This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2440264, IEEE Transactions on Affective Computing

Self-reported symptoms of depression and PTSD are associated with reduced vowel space in screening interviews

Stefan Scherer, *Member, IEEE,* Gale Lucas, Jonathan Gratch, *Member, IEEE,* Albert Rizzo *Member, IEEE,* and Louis-Philippe Morency, *Member, IEEE*

Abstract—Reduced frequency range in vowel production is a well documented speech characteristic of individuals with psychological and neurological disorders. Affective disorders such as depression and post-traumatic stress disorder (PTSD) are known to influence motor control and in particular speech production. The assessment and documentation of reduced vowel space and reduced expressivity often either rely on subjective assessments or on analysis of speech under constrained laboratory conditions (e.g. sustained vowel production, reading tasks). These constraints render the analysis of such measures expensive and impractical. Within this work, we investigate an automatic unsupervised machine learning based approach to assess a speaker's vowel space. Our experiments are based on recordings of 253 individuals. Symptoms of depression and PTSD are assessed using standard self-assessment questionnaires and their cut-off scores. The experiments show a significantly reduced vowel space in subjects that scored positively on the questionnaires. We show the measure's statistical robustness against varying demographics of individuals and articulation rate. The reduced vowel space for subjects with symptoms of depression can be explained by the common condition of psychomotor retardation influencing articulation and motor control. These findings could potentially support treatment of affective disorders, like depression and PTSD in the future.

Index Terms—Depression, Post-traumatic stress, psychomotor retardation, vowel space, unsupervised learning

1 Introduction

Reduced frequency range in vowel production is a well documented speech characteristic of individuals suffering from psychological and neurological disorders, including but not limited to depression [1], [2], cerebral palsy [3], amyotrophic lateral sclerosis [4], and Parkinson's disease [5]. The assessment and documentation of reduced vowel space often either rely on subjective assessments or on analysis of speech under constrained laboratory conditions (e.g. sustained vowel production, designed reading tasks), rendering analysis impractical and expensive [6]. Such limited and constrained approaches are at present the only ways to assess such acoustic characteristics, that would otherwise be inaccessible to the clinician. Hence, we aim towards an automatic approach to support clinicians and healthcare providers with much needed additional, quantified, and objective measures of nonverbal behavior and in particular voice characteristics to allow for a more informed and objective assessment of an individual's health status [7], [8].

In particular, analysis of acoustic characteristics of speech in depression, an affective disorder that is one of the leading causes of disability worldwide[1], has received considerable attention in the past [9], [10], [11]; a detailed review of speech characteristics of depression and suicidality is provided in [12]. Specifically, prior investigations revealed that depressed patients often display flattened or negative affect [11], [13], reduced speech variability and monotonicity in loudness and pitch [1], [14], [15], [16], reduced speech [17], reduced articulation rate [18], increased pause duration [6], [9], and varied switching pause duration [19]. Further, depressed speech was found to show increased tension in the vocal tract and the vocal folds [1], [20].

In the present work, we aim to extend the existing body of related work and investigate vowel space, a measure of frequency range, extracted from unconstrained conversational speech and its relationship to self-reported symptoms of depression and post-traumatic stress disorder (PTSD); two conditions that have been found to be highly co-morbid [8], [21], [22], [23]. In particular, we investigate an automatic unsupervised machine learning approach to assess a speaker's vowel space - defined as the frequency range spanned by the first and second formant of the vowels /i/ (as in *heed*), /a/ (as in *hod*), and /u/ (as in *who'd*) with respect to the reference sample - within unconstrained screening interviews. Our approach is based on an accurate voiced-speech detector, a robust formant tracker, and a subsequent vector quantization step using the k-means algorithm.

We evaluate the automatically assessed vowel space in experiments with a sample of 253 individuals and show that the novel measure reveals a significantly reduced vowel space in subjects that reported symptoms of depression and PTSD. We show that the measure is robust when analyzing only parts of the full interactions or limited amounts of speech data, which has implications on the approach's practicality. Lastly, we successfully show the measure's statistical robustness against varying demographics of individuals and articulation rate. We are convinced that such a robust automatic measure characterizing an individual's vowel space is viable for manifold applications, without requiring highly constrained recording conditions.

1. http://www.who.int/mediacentre/factsheets/fs369/en/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2440264, IEEE Transactions on Affective Computing

2 Research Hypotheses

Motivated by findings of prior and related work discussed in Section 1, we investigate automatically assessed vowel space and its relationship to reported symptoms of depression and PTSD. Specifically, for this work we identify three research hypotheses:

H1: We hypothesize that the automatically assessed vowel space of subjects with self-reported symptoms of depression or PTSD is significantly reduced, when compared to those of subjects without the respective symptoms. We hypothesize that the vowel space for subjects with depression or PTSD is reduced based on the findings and characterizations of prior work. In particular, psychomotor retardation is hypothesized to have an impact on the individuals' vowel space due to its effect on motor control and speech production.

H2: We hypothesize that our findings for vowel space are robust even when only limited amounts of data are available. Specifically, we investigate the total amount of conversation length and speech time required to significantly discriminate subjects with and without symptoms of depression or PTSD. This investigation is of particular importance when considering practical applications of the approach to characterize an individual's vowel space in the healthcare context.

H3: We further hypothesize that the observed differences are associated with the underlying psychological conditions and the speakers' affective state rather than other factors such as demographics (e.g. gender, education, ethnicity) and articulation rate.

3 Related Work

3.1 Speech Characteristics of Depression and PTSD

As mentioned speech characteristics of depression have been investigated extensively in the past [12]. More specifically, researchers for example investigated the speech characteristics of 13 male subjects with major depressive disorder and six male control subjects [1]. The evaluation and characterization of the speakers were conducted subjectively by one experienced judge following the Mayo clinic dysarthria scale [24], [25]. The Mayo clinic dysarthria scale is a standard 40 item scale which is well documented for a large variety of neurologic disorders and covers a large variety of speech characteristics. The analyzed speech samples (about 30-40 minutes in length) comprise conversational speech, spontaneous monologues, read speech, as well as sustained vowels, and phoneme repetitions. Total scores of subjects on the assessed scale were significantly higher for those with depression than those for healthy controls, both in variance ($p < .01$) and mean ($p < .001$). In particular, the ratings revealed the following characteristics in depressed patients: reduced stress patterns, reduced pitch and intensity, increased harshness of the voice, as well as lack of pitch and intensity variability.

Speech characteristics of psychomotor retardation - a common finding in depression were investigated in [26]. They assessed the speech of 30 depressed subjects, 30 subjects with Parkinson's disease, and 31 control subjects. The subjects were repeatedly asked to read four specifically designed sentences, which formed the basis of analysis for three investigated features, namely second formant transition, voice onset time between consonants and vowels that follow, and spirantization referring to the presence of noise - not attributable to background noise - during closure of the vocal tract. Subjects with depression have significantly reduced second formant transitions ($p \leq .05$), reduced voice onset times ($p < .04$), and increased spirantization ($p = .02$) when compared to healthy controls. Increased spirantization can be perceived as a more breathy voice quality, which somewhat is a contradiction with respect to the previously noted harsher or tenser voice qualities in depression. However, this finding can be explained by the speech task (reading vs. free speech) as well as with the investigation of psychomotor retardation, as a specific symptom of depression. No significant differences were found between subjects with depression and those with Parkinson's disease.

Acoustic spectral measures associated with psychomotor retardation at different time resolutions are investigated [27], [28] in two international challenges to identify depression severity in subject's voice characteristics, namely the Audiovisual Emotion Challenge AVEC 2013 and 2014 [29], [30]. The investigations aimed to exploit changes in coordination across articulators as reflected in observed formant frequencies. Specifically, the authors investigated changes in correlation that occur at different time scales across formant frequencies and also across channels of the delta-mel-cepstrum. The approach is motivated by the observation that auto- and cross-correlations of measured signals can reveal hidden parameters in the stochastic-dynamical systems that generate the signals. The approach was further tuned and extended for AVEC 2014. The phonetic-based features were expanded to include phoneme-dependent pitch dynamics. New coordination-based features were also added, including the correlation between formant frequencies and a cepstral-peak-prominence measure [31], reflecting coordination between articulators and the vocal source. The challenge was won by the researchers with an achieved root mean square error (RMSE) of 8.12 [28].

In our prior work, we complement investigations on acoustic characteristics of individuals reporting symptoms of depression with those reporting symptoms of post-traumatic stress disorder (PTSD). In particular, we investigate characteristics related to voice quality, i.e. the timbre or perceptual coloring of the voice, in individuals with and without the respective symptoms [8], [20], [32]. As in the present work, the conditions were assessed using standard self-assessment questionnaires. Specifically, 18 subjects scored positively for symptoms of depression and 20 scored positively for symptoms of PTSD [20]. A high overlap of the groups was observed. We focused on speaker-independent vocal tract features characterizing the speech on a breathy to tense voice quality dimension [33], [34]. Using this approach, we observed significant differences in the speakers' voice quality and vocal tract source parameters with respect to symptoms of depression and PTSD when compared to control participants. In particular, speakers with symptoms of psychological disorders exhibit more tense voice quali-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2440264, IEEE Transactions on Affective Computing

ties, confirming previous results [1], [16], [26]. For example, participants with symptoms of depression or PTSD show a significantly reduced opening phase of the vocal fold vibration ($p < .02$ for depression and $p = .024$ for PTSD) which is correlated with tenser voice quality, as measured with a novel neural network based approach [35].

In the work closest to the present work, researchers investigated reduced spectral variability using Monte Carlo methods to assess the probabilistic acoustic volume of a speaker and its relationship to depression [2]. The identified acoustic volume was significantly reduced for subjects with depression ($p<0.01$) and was strongly correlated with depression severity. The utilized dataset was the same as in AVEC 2013 [29]. These findings are closely related to the investigated vowel space of the present work that only focuses on the first two formant frequencies and their distribution in a two dimensional frequency space rather than the entire spectrum.

3.2 Vowel Space Assessment in the Literature

Within the present work we focus on vowel space and its relationship to symptoms of depression and PTSD. Unlike the above introduced investigations, vowel space measures themselves have not been directly investigated for depression and PTSD. However, some researchers previously assessed vowel space to characterize other clinical conditions including Parkinson's disease [5] and cerebral palsy [3].

In particular, the vowel space of speakers with Parkinson's disease was compared to that of healthy controls in reading tasks [5]. Thirteen subjects and controls read a passage out loud at three different rates, i.e. habitual, fast, and slow rates. The acoustic characteristics of the vowels /i/, /a/, /u/, and /æ/ were investigated along with those of two fricatives /s/ and /ʃ/. The tokens for each of the investigated vowels and fricatives were manually selected from the recordings and spectrally analyzed. The observed average vowel space for subjects with Parkinson's disease was significantly smaller than that of healthy controls ($p = .019$). Further, the articulation rate of subjects with Parkinson's was significantly higher ($p = .024$).

The reduced vowel space of young adults with cerebral palsy, for example, was investigated with respect to the intelligibility of Mandarin [3]. In that work vowel space is characterized as "an index of the accuracy of vowel articulation, which signifies gross motor control ability of the tongue and jaw coordination", which poses a major challenge for patients with cerebral palsy [36], [37], [38]. The researchers found that vowel space has been significantly reduced for subjects with cerebral palsy when compared to healthy controls ($p < .001$) and directly correlated with word and vowel intelligibility ($r > 0.6$; $p < .005$). The researchers defined the investigated vowel space as the frequency range triangle of the first and second formant, i.e. the resonance frequencies of the vocal tract, spanned by the vowels /i/, /a/, and /u/. Within the present work, we adopt the same definition for consistency. However, here we opt to evaluate vowel space as a ratio between an individual's vowel space and that of a reference sample rather than the actual area as measured in Hz^2 in order to render the method gender independent and more comparable. Gender based vowel space differences have been identified and investigated in the past [39].

4 Materials and Methods

4.1 Ethics Statement

The purpose of this interview study - approved by the University of Southern California Institutional Review Board (UP-11-00342) - is to collect behavioral data that will be used to train computer techniques for recognizing human mental state factors (such as emotion, depressive-tendencies, social anxiety) from behavioral cues such as head nods, head shakes, posture shifts, eye gaze, facial expression, speech prosody and breathing patterns. All participants in this research were treated in accordance with APA guidelines for the ethical treatment of research participants.

4.2 Distress Assessment Interview Corpus

Within this work we utilize the Distress Analysis Interview Corpus (DAIC), a large multimodal collection of semi-structured clinical interviews [23], [40]. These interviews are designed to simulate standard protocols for identifying people at risk for post-traumatic stress disorder (PTSD) or major depression and to elicit nonverbal and verbal behavior indicative of such psychological distress. In order to increase the comparability of behaviors between individuals, we use a virtual human as an interviewer[2]. A virtual human, i.e. a digital graphical representation of a human, in the present work allows for a higher level of control for the administration of stimuli (e.g. asking questions of varying levels of intimacy or acoustic parameters of the interviewer). It is known that with human interviewers accommodation effects or mirroring is persistent in human interactants [41], [42], [43] and could lead to biases in the observed results [44]. The interviews were collected as part of a larger effort named *SimSensei* to create a virtual agent that interviews people and identifies verbal and nonverbal indicators of mental illness [45].

4.2.1 Participants

The DAIC was recorded at the USC Institute for Creative Technologies (ICT). Participants are drawn from two distinct populations: veterans of the U.S. armed forces and U.S. general population. They are coded for depression and PTSD based on accepted psychiatric questionnaires. In total 253 subjects interacted with the automatic SimSensei system. Overall, 186 male subjects and 67 female subjects with an average age of 44.7 (SD = 12.37) years were recorded. Out of the 253 subjects 40.3% have some college education (N = 102), 42.2% are African American (N = 107), and 11.6% are of Hispanic ethnicity (N = 30). The sessions all followed the same general procedure introduced below. On average each conversation lasted for 18.76 minutes with a standard deviation of SD = 7.85.

The sample of subjects consisted of individuals recruited from Craigslist and the direct recruitment of veterans at a US Vets facility in Long Beach. One posting on Craigslist asked

2. Sample interaction between the virtual agent and a human actor can be seen here: http://youtu.be/ejczMs6b1Q4

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2440264, IEEE Transactions on Affective Computing

for participants who had been previously diagnosed with depression, PTSD, or traumatic brain injury, while another asked for any subjects between the ages of 18 and 65. All subjects who met requirements (i.e. age requirements, adequate eyesight) were accepted. Some subjects were connected to a BIOPAC[3] to measure psychophysiological signals.

When participants were asked about their history of particular psychological disorders, 54% reported that they have been diagnosed with depression in their past and 32% reported PTSD. Following the assessment using the self-assessment questionnaires introduced in the following sections, 18.6% scored positive for depression (N = 47; 33 male and 14 female) and 34.6% for PTSD (N = 88; 58 male and 29 female). The self-reported conditions for PTSD and depression are significantly correlated for both the categorical (i.e. positive vs. negative) as well as the score assessments (i.e. assessed severity scores). In particular, the observed categorical correlation is r = .494 ($p < .001$) and the continuous correlation r = .814 ($p < .001$). In our previous investigations we have observed similar correlations between these conditions [23]. Here, it is important to note that the self-assessment questionnaires do not constitute diagnoses of depression or PTSD and that participants who scored positively for both conditions were included in both the PTSD and depression groups. Sample distribution across distress groups are summarized in Table 1.

TABLE 1
Participant distribution based on conditions. All values are numbers of participants.

Distress Group:	Gender	College	African Am.	Hispanic
Depression	33/14	16/31	17/30	1/46
No Depression	153/52	88/117	88/117	28/177
PTSD	58/30	33/55	35/53	11/77
No PTSD	128/37	71/94	70/95	18/147

Sample distribution over distress groups as assessed using self-assessment questionnaires. Gender is reported as *male/female*; College as *has college degree/no college degree*; African American (African Am.) as *is African American/not African American*; and Hispanic as *is Hispanic/not Hispanic*.

4.2.2 Procedure

For the recording of the dataset we adhered to the following procedure: after a short explanation of the study and giving consent, participants were left alone to complete a series of questionnaires at a computer. Questionnaires included the following: basic demographic information, the PTSD Checklist-Civilian version (PCL-C), and the Patient Health Questionnaire depression module (PHQ-9). This process took from 30-60 minutes, depending on the participant.

Upon completion of the questionnaires, the participants were asked to sit down in a chair facing the virtual human interviewer directly, which was displayed on a large 50 inch monitor at about 1.5 meter distance. Within this work we utilize the SimSensei virtual human platform designed to create an engaging interaction through both verbal and nonverbal communicative channels [45]. For verbal processing, the platform integrates modules to recognize spoken words (e.g., using CMU's PocketSphinx recognizer [46]), analyze the spoken responses [47] and decide on the proper response or question using the Flores dialogue manager [48]. For nonverbal processing, acoustic and visual signals (e.g., facial expressions, gaze and voice quality) are automatically recognized using MultiSense[4] before being integrated with the verbal responses in the dialogue manager [49].

The participants are video recorded with an HD webcam (Logitech 720p) and a depth sensor (Microsoft Kinect). A confederate helped the participant set up the head mounted microphone (Sennheiser HSP 4-EW-3) and then the virtual human appeared and proactively started the semi-structured conversation. The audio is recorded at 16 kHz and a 16 bit resolution. The interaction between the participants and the fully automatic virtual human was designed as follows: the virtual human explains the purpose of the interaction and that it will ask a series of questions. It further, tries to build rapport with the participant in the beginning of the interaction with a series of ice-breaker questions about Los Angeles, the location of the recordings. Then a series of more personal questions with varying valence polarity follow. The positive phase included questions like: "What would you say are some of your best qualities?" or "What are some things that usually put you in a good mood?". The negative phase included questions such as: "Do you have disturbing thoughts?" or "What are some things that make you really mad?". Neutral questions included: "How old were you when you enlisted? " or "What did you study at school?". The recordings of the entire interviews are used in the present study. The questions were pre-recorded and animated using the SmartBody architecture [50]. In addition to SmartBody, ICT's Cerebella software automates the generation of physical behaviors for virtual humans, including nonverbal behaviors accompanying the virtual human interaction, responses to perceptual events, as well as listening behaviors [51], [52].

Finally, the participant was asked to complete a final set of questionnaires, which took between 10 and 20 minutes. Participants were then debriefed, paid $35, and escorted out.

4.2.3 Measures

Standard clinical screening measures were used to assess symptoms of PTSD and depression.

Post-traumatic stress disorder checklist-civilian (PCL-C). The PTSD Checklist-Civilian version (PCL-C) [53] is a self-report measure that evaluates all 17 PTSD criteria using a 5-point Likert scale[5]. It is based on the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV; American Psychiatric Association, 1994). Scores range from 17-85, and PTSD severity is reflected in the size of the score, with larger scores indicating greater severity. Sensitivity and specificity are reportedly 0.82 and 0.83, respectively for detecting DSM PTSD diagnoses. The PCL-C is scored based on the DSM-IV schema, with symptomatic responses (moderately or above) to at least six items from three categories. The scores are added to assess the severity of symptoms. The PCL is widely used in PTSD research [54],

3. http://www.biopac.com/

4. http://multicomp.ict.usc.edu
5. http://tinyurl.com/boa6zar

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2440264, IEEE Transactions on Affective Computing

[55]. Within our investigations, we follow the standardized guidelines that at least six of the 17 items in the PCL-C need to be scored at *moderately or above* for an individual to be considered as scoring positively.

Patient Health Questionnaire-Depression 9 (PHQ-9). The Patient Health Questionnaire-Depression 9 (PHQ-9) is a ten-item self-report measure based directly on the diagnostic criteria for major depressive disorder in the DSM-IV [56]. The PHQ-9 is typically used as a screening tool for assisting clinicians in diagnosing depression as well as selecting and monitoring treatment. Further, it has been shown to be a reliable and valid measure of depression severity [57]. Scores range from 0-27, with higher scores indicating higher depression severity. Due to IRB requirements, we used a 9-question PHQ-9 instrument, excluding question 9 about suicidal thoughts. When scoring the PHQ-9, response categories 2-3 (More than half the days or above) are treated as symptomatic and responses 0-1 (Several days or below) as non-symptomatic. At least five of the first eight questions must be checked as symptomatic, including at least one of the first two questions. Additionally, the last question must be checked as at least somewhat difficult[6]. Severity is calculated by totaling the answers to all of the questions. A PHQ-9 score of at least 10 was used to determine a positive assessment, in addition to the previous requirements. PHQ-9 score of at least 10 (i.e. moderate depression) results in a specificity and sensitivity of 88% for depression as reported in [57].

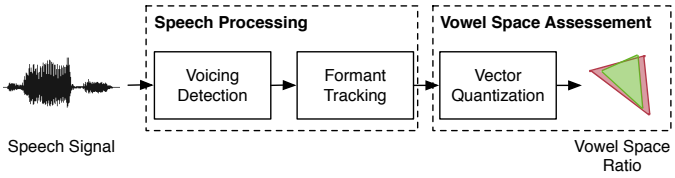

Fig. 1. **Algorithm overview figure.** Basic overview figure of the approach to automatically assess vowel space ratio. The process is separated into two major steps including speech processing (i.e. voicing detection and vowel tracking) and the vowel space assessment (i.e. vector quantization using k-means clustering and vowel space ratio calculation). The output of the algorithm is the ratio between the reference sample vowel space (depicted as a red triangle) and the individual's vowel space (depicted as a green triangle). The larger the ratio the larger the individual's vowel space with respect to the reference.

4.3 Speech Processing and Formant Tracking

For the processing of the speech signals, we use the freely available COVAREP toolbox (v 1.1.0), a collaborative speech analysis repository available for Matlab and Octave [58][7]. COVAREP provides an extensive selection of open-source robust and tested speech processing algorithms enabling comparative and cooperative research within the speech community[8].

In particular, we employ the following steps for speech processing: First, we utilize a robust fundamental frequency tracker and voicing detection algorithm to identify regions of interest for our vowel space analysis [59]. While formants can be tracked throughout unvoiced (i.e. the vocal folds are not vibrating) speech [60], we are primarily interested in the characteristics of the spoken vowels, which are always voiced. Next, based on the identified voiced regions, we track the first two formants F1 and F2 (i.e. the vocal tract resonance frequencies) using a robust formant tracker based on the so-called differential phase spectrum [61]. The first two formants (i.e. the two spectral peaks with the lowest frequencies) of the speech signal are in large responsible for the identification and characterization of vowels [62]. Their formant frequencies are characterized to a large part by the tongue position and the overall shape of the vocal tract producing the vowel. These two steps are applied for each individual speech recording. Below, we describe the approach in more detail. The entire algorithm is shown in Figure 1.

4.3.1 Voicing Detection and Fundamental Frequency Tracking

In [59], a method for fundamental frequency f_0 tracking and simultaneous voicing detection based on residual harmonics is introduced. The method is especially suitable in noisy and unconstrained conditions. The residual signal $r(t)$ is calculated from the speech signal $s(t)$ for each frame using inverse filtering, for all times t. In particular, we utilize a linear predictive coding (LPC) filter of order $p = 12$ estimated for all Hann windowed speech segments. Each speech segment has the length of 25 ms and is shifted by 5 ms. This process removes strong influences of noise and vocal tract resonances. For each $r(t)$ the amplitude spectrum $E(f)$ is computed, revealing peaks for the harmonics of f_0, the fundamental frequency. Then, the summation of residual harmonics (SRH) is computed as follows [59]:

$$SRH(f) = E(f) + \sum_{k=2}^{N_{harm}} [E(k \cdot f) - E((k - \frac{1}{2}) \cdot f)], \quad (1)$$

for $f \in [f_{0,\min}, f_{0,\max}]$, with $f_{0,min} = 50$ and $f_{0,max} = 500$, and $N_{harm} = 5$. The frequency f for which $SRH(f)$ is maximal $f_0 = \arg\max_f(SRH(f))$ is considered the fundamental frequency of the investigated speech frame. By using a simple threshold $\theta = 0.07$, the unvoiced frames can be discarded as in [59].

4.3.2 Formant Tracking

The formant tracker used in this approach is introduced in detail in [61]. Initially, the speech signal $s(t)$ is windowed using a Blackman window and differentiated [63]. Each analyzed speech segment has the size of 25 ms and shifted by 5 ms. For each segment we then remove the phase information. Subsequently we calculate the chirp z-transformation and compute the differential phase spectrum as in [61]. For the differential phase spectrum it has been shown that observable peaks caused by the vocal tract resonance frequencies are more prominent than in the amplitude spectrum, which renders the approach more robust and accurate. Lastly, we identify the peaks within the differential phase spectrum to identify the formants of the observed sample. In particular, we are interested in the first and second formants F1 and F2. In order to remove small

6. http://tinyurl.com/mjddf7r
7. http://covarep.github.io/covarep/
8. The vowel space assessment algorithm presented within this work will be made publicly available within COVAREP after publication.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2440264, IEEE Transactions on Affective Computing

fluctuations we apply a median filter with a filter length n = 15 to the tracked formants. Here it is important to acknowledge that formant tracing in general can be noisy or inaccurate [60], [61], [64]. Hence, we apply a median filter after formant tracking as well as the subsequent vector quantization to allow for a robust and accurate assessment of the vowel space (cf. Figure 1). Formants are tracked for all voiced regions, i.e. not only vowels.

4.3.3 Articulation Rate (manual and automatic)

We further assess articulation rate approximated by number of words spoken per second based on *manual* transcriptions of 95 of the 253 interactions. Due to the high cost of precise manual annotations we only annotated a subset of the conversations. Based on manual transcriptions, we have conducted further evaluations comprising the valence of the spoken words, articulation rate, answer onset timings, and overall answer lengths with respect to the here evaluated psychological conditions [47]. As an exemplary result, we find that speakers that scored positively for depression or PTSD take significantly longer time to respond to the positive question "When was the last time you felt really happy?" than those without. Further, their responses are significantly shorter and include less positively valenced words. Within the present study we only utilize articulation rate to assess its influence on the automatically measured vowel space. In addition, we *automatically* assessed articulation rate within all 253 interactions. For this purpose we utilize an algorithm developed in Praat for the detection of syllable nuclei that relies on an intensity peak detection and subsequent voicing detection algorithm [65]. We observed some strong outliers after automatically extracting articulation rate from all 253 interactions and removed those for which the articulation rate was below one syllable per second, which is far below the expected rate [66]. We report results for both the manual and automatic approach.

4.4 Vowel Space Assessment

Based on the tracked formants F1 and F2 for the voiced regions of speech we compute the vowel space for each recorded subject individually. Figure 2 shows an example of the assessed vowel space for two subjects. In particular, the observed formant frequency pairs (gray dots), the reference vowel space (red triangle), and the subject's vowel space (green triangle) are seen. We define the vowel space, as seen in Figure 2, as the frequency region covered by the triangle in the two dimensional frequency space spanned by F1 and F2 for the vowels /i/ (as in *heed*), /a/ (as in *hod*), and /u/ (as in *who'd*) following [3]. These three vowels represent the vowels with the most extreme positions of the tongue and are therefore located in the extremes of this triangularly shaped two-dimensional frequency space [62], [67] (cf. Figure 2).

As we do not precisely know when the recorded subjects produced these vowels, we propose to apply a vector quantization approach, namely k-means clustering, to identify the prototypical locations of /i/, /a/, and /u/ for each speaker to automatically assess the individual's vowel space [68] (cf. Figure 1). We closely follow a recently proposed approach to automatically identify the vowel space in speech, that has been validated to highly correlate with manual measures of vowel space ($r > .7$) for both male and female speakers [69].

In detail the approach comprises the following steps: (1) To assess an individual's vowel space using k-means, we first initialize the k = 12 cluster centers c_i with $i = 1, \ldots, 12$ with the prototypical formant frequencies of F1 and F2 for the investigated individual's gender as proposed in [70] and reported in Table 2. (2) We adapt the cluster centers c_i based on the observed formant frequencies $x_m \in \mathbb{R}^2$ for the investigated individual using the basic k-means algorithm. The algorithm iteratively minimizes the within cluster sum of squares and yields prototypical locations for all k cluster centers. (3) After optimization we identify the three cluster centers $c_{/i/}$, $c_{/a/}$, and $c_{/u/}$ *closest* to the average formant locations of the vowels /i/, /a/, and /u/ using Euclidian distance, as listed in Table 2. At this point we would like to note that the three cluster centers $c_{/i/}$, $c_{/a/}$, and $c_{/u/}$ are not necessarily located near the formant locations of the vowels /i/, /a/, and /u/. (4) After identifying the cluster centers $c_{/i/}$, $c_{/a/}$, and $c_{/u/}$, we compute the area A of the spanned triangle using Heron's formula $A = \sqrt{s(s-a)(s-b)(s-c)}$ with $s = \frac{a+b+c}{2}$ and a, b, c the lengths of the triangle's three sides. We then compute the vowel space ratio $vs_{ratio} = \frac{A_{ind}}{A_{ref}}$ of the individual's vowel space area A_{ind} and the reference vowel space area A_{ref} to characterize how large the individual's vowel space is to the reference sample vowel space with respect to the individual's corresponding gender. The reported values in Section 5 are vowel space ratios.

5 Results

Here, we report the statistical findings with respect to the three hypotheses stated in Section 2. It is important to note that if participants scored positively for both conditions they were utilized in both the PTSD and Depression groups.

5.1 Psychological Condition Group Differences

We report statistical evaluation results below with M denoting the arithmetic mean. Additionally, we present the p-values of two-tailed t-tests and Hedges' g values as a measure of the effect size. The g value denotes the estimated difference between means of the two samples in magnitudes of standard deviations [71]. Hedges' g is a commonly used standardized mean difference measure that can be transferred into other measures like Cohen's d [72]. The observed mean vowel space measure per condition and the standard errors are visualized in Figure 3. The observed mean values M, standard deviations SD, and Hedges' g are summarized in Table 3.

We first consider differences in vowel space by distress group membership, namely depression and PTSD. Participants categorized as having depression by the PHQ-9 exhibited smaller vowel space than those not categorized as having depression (depressed M = 0.49, non-depressed M = 0.55, t(251) = 2.69, p = .008, Hedges' g = -0.43). Likewise, those categorized as having PTSD by the PCL-C had smaller vowel space than those not categorized as having PTSD (PTSD M = 0.51, non-PTSD M = 0.56, t(251) = 2.55, p = .01, Hedges' g = -0.34).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2440264, IEEE Transactions on Affective Computing

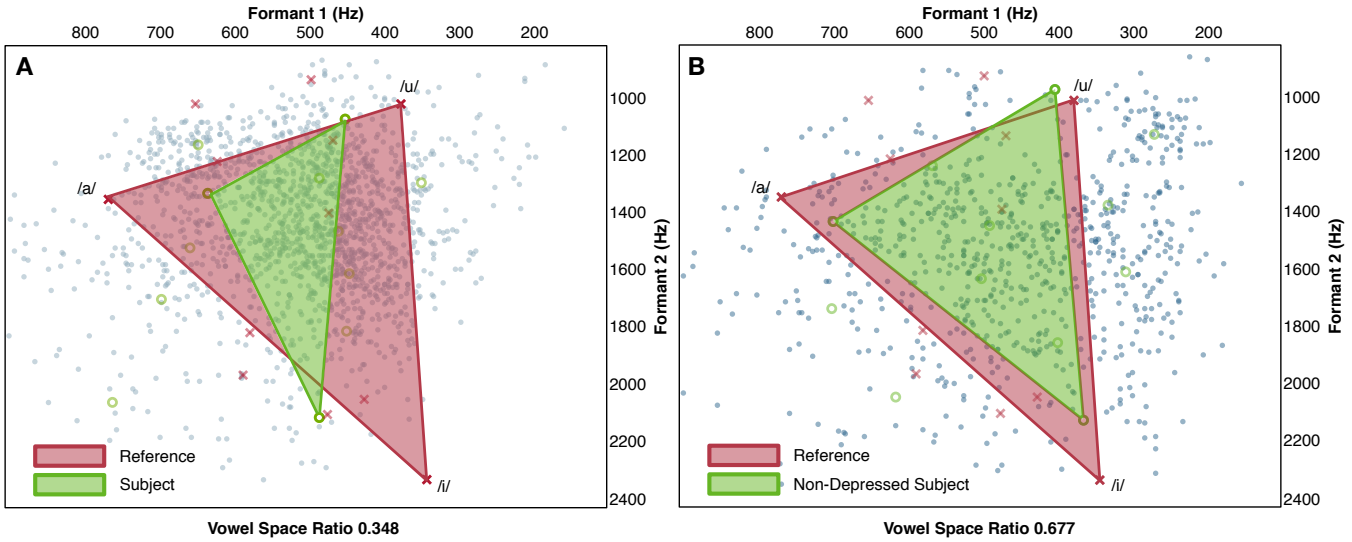

Fig. 2. **Example vowel space assessment for two male subjects.** The male reference sample vowel space (i.e. /i/, /a/, /u/) depicted in red is compared to the subjects' vowel spaces depicted in green, for a subject that scored positively for depression using the self-assessment questionnaires (A) and a subject that scored negatively (B). The vowel spaces are visualized on a two-dimensional plot with Formant 1 on the x-axis and Formant 2 on the y-axis (both in Hz). Additional two-dimensional vowel centers are displayed for both the male reference sample (red x-symbols) and the investigated subjects' vowel space cluster centroids (green circles). The corners of the triangular vowel space for both subjects are determined through minimal distance of cluster centroids to the reference locations of /i/, /a/, and /u/. The grey dots depict all observations of the first two formants across an entire interview. The subject's vowel space scoring positively (A) is visibly smaller than the non-depressed subject's vowel space (B) resulting in a smaller vowel space ratio value.

TABLE 2
Average formant frequencies of F1 and F2 for American English vowels as reported in [67].

Formant	Gender	/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɑ/	/ɔ/	/o/	/ʊ/	/u/	/ʌ/	/ɜ/
F1	**Male**	342	427	476	580	588	768	652	497	469	378	623	474
F1	**Female**	437	483	536	731	669	936	781	555	519	459	753	523
F2	**Male**	2322	2034	2089	1799	1952	1333	997	910	1122	997	1200	1379
F2	**Female**	2761	2365	2530	2058	2349	1551	1136	1035	1225	1105	1426	1588

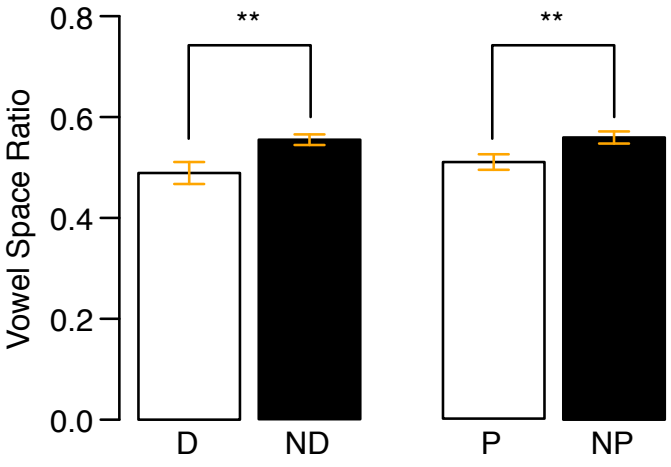

Fig. 3. **Vowel space ratio across conditions.** Observed mean vowel space ratios across conditions depression (D) vs. no-depression (ND) and PTSD (P) vs. no-PTSD (NP). The displayed whiskers signify standard errors and the brackets show significant results with ** ... $p < .01$.

TABLE 3
Distress group evaluation of investigated vowel space measure, F2 interquartile range, and articulation rate.

Distress Group: Feature	M (SD)	M (SD)	**Hedges' g**
	Depression	**No Depression**	
Vowel space	0.49 (0.15)	0.55 (0.15)	**-0.43****
F2 IQR	375.72 (37.26)	381.32 (44.90)	-0.13
Art. rate	3.15 (0.45)	3.09 (0.35)	0.20
Art. rate (auto)	2.74 (0.94)	2.77 (0.87)	-0.03
	PTSD	**No PTSD**	
Vowel space	0.51 (0.14)	0.56 (0.15)	**-0.34****
F2 IQR	376.85 (41.93)	382.07 (44.39)	-0.12
Art. rate	3.06 (0.42)	3.12 (0.34)	0.14
Art. rate (auto)	2.65 (0.93)	2.83 (0.86)	-0.19

Distress group differences with respect to observed acoustic features, namely the proposed vowel space measure, the standard F2 interquartile range (F2 IQR), and articulation rate (Art. rate) both manual and automatic (auto). The arithmetic mean M and the standard deviations SD (in brackets) are shown along with Hedges' g a measure for effect size. ** ... indicate significant difference with p-values $< .01$.

None of the observed differences in vowel space, however, can be explained by articulation rate (both automatically and manually assessed), as articulation rate did not differ based on distress. Participants did not significantly differ in articulation rate based on being categorized with depression (manual: depressed M = 3.15 non-depressed M

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2440264, IEEE Transactions on Affective Computing

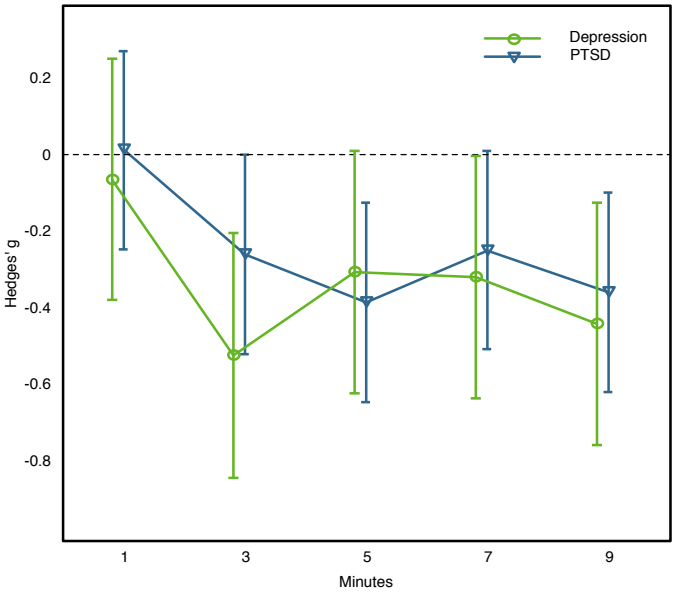

Fig. 4. **Development of effect size Hedges' g over varying intervals of the virtual human interviews.** Observed Hedges' g value for the investigated conditions over the initial intervals of the virtual human interviews, namely after the first, third, fifth, seventh, and ninth minute of the interaction. Hedges' g values are depicted (symbols) with their 95% confidence intervals. Significant differences in the investigated groups are found if the entire confidence interval is below zero for a two-tailed test.

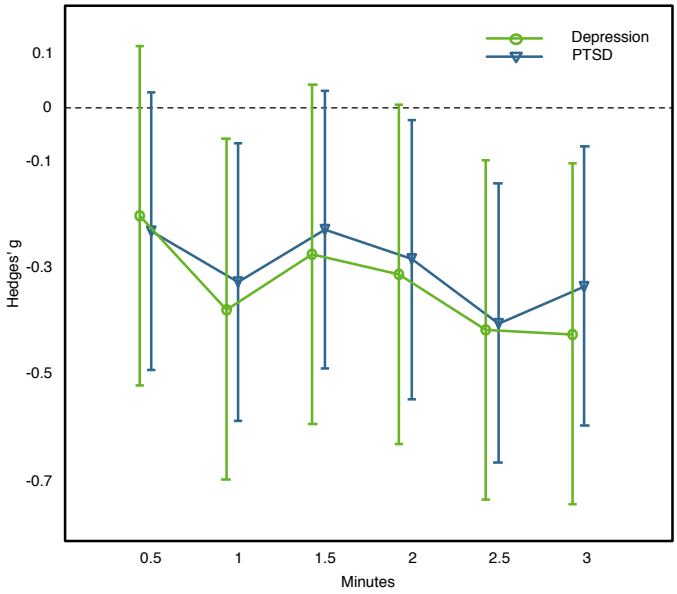

Fig. 5. **Development of effect size Hedges' g over intervals of the individuals' actual voiced speech.** Observed Hedges' g value for the investigated conditions over the initial observed time intervals of the individuals' actual voiced speech, namely 30 sec, 1 min, 1.5 min, 2 min, 2.5 min, and 3 min. Hedges' g values are depicted (symbols) with their 95% confidence intervals. Significant differences in the investigated groups are found if the entire confidence interval is below zero for a two-tailed test.

= 3.09, t(92) = -0.65, p = .52, Hedges' g = 0.20; automatic: depressed M = 2.74 non-depressed M = 2.77 t(188) = 0.21, p = .83, Hedges' g = -0.03) or PTSD (manual: PTSD M = 3.06, non-PTSD M = 3.12, t(92) = 0.80, p = .43, Hedges' g = 0.14; automatic: PTSD M = 2.65 non-PTSD M = 2.83 t(188) = 1.32, p = .19, Hedges' g = -0.19).

While our measure of vowel space differentiated all groups of distressed participants, F2 interquartile distance showed no differences by distress. Participants did not significantly differ in F2 interquartile distance based on being categorized with depression (depressed M = 375.72, non-depressed M = 381.32, t(251) = 0.80, p = .42, Hedges' g = -0.13) or PTSD (PTSD M = 376.85, non-PTSD M = 382.07, t(251) = 0.91, p = .37, Hedges' g = -0.12)

5.2 Temporal analysis of vowel space ratio

We next consider how differences in vowel space by distress group membership are stable over different lengths of speech. The observed effect sizes as measured using Hedges' g and the 95% confidence intervals of g for the different analyzed interaction lengths are shown in Figure 4. First, using repeated measures ANOVA, we examine differences by distress group at different lengths of conversation including the first minute of conversation after the introduction by our virtual human, the first three minutes after this introduction, the first five minutes after, the first seven after, and the first nine after. When depression was entered into this repeated measures ANOVA, beyond the significant main effect (F(1, 250) = 6.86, p = .009) reflecting the difference in vowel space by depression described above, there was a trend for a main effect of conversation length (F(4, 1000) = 1.97, p = .10) such that vowel space peaks at first five minutes of conversation. The difference in vowel space by group membership did not depend on conversation length (interaction F(4, 1000) = 1.60, p = .17). Likewise, when PTSD was entered, analysis revealed a significant main effect of PTSD (F(1, 250) = 5.40, p = .02), and there was a trend for a main effect of conversation length (F(4, 1000) = 1.99, p = .09). However, there was no interaction between PTSD group status and conversation length (F(4, 1000) = 210, p = .08).

Next, we examine differences by distress group at different lengths of actual participants' voiced speech including the first thirty seconds of speech, the first minute of speech, the first minute and a half of speech, the first two minutes of speech, the first two and a half minutes of speech, and the first three minutes of observed voiced speech. The observed effect sizes as measured using Hedges' g and the 95% confidence intervals of g for the varying amounts of actual speech analyzed are shown in Figure 5. When depression was entered into this repeated measures ANOVA it revealed only a significant main effect of depression group (F(1, 250) = 6.04, p = .02), barely a trend for a main effect of speech length (F(4, 1000) = 1.71, p = .13) such that vowel space peaks at one and a half minutes, and no interaction (F(4, 1000) = 0.54, p = .74). When PTSD was entered, we again saw a significant main effect of PTSD (F(1, 250) = 7.31, p = .007), and this time a significant main effect of speech length (F(4, 1000) = 2.22, p = .05), but no interaction (F(4, 1000) = 0.43, p = .83).

5.3 Demographic differences

We next consider differences in vowel space by demographic variables including gender, race, ethnicity, and ed-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2440264, IEEE Transactions on Affective Computing

TABLE 4
Demographic evaluation of investigated vowel space measure, F2 interquartile range, and articulation rate.

Demographics: Feature	M (SD)	M (SD)	**Hedges' g**
	Male	**Female**	
Vowel space	0.55 (0.16)	0.51 (0.12)	0.25
F2 IQR	372.7 (39.66)	401.22 (47.14)	**-0.68*****
Art. rate	3.12 (0.35)	3.04 (0.42)	0.10
Art. rate (auto)	2.79 (0.87)	2.70 (0.95)	0.08
	African Am.	**Other Race**	
Vowel space	0.55 (0.14)	0.54 (0.16)	0.04
F2 IQR	383.85 (42.51)	377.66 (44.23)	0.14
Art. rate	3.06 (0.37)	3.12 (0.37)	0.27
Art. rate (auto)	2.68 (0.91)	2.82 (0.87)	-0.16
	Hispanic	**Other Ethnicity**	
Vowel space	0.57 (0.18)	0.54 (0.15)	0.19
F2 IQR	382.14 (39.70)	380.01 (44.09)	0.05
Art. rate	3.22 (0.34)	3.08 (0.37)	0.32
Art. rate (auto)	2.97 (0.72)	2.73 (0.91)	0.17
	Some College	**No College**	
Vowel space	0.56 (0.14)	0.53 (0.16)	0.17
F2 IQR	386.2 (46.53)	376.97 (40.97)	0.23
Art. rate	3.07 (0.41)	3.12 (0.34)	0.02
Art. rate (auto)	2.79 (0.90)	2.75 (0.88)	0.05

Demographic differences with respect to observed acoustic features, namely the proposed vowel space measure, the standard F2 interquartile range (F2 IQR), and articulation rate (Art. rate) both manual and automatic (auto). African American is abbreviated as *African Am.*. The arithmetic mean M and the standard deviations SD (in brackets) are shown along with Hedges' g a measure for effect size. Significant results are marked with the following symbols: *** ... indicate significant difference with p-values $< .001$.

ucation. We present the p-values of two-tailed t-tests and Hedges' g values as a measure of the effect size.

Within our analysis, male and female vowel spaces are not significantly different (Female M = 0.51, Male M = 0.55, t(251) = -1.73, p = .09, Hedges' g = 0.25). African-American participants do not differ from other races (African-American M = 0.55, other race M = 0.54, t(251) = -0.28, p = .78, Hedges' g = 0.04), nor do participants of Hispanic ethnicity differ from non-Hispanics (Hispanic M = 0.57, non-Hispanic M = 0.54, t(251) = -0.97, p = .33, Hedges' g = 0.19). Additionally, participants with some college education do not differ from those who never attended college (college M = 0.56, no college M = 0.53, t(251) = -1.33, p = .19, Hedges' g = 0.17).

Although vowel space did not tend to differ by these demographic characteristics, F2 interquartile distance showed more differences. Females show a significantly larger F2 interquartile distance than men (Female M = 401.22, Male M = 372.70, t(251) = 4.79, p $< .001$, Hedges' g = -0.68). Like vowel space, African-American participants do not differ from other races in F2 interquartile distance (African-American M = 383.85, other race M = 377.66, t(251) = -1.12, p = .27, Hedges' g = 0.14), nor do participants of Hispanic ethnicity differ from non-Hispanics (Hispanic M = 382.14, non-Hispanic M = 380.01, t(251) = -0.25, p = .81, Hedges' g = 0.05). However, participants with some college education show marginally larger F2 interquartile distance those who never attended college (college M = 386.20, no college M = 376.97, t(251) = -1.82, p = .07, Hedges' g = 0.23).

None of the observed differences in vowel space or F2 interquartile distance, however, are likely due to articulation rate, as articulation rate did not differ based on demographic characteristics (both manual and automatically assessed). Among the 95 participants for whom *manual* transcriptions, and therefore articulation rate as words per second, were available, participants did not significantly differ in articulation rate based on gender (Female M = 3.04, Male M = 3.12, t(92) = -0.82, p = .42, Hedges' g = 0.10), race (African-American M = 3.06, other race M = 3.12, t(92) = 0.81, p = .42, Hedges' g = -0.27), ethnicity (Hispanic M = 3.22, non-Hispanic M = 3.08, t(92) = -1.01, p = .32, Hedges' g = 0.18), or education level (college M = 3.07, no college M = 3.12, t(92) = 0.61, p = .55, Hedges' g = 0.02).

Similarly, participants did not significantly differ in *automatically* assessed articulation rate based on gender (Female M = 2.70, Male M = 2.79, t(188) = -0.58, p = .56, Hedges' g = 0.08), race (African-American M = 2.68, other race M = 2.82, t(188) = 1.09, p = .28, Hedges' g = -0.16), ethnicity (Hispanic M = 2.97, non-Hispanic M = 2.73, t(188) = -1.19, p = .24, Hedges' g = 0.17), or education level (college M = 2.79, no college M = 2.75, t(188) = -0.33, p = .74, Hedges' g = 0.05). The observed mean values M, standard deviations SD, and Hedges' g are summarized in Table 4.

5.4 Interaction of depression and PTSD

We also consider the combined influence of depression and PTSD, as assessed using self-assessment questionnaires, while controlling for demographic characteristics that are related to our dependent variables (gender, college education). A two-way ANOVA revealed a trend for participants in the depressed group to have smaller vowel space (depressed M = 0.49, non-depressed M = 0.55, F(1, 246) = 2.54, p = .11, Hedges' g = -0.20), whereas vowel space did not differ based on PSTD (PTSD M = 0.51, non-PTSD M = 0.53, F(1, 246) = 0.60, p = .44, Hedges' g = -0.09) or the interaction of depression and PTSD (F(1, 246) = 0.02, p = .90, Hedges' g = 0.04). In this analysis, gender was a marginally significant covariate (F(1, 246) = 3.43, p = .07, Hedges' g = 0.24) and there was a trend for college attendance (F(1, 246) = 2.26, p = .13, Hedges' g = 0.19).

Again, the trend for depression to be associated with reduced vowel space cannot be attributed to articulation rate. Articulation rate (manual) did not differ based on depression (depressed M = 3.14, non-depressed M = 3.10, F(1, 59) = 0.05, p = .83, Hedges' g = 0.06), PTSD (PTSD M = 3.13, non-PTSD M = 3.11, F(1, 59) = 0.02, p = .90, Hedges' g = 0.02), their interaction (F(1, 59) = 0.12, p = .74, Hedges' g = 0.09), gender (F(1, 184) = 0.14, p = .72, Hedges' g = 0.09) or college attendance (F(1, 59) = 0.54, p = .46, Hedges' g = 0.19). Using the automatically derived measure of articulation rate, articulation rate did not differ based on depression (depressed M = 2.92, non-depressed M = 2.73, F(1, 184) = 0.83, p = .36, Hedges' g = 0.13), PTSD (PTSD M = 2.66, non-PTSD M = 2.99, F(1, 184) = 2.26, p = .14, Hedges' g = -0.22), their interaction (F(1, 184) = 0.59, p = .44, Hedges' g = 0.11), gender (F(1, 184) = 0.09, p = .76, Hedges' g = 0.06) or college attendance (F(1, 184) = 0.13, p = .73, Hedges' g = 0.06).

While our measure of vowel space showed a trend to differentiate distressed participants, F2 interquartile distance

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2440264, IEEE Transactions on Affective Computing

still showed no differences when both depression and PTSD were considered in the model with demographic covariates. In this model, only gender was a significant predictor of F2 interquartile distance (F(1, 246) = 22.30, $p < .001$, Hedges' g = -0.60). F2 interquartile distance did not differ based on depression (depressed M = 372.02, non-depressed M = 378.72, F(1, 246) = 0.58, p = .45, Hedges' g = -0.09), PTSD (PTSD M = 375.38, non-PTSD M = 375.36, F(1, 246) = 0.00, p = .99, Hedges' g = -0.01), their interaction (F(1, 246) = 1.39, p = .24, Hedges' g = 0.16), or college attendance (F(1, 246) = 0.82, p = .37, Hedges' g = 0.11).

6 Discussion

6.1 Hypothesis 1 - Effect of Psychological Conditions on Vowel Space

Our experiments indeed reveal that the observed vowel space for individuals that scored positively for depression, as categorized by the utilized PHQ-9 questionnaire, are significantly smaller than those that scored negatively (cf. Table 3 and Figure 3). The measure assesses the longitudinal frequency coverage of the first and second formant for an individual in an unconstrained interaction. Specifically, the measure captures the range and extremes of a speaker's vowel articulation and aims to capture assessments of psychomotor retardation, a commonly found symptom of depression and Parkinson's disease [10], [26]. While prior work often focused on the analysis of fundamental frequency, pitch variations, and inflection for the purpose of automatically quantifying lack of expressivity, the present work aims to assess a more holistic measure of vowel articulation over longer periods of time for this purpose [6], [18], [73]. One notable exception is recent work investigating a measure of probabilistic acoustic volume [2]. Similar to the present work the researchers found a reduced volume for individuals that suffer from depression.

Vowel space assessment for the characterization of speech motor control in general has been investigated for various conditions including cerebral palsy [3], amyotrophic lateral sclerosis [4], and Parkinson's disease [5]. However, the present work is the first to automatically identify reduced vowel space in conversational speech for individuals that scored positively for symptoms of depression. While we expect that psychomotor retardation is correlated with the assessed vowel space measure further investigations are required to draw the direct link. Within the present study, we do not have access to diagnosis and expert assessments of psychomotor retardation, which we plan to accomplish in the near future.

As for individuals scoring positively for depression, the vowel space for individuals within our study that scored positively for PTSD are also found to be significantly reduced. This finding can be explained as a characteristic of PTSD or by the high overlap and correlation between conditions of PTSD and depression within the investigated sample. Indeed the comorbidity between PTSD and depression has been previously identified in the literature [21], [22] and the observed strong correlation between conditions has been further discussed in our prior work, where we have identified the more generic condition of *general distress* as a common denominator of the investigated instruments [23]. Overall, speech characteristics of PTSD have been widely understudied in the past, which renders these results interesting and promising. Specifically, future applications for PTSD screening, diagnosis, and symptoms monitoring could highly benefit from our findings if confirmed and verified in subsequent investigations.

In order to confirm our investigations with the DAIC corpus, we analyzed the vowel space measure with two additional datasets of depressed and suicidal speech in a separate study [74]. Specifically, we analyze the AVEC 2013 audio-visual depression corpus (AVEC) read speech portions [29]. We found that the vowel space ratio is again reduced for depressed subjects. While the effect is not significant (depressed M = 0.47, non-depressed M = 0.51, t(66) = 1.12, p = .268, Hedges' g = -0.27), several factors might have influenced the findings: Read speech is articulated differently from conversational speech, reading proficiency might be a confounding factor, and the individuals spoke German. We are planning to further investigate this as the probabilistic acoustic volume was found to be reduced for this sample [2]. In addition, we expanded our investigations to an interview dataset of suicidal and non-suicidal adolescents recorded at the Cincinnati Children's Hospital Medical Center [44]. In fact this separate study reveals that suicidal adolescents showed a reduced vowel space when compared to their non-suicidal peers (suicidal M = 0.36, non-suicidal M = 0.42, t(57) = 2.14, p = .037, Hedges' g = -0.55). This finding is aligned with prior work reporting high comorbidity between PTSD as well as suicidality and depression [21], [75].

Further, we compare our measure of vowel space with the commonly used interquartile range of F2 to assess articulatory motility within longer segments of speech [76], [77]. Overall, we found no significant effect for any of the assessed conditions and F2 interquartile range (cf. Table 3). In general, F2 is gender dependent due to the anatomical differences of the vocal tract length [70], [78], which could have possibly influenced the finding (cf. Table 4). This suggests, that the interquartile range of F2 is less robust against demographic influences and would require subsequent normalization steps. Further, we are convinced that the incorporation of F1 measures into the assessment of the vowel space adds to the robustness and the holistic assessment of articulatory characteristics.

One of the clear benefits of the present approach is the possibility to assess an individual's vowel space in an unconstrained automatic fashion during conversational speech instead of under laboratory conditions, allowing for much needed objective assessments of conversational speech that can be of great benefit to healthcare personnel [6]. For example, tele-health interviews with mental health providers could add such objective acoustic measures to better assess a patient's mental distress and compare it from visit to visit. This is possible because in our evaluation, no specifically designed and structured tasks, such as prolonged articulation of vowels, or specifically designed reading tasks are necessary. In fact conversational speech might reveal clearer voice characteristics of depression than constrained reading tasks as it can be assessed in a less obtrusive and more naturalistic manner.

Our approach could be of advantage for the assessment

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2440264, IEEE Transactions on Affective Computing

of an individual's condition over long periods of time. As this approach allows the assessment of unconstrained conversational speech, the biasing effect of boredom in highly constrained tasks over repeated measures is minimized.

Automatic assessment approaches, as the proposed one, are of additional value in a wide range of speech related research including manifold affective computing applications, as the automatic setup allowed us to assess the vowel space of a much larger sample than commonly investigated with tedious and expensive manual assessments. In total, here we investigate the vowel space of over 250 individuals. Such large scale investigations have a lot of potential to understand the connection between nonverbal behaviors and various affective, psychological, and neurological disorders. Further, the analysis of nonverbal behavior with respect to such disorders is not limited to speech alone. In fact, researchers have started to approach the characterization of depression and other psychological conditions using audiovisual or multimodal behavior quantization approaches [79], [80], [81], [82], [83], [84], [85], [86].

6.2 Hypothesis 2 - Robustness of Vowel Space Measure Based on Limited Data

It is known that formant tracking in general can be noisy or inaccurate [60], [61], [64], however, as proposed here the applied median filter after formant tracking as well as the subsequent vector quantization step allow for a robust and accurate assessment of the vowel space. In order to assess the required amount of data to obtain a robust assessment of vowel space, we investigate the measure on segments of the available conversational data. Within our experiments we found that conversation length is not significantly influencing the discriminative faculty of the vowel space measure for depression. For the condition PTSD we observed a minor interaction with conversation segment length ($p = .08$), which can be explained by a reversal of observed vowel space when analyzing the first minute of conversation only. This underlines the robustness of the measure after only several minutes of interaction. Overall, we see that the vowel space measure stabilizes for each condition within the first five minutes of the conversation. The measure shows significant differences in observed vowel space for both conditions at five minutes of conversation (cf. Figure 4), which corresponds to a quarter the average length (i.e. 18 minutes) of the interactions in this study.

As we cannot guarantee the amount of speech produced by the investigated individual is adequate when segmenting the data based on conversation segments of one or more minutes, we further investigate the observed vowel space measure based on parts of actual individual speech. In particular, we analyze the observed vowel space measure for each individual on 30 seconds up to 3 minutes of actual speech in the interaction. Here, we found that speech length is not significantly influencing the discriminative faculty of the vowel space measure for depression. Again, for the condition PTSD we observed a marginally significant effect of speech length ($p = .05$). Overall, the observed discriminative faculty of the vowel space measure stabilizes around only 2 minutes of actual required speech (cf. Figure 5). This finding suggests that as little as 2-3 minutes of speech per individual is enough to characterize an individual's vowel space robustly, rendering the proposed method valuable for manifold applications, such as distress call center hotlines or mobile health applications monitoring the severity or changes of an individual's psychological condition over time.

The validity of our findings is further supported by prior research, where the automatic measure of vowel space was first validated against actual manual assessments. The researchers could find strong correlations that support the present investigations based on only ten sentences per speaker [69], which resembles a similarly small amount of required data as in the present study.

6.3 Hypothesis 3 - Robustness of Vowel Space Measure with Respect to Demographic Variables and Articulation Rate

Lastly, we investigate the approach's robustness against other factors such as demographics (i.e. gender, race, ethnicity, and education), as well as articulation rate, which can reportedly have an influence on vowel space [67]. Overall, we could not find any significant differences between the automatically assessed vowel space measure and gender, race, ethnicity, or education, which suggests that the measure is quite robust against such factors. In particular, ethnicity, race, and accordingly varying dialects (e.g. African American Vernacular English) have no observed effect on the assessed measure. In fact, the identified cluster centers for the vowels might be dependent on dialect and gender, however, the overall measure of vowel space is not affected by this. The approach using vector quantization and a subsequent ratio calculation allows for a much wanted and needed generalization capability.

In addition, the initialization of the cluster centers (cf. Table 2) renders the approach flexible for future adaptations and investigations. For example, in the present study we initialize the cluster centers based on gender and calculate the vowel space ratio with respect to the reference gender. This allows the evaluated measure to remain gender independent, in contrast to the reference approach (i.e. interquartile range of F2), which is highly gender dependent (cf. Table 4). Further, the approach could easily be extended to other languages, e.g. German formant frequencies [87], or age groups, e.g. average vowel frequencies of children [70].

As suggested by prior work, articulation rate can have a significant impact on the size of the observed vowel space [67]. Our investigations reveal that the vowel space measure negligibly correlates negatively with the articulation rate with $r = -.12$ ($p = .25$). Nor can the effect of observed reduced vowel space for subjects with psychological conditions be explained by articulation rate (cf. Table 3). This suggests that the assessed psychological conditions of the individuals dictate the reduction of the vowel space.

6.4 Why Virtual Human Interviewers?

The investigated unconstrained speech samples in this work are recorded using a fully automatic virtual human interviewer [45]. We chose this approach as virtual humans hold several advantages over their natural counterparts [88]: the involvement and use of virtual humans increases the

available level of control for the investigators or clinical personnel over the assessment process and the presentation of stimuli (e.g. questions with positive or negative affect); the virtual human's behavior can be pre-programmed to the slightest detail and no behavioral bias is introduced into the interview process. This enables comparability between each session, as it reduces contagion effects that have been observed in human-human interaction [19], [32], [44]. Further, findings suggest that virtual humans can reduce the stress and fear associated with the perception of being judged, and thereby, lower emotional barriers to seeking help [88]. Virtual humans have also been studied within the context of schizophrenia, depression, and autism [20], [89], [90], [91].

Another potential benefit of using virtual human interviewers is that researchers may be able to get more, or richer samples of speech than with real human interviewers. Interacting with a virtual human can increase participants' willingness to say more. In particular, an investigation of the effects of framing the character as human-controlled or autonomous showed that participants felt more comfortable disclosing personal information with a character that was framed as autonomous than when it was framed as human-controlled [92], [93]. Specifically, participants reported experiencing lower fear of negative evaluation and engaged in less impression management when the character was framed as autonomous than when it was framed as human-controlled [92], [93]. In fact, actual method of data collection (human-controlled versus automated agent interviews) had no impact on fear of negative evaluation or impression management, but participants who believed they were interacting with human versus computer effected both fear of negative evaluation and impression management.

7 Concluding Remarks

Overall, we showed that the proposed method reveals promising results that are robust against varying factors including demographic variables, articulation rate, as well as only small amounts of data. Our investigations show that the assessed reduced vowel space indeed is associated with conversational speech of individuals with symptoms related to depression or PTSD, as assessed with self-assessment questionnaires. The possibly largest caveat of our investigations is the lack of gold standard clinical assessments of the individuals' psychological conditions, which we are planning to investigate in the near future. We would like to further acknowledge that the proposed measure of vowel space is not specific to depression or PTSD, but should also be investigated in the context of other conditions. Hence, we would like to expand our investigations to conditions such as Parkinson's disease and schizophrenia, for which speech deficits have also been reported [94].

Further, we plan to extend our investigations towards the longitudinal assessment of vowel space in conversational speech. With this we aim towards creating a measure that could be of help in identifying various psychological conditions, affective states, and related symptoms at an early stage or assess therapeutic success for different conditions. We are convinced that automatically assessed vowel space from conversational data could become an essential piece for the objective analysis and assessment by healthcare providers for a wide range of psychological or neurologic conditions.

Acknowledgments

The work depicted here is sponsored by the U.S. Army Research Laboratory (ARL) under contract number W911NF-14-D-0005 and was partially sponsored by DARPA under contract number W911NF-04-D-0005. Statements and opinions expressed and content included do not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

[1] J. K. Darby, N. Simmons, and P. A. Berger, "Speech and voice parameters of depression: a pilot study," *Journal of Communication Disorders*, vol. 17, no. 2, pp. 75–85, 1984.

[2] N. Cummins, S. Vidhyasaharan, J. Epps, and J. Krajewski, "Probabilistic acoustic volume analysis for speech affected by depression," in *Proceedings of Interspeech 2014*, 2014, pp. 1238–1242.

[3] H.-M. Liu, F.-M. Tsao, and P. K. Kuhl, "The effect of reduced vowel working space on speech intelligibility in mandarin-speaking young adults with cerebral palsy," *Journal of the Acoustical Society of America*, vol. 117, no. 6, pp. 3879–3889, 2005.

[4] G. S. Turner, K. Tjaden, and G. Weismer, "The influence of speaking rate on vowel space and speech intelligibility for individuals with amyotrophic lateral sclerosis," *Journal of Speech and Hearing Research*, vol. 38, no. 5, pp. 1001–1013, 1995.

[5] P. A. McRae, K. Tjaden, and B. Schoonings, "Acoustic and perceptual consequences of articulatory rate change in parkinson disease," *Journal of Speech, Language, and Hearing Research*, vol. 45, no. 1, pp. 35–50, 2002.

[6] M. Alpert, E. R. Pouget, and R. R. Silva, "Reflections of depression in acoustic measures of the patient's speech," *Journal of Affective Disorders*, vol. 66, no. 1, pp. 59–69, 2001.

[7] J. Gratch, L.-P. Morency, S. Scherer, G. Stratou, J. Boberg, S. Koenig, T. Adamson, and A. Rizzo, "User-state sensing for virtual health agents and telehealth applications." *Studies in health technology and informatics*, vol. 184, pp. 151–157, 2012.

[8] S. Scherer, G. Stratou, G. Lucas, M. Mahmoud, J. Boberg, J. Gratch, A. Rizzo, and L.-P. Morency, "Automatic audiovisual behavior descriptors for psychological disorder analysis," *Image and Vision Computing Journal, Special Issue on Best of Face and Gesture 2013*, vol. 32, no. 10, pp. 648–658, 2014.

[9] C. Sobin and H. A. Sackheim, "Psychomotor symptoms of depression," *American Journal of Psychiatry*, vol. 154, no. 1, pp. 4–17, 1997.

[10] D. Schrijvers, W. Hulstijn, and B. G. Sabbe, "Psychomotor symptoms in depression: a diagnostic, pathophysiological and therapeutic tool," *Journal of Affective Disorders*, vol. 109, no. 1-2, pp. 1–20, 2008.

[11] L. M. Bylsam, B. H. Morris, and J. Rottenberg, "A meta-analysis of emotional reactivity in major depressive disorder," *Clinical Psychology Review*, vol. 28, pp. 676–691, 2008.

[12] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 17, pp. 10–49, 2015.

[13] J. E. Perez and R. E. Riggio, *Nonverbal social skills and psychopathology*, ser. Nonverbal behavior in clinical settings. Oxford University Press, 2003, pp. 17–44.

[14] A. Nilsonne, "Speech characteristics as indicators of depressive illness," *Acta Psychiatrica Scandinavica*, vol. 77, no. 3, pp. 253–263, 1988.

[15] J. Leff and E. Abberton, "Voice pitch measurements in schizophrenia and depression," *Psychological Medicine*, vol. 11, no. 4, pp. 849–852, 1981.

[16] D. J. France, R. G. Shiavi, S. E. Silverman, M. K. Silverman, and D. M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 7, pp. 829–837, 2000.

[17] J. A. Hall, J. A. Harrigan, and R. Rosenthal, "Nonverbal behavior in clinician-patient interaction," *Applied and Preventive Psychology*, vol. 4, no. 1, pp. 21–37, 1995.

[18] M. Cannizzaro, B. Harel, N. Reilly, P. Chappell, and P. J. Snyder, "Voice acoustical measurement of the severity of major depression," *Brain and Cognition*, vol. 56, no. 1, pp. 30–35, 2004.

[19] Y. Yang, C. Fairbairn, and J. F. Cohn, "Detecting depression severity from vocal prosody," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 142–150, 2013.

[20] S. Scherer, G. Stratou, J. Gratch, and L.-P. Morency, "Investigating voice quality as a speaker-independent indicator of depression and ptsd," in *Proceedings of Interspeech 2013*. ISCA, 2013, pp. 847–851.

[21] D. Campbell, B. Felker, C.-F. Liu, E. Yano, J. Kirchner, D. Chan, L. Rubenstein, and E. Chaney, "Prevalence of depression-ptsd comorbidity: Implications for clinical practice guidelines and primary care-based interventions," *Journal of General Internal Medicine*, vol. 22, pp. 711–718, 2007, 10.1007/s11606-006-0101-4. [Online]. Available: http://dx.doi.org/10.1007/s11606-006-0101-4

[22] G. N. Marshall, T. L. Schell, and J. N. V. Miles, "All ptsd symptoms are highly associated with general distress: ramifications for the dysphoria symptom cluster," *Journal of Abnormal Psychology*, vol. 119, no. 1, pp. 126–135, 2010.

[23] S. Scherer, G. Stratou, M. Mahmoud, J. Boberg, J. Gratch, A. Rizzo, and L.-P. Morency, "Automatic behavior descriptors for psychological disorder analysis," in *Proceedings of IEEE Conference on Automatic Face and Gesture Recognition*. IEEE, 2013.

[24] F. L. Darley, A. E. Aronson, and J. R. Brown, "Differential diagnostic patterns of dysarthria," *Journal of Speech and Hearing Research*, vol. 12, pp. 246–269, 1969.

[25] ——, "Clusters of deviant speech dimensions in the dysarthrias," *Journal of Speech and Hearing Research*, vol. 12, pp. 462–496, 1969.

[26] A. J. Flint, S. E. Black, I. Campbell-Taylor, G. F. G. Gailey, and C. Levinton, "Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression," *Journal of Psychiatric Research*, vol. 27, no. 3, pp. 309–319, 1993.

[27] J. R. Williamson, T. F. Quatieri, B. S. Helfer, R. Horwitz, B. Yu, and D. D. Mehta, "Vocal biomarkers of depression based on motor incoordination," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, ser. AVEC '13. ACM, 2013, pp. 41–48. [Online]. Available: http://doi.acm.org/10.1145/2512530.2512531

[28] J. Williamson, T. Quatieri, B. Helfer, G. Ciccarelli, and D. Mehta, "Vocal and facial biomarkers of depression based on motor incoordination and timing," in *Proceedings of the 4th ACM International Workshop on Audio/Visual Emotion Challenge (AVEC '14)*, 2014, pp. 65–72.

[29] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "Avec 2013 - the continuous audio / visual emotion and depression recognition challenge," in *Proceedings of ACM international workshop on Audio/visual emotion challenge*, 2013, pp. 3–10.

[30] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "Avec 2014: 3d dimensional affect and depression recognition challenge," in *Proceedings of the 4th ACM International Workshop on Audio/Visual Emotion Challenge (AVEC '14)*, 2014, pp. 3–10.

[31] Y. Heman-Ackah, R. Heuer, D. Michael, R. Ostrowski, M. Horman, M. Baroody, J. Hillenbrand, and R. Sataloff, "Cepstral peak prominence: A more reliable measure of dysphonia," *Annals of Otology, Rhinology and Laryngology*, vol. 112, pp. 324–333, 2003.

[32] S. Scherer, Z. Hammal, Y. Yang, L.-P. Morency, and J. F. Cohn, "Dyadic behavior analysis in depression severity assessment interviews," in *Proceedings of International Conference on Multimodal Interaction*. ACM, 2014.

[33] C. Gobl and A. Ní Chasaide, "Acoustic characteristics of voice quality," *Speech Communication*, vol. 11, pp. 481–490, 1992.

[34] S. Scherer, J. Kane, C. Gobl, and F. Schwenker, "Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification," *Computer Speech and Language*, vol. 27, no. 1, pp. 263–287, 2013.

[35] J. Kane, S. Scherer, L.-P. Morency, and C. Gobl, "A comparative study of glottal open quotient estimation techniques," in *Proceedings of Interspeech 2013*. ISCA, 2013, pp. 1658–1662.

[36] R. D. Kent, R. Netsell, and J. H. Abbs, "Articulatory abnormalities in athetoid cerebral palsy," *Journal of Speech and Hearing Disorders*, vol. 43, pp. 353–373, 1978.

[37] L. J. Platt, G. Andrews, M. Young, and P. T. Quinn, "Dysarthria of adult cerebral palsy. i. intelligibility and articulatory impairment," *Journal of Speech and Hearing Research*, 1980.

[38] K. Yorkston, D. R. Beukelmann, and K. R. Bell, *Clinical Management of Dysarthric Speakers*. Pro ed, 1988.

[39] A. P. Simpson and C. Ericsdotter, "Sex-specific differences in f0 and vowel space," in *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS)*, J. Trouvain and W. J. Barry, Eds., 2007, pp. 933–936.

[40] J. Gratch, R. Artstein, G. Lucas, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, A. Rizzo, and L.-P. Morency, "The distress analysis interview corpus of human and computer interviews," in *Proceedings of Language Resources and Evaluation Conference (LREC)*, 2014, pp. 3123–3128.

[41] C. De Looze, S. Scherer, B. Vaughan, and N. Campbell, "Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction," *Speech Communication*, vol. 58, pp. 11–34, 2014.

[42] R. Levitan and J. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions," in *Proceedings of Interspeech 2011*. ISCA, 2011, pp. 3081–3084.

[43] C. A. Shepard, H. Giles, and B. A. Le Poired, *Communication Accommodation Theory*, P. Robinson and H. Giles, Eds. Wiley, 2001.

[44] S. Scherer, J. P. Pestian, and L.-P. Morency, "Investigating the speech characteristics of suicidal adolescents," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2013, pp. 709–713.

[45] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgilia, J. Gratch, A. Hartholt, M. Lhommet, G. Lucas, S. Marsella, F. Morbini, A. Nazarian, S. Scherer, G. Stratou, A. Suri, D. Traum, R. Wood, Y. Xu, A. Rizzo, and L.-P. Morency, "Simsensei kiosk: A virtual human interviewer for healthcare decision support," in *Proceedings of Autonomous Agents and Multiagent Systems (AAMAS)*, 2014, pp. 1061–1068.

[46] H. D. Daines, M. Kumar, A. Chan, A. W. Black, M. Ravishankar, and A. I. Rudnicky, "Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices," in *Proceedings of Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 2006.

[47] D. DeVault, K. Georgilia, R. Artstein, F. Morbini, D. Traum, S. Scherer, A. Rizzo, and L.-P. Morency, "Verbal indicators of psychological distress in interactive dialogue with a virtual human," in *Proceedings of SigDial 2013*. Association for Computational Linguistics, 2013, pp. 193–202.

[48] F. Morbini, D. DeVault, K. Sagae, J. Gerten, A. Nazarian, and D. Traum, "FLoReS: a forward looking, reward seeking, dialogue manager," in *Proceedings of 4th International Workshop on Spoken Dialog Systems*, 2012.

[49] S. Scherer, S. Marsella, G. Stratou, Y. Xu, F. Morbini, A. Egan, A. Rizzo, and L.-P. Morency, "Perception markup language: Towards a standardized representation of perceived nonverbal behaviors," in *Proceedings of Intelligent Virtual Agents (IVA'12)*, ser. LNAI 7502. Springer, 2012, pp. 455–463.

[50] M. Thiebaux, S. Marsella, A. N. Marshall, and M. Kallmann, "Smartbody: behavior realization for embodied conversational agents," in *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems - Volume 1*, ser. AAMAS '08. International Foundation for Autonomous Agents and Multiagent Systems, 2008, pp. 151–158.

[51] S. Marsella, A. Shapiro, A. Feng, Y. Xu, M. Lhommet, and S. Scherer, "Towards higher quality character performance in previz," in *Proceedings of Digital Production Symposium 2013 (DigiPro2013)*. ACM, 2013, pp. 31–35.

[52] S. Marsella, Y. Xu, M. Lhommet, A. Feng, S. Scherer, and A. Shapiro, "Virtual character performance from speech," in *Proceedings of Symposium on Computer Animation 2013 (SCA2013)*. ACM, 2013, pp. 25–35.

[53] E. B. Blanchard, J. Jones-Alexander, T. Buckley, and C. Forneris, "Psychometric properties of the ptsd checklist (pcl)," *Behaviour Research and Therapy*, vol. 34, no. 8, pp. 669–673, 1996.

[54] E. E. Bolton, M. J. Gray, and B. T. Litz, "A cross-lagged analysis of the relationship between symptoms of ptsd and retrospective reports of exposure," *Journal of Anxiety Disorders*, vol. 20, no. 7, pp. 877–895, 2006.

[55] C. W. Hoge, C. A. Castro, S. C. Messer, D. McGurk, D. I. Cotting, and R. L. Koffman, "Combat duty in iraq and afghanistan, mental health problems, and barriers to care," *New England Journal of Medicine*, vol. 351, no. 1, pp. 13–22, 2004.

[56] K. Kroenke and R. L. Spitzer, "The phq-9: A new depression and diagnostic severity measure," *Psychiatric Annals*, vol. 32, pp. 509–521, 2002.

[57] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, "The phq-9," *Journal of General Internal Medicine*, vol. 16, no. 9, pp. 606–613, 2001.

[58] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarep - a collaborative voice analysis repository for speech technologies," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014)*, 2014, pp. 960–964.

[59] T. Drugman and A. Abeer, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Proceedings of Interspeech 2011*. ISCA, 2011, pp. 1973–1976.

[60] D. D. Mehta, D. Rudoy, and P. J. Wolfe, "Karma: Kalman-based autoregressive moving average modeling and inference for formant and antiformant tracking," *The Journal of the Acoustical Society of America*, vol. 132, no. 3, pp. 1732–1746, 2012.

[61] B. Bozkurt, B. Doval, C. d'Alessandro, and T. Dutoit, "Improved differential phase spectrum processing for formant tracking," *Proceedings of Interspeech - ICSLP*, pp. 2421–2424, 2004.

[62] P. Ladefoged, *Elements of Acoustic Phonetics*, 2nd ed. Chicago University Press, 1996.

[63] R. B. Blackman and J. W. Tukey, *Particular Pairs of Windows*, ser. The Measurement of Power Spectra, From the Point of View of Communications Engineering. Dover, 1959, pp. 98–99.

[64] K. Mustafa and I. C. Bruce, "Robust formant tracking for continuous speech with speaker variability," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 435–444, 2006.

[65] N. H. De Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behavior research methods*, vol. 41, no. 2, pp. 385–390, 2009.

[66] T. H. Crystal and A. S. House, "Articulation rate and the duration of syllables and stress groups in connected speech," *Journal of The Acoustical Society of America*, vol. 88, no. 1, pp. 101–112, 1990.

[67] B. Lindblom, "Explaining phonetic variation: A sketch of the h&h theory," *Speech Production and Speech Modeling*, pp. 403–439, 1990.

[68] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, 1967, pp. 281–297.

[69] S. Sandoval, V. Berisha, R. L. Utianski, J. M. Liss, and A. Spanias, "Automatic assessment of vowel space area," *Journal of the Acoustical Society of America*, vol. 134, no. 5, pp. 477–483, 2013.

[70] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic characteristics of american english vowels," *Journal of the Acoustical Society of America*, vol. 97, no. 5, pp. 3099–3111, 1995.

[71] L. V. Hedges, "Distribution theory for glass's estimator of effect size and related estimators," *Journal of Educational Statistics*, vol. 6, no. 2, pp. 107–128, 1981.

[72] J. A. Durlak, "How to select, calculate, and interpret effect sizes," *Journal of Pediatric Psychology*, vol. 34, no. 9, pp. 917–928, 2009.

[73] H. H. Stassen, G. Bomben, and E. Günther, "Speech characteristics in depression," *Psychopathology*, vol. 24, pp. 88–105, 1991.

[74] S. Scherer, L.-P. Morency, J. Gratch, and J. P. Pestian, "Reduced vowel space is a robust indicator of psychological distress: A cross-corpus analysis," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015.

[75] M. K. Nock, I. Hwang, N. A. Sampson, and R. C. Kessler, "Mental disorders, comorbidity and suicidal behavior: Results from the national comorbidity survey replication," *Molecular Psychiatry*, vol. 15, pp. 868–876, 2010.

[76] Y. Yunusova, G. Weismer, R. D. Kent, and N. M. Rusche, "Breath-group intelligibility in dysarthria characteristics and underlying correlates," *Journal of Speech, Language, and Hearing Research*, vol. 48, pp. 1294–1310, 2005.

[77] K. Tjaden and G. E. Wilding, "Rate and loudness manipulations in dysarthria: Acoustic and perceptual findings," *Journal of Speech, Language, and Hearing Research*, vol. 47, pp. 766–783, 2004.

[78] W. T. Fitch and J. Giedd, "Morphology and development of the human vocal tract: A study using magnetic resonance imaging," *Journal of the Acoustical Society of America*, vol. 106, no. 3, pp. 1511–1522, 1999.

[79] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Ying, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. De la Torre, "Detecting depression from facial actions and vocal prosody," in *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, 2009, pp. 1–7.

[80] S. Scherer, G. Stratou, and L.-P. Morency, "Audiovisual behavior descriptors for depression assessment," in *Proceedings of International Conference on Multimdoal Interaction*. ACM, 2013.

[81] Y. Zhou, S. Scherer, D. Devault, J. Gratch, G. Stratou, L.-P. Morency, and J. Cassell, "Multimodal prediction of psychological disorder: Learning nonverbal commonality in adjacency pairs," in *Proceedings of Workshop Series on the Semantics and Pragmatics of Dialogue*, 2013.

[82] J. Joshi, R. Goecke, S. Alghowinem, A. Dhall, M. Wagner, J. R. Epps, G. Parker, and M. Breakspear, "Multimodal assistive technologies for depression diagnosis and monitoring," *Journal on Multimodal User Interfaces*, vol. 7, pp. 217–228, 2013.

[83] V. Venek, S. Scherer, L.-P. Morency, A. Rizzo, and J. P. Pestian, "Adolescent suicidal risk assessment in clinician-patient interaction: A study of verbal and acoustic behaviors," in *Proceedings of Spoken Langugae Technologies (SLT)*, 2014.

[84] M. Chatterjee, G. Stratou, S. Scherer, and L.-P. Morency, "Context-based signal descriptors of heart-rate variability for anxiety assessment," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2014, pp. 3659–3663.

[85] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. M. Mavadati, Z. Hammal, and D. P. Rosenwald, "Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses," *Image and Vision Computing Journal*, 2014.

[86] G. Stratou, S. Scherer, J. Gratch, and L.-P. Morency, "Automatic nonverbal behavior indicators of depression and ptsd: the effect of gender," *Journal on Multimodal User Interfaces*, pp. 1–13, 2014.

[87] M. Pätzold and A. P. Simpson, "Acoustic analysis of German vowels in the *Kiel Corpus of Read Speech*," in *The Kiel Corpus of Read/Spontaneous Speech — Acoustic data base, processing tools and analysis results*, ser. Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel (AIPUK) 32, A. P. Simpson, K. J. Kohler, and T. Rettstadt, Eds., 1997, pp. 215–247.

[88] J. Hart, J. Gratch, and S. Marsella, *How Virtual Reality Training Can Win Friends and Influence People*, ser. Human Factors in Defence. Ashgate, 2013, ch. 21, pp. 235–249.

[89] J. Ku, H. J. Jang, K. U. Kim, S. H. Park, J. J. Kim, C. H. Kim, S. W. Nam, I. Y. Kim, and S. I. Kim, "Pilot study for assessing the behaviors of patients with schizophrenia towards a virtual avatar," *CyberPsychology & Behavior*, vol. 9, no. 5, pp. 531–539, 2006.

[90] B. Robins, K. Dautenhahn, R. Te Boekhorst, and A. Billard, "Robotic assistants in therapy and education of children with autism: Can a small humanoid robot help encourage social interaction skills?" *Universal Access in the Information Society*, vol. 4, no. 2, pp. 105–120, 2005.

[91] E. Kim, J. Ku, J.-J. Kim, H. Lee, K. Han, S. I. Kim, and H.-S. Cho, "Nonverbal social behaviors of patients with bipolar mania during interactions with virtual humans," *The Journal of nervous and mental disease*, vol. 197, no. 6, pp. 412–418, 2009.

[92] J. Gratch, G. M. Lucas, A. King, and L.-P. Morency, "It's only a computer: The impact of human-agent interaction in clinical interviews," in *Proceedings of the 13th annual conference on Autonomous Agents and Multi-Agent Systems*, 2014, pp. 85–92.

[93] G. Lucas, J. Gratch, A. King, and L.-P. Morency, "It's only a computer: Virtual humans increase willingness to disclose," *Computers in Human Behavior*, vol. 37, pp. 94–100, 2014.

[94] A. Cohen, M. Alpert, T. Nienow, T. Dinzeo, and N. Docherty, "Computerized measurement of negative symptoms in schizophrenia," *Journal of psychiatric research*, vol. 42, pp. 827–836, 2008.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2440264, IEEE Transactions on Affective Computing

Stefan Scherer (https://schererstefan.net) is a Research Assistant Professor in the Department of Computer Science at the University of Southern California (USC) and Research Associate at the USC Institute for Creative Technologies where he leads research projects funded by NSF and ARL. He received the degree of Dr. rer. nat. from the faculty of Engineering and Computer Science at Ulm University in Germany with the grade summa cum laude (i.e. with distinction). His research aims to automatically identify, characterize, model, and synthesize individuals' multimodal nonverbal behavior within both human-machine as well as machine-mediated human-human interaction.

Albert "Skip" Rizzo is a clinical psychologist and Director of Medical Virtual Reality at the University of Southern California Institute for Creative Technologies. He is also a research professor with the USC Dept. of Psychiatry and at the USC Davis School of Gerontology. Skip conducts research on the design, development and evaluation of Virtual Reality systems targeting the areas of clinical assessment, treatment and rehabilitation across the domains of psychological, cognitive and motor functioning in both healthy and clinical populations. This work has focused on PTSD, TBI, Autism, ADHD, Alzheimer's disease, stroke and other clinical conditions. In spite of the diversity of these clinical R&D areas, the common thread that drives all of his work with digital technologies involves the study of how interactive and immersive Virtual Reality simulations can be usefully applied to address human healthcare needs beyond what is possible with traditional tools and methods. In 2010, he received the American Psychological Association Award for Outstanding Contributions to the Practice of Trauma Psychology for his R&D work on VR exposure therapy and in 2015 he received the Society for Brain Mapping and Therapeutics *Pioneer in Medicine* award.

Gale M. Lucas received her B.A. from Willamette University, a small liberal arts college in her home state of Oregon. She went on to earn her Ph.D. in social psychology from Northwestern University. After completing her doctorate, Lucas spent two years teaching at small liberal arts universities and subsequently worked as a post-doc at the USC Marshall School of Business. She transferred over to USCs Institute for Creative Technologies to continue her post-doctoral work with Jon Gratch in the Virtual Humans Group, and now works for the group as a Senior Research Associate.

Jonathan Gratch (http://www.ict.usc.edu/~gratch) is Director for Virtual Human Research at the University of Southern Californias (USC) Institute for Creative Technologies, a Research Full Professor of Computer Science and Psychology at USC and director of USCs Computational Emotion Group. He completed his Ph.D. in Computer Science at the University of Illinois in Urban-Champaign in 1995. Dr. Gratchs research focuses on computational models of human cognitive and social processes, especially emotion, and explores these models role in shaping human-computer interactions in virtual environments. He studies the relationship between cognition and emotion, the cognitive processes underlying emotional responses, and the influence of emotion on decision making and physical behavior. He is the founding and current Editor-in-Chief of IEEEs Transactions on Affective Computing (3.5 impact factor in 2013), Associate Editor of Emotion Review and the Journal of Autonomous Agents and Multiagent Systems, and former President of the Association for the Advancement of Affective Computing (AAAC). He is a AAAI Fellow, a SIGART Autonomous Agents Award recipient, a Senior Member of IEEE, and member of the International Society for Research on Emotion (ISRE). Dr. Gratch is the author of over 200 technical articles.

Louis-Philippe Morency is Assistant Professor in the Language Technology Institute at the Carnegie Mellon University where he leads the Multimodal Communication and Machine Learning Laboratory (MultiComp Lab). He received his Ph.D. and Master degrees from MIT Computer Science and Artificial Intelligence Laboratory. His research focuses on creating algorithms and probabilistic models to analyze, recognize and predict human subtle communicative behaviors during social interactions. In 2008, Dr. Morency was selected as one of "AI's 10 to Watch" by IEEE Intelligent Systems. He has received 7 best paper awards in multiple ACM- and IEEE-sponsored conferences for his work on context-based gesture recognition, multimodal probabilistic fusion and computational models of human communication dynamics.