
Mundt, J.C., Vogel, A.P., Feltner, D.E., Lenderking, W.R., 2012. Vocal acoustic biomarkers of depression severity and treatment response. Biol. Psych. 72 (2), 580–587.
Nilsonne, A., 1987. Acoustic analysis of speech variables during depression and after improvement. Acta. Psychiatr. Scand. 76 (3), 235–245.
Nilsonne, A., Sundberg, J., Ternström, S., Askenfelt, A., 1988. Measuring the rate of change of voice fundamental frequency in fluent speech during mental depression. J. Acoust. Soc. Am. 83 (2), 716–728.
Perepa, L.S., 2017. Psychogenic voice disorders. Global J. Otolaryngol. 5 (3), 1–10.
Roy-Byrne, P.P., Weingartner, H., Bierer, L.M., Thompson, K., Post, R.M., 1986. Effortful and automatic cognitive processes in depression. Arch. Gen. Psychiatry 43, 265–267.
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.
Rush, A.J., Trivedi, M.H., Ibrahim, H.M., Carmody, T.J., Arnow, B., Klein, D.N., Markowitz, J.C., Ninan, P.T., Kornstiein, S., Manber, R., Thase, M.E., Kocsis, J.H., Keller, M.B., 2003. The 16-item quick inventory of depressive symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. Biol. Psych. 54 (5), 573–583.
Sadjadi, S.O., Hansen, J.H.L., 2013. "Unsupervised speech activity detection using voicing measures and perceptual spectral flux. IEEE Signal Process Lett. 20 (3), 197–200.
Sahu, S., Espy-Wilson, C., 2016. Speech features for depression detection. Proc. Interspeech 1928–1932.
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 narration. Glossa 2 (1), 1–18.
Scherer, S., Hammal, Z., Yang, Y., Morency, C., Cohn, J., 2014. Dyadic behavior analysis in depression severity assessment interviews. In: Proc. of ACM Intern. Conf. Multimodal Interact., pp. 112–119.
Segrin, C., Flora, J., 1998. Depression and verbal behavior in conversations with friends and strangers. J. Lang. Social Psychol. 17 (4), 492–503.
Sheehan, D.V., Lecrubier, Y., Sheehan, K.H., Amorim, P., Janavs, J., Weiller, E., Hergueta, T., Baker, R., Dunbar, G.C., 1998. The mini-international neuropsychiatric interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J. Clin. Psych. 59, 22–33.
Shriberg, E.E., 1994. Preliminaries to a theory of speech disfluencies PhD Thesis. University of California, Berkeley, CA, USA.
Shriberg, E.E., 2005. Spontaneous speech: how people really talk and why engineers should care. In: Proc. of INTERSPEECH, Lisbon, Portugal, pp. 1781–1784.
Silberman, E.K., Weingartner, H., Post, R.M., 1983. Thinking disorder in depression. Arch. Gen. Psych. 40, 775–783.
Stasak, B., Epps, J., Goecke, R., 2017. Elicitation design for acoustic depression classification: an investigation of articulation effort, linguistic complexity, and word affect. In: INTERSPEECH 2017 Conf., Stockholm, Sweden, pp. 834–838.
Stasak, B., Epps, J., Goecke, R., 2018a. An investigation of linguistic stress and articulatory vowel characteristics for automatic depression classification. Comput. Speech Lang. 53, 1–16.
Stasak, B., Epps, J., Lawson, A., 2018b. Pathologic speech and automatic analysis for healthcare applications (batteries not included?). In: Proc. of Speech and Science Technology, Sydney, Australia, pp. 161–164.
Stassen, H.H., Kuny, S., Hell, D., 1998. The speech analysis approach to determining onset of improvement under antidepressants. Eur. Neuropsychopharmacol. 8 (4), 303–310.
Szabadi, E., Bradshaw, C.M., Besson, J.A., 1976. Elongation of pause-time in speech: a simple, objective measure of motor retardation in depression. Brain J. Psych 129 (6), 592–597.
Trouvain, J., Koreman, J., Erriquez, A., Braun, B., 2001. Articulation rate measures and their relation to phone classification in spontaneous and read German speech. In: Proc. of ITRW on Adaptation Methods for Speech Recognition, Sophia Antipolis, France, pp. 155–158.
Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, M., Torres-Torres, M., Scherer, S., Stratou, G., Cowie, R., Pantic, M., 2016. AVEC 2016 – depression, mood, and emotion recognition workshop and challenge. In: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, Amsterdam, The Netherlands, pp. 3–10.
Wechsler, D., 2001. Wechsler Test of Adult Reading: WTAR, The Psychological Corporation. San Antonio, USA.
Weingartner, H., Cohen, R.M., Martello, J.D.I., Gerdt, C., 1981. Cognitive processes in depression. Arch. Gen. Psych. 38, 42–47.