An Investigation of Linguistic Stress and Articulatory Vowel Characteristics for Automatic Depression Classification

Brian Stasak [1,2], Julien Epps [1,2], and Roland Goecke[3]

[1] School of Elec. Eng. & Telecom., University of New South Wales,
Sydney, Australia
[2] Data61-CSIRO, Sydney, Australia
[3] Human-Centred Technology, University of Canberra, Canberra, Australia
b.stasak@unsw.edu.au, j.epps@unsw.edu.au, roland.goecke@ieee.org

ABSTRACT

The effects of psychomotor retardation associated with clinical depression have been linked with a reduction in variability in acoustic parameters in previous work. However, despite opportunities for exploring this reduction through tightly coupled linguistic-acoustic analyses that are afforded by contemporary automatic systems, linguistic stress differences between non-depressed and clinically depressed individuals have yet to be investigated. In this paper, by examining regions within the vowel space corresponding to articulatory parameters, statistically significant differences in articulatory characteristics were discovered at a paraphonetic level. Considering linguistic stress components, depressed speakers exhibited shorter vowel durations and reductions in loudness, which were statistically significant for several articulatory characteristics, accompanied by less variance, especially for 'mid' positioned vowels. Results using a small set of linguistic stress based features derived from multiple vowel articulatory parameter sets generated gains of 7% in two-class depression classification performance of baseline approaches, for the DAIC-WOZ dataset. Further, linguistic stress feature results indicate that specific vowel set analysis provides better discrimination of clinically depressed and non-depressed speakers. Apart from improved classification, knowledge gleaned from this research can also be used for designing more effective depressed speech elicitation methods to be used in conjunction with automatic depression disorder speech assessment systems.

Keywords: Hypoarticulation; Paralinguistics; Psychomotor retardation; Vowel quadrilateral

1. INTRODUCTION

Although spoken language sounds effortless, it is a well-organized cerebral-physiological action involving a cognitively demanding series of complex movements. At a pre-audible stage, multiple areas of the brain simultaneously activate (e.g. Wernicke's, Broca's, prefrontal cortex, supramarginal gyrus), wherein each quickly labors to access meaningful streams of arranged words (Miller, 1963; Edwards et al., 2010; Tremblay et al., 2016). The articulation of a spoken phrase requires a sophisticated degree of simultaneously memorized physiological musculature coordination involving the respiratory system, laryngeal muscles, supra-laryngeal muscles, and fine articulatory positions. Over 100 independently innervated muscles are utilized to produce naturally connective speech sounds (Lenneberg, 1967). From the originating neurological stage to final articulated stage, the overall propagation time interval is extraordinarily brief, for instance, in conversational speech, a person can intelligibly generate up to nine syllables per second (Kent, et al., 2000). What is perhaps more astonishing is the overall degree of targeted articulatory precision exhibited by both adolescent and adult speakers. During spontaneous discourse, speakers produce approximately one speech or language error per 900 words (Garnham et al., 1981). Indeed, there are very few other deliberate cerebral-physiological actions people undertake on a daily basis with such keen accuracy.

These many hidden cognitive-motor intricacies related to speech production are only ordinarily revealed to others when an individual bears a disorder and/or neurological disease in which a disturbance impedes the body's ability to properly verbally communicate. Observed speech behaviors and communicative defects are frequently indicators of common illness and neurological concerns (Hirschberg et al., 2010). Consequently, during clinical assessments, it is unsurprising that current diagnoses of many prevalent diseases/disorders encompass some degree of subjective and/or objective speech-language behavioral evaluation analyses (Chevrie-Muller et al., 1985; Bennabi et al., 2013). Precluding obvious isolated speech-language disorders (e.g. aphasia, apraxia, stammering), studies have shown discernable speech-articulation patterns and/or motor impairments for individuals diagnosed with the following illnesses: Amyotrophic Lateral Sclerosis (ALS) (Kent et al., 1990), Alzheimer's disease (McKhann et al., 2011), autism (Boucher, 1976), depression (Cummins et al., 2015), Parkinson's disease (Harel et al., 2004), schizophrenia (Leff et al., 1981), and Systemic Lupus Erythematosus (SLE) (de Macedo et al., 2017).

Long before automatic methods were explored as tools for depression diagnosis, early subjective studies (Kraepelin, 1921; Stinchfield, 1933; Newman & Mather, 1938; Moses, 1954; Eldred & Price, 1958) examined the speech patterns found in clinically depressed speakers. The primary speech indicators for depression exhibited by depressed patients in the aforementioned studies included differences in prosodic vocal loudness, pitch range, rate-of-speech, and voice quality. In the decades following, many studies by Otswald (1965), Szabadi et al. (1976), Darby and Hollien (1977), Hollien (1980), Greden and Carroll (1981), and Darby et al. (1984) more extensively examined speech from individuals with depression disorders using recorded speech and automatic speech analysis methods. Again, researchers in these studies found that fundamental cues for depression can be derived, at least to some fair extent in spoken English, from similar prosodic acoustic elements, such as fundamental frequency (F0), intensity, and duration.

More recently, with the rise in global depression disorders (WHO, 2017) and further advancements in machine learning, new automatic depression recognition systems have been proposed in (Mundt et al., 2007; DeVault et al., 2013; 2014; Scherer et al., 2014) as an assessment device for clinicians. While there currently is no agreed upon state-of-the-art depression feature set or recognition system, over the last five years gains in this area have been made using a variety of statistical methods, such as Support Vector Machines (SVM) (Algohowinem et al., 2013; Helfer et al., 2013), Gaussian Mixture Modeling with Universal Background Model (GMM-UBM) (Cummins et al., 2013), and neural networks (Stolar, 2016). Additionally, a variety of acoustic speech features (e.g. formants, Mel-cepstral coefficients, Teager energy operator, vocal tract coordination) have been experimented with for depression classification (Cummins et al., 2015). In addition, linguistic, articulatory, and affect related features along with data selection measures have also been examined and advocated in Williamson et al. (2016) and Stasak et al. (2016; 2017a; 2017b).

For linguistic and articulatory related methods in general, there is still good opportunity for the exploration and exploitation of new discriminative information for automatic depression classification. Recently, the Audio Visual Emotion Challenge (AVEC) (Ringeval et al., 2017) has motivated research in the area of speech and depression, including new linguistic text-based approaches, such as topic modeling (Gong & Poellabauer, 2017) and natural language processing (Dang et al., 2017). It can be noted that these approaches largely treat acoustic and linguistic information separately (e.g. fusing the outputs of two independent subsystems), while there is still scope for acoustic analyses that are dependent on the linguistic transcript. Gábor & Klára (2014) suggested that acoustic-based depression classification should put more priority on discovering a correlation between depression severity and changes in articulatory acoustic phoneme parameters. For instance, Stasak et al. (2017a) used articulation effort measures based on age of articulatory mastery to help improve depression classification performance. Additionally, Stasak et al. (2017b) later investigated speech gestural measures based on phonetic markedness (e.g. phoneme transitions) to reveal more discriminate speech segments for depression classification.

During diagnostic evaluations, clinicians have repeatedly referred to depressed speakers' speech using subjective auditory descriptors such as 'flat', 'monotonous', and 'monoloud' (Newman & Mather, 1938; Ostwald, 1965; Darby & Hollien, 1977; Cummins et al., 2015). For individuals with depression disorders, psychomotor retardation (Mayer-Gross et al., 1969) is a key sub-symptom that encompasses a measurable decline in neural planning and control of motor movements. Clinical depression studies regarding depressed speakers displaying psychomotor retardation (Szabadi et al., 1976; Darby & Hollien, 1977; Flint et al., 1993; Cannizzaro et al., 2004; Buyukdura et al., 2011; Bennabi et al., 2013) have discovered abnormal recurrent speech production indicators, such as greater muscle tension and respiratory rate, especially as an individual's depression severity increases (Scherer, 1986; Kreibig, 2010). The increase in overall muscle tension directly impacts the dynamic function and range of the vocal folds. In depressed individuals exhibiting psychomotor retardation, Roy et al. (2009) found that constraints in the vocal folds also similarly impact the jaw and facial muscles in a gross manner. This global manifestation of fine motor strain adversely impacts speech production leading to an increase in speech errors, decrease in speaking rate, and more hesitant speech patterns (Szabadi et al., 1976; Darby et al., 1984; Nilsonne, 1987, 1988; Ellgring & Scherer, 1996; Sobin & Seckbim, 1997; Fossati et al, 2003; Cannizzaro et al, 2004).

The deterioration of fine motor control due to psychomotor retardation ordinarily results in under-articulation (Darby et al., 1984; Scherer et al., 2015), which is also referred to as hypoarticulation (Lindblom, 1990). In brief, hypoarticulation is a uniform non-dynamic speech production approach that tends to minimize the overall degree of articulatory effort and variability. Therefore, hypoarticulation causes greater perceptual auditory blur between dissimilar sounds while also affecting elements of prosody across syllables. On the contrary, hyperarticulation is a highly dynamic speech production manner, maximizing contrast and variability between individual sounds (Trager & Smith, 1951; Jones, 1960; Beckman, 1986; Kent & Netsell, 1971; Engstrand, 1988; Lindblom, 1990). During hyperarticulation, speakers often utilize more variable loudness and greater elocution as articulatory strategy to help increase intelligibility (Lindblom, 1990; de Jong, 1995, 1998). Although both of these hyperarticulation strategies are effective, it is known that the greater degree of enunciation requires more kinematic effort than simply speaking louder (de Jong, 1995, 1993, 1998). It has been documented in several speech and depression studies (Ostwald, 1965; Gruenwald & Zuberbier, 1960; Hargreaves & Starkweather, 1964; Greden et al., 1981; Scherer & Zei, 1988; Kuny & Stassen, 1993; Cannizzaro et al., 2004; Mundt et al., 2007; 2012; Helfer et al., 2013) that clinically depressed speakers exhibit a reduction in vocal emphasis quality and articulatory precision that results in poorer speech intelligibility than what is found in healthy populations.

In the literature, to our knowledge thus far, no research has investigated specific vowel sets with regards to linguistic stress. From a paraphonic (i.e. individual phoneme) standpoint, in English and most other spoken languages (de Jong & Zawaydeh, 1998), linguistic stress is a perceptual observation of rapid spoken fluctuations in the following: duration (length), loudness, pitch (F0), and quality (Fry, 1955, 1958, 1965; Morton & Jassem, 1965). In linguistic terms, the speech modulation can be understood as being composed of a mixture of stressed and non-stressed sounds, which demonstrates an effect of higher speech entropy in informational theoretic terms. In natural speech, linguistic stress functions at a phoneme unit level to permit greater segmental distinction between streams of interlinked phonemes (e.g. syllables, words, phrases). Fundamentally, linguistic stress improves speech intelligibility by emphasizing which sound units differ from each other, while also simultaneously providing audible cues as to which sound units carry the most important informational content (Hockett, 1958; Miller, 1963; Ladefoged, 1967). In Hitchcock & Greenberg (2001) and Greenburg (2002), it was shown that syllable stress perceptually influences an individual's ability to identify phonetic segments in spontaneous speech, especially temporal aspects of the vocalic nucleus. It is believed that depressed speakers with hypoarticulation will exhibit an overall reduction in intensity and length variation across all syllable types (e.g. stressed, unstressed), whereas non-depressed speakers will demonstrate a greater variability. Additionally, due to tongue mobility limitations and neutral tongue placement found in speakers with hypoarticulation, it is suspected that more severely depressed speakers will have shorter vowel durations when compared with non-depressed speakers.

A few studies have focused on the effects from depression on English vowel production (Scherer, et al., 2016; Vlasenko et al., 2017). Scherer (2016) found that depression affected vowel frequencies F1 and F2 and calculated Vowel Space Area (VSA) (Liu, et al., 2003), which was previously suggested as a good measure for speech clarity (Bradlow et al., 1996). However, results in Scherer et al. (2016) found the largest difference in VSA between clinically suicidal and non-suicidal speakers, whereas the VSA differences between 'depressed' and 'non-depressed' speakers indicated only small changes. Scherer et al. (2016) did not extend experimental specificity into possible gender differences in VSA. But, recently, using a similar F1 and F2 VSA-based approach to that of Scherer et al. (2016), Vlasenko et al. (2017) used a phonetic recognizer to compare recorded vowels of clinically depressed and non-depressed speakers, but with further implementation of gender-dependent modeling. In both Scherer et al. (2016) and Vlasenko et al. (2017), their experimental approaches did not investigate articulatory characteristics apart from just the amount of general vowel-plane variability; furthermore, they did not evaluate linguistic stress within particular vowel sets.

In general, Fig. 1 contains important details regarding how the placement of the tongue should normally operate during English vowel sounds. As shown in Fig. 1, the American English vowel space comprises different articulatory positions that affect vowel quality and allows greater distinctions between different vowel sounds. There are three major vocoid articulations that have a significant impact on the shape of the oral cavity, and in turn, vowel quality: tongue-height, tongue advancement, and lip position (Hockett, 1958). The tongue height is based on the vertical positioning of the tongue along with the upper and lower jaw positions. The opening of the jaw aids in allowing the tongue to reach its correct placement. In Fig. 1, on the *y*-axis, tongue height is described in terms of its vertical movement: high, mid, and low. The *x*-axis indicates tongue advancement, which is related to the horizontal positioning of the tongue in terms of tongue area predominantly involved during vowel production. For example, in the sound /*iy*/, the whole upper portion of the tongue is high from dorsum to blade, whereas with /*uw*/ only the dorsum has high placement. The lip position relates to the shape of the lips during articulation. Both the /*uw*/ and /*ow*/ vowel sounds are considered rounded, whereas /ah/ for instance is unrounded. An additional aspect of vowel production is lax and tense. Lax vowels (/*ih*, *eh*, *ah*, *ae*/) tend to be shorter in duration than tense vowels and are produced in a less constrained muscular manner. On the contrary, tense vowels (/*iy*, *uw*, *ow*, *aa*/) are usually longer in duration than lax vowels, and are produced with more lip rounding involving greater muscular tension (Hockett, 1958). A monophthong is a fixed pure vowel sound within a single syllable, whereas a diphthong is a single syllable containing a vowel transformation into another adjacent vowel.

We hypothesize that certain articulatory vowel parameters and/or movements as illustrated in Fig. 1 are more affected by depression than others. This hypothesis is semi-based on studies (Tolkmitt et al., 1982; Flint et al., 1993) that have previously inferred reduced articulatory effort effects based on noticeable F2 reductions in depressed speakers. We anticipate that depressed individuals will demonstrate unusual articulatory characteristics akin to hypoarticulation, which in turn affects the dynamic nature of linguistic stress strategies (e.g. duration, loudness, pitch) triggering shorter, more unified vowel durations and less dynamic vocal intensity.

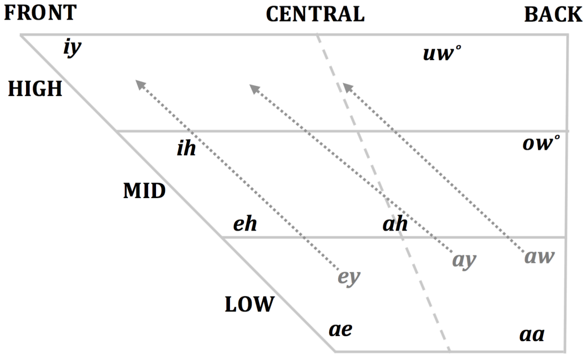

Figure 1: Illustration of the vowel quadrilateral and various tongue parameters based on North American English studies (Hockett, 1958; Ladefoged, 1967, 1975). Note that only the 8 monophthong (in black) and 3 diphthong (in grey) vowels investigated in this paper are shown. The arrows indicate the approximate starting and ending positions for the diphthongs. The superscript ° indicates that a vowel is rounded.

The primary motivation behind research presented in this paper is to link known articulatory norms stemming from linguistic stress to the hypoarticulation effects found in speakers with a depression disorder. It is proposed that by investigating English linguistic stress components at a fine-grained paraphonetic level, acoustic differences between depressed and non-depressed speakers will become more evident. This paper is organized into the following subsections: Section 1 includes a discussion concerning the existing literature on spoken language production, affects of depression disorder on speech, linguistic stress components, articulatory vowel space, and parallels between psychomotor retardation and hypoarticulation, which results in a verbal overall reduction in linguistic stress. Section 2 details the experimental database. Section 3 presents the methodologies applied for analyses, specifically phonetic segmentation, feature extraction, classifier settings, and a system configuration summary. Section 4 presents various vowel set investigations along with a discussion pertaining to the results. Section 5 summarizes experimental findings and suggests future directions for additional research.

2. DATABASE

For all experiments herein, an audio subset of the training and development from the Distress Analysis Interview Corpus – Wizard of Oz (DAIC-WOZ) (Gratch et al., 2014) was used due to its large size and previously published speech depression disorder analysis (Valstar et al., 2016; Stasak, et al, 2017a; Stasak, et al., 2017b). The DAIC-WOZ was created to examine a series of various language related behaviors, such as speech patterns, kinesics, psychophysiology, and assisted human-computer spoken dialog. Unlike a human interviewer, the virtual human-computer interviewer provides neutral unbiased emotion and a limited number of question/responses. Shown in Fig. 2, the experimental data subset has a total of 82 male and female speakers that were recorded using a high-quality close-talking microphone. All recordings contain naturally spoken North American English in a clinical styled environment. The average length per speaker file excluding silence was approximately 7 minutes.

Every speaker had a Patient Health Questionnaire (PHQ-8) score provided, which indicates his/her severity level of clinical depression. The PHQ-8 is a commonly referenced self-administered mental health assessment tool, which is frequently used by clinicians during depression disorder diagnosis (Kroenke et al., 2001, 2009). Each of the questions in the PHQ-8 has a qualitative answer value between 0-3. The total score for the PHQ-8 has a scale of 0 to 24, wherein larger scores imply greater depression severity. Similarly to studies that precluded speakers with clinically ‘mild’ to ‘moderate’ depression (Solomon, 2015; Liu, 2016; Stasak et al., 2017a; Stasak et al., 2017b), experiments herein also omitted speakers within these ranges. Therefore, only speakers from the DAIC with PHQ-8 scores of 0-4 (i.e. “no significant depression” symptoms) and 15-24 (“moderately severe” to “severe” symptoms) were evaluated. Furthermore, speakers with “moderately severe” to “severe” PHQ-8 symptoms are also the most likely to exhibit psychomotor retardation (Shah et al., 1997; Loo et al., 2008; Yorbik, et al., 2014).

Of the total 82 speakers in the DAIC subset, approximately 20% were labeled as ‘Depressed’. While this percentage is a higher representation than what is typically found in a primary care setting, where ~10% of patients meet the diagnostic criteria for depression (Luber et al., 2000), it is also further estimated that nearly two-thirds of individuals with clinical depression go clinically undiagnosed in a primary setting (Ani et al., 2008). Hence, it is very likely that the percentage of depression disorders among general populations is higher than originally thought. In terms of automatic modeling, the majority of systems require an adequate number of training files to properly generate meaningful decision outputs and increase robustness across large speaker populations.

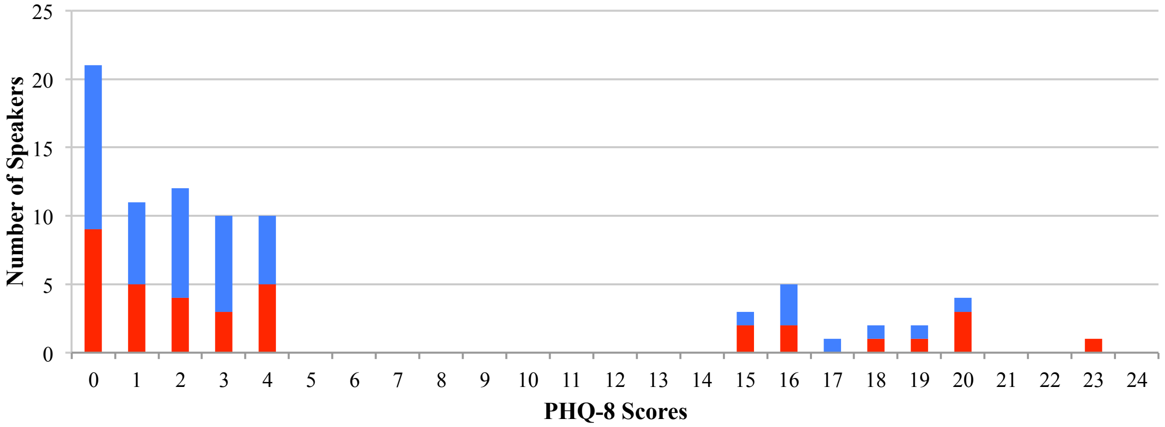

Figure 2: Distribution of PHQ-8 scores for the experimental DAIC-WOZ subset. In total, there were 82 speakers; 36 females (red); and 46 males (blue). PHQ-8 ranges from 0-4 were labeled as ‘Non-Depressed’ (~80% of speakers), whereas ranges 15-24 were labeled as ‘Depressed’ (~20% of speakers).

3. EXPERIMENTAL SETTINGS

3.1 Phonetic Segmentation for Detection of Linguistic Stress

To automatically segment phonemes from the original audio files, the Brno Phoneme Recognizer was used (Schwarz et al., 2006). While the Brno recognizer still has a moderate degree of error (~30%), it has been widely applied in speech processing research and can be considered a de facto standard recognizer (Trevino et al., 2011). Moreover, studies on human-annotated transcripts have demonstrated a degree of error roughly as high as contemporary automatic methods, especially for conversational speech (Hayden, 1950; Mines, 1978). For phoneme modeling, a North American English model was used because all speech data analyzed comprised a similar dialect origin. Prior to phonetic segmentation, all speaker files had voice activity detection (VAD) (Kinnunen et al., 2013) applied to remove undesirable silence and noise. Afterwards, each speaker file was processed using the phonetic recognizer with the standard parameter settings. The phonetic recognizer generated a metadata file with the proposed phoneme output per speaker file along with the start and end times in milliseconds per phoneme.

The experiments herein focused specifically on the most frequently occurring vowels across all speakers, as listed in In Table 1. a few vowels and diphthongs were omitted because of their low occurrence or insufficient examples across all 82 speakers. It should be noted that the histogram of total phoneme outputs (e.g. phoneme percentage distribution for *all* sounds including consonants) generated by the automatic phoneme recognizer was consistent with prior large corpus human-transcription phoneme study distributions (French et al., 1930; Voelker, 1935; Hayden, 1950).

Table 1: Summary of North American English vowels extracted from DAIC-WOZ using a phonetic recognizer, along with articulatory parameter descriptions based on Fig. 1. Each vowel sound has a word example with its vowel pronunciation highlighted in red. A total vowel count across all speakers is provided; as expected, some vowels occur more frequently than others.

	Phonetic Symbol	Example	Tongue Height	Tongue Advancement	Lip Position	Contrast (duration)	Transition	# Total
ih	/ɪ/	*sit*	High	Front	Unrounded	Lax (short)	Monophthong	16,250
iy	/i/	*eat*	High	Front	Unrounded	Tense (long)	Monophthong	14,969
uw	/u/	*boot*	High	Back	Rounded	Tense (long)	Monophthong	6,324
eh	/ɛ/	*bet*	Mid	Front	Unrounded	Lax (short)	Monophthong	4,453
ah	/ʌ/	*cut*	Mid	Central	Unrounded	Lax (short)	Monophthong	7,275
ow	/o/	*over*	Mid	Back	Rounded	Tense (long)	Monophthong	4,035
ae	/æ/	*cat*	Low	Front	Unrounded	Lax (short)	Monophthong	4,935
aa	/ɑ/	*hot*	Low	Back	Unrounded	Tense (long)	Monophthong	3,906
ey	/eɪ/	*bay*	Low-Front	High-Front	Unrounded	Tense (long)	Diphthong	3,426
ay	/ɑɪ/	*hide*	Low-Back	High-Front	Unrounded	Tense (long)	Diphthong	9,346
aw	/ɑʊ/	*brown*	Low-Back	High-Central	Unrounded to Rounded	Tense (long)	Diphthong	2,754

3.2 Feature Extraction Based on Articulatory Characteristics

For experiments on articulatory characteristics and linguistic stress herein, vowel duration features were computed per speaker based on the mean and standard deviation of various articulatory parameter sets. For acoustic feature extraction, the open-source openSMILE speech toolkit was used to extract 88 low-level eGeMAPS (Eyben, et al., 2016) acoustic speech values (i.e. F0, loudness, formants, Mel-cepstral coefficients) by aggregating valid 20-ms frame-level features across a particular speaker's segmented vowels. The eGeMAPS feature set was chosen because it has been used previously for emotion and speech-based depression research (Valstar et al., 2016; Cummins et al., 2016; Stasak, ACII, 2017). As an experimental baseline, the mean 88 low-level eGeMAPS features were extracted from the whole file (e.g. consonants and vowels) and also from individual whole vowels (e.g. only 11 vowels).

In addition, two other features based on vowel sets were proposed based on the 88 eGeMAPS features: Articulatory Characteristic (*AC*) and Linguistic Stress (*LS*). These vowel set features were constrained and only contained feature from vowels within a specific vowel set as detailed previously in Table 1 (e.g. tongue position, lip shape). The *AC* features were obtained by combining all eGeMAPS vowel sound features within a particular vowel set per speaker and then calculating the mean per speaker. Therefore, for each of the 12 different vowel sets the *AC* feature vector per individual speaker comprised of 88-dimensional eGeMAPS feature **V** together with a duration mean dimension $\bar{L}_n$.

$$AC = \left[\mathbf{V}_1{}^T\ \bar{L}_1\ \mathbf{V}_2{}^T\ \bar{L}_2 \ldots \mathbf{V}_n{}^T\ \bar{L}_n\right]^T \tag{1}$$

For the linguistic stress experiments, an *LS* feature was proposed by using only the eGeMAPS loudness mean $\bar{L}_n$ and pitch mean $\bar{P}_n$ features along with duration mean $\bar{D}_n$ derived from the phonetic recognizer timestamps, wherein their mean and standard deviation was computed per *n*th (of 12 total) vowel parameter set. Thus, the complete *LS* features comprised a more compact feature vector size of 24 dimensions per speaker.

$$LS = \left[\bar{L}_1\ \bar{P}_1\ \bar{D}_1\ \ \sigma_{L_1}\sigma_{P_1}\sigma_{D_1}\ \ \bar{L}_2\ \bar{P}_2\ \bar{D}_2\ \ \sigma_{L_2}\sigma_{P_2}\sigma_{D_2} \ldots \bar{L}_n\ \bar{P}_n\ \bar{D}_n\ \ \sigma_{L_n}\sigma_{P_n}\sigma_{D_n}\right]^T \tag{2}$$

3.3 Classifier and Performance Metrics

Similarly to (Mitra et al., 2014; Stasak ACII, 2017), depression classification was conducted using decision trees, which performed well in preliminary experiments across different classifier types. All experiments used the medium decision tree classifier from the MATLAB toolkit using a few leaves and a maximum of 20 splits. Experiments utilized 10-fold cross validation using a 90/10 training/test split to help maximize data available for training.

Classification performance was determined using overall accuracy and individual class F1 scores (similar to Valstar et al., 2016; Stasak ACII, 2017; Cummins, 2017). The F1 score is a common metric that combines precision and recall, thus allowing further evaluation of specific class performance based on true/false positives and true/false negatives, and is a helpful evaluation criterion for unbalanced classification problems. The F1 score is computed by the following equation (a large F1 score implies better discrimination):

$$F1 = \frac{precision \cdot recall}{precision + recall} \tag{3}$$

3.4 System Configuration

The system employed in the depression classification experiments is shown in Fig. 3. All experiments evaluated baseline **V**, articulatory vowel parameter *AC*, and linguistic stress *LS* features together with variants of these. In addition, the effects of data selection were examined specifically for articulatory vowel parameter and linguistic features. The speech data were preprocessed using VAD and then automatically segmented using a phonetic recognizer. Data selection was employed for articulatory characteristic and linguistic stress experiments, in which specific vowel sets were selected for analysis on a per-vowel set basis. A decision classifier was then used to determine a score prediction output and compare the ground truth labels to the test labels.

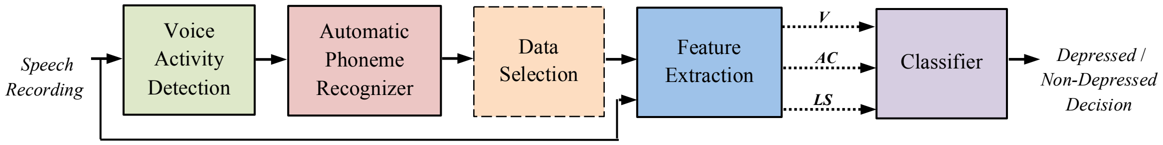

Figure 3: System configuration for the experiments. Dashed lines indicate data selection based on articulatory vowel parameters (e.g. 'front', 'central, 'back').

4. EXPERIMENTAL RESULTS

4.1 Articulatory Characteristics Sensitive to Depression

In Section 1, it was hypothesized that specific articulatory vowel parameters and/or movements will be more affected by depression than others. Prior to experimentation, it was also suggested that the effects of psychomotor retardation and hypoarticulation in depressed speakers would mostly impact vowels that require more kinematic effort. For instance, vowels with pertaining to positions with further proximity from a neutral or 'central' vowel position, entailing lip rounding, or transitions found in diphthongs were believed to have the most discriminative depression characteristics.

Fig. 4 shows the results for various articulatory characteristic (*AC*) feature sets. Based on Fig. 1 and Table 1, experiments were conducted evaluating how articulatory vowel parameters influence depression classification. For the tongue height position vowel set, the 'mid' set yielded the best depression classification and F1 scores. This is likely because the 'mid' tongue position is the only set that has three different tongue activation placements – /*eh*/ (front), /*ah*/ (central), and /*ow*/ (back). It is suggested that the increased range of the tongue activation within this set improves its performance. Among the tongue advancement sets, the 'front' vowels perform better than 'central' or 'back'. It should be noted that the 'front' vowels have four vowel sounds (two of which occur most frequently), whereas the 'central' and 'back' have fewer. As predicted earlier, the 'central' set, due to its neutral positioning, mild kinematic demand, and generally short duration, produced the lowest depression classification accuracy and poor F1 scores when compared to the majority of other vowel sets.

The 'rounded' lip position set results were among the best recorded for the articulatory vowel parameter experiments. This was surprising because the 'rounded' set only consists of two vowel sounds (/*ow*, *uw*/). Yet, despite the small pool of 'rounded' vowels, based on Hockett (1958) and Flint et al. (1993) it was suspected that the auxiliary bilabial muscular demand would lead to possible depression discriminatory characteristics. These results may indicate lip articulatory differences between non-depressed and depressed speakers, which in turn affects vowel quality aspects captured by the eGeMAPS feature set. The aforementioned literature (Scherer, 1986; Roy et al., 2009; Kreibig, 2010) noted tightening of the facial muscles and additional visual studies on depression have noted flat facial responses (Widlocher, 1983; Parker & Hadzi-Pavlovic, 1996; Gehricke & Shapiro, 2000). The effects of psychomotor retardation might contribute to a decrease in lip rounding within more severely depressed speakers than non-depressed populations. On the contrary, the 'unrounded' lip position set is believed to have generally performed well because of the large number of vowel sounds contained within it - including the 'mid' set that performed the best overall for the fixed vowels.

As anticipated, the contrast results show that the 'tense' vowel set surpasses the 'lax' vowel set in performance, especially in terms of F1 depressed classification (0.11 absolute gain). This is likely due to the greater degree of kinematic effort involved when producing tense relative to lax vowels. Based on English linguistic-phonetic studies (Knight, 2012; Flemming, 2007), most lax vowels or unstressed syllable vowel components are produced as a generic schwa vowel sound, whereas tense vowels typically do not follow in this manner. In addition, tense vowels generally tend to be longer than lax vowels; thus, tense vowels provide more syllabic nucleus informational content due to their longer duration. Transitional vowels in the form of the 'diphthong' set generated better classification performance relative to fixed pure vowels in the 'monophthong' set. Diphthongs are inherently more dynamic than monophthongs due to their transition from one vowel sound to another within a single syllable, as well as being longer in duration than monophthongs (Hockett, 1958). It is worth mentioning that the quantity of training data may have influenced results shown in Fig. 4.

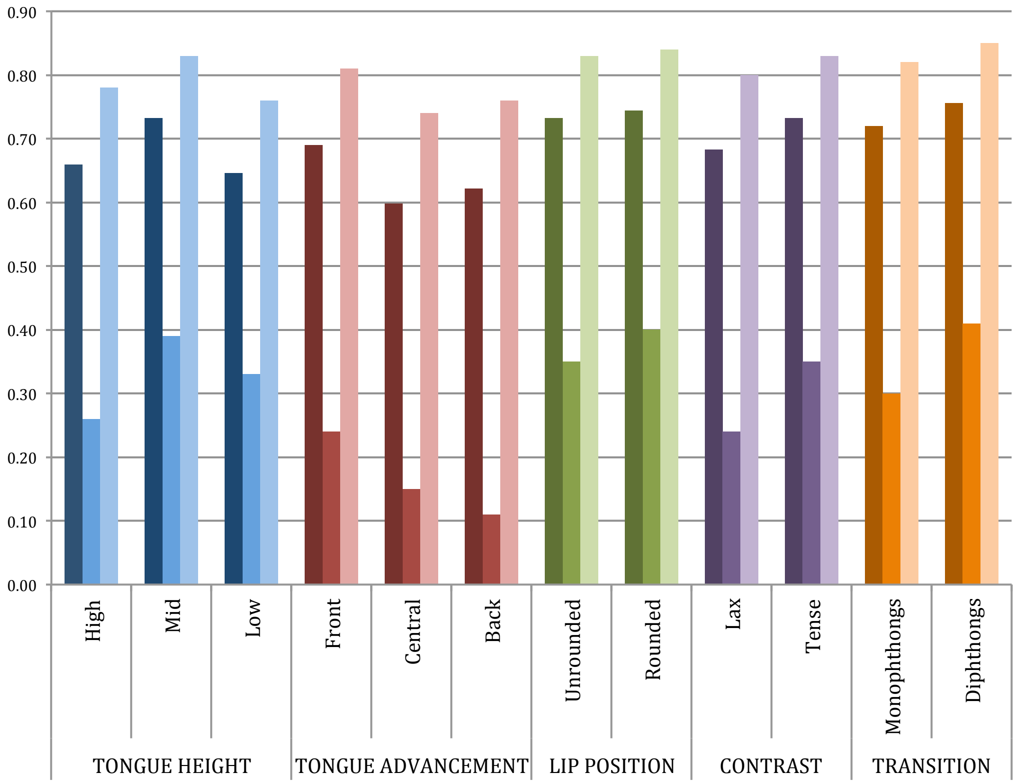

Figure 4: 'Non-Depressed' and 'Depressed' classification results for combined vowel articulatory characteristics (*AC*) features. The colors represent classification accuracy (dark shade); depressed F1 (medium shade); and non-depressed (light shade). Note that apart from the diphthong set, all others included only monophthong fixed pure vowels.

4.2 Articulatory Vowel Parameters Using Linguistic Stress Features

In Section 1, it was hypothesized that the depressed speakers would exhibit hypoarticulation and overall a decrease in linguistic stress. Fig. 5(a) and 5(c) show that the median duration values for depressed speakers are shorter for all vowel sets. Further, Fig. 5(a) shows considerable reductions in depressed speakers' mean duration ranges for 'mid', ‘front’, ‘back’, ‘rounded’, ‘tense’, and ‘monophthong’ vowel sets when compared with the non-depressed speakers. Also, as predicted based on articulatory characteristics, the median lengths for the ‘central’ vowel set were the shortest, whereas the ‘diphthong’ set contained the longest. Fig. 5(b) also shows statistically significant duration standard deviation differences between vowel sets, especially for the ‘mid’, ‘back’, ‘low’, ‘unrounded’, ‘rounded’, ‘tense’, ‘monophthong’, and ‘diphthong’. A surprising result is seen in Fig. 5(a) and Fig. 5(b), for the ‘diphthong’ set, wherein the depressed speakers exhibited a wider range of mean and standard deviation durations than non-depressed speakers. The increase in duration range for depressed speakers is possibly related to the transitioning nature of diphthongs together with the effects of psychomotor retardation, which slow motor planning and reliable execution precision (Mayer-Gross et al., 1969; Kuny & Stassen, 1993; Cannizzaro et al., 2004).

With regard to the loudness mean and standard deviations, shown in Fig. 5(c) and 5(d), again the overall median values were less for depressed speakers than for non-depressed speakers. The loudness mean ranges containing the greatest difference between depressed and non-depressed speakers in Fig. 5(c) were the ‘high’, ‘front’, ‘back’, ‘rounded’, ‘lax’, and ‘diphthong’ vowel sets. With respect to the loudness standard deviations, Fig. 5(d) shows minor differences in ranges for ‘high’, ‘front’, and ‘diphthong’ vowel sets. As indicated by Fig. 5(d), according to paired *t*-tests, the loudness standard deviation differences across vowel parameters were generally not as statistically significant when compared to its mean or loudness. The pitch mean and standard deviation plots were not shown because these also showed little statistical significance for duration and loudness. The overall median for pitch showed a trend of depressed speakers having an increase over non-depressed speakers, which was likely due to having more females in ‘depressed’ group.

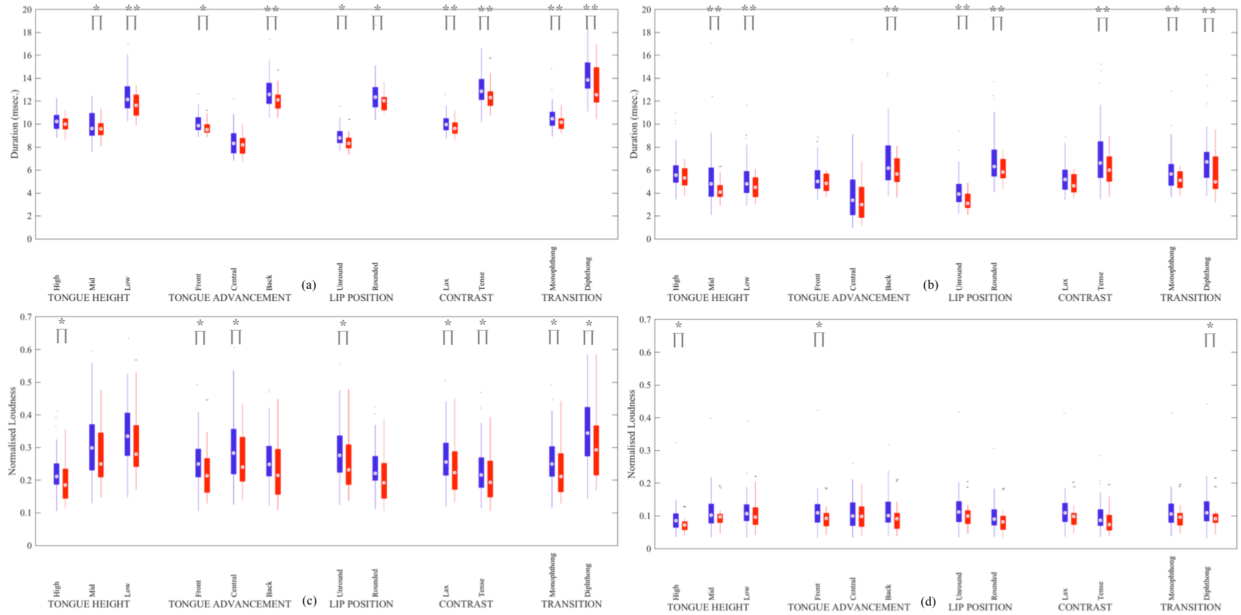

Figure 5: Speaker *mean* distributions for duration (a) and loudness (c) along with *standard deviation* distributions for duration (b) and loudness (d); ‘non-depressed’ (blue) and ‘depressed’ (red). The circle and color bar edges indicate the median and 25th to 75th percentile ranges of each vowel set respectively, whereas the narrower lines indicate the outer ranges and outliers are shown as

dots. A starred and double starred bracket indicates pairs of results that were statistically significantly different using a paired t-test with α = 0.05 and α = 0.01 setting, respectively.

Initially, the good performance shown previously in Fig. 4 for the 'mid' vowel set was difficult to explain because it was believed that the 'high' and 'low' sets would generate better results due to their further proximity from the 'mid' tongue position. It was believed that a greater kinematic effort on the basis of locations further form the 'mid' position location would yield better depression and non-depression discrimination. However, with further linguistic stress duration analysis shown in Fig. 5(b), it is apparent that the 'mid' vowel set duration standard deviation range for depressed speakers is much narrower than their 'high' and 'low' sets. Furthermore, the duration difference between 'mid' for depressed and non-depressed speakers is quite large. This range difference is likely explained due to non-depressed speakers varying 'mid' vowels more with regards to duration and loudness. Moreover, it is possible that depressed speakers having hypoarticulation tendencies to generalize 'mid' sounds to a generic schwa vowel sound than non-depressed speakers do.

Classification results using the linguistic stress (*LS*) features are shown in Table 2. Based on the depressed and non-depressed comparisons shown above in Fig. 5(a) and 5(b), it was expected that the duration *LS* features would perform the best. The duration mean obtained a high of 76.8% classification accuracy with 0.35 (0.86) F1 scores, while the duration standard deviation achieved an *LS* feature performance of 80.5% classification accuracy with 0.50 (0.88) F1 scores. The latter represents the best accuracy and F1 scores amongst all other *LS* feature combinations.

Loudness *LS* features achieved slightly lower classification performance than the duration *LS* features. These results appear to indicate that the combined mean and standard deviation are complementary for both loudness and pitch as each improves when these *LS* features are combined. It should be noted that among the various *LS* features, the variation in performance accuracy and F1 scores is wide; differences of almost 20% absolute for classification accuracy and up to 0.33 absolute for F1 depressed can be seen.

Results in Table 2 broadly concur with previous studies (Ostwald, 1965; Darby & Hollien, 1977; Mundt et al., 2007; 2013) that indicate reductions in overall duration and intensity dynamics in clinically depressed speakers. However, results shown herein provide a more insightful analysis in regards to psychomotor retardation and hypoarticulation in connection with kinematic expectations based on articulatory vowel characteristics. The statistical significance results shown in Fig. 5 across several vowel sets reinforces that vowels with greater kinetic demand carry more discriminant information for depression classification. While Trevino et al. (2011) examined statistical correlations between individual phoneme length/intensity and major depression sub-symptom clinician scores; experiments therein used a much smaller number of speakers that did not extend to evaluate phonetic trends in standard deviation and articulatory vowel parameters for depression classification. Moreover, Trevino et al., (2011) provided minimal articulatory production-based evidence to support why some phonemes performed better than others. To our knowledge, this is possibly one of the first papers to examine and utilize articulatory vowel groupings with linguistic stress components for depressed speech classification. The depressed and non-depressed speaker comparisons based on our newly proposed articulatory vowel sets in Fig. 5 and *LS* feature results show advantages to utilizing articulatory acoustic phoneme parameters, the need for which was proposed by Gábor & Klára (2014).

Table 2: Summary of ‘Depressed’ and ‘Non-Depressed’ classification accuracy results and F1 scores for mean/standard deviation linguistic stress (*LS*) feature combinations. These results are from *LS* features based on using all 12 possible vowel parameters listed previously in Fig. 4 and 6. The total number of *LS* features (feature dimension) is shown in parenthesis.

	Linguistic Stress (*LS*) Features	**%**	**F1 Depressed**	**F1 Non-Depressed**
Mean	Duration (12)	76.8	0.35	0.86
	Loudness (12)	64.6	0.17	0.78
	Pitch (12)	61.0	0.20	0.74
	Duration + Loudness (24)	69.5	0.32	0.80
	Duration + Pitch (24)	75.6	0.38	0.85
	Duration + Loudness + Pitch (36)	68.3	0.28	0.80
Standard Deviation	Duration (12)	**80.5**	**0.50**	**0.88**
	Loudness (12)	69.5	0.26	0.78
	Pitch (12)	69.5	0.32	0.80
	Duration + Loudness (24)	76.8	0.42	0.86
	Duration + Pitch (24)	73.2	0.35	0.85
	Duration + Loudness + Pitch (36)	76.8	0.42	0.86
Mean & Standard Deviation	Duration (12)	78.0	0.44	0.86
	Loudness (12)	69.5	0.24	0.81
	Pitch (12)	74.4	0.32	0.84
	Duration + Loudness (24)	70.7	0.40	0.81
	Duration + Pitch (24)	72.0	0.43	0.82
	Duration + Loudness + Pitch (36)	72.0	0.30	0.82

4.3 Baseline and Articulatory Vowel Parameters with Linguistic Stress

Experiments using the eGeMAPS (all sounds) features together with the *LS* features are shown in Table 3. When compared with the baseline, the duration standard deviation + eGeMAPS combination achieved a classification accuracy improvement of ~4% in absolute terms, with similar F1 scores. Further, in Table 3 (indicated in bold), several other feature combinations also performed slightly better in terms of classification accuracy, suggesting that complementary information is provided by the linguistic stress features. While the combined eGeMAPS and *LS* features only showed a small improvement in classification accuracy over the eGeMAPS baseline, and did not approach the much lower-dimension stand-alone duration *LS* standard deviation accuracy result (80.5%) in Table 2; it can also be observed that the combined feature sets were more consistent in terms of F1 scores.

Table 3: Summary of ‘Depressed’ and ‘Non-Depressed’ classification accuracy results and F1 scores for mean/standard deviation feature combinations per linguistic stress types combined with eGeMAPS. The combined results are based on using eGeMAPS (all sounds) and all 12 vowel parameters listed previously in Fig. 4 and 6.

	Feature Combination	**%**	**F1 Depressed**	**F1 Non-Depressed**
	eGeMAPS (all sounds)	*73.2*	*0.48*	*0.82*
	eGeMAPS (11 vowels)	*72.5*	*0.42*	*0.82*
Mean	eGeMAPS + Duration	74.4	0.43	0.84
	eGeMAPS + Loudness	70.7	0.40	0.81
	eGeMAPS + Pitch	73.2	0.45	0.82
	eGeMAPS + Duration + Loudness + Pitch	74.4	0.46	0.83
Standard Deviation	eGeMAPS + Duration	**76.8**	**0.46**	**0.85**
	eGeMAPS + Loudness	72.0	0.40	0.81
	eGeMAPS + Pitch	74.4	0.46	0.83
	eGeMAPS + Duration + Loudness + Pitch	72.0	0.44	0.81
Mean & Standard Deviation	eGeMAPS + Duration	**74.4**	**0.43**	**0.84**
	eGeMAPS + Loudness	**76.8**	**0.49**	**0.85**
	eGeMAPS + Pitch	68.3	0.38	0.79
	eGeMAPS + Duration + Loudness + Pitch	73.2	0.39	0.83

As noted in Howe et al. (2014) and Stasak et al. (2017a), the quantity and quality of information gathered during clinical depression assessments is greatly dependent on the elicitation method/s used. Experts in the field of speech related depression analysis have yet to agree on which elicitation methods are most discriminative pertaining to depression classification. However, by focusing greater attention on a number of isolated aspects of spoken language, such as the articulatory parameters and linguistic stress found in the study herein, a better understanding concerning the most useful type/s information during clinical assessment can be further discovered.

5. CONCLUSION

In this research, articulatory characteristics and features associated with linguistic stress were evaluated for the automatic analysis of clinically depressed and non-depressed speakers. An analysis of vowel articulatory characteristics using vowel sets based on shared articulatory parameters indicates considerable differences between depressed and non-depressed speakers. In particular, clinically depressed speakers demonstrated statistically significant reductions in duration for the ‘mid’, ‘back’, ‘rounded’, and ‘tense’ vowel sets, when compared with other sets. Further, it can be argued that psychomotor retardation affects a depressed speaker’s articulatory ability, causing hypoarticulation, which influences the degree of his/her linguistic stress. Our experimental results across various vowel sets indicate that depressed speakers have a reduction in linguistic stress duration and loudness components. Moreover, we provided evidence that by utilizing various vowel set linguistic stress components as a compact and interpretable feature set, increases in depression classification can be achieved over baseline approaches. It is believed that knowledge gleaned from vowel sets will be useful for designing clinical elicitation protocols to provide increased discrimination between depressed and non-depressed speakers, which in turn will help to improve automatic diagnosis and monitoring of depression.

6. ACKNOWLEDGEMENTS

The work of Brian Stasak, Julien Epps, and Roland Goecke was partly supported by ARC Discovery Project DP130101094, and Brian Stasak and Julien Epps were partly supported by ARC Linkage Project LP160101360, Data61, CSIRO.

7. REFERENCES

Ani, C., Bazargan, M., Hindman, D., Bell, D., Farooq, M.A., Akhanjee, L., Yemofio, F., Baker, R., & Rodriguez, M., 2008. Depression symptomatology and diagnosis: discordance between patients and physicians in primary care settings, BMC Family Practice, Vol. 9(1).

Alghowinem, S., Goecke, R., Wagner, M., Epps, J., Parker, G., & Breakspear, M., 2013. Characterising depressed speech for classification, InterSpeech 2013, Lyon – France, pp. 2534-2538.

Beckman, M.E., & Pierrehumbert, J.B., 1986. Intonational structure in Japanese and English, Phonology Yearbook, Vol. 3, pp. 255-310.

Bennabi, D., Vandel, P., Papaxanthis, C., Pozzo, T., & Haffen, E., 2013. Psychomotor retardation in depression: a systematic review of diagnosis, pathophysiologic, and therapeutic implications, BioMed Research International, Vol. 2013, pp. 1-18.

Boucher, J., 1976. Articulation in early childhood autism, Journal of Autism and Childhood Schizophrenia, Vol. 6, No. 4.

Buyukdura, J.S., McClintock, S.M., & Croarkin, P.E., 2011. Psychomotor retardation in depression: biological underpinnings, measurement, and treatment, Progress in Neuro-Psychopharmacology & Biological Psychiatry, Vol. 35, pp. 395-409.

Cannizzaro, M., Harel, B., Reilly, N. Chappell, P., & Snyder, P.J., 2004. Voice acoustical measurement of the severity of major depression, Brain and Cognition, Vol. 56, pp. 30-35.

Carterette, E.C., & Jones, M.H., 1974. Informal speech/alphabetic and phonemic texts with statistical analyses and tables, University of California Berkeley, CA – USA.

Chevrie-Muller, C., Sevestre, P., & Seguier, N., 1985. Speech and psychopathology, Language & Speech, Vol. 28, Part I, pp. 57-79.

Cummins, N., Joshi, J., Dhall, A., Sethu, V., Goecke, R., & Epps, J., 2013. Diagnosis of depression by behavioural signals: a multimodal approach, AVEC 2013, Proc. of the 3rd ACM Intern. Workshop on Audio/Visual Emotion Challenge, Barcelona – Spain, pp. 11-20.

Cummins, N., Scherer, S., Krajewski, Schnieder, S., Epps, J., & Quatieri, T.F., 2015. A review of depression and suicide risk assessment using speech analysis, Speech Communication, Vol. 71, pp. 10-49.

Cummins, N., Vlasenko, B., Sagha, H., & Schuller, B., 2017. Enhancing speech-based depression detection through gender dependent vowel-level formant features', Conf. on Artificial Intelligence in Medicine Europe, AIME 2017, pp. 209-214.

Darby, J.K., & Hollien, H., 1977. Vocal and speech patterns of depressive patients, Folia Phoniatrica, Vol. 29, pp. 75-85.

Darby, J.K., Simmons, N., & Berger, P.A., 1984. Speech and voice parameters of depression: a pilot study, J. Commun. Disord., Vol. 17, pp. 75-85.

Dang, T., Stasak, B., Huang, Z., Jayawardena, S., Atcheson, M., Hayat, M., Le, P., Sethu, V., Goecke, R., & Epps, J., 2017. Investigating word affect features and fusion of probabilistic predictions incorporating uncertainty in AVEC 2017, AVEC'17, AVEC'17, Proc. of the 7th Annual Workshop on Audio/Visual Emotion Challenge, Mountain View, CA – USA.

de Jong, K., Beckman, M.E., & Edwards, J., 1993. The interplay between prosodic structure and coarticulation, Lang. Speech, Vol. 36(2), pp. 197-212.

de Jong, K., 1995. The supraglottal articulation of prominence in English: linguistic stress as localized hyperarticulation, J. of the Acoustical Society of America, Vol. 97, pp. 491-504.

de Jong, K., 1998. Stress-related variation in the articulation of coda alveolar stops: flapping revisited, J. of Phonetics, Vol. 26, pp. 283-310.

de Macedo, M.S.F.C., Costa, K.M., & da Silva Filho, M., 2017. Voice disorder in systemic lupus erythematosus", PLoS One, Vol. 12(4).

DeVault, D., Georgila, K., Artstein, R., Morbini, F., Traum, D., Scherer, S., Rizzo, A., & Morency, L., 2013. Verbal indicators of psychological distress in interactive dialogue with a virtual human, in Proc. of the SIGDIAL Conf., pp. 193-202.

DeVault, D., Artstein, R., Benn, G., et al. 2014. SimSensei kiosk: a virtual human interviewer for healthcare decision support, Proc. of the 2014 Intern. Conf. on Auton. Agents and Multi-agent Sys., Paris – France, pp. 1961-1968.

Edwards, E., Nagarajan, S.S., Dalal, S.S., Canolty, R.T., Kirsch, H.E., Barbaro, N.M., & Knight, R.T., 2010. Spatiotemporal imaging of cortical activation during verb generation and picture naming, NeuroImage, Vol. 50(1), pp. 291-301.

Eldred, S.H. & Price, D.B., 1958. A linguistic evaluation of feeling states in psychotherapy, Vol. 21, pp. 115-121.

Ellgring, H., & Scherer, K., 1996. Vocal indicators of mood change in depression, J. Nonverbal Behav., Vol. 20, pp. 83-110.

Engstrand, O., 1988. Articulatory correlates of stress and speaking rate in Swedish CV utterances, J. of the Acoustical Society of America, Vol. 83, pp. 1863-1875.

Eyben, F., Scherer, K., Schuller, B., Sundberg, J., André, E., Busso, C., Devillers, L., Epps, J., Laukka, P., Narayanan, S., and Truong, K., 2016. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing, IEEE Transactions on Affective Computing, Vol. 7(2), pp. 190-202.

Flemming, E., & S. Johnson, 2007. Rosa's roses: reduced vowels in American english. Journal of the International Phonetic Association, Vol. 37, pp. 83-96.

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. Psychiat. Res., Vol. 27, No. 3, pp. 309-319.

Fossati, P., Guilllaume, L., Ergis, A., & Allilaire, J., 2003. Qualitative analysis of verbal fluency in depression, Psychiatry Research, Vol. 117, pp. 17-24.

French, N.R., Carter, C.W., Jr., & Koenig, W., 1930. The words and sounds of telephone conversations, Bell System Tech. J., Vol. 9, pp. 290-324.

Fry, D.B., 1955. Duration and intensity as physical correlates of linguistic stress, J. Acoust. Soc. Amer., Vol. 27, pp. 765.

Fry, D.B., 1958. Experiments in the perception of stress, University of College – London, Vol. 1(2), pp. 126-152.

Fry, D.B., 1965. The dependence of stress judgments on vowel formant structure, in Zwirner, E. and Bethge, W. (eds): Phonetic Sciences, 5th International Congress, Munster, pp. 306-311.

Gábor, K. & Klára, V., 2014. Physiological and cognitive status monitoring on the base of acoustic-phonetic speech parameters, in L. Besacier, A.H. Dediu, & C. Martin-Vide (eds), Statistical Lang. and Speech Processing.

Garnham, A., Shillcock, R.C., Brown, G.D.A., Mill, A.I.D., & Culter, A., 1981. Slips of the tongue in the London-lund corpus of spontaneous conversation, Linguistics, Vol. 19, pp. 805-817.

Gehricke, J.G., & Shapiro, D., 2000. Reduced facial expression and social context in major depression: discrepancies between facial muscle activity and self-reported emotion, Psych. Research, Vol. 95(2), pp. 157-167.

Gong, Y., & Poellabauer, C., 2017. Topic modeling based on multi-modal depression detection, AVEC'17, Proc. of the 7th Annual Workshop on Audio/Visual Emotion Challenge, Mountain View, CA – USA, pp. 69-76.

Gratch, J., Artstein, R., Lucas, G., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., Traum, D., Rizzo, S., & Morency, L., 2014. The distress analysis interview corpus of human and computer interviews, *LREC*, pp. 3123-3128.

Greden, J.F., & Carroll, B.J., 1980. Decrease in speech pause times with treatment of endogenous depression, Biol. Psychiatry, Vol. 15(4), pp. 575-587.

Greden, J.F., 1993. Psychomotor monitoring: a promise being fulfilled?, J. Psychiatr. Res., Vol. 27(3), pp. 285-287.

Hayden, R.E., 1950. The relative frequency of phonemes in general American english", Word, Vol. 6(3), pp. 217-223.

Harel, B., Cannizzaro, M.S., Cohen, H., Reilly, N., & Syder, P., 2004. Acoustic characteristics of Parkinsonian speech: a potential biomarker of early disease progression and treatment, Journal of Neurolinguistics, Vol. 17, pp. 439-453.

Helfer, B.S., Quatieri, T.F., Williamson, J.R., Mehta, D.D., Horwitz, R., & Yu, B., 2013. Classification of depression state based on articulatory precision, Proc. of Interspeech, ISCA, Lyon – France, pp. 2172-2176.

Hirschberg, J., Hjalmarsson, A., & Elhadad, N., 2010. Your as sick as you sound: using computational approaches for modeling speaker state to gauge illness and recovery, A. Neustein (ed.), Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics, Springer Science + Business Media, pp. 305-322.

Hockett, C.F., 1958. A Course in Modern Linguistics, Oxford & IBH Publishing Co., New Delhi – India.

Hollien, H., 1980. Vocal indicators of psychological stress, Annals of the New York Academy of Sciences – Issue of Psychological Stress, pp. 47-72.

Howes, C., Purver, M., & McCabe, R., 2014. Linguistic indicators of severity and progress online text-based therapy for depression, Workshop on Comp. Ling. and Clinical Psych., from Ling. Signal to Clinic. Reality, Baltimore, MD – USA, pp. 7-16.

Jones, D., 1960. An Outline of English Phonetics (9th Ed.), Cambridge University Press, Cambridge – England.

Kent, R.D., & Netsell, R., 1971. Effects of stress contrasts on certain articulatory parameters, Phonetica, Vol. 24, pp. 23-44.

Kent, R.D., Kent, J.F., Weismer, G., Sufit, R.L., Rosenbek, J.C., Martin, R.E., & Brooks, B.R., 1990. Impairment of speech intelligibility in men with amyotrophic lateral sclerosis', Journal of Speech and Hearing Disorders, Vol. 55, pp. 721-728.

Kent, R.D., 2000. Research on speech motor control and its disorders: a review and prospective, J. of Communication Disorders, Vol. 33(5), pp. 391-427.

Kinnunen, T., & Rajan, P., 2013. A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data, Acoustics, Speech and Signal Processing Conf., Vancouver, B.C. – Canada.

Knight, R.A., 2012. Phonetics: A Course Book, Cambridge University Press, Cambridge - England.

Kraepelin, E., 1921. Manic-depressive insanity and paranoia, Livingston, Edinburgh – Scottland.

Kreibig, S.D., 2010. Automatic nervous system activity in emotion: a review, Biol. Psychol., Vol. 84, pp. 394-421.

Kroenke, K., Spitzer, R., & Williams, J., 2001. The PHQ-9: validity of a brief depression severity measure, Gen. Intern. Med., Vol. 16(9), pp. 606-613.

Kroenke, K., Strine, T., Spitzer, R., Williams, J., Berry, J., & Mokdad, A., 2009. The PHQ-8 as a measure of current depression in general population, Journal of Affective Disorders, Vol. 114, pp. 163-173.

Ladefoged, P., 1967. Three Areas of Experimental Phonetics, Oxford University Press, London - England.

Ladefoge, P., 1975. A Course in Phonetics, Harcourt Brace, Orlando – USA.

Leff, J., & Abberton, E., 1981. Vocal pitch measurements in schizophrenia and depression, Psychological Medicine, Vol. 11, pp. 849-852.

Lenneberg, E.H., 1967. Biological Foundations of Language, John Wiley, New York City, NY - USA.

Lindblom, B., 1990. Explaining phonetic variation: a sketch of the H&H theory, Speech Production and Speech Modeling, (ed.) by W.J. Hardcastle and A. Marchal, Kluwer, Dordrecht – The Netherlands, pp. 403-409.

Liu, Z., Liu, F., Kang, H., Li, X., Yan, L., & Wang, T., 2016. Evaluation of depression speech severity in speech, Brain Informatics and Health 2016, pp. 312-321.

Loo, C.K., Sachdev, P., Mitchell, P.B., Gandevia, S.C., Malhi, G.S., Todd, G., et al., 2008. A study using transcranial magnetic stimulation to investigate motor mechanisms in psychomotor retardation in depression, Int. J. Neuropsychopharmocol., Vol. 11(7), pp. 935-946.

Luber, M.P., Hollenberg, J.P., Williams-Russo, P., DiDomenico, T.N., Meyers, B.S., Alexopoulos, G.S., & Charlson, M.E., 2000. 'Diagnosis, treatment, comorbidity, and resource utilization of depressed patients in a general medical practice', Int. J. Psychiatry Med., Vol. 30(1), pp. 1-13.

Mayer-Gross, W., Slater, E. & Roth, M., 1969. Clinical Psychiatry, Bailliere, Tindall and Cassell, London - England.

McKhann, G.M., Knopman, D.S., Chertkow, H., Hyman, B., Jack Jr., C.R., Kawas, C.H., Klunk, W.E., Koroshetz, W.J., Manly, J.J., Mayeux, R., Mohs, R.C., Morris, J.C., Rossor, M.N., Scheltens, P., Carrillo, M.C., Thies, B., Weintraub, S., & Phelps, C.H., 2011. The diagnosis of dementia due to Alzheimer's disease: recommendations from the national institute on aging-Alzheimer's association workgroups on diagnostic guidelines for Alzheimer's disease, Alzheimer's & Dementia, Vol. 7, pp. 263-269.

Miller, G., 1963. Language and Communication, McGraw Hill, New York City, NY - USA.

Mines, A., Hanson, B.F., & Shoup, J.E., 1978. Frequency of occurrence of phonemes in conversational english, Language and Speech, Vol. 21(3), pp. 221-241.

Mitra, V., Shriberg, E., McLaren, M., Kathol, A., Richey, C., Vergyri, D., & Graciarena, M., 2014. The SRI AVEC-2014 evaluation system, Proc. AVEC '14, pp. 93-101.

Moron, J., & Jassem, W., 1965. Acoustic correlates of stress, Language and Speech, London, Vol. 8(3), pp. 159-181.

Moses, P., 1954. The Voice of Neurosis, Grune and Stratton, New York, NY – USA.

Mundt, J.C., Snyder, P.J., Cannizzaro, M.S., Chappie, K., & Geralts, D.S., 2007. 'Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology', J. Neurolinguistics, Vol. 20(1), pp. 50-64.

Mundt, J.C., Vogel, A.P., Feltner, D.E., & Lenderking, W.R., 2012. Vocal acoustic biomarkers of depression severity and treatment response, Biol. Psychiatry, Vol. 72(2), pp. 580-587.

Newman, S., & Mather, V., 1938. Analysis of spoken language of patients with affective disorders, American Journal of Psychiatry, Vol. 94, No. 4, pp. 913-942.

Nilsoone, A., 1987. Acoustic analysis of speech variables during depression and after improvement, Acta. Psychiatr. Scand., Vol. 76, pp. 235-245.

Nisonne, A., 1988. Speech characteristics as indicators of depressive illness, Acta. Psychiatr. Scand., Vol. 77, pp. 253-263.

Ostwald, P.F., 1965. Acoustic methods in psychiatry, Scientific American, Vol. 212, No. 3, pp. 82-92.

Parker, G., & Hadzi-Pavlovic, 1996. Melancholia: a disorder of movement and mood: a phenomenological and neurobiological review, Cambridge University Press, Cambridge: New York, NY – USA.

Ringeval, F., Schuller, B., Valstar, M., Gratch, J., Cowie, R., Scherer, S., Mozgai, S., Cummins, N., Schmitt, M., & Pantic, M., 2017. AVEC 2017 – real-life depression and affect recognition workshop and challenge, AVEC'17, Mountain View, CA – USA, 1-7.

Roy, N., Nissen, S.L., Dromey, C., & Sapir, S., 2009. Articulatory changes in muscle tension dysphonia: evidence of vowel space expansion following manual circumlaryngeal therapy, J. Commun. Disord., Vol. 42, pp. 124-135.

Scherer, K.R. 1986. Vocal affect expression: a review and a model for future research, Psych. Bulletin, Vol. 99, pp. 143-165.

Scherer, S., Pestian, J., & Morency, L., 2013. Investigating the speech characteristics of suicidal adolescents, In: IEEE (ed.), Proceedings of ICASSP, Vancouver – Canada, pp. 709-713.

Scherer, S., Morency, L., Gratch, J., & Petisan, J., 2015. Reduced vowel space is a robust indicator of psychological distress: a cross-corpus analysis, IEEEE Intern. Conf. on Acoustic, Speech and Signal Processing, pp. 4789-4793.

Schwarz, P., Matevjka, P., Burget, L., and Glembek, O., 2006. Phoneme recognizer based on long temporal context, online available: accessed Oct. 2017.

Shah, P.J., Ogilive, A.D., Goodwin, G.M., & Ebmeier, K.P., 1997. Clinical and psychometric correlates of dopamine D2 binding in depression, Psychol. Med., Vol. 27(6), pp. 123-133.

Sobin, C. & Seckbim, H.A., 1997. Psychomotor symptoms of depression, Am. J. Psych., Vol. 154, pp. 4-17.

Soloman, C., Valstar, M.F., Morris, R.K., & Crowe, J., 2015. Objetive methods for reliable detection of concealed depression, Frontiers in ICT, Vol. 2(5), pp. 1-16.

Stasak, B., Epps, J., Cummins, N., & Goecke, R., 2016. An investigation of emotional speech in depression classification, InterSpeech 2016, San Francisco, CA – USA, pp. 485-489.

Stasak, B., Epps, J., & Goecke, R., 2017a. Elicitation design for acoustic depression classification: an investigation of articulation effort, linguistic complexity, and word affect, Proc. InterSpeech 2017, pp. 834-838.

Stasak, B., Epps, J., & Lawson, A., 2017b. Analysis of phonetic markedness and gestural effort measures for acoustic speech-based depression classification, ACII Conf., pp. 1-6.

Stinchfield, S.M., 1933. Speech disorders: a psychoanalytical study of the various defects in speech, New York, NY – USA.

Stolar, M., 2016. Acoustic and conversational speech analysis of depressed adolescents and their parents, PhD thesis School of Engineering, RMIT University, Melbourne – Australia.

Szabadi, E., Bradshaw, C.M., Besson, J.A., 1976. Elongation of pause-time in speech: a simple, objective measure of motor retardation in depression, Br. J. Psychiatry, Vol. 129, pp. 592-597.

Trager, G.L. & Smith, H.L. 1951. An Outline of English Structure Studies in Linguistics, No. 3, Battenberg Press, Norman, OK – USA.

Tremblay, P., Dechamps, I., & Gracco, V.L., 2016. Neurobiology of speech production: a motor control perspective, Neurology of Language, pp. 741-750.

Trevino, A., Quatieri, T., & Malyska, N., 2011. Phonologically-based Biomarkers for major depressive disorder, *EURASIP* Journal on Advances in Signal Processing, Vol. 42.

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 Proc. 6th aCM AVEC'16, Amsterdam – The Netherlands, pp. 3-10.

Voelker, C.H., 1937. A comparative study of investigation of phonetic dispersion in connected American speech, Arch. Neerl. De Phon. Exper., Vol. 13, pp. 138-157.

Widlocher, D.J., 1983. Psychomotor retardation: clinical, theoretical, and psychometric aspects, Psychiatr. Clin. North. Amer., Vol. 6, pp. 27-40.

Williamson, J.R., Godoy, E., Cha, M., Schwarzentruber, A., Khorrami, P., Gwon, Y., Kung, H.T., Dagli, C., & Quatieri, T.F., 2016. Detecting depression using vocal, facial, and semantic communication cues', AVEC'16, Amsterdam – The Netherlands, pp. 11-18.

World Health Organization (WHO), 2017. Depression: let's talk says WHO, as depression tops list of causes of ill health, downloaded news release, March 30th 2017: http://www.who.int/mediacentre/news/releases/2017/world-health-day/en/

Yorkbik, O., Birmaher, B., Axelson, D., Williamson, D.E., & Ryan, N.D., 2014. Clinical characteristics of depressive symptoms in children and adolescents with major depressive disorder", J. Clin. Psychiatry., Vol. 65(12), pp. 1654-1659.