Published in final edited form as:

J Neurolinguistics. 2007 January ; 20(1): 50–64. doi:10.1016/j.jneuroling.2006.04.001.

Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology

James C. Mundt¹, Peter J. Snyder^{2,3}, Michael S. Cannizzaro^{2,4}, Kara Chappie², and Dayna S. Geraltz¹

¹Healthcare Technology Systems, Inc

²Pfizer Global Research & Development – Groton Laboratories (Groton, CT)

³Department of Psychology, University of Connecticut (Storrs, CT)

⁴Department of Communication Sciences, University of Vermont (Burlington, VT)

Abstract

Efforts to develop more effective depression treatments are limited by assessment methods that rely on patient-reported or clinician judgments of symptom severity. Depression also affects speech. Research suggests several objective voice acoustic measures affected by depression can be obtained reliably over the telephone. Thirty-five physician-referred patients beginning treatment for depression were assessed weekly, using standard depression severity measures, during a six-week observational study. Speech samples were also obtained over the telephone each week using an IVR system to automate data collection. Several voice acoustic measures correlated significantly with depression severity. Patients responding to treatment had significantly greater pitch variability, paused less while speaking, and spoke faster than at baseline. Patients not responding to treatment did not show similar changes. Telephone standardization for obtaining voice data was identified as a critical factor influencing the reliability and quality of speech data. This study replicates and extends previous research with a larger sample of patients assessing clinical change associated with treatment. The feasibility of obtaining voice acoustic measures reflecting depression severity and response to treatment using computer-automated telephone data collection techniques is also established. Insight and guidance for future research needs are also identified.

Keywords

Depression assessment; methodology; speech; voice acoustics; telephone; interactive voice response (IVR)

1. Introduction

1.1 Depression

Depression is a common, disabling, and life shortening condition that occurs throughout the lifespan. More than 1 in 6 persons in the US experience a major depressive episode in their lifetime (Kessler, et al., 1994). Depression is associated with half of all suicides and present

an annual national economic burden exceeding \$43 billion (Mann, 2002; Greenberg, Stiglin, Kinkelstein, & Berndt, 1993). Billions of dollars are invested developing more effective and faster acting treatments with fewer adverse side effects. However, assessment tools currently used to measure depression severity and clinical change present limitations.

1.2 Assessment of depression by clinicians

Research in any domain is limited by the accuracy and precision of available measurement instruments. Despite recent advances in understanding the neurophysiology associated with depression, the clinical tools used in research of depression treatments have remained basically unchanged over the past 40 years.

Depression measures have historically relied upon either patient self-reported or clinician judgments of symptom severity. Both approaches attempt to derive reliable, objective measures of depression severity using subjective judgments that depend upon the training and experience of the clinician and patient. Such measures are critical for patient screening, severity assessment, measurement of change over time, identification of treatment benefits and remission, and detection of relapse or recurrence (Rush & Ryan, 2002).

The Hamilton Depression Rating Scale (HAMD; Hamilton, 1960) is the most widely used assessment of depression in clinical trial research. Despite noted psychometric weaknesses of the HAMD (e.g., Bech, Grosby, Husum, & Rafaelsen, 1984; Gibbons, Clark, & Kupfer, 1993; Faries et al., 2000; Santor & Coyne, 2001; Zimmerman, Posternak, & Chelminski, 2005), it is the most common ‘gold standard’ depression assessment. Efforts to standardize HAMD assessment (e.g., Williams, 1988), as well as intensive training of HAMD raters to improve rating reliability (Demitrack, Fairies, Herrera, DeBrola, & Potter, 1998), have shown limited benefits in reducing measurement error. Other widely used clinician-based rating scales include the Montgomery-Asberg Depression Rating Scale (Montgomery & Asberg, 1979), the Inventory of Depressive Symptomatology (Rush, Gullion, Basco, Jarrett, & Trivedi, 1996), and the Quick Inventory of Depressive Symptomatology (Rush, et al., 2003). Each requires rater training, practice, and certification to produce acceptable inter-rater reliabilities.

In addition to ‘simple’ measurement error problems inherent in clinician-based assessments of depression, systematic bias of such ratings in clinical trial research has become evident in recent years. DeBrola et al. (1999) found evidence of substantial inflation of baseline scores resulting in the inclusion of inappropriate study participants. Study sponsors often pressure study sites to accelerate patient enrollment, and comparisons between sites may contribute to bias in subsequent ratings (Byrom & Mundt, 2005). Such issues may have contributed to increased placebo response and failed or negative trials in recent years (Uhlenhuth, Matuzas, Warner, & Thompson, 1997; Robinson & Rickels, 2000; Walsh, Seidman, Sysko, & Gould, 2002; Khan, Leventhal, Khan, & Brown, 2002). Petkova, Quitkin, McGrath, Stewart, and Klein (2000) suggested that patient-reported measures of depression might be used to estimate the extent to which clinician-based ratings are influenced by rater bias.

1.3 Assessment of depression by patient self-report

Patient-reported measures of depression are commonplace, but pose different challenges for measuring depression severity or response to treatment. Established instruments such as the Beck Depression Inventory (Beck, Ward, Mendelson, Mock, & Erlbaum, 1961), Zung Self-Rating Scale (Zung, 1965), and the Hospital Anxiety and Depression Scale (Zigmond & Snaith, 1983) are readily available, but rely on patients’ knowledge and experience with depressive symptoms, which are idiosyncratic. Most previous research suggests that clinician-based assessments result in larger treatment effect sizes (e.g., Edwards, et al.,

1984; Lambert, Hatch, Kingston, & Edwards, 1986; Kobak, Greist, Jefferson, Mundt, & Katzelnick, 2001), but evidence that patient-reported measures can provide equivalent or superior drug-placebo separation has also been found (Rayamajhi, Lu, DeBrot, Demitrack, & Greist, 2002; DeBrot, et al., 2005). The FDA acknowledges patient-reported depression measures as acceptable and valid as a primary outcome in adult outpatient antidepressant registration trials.

1.4 Biomarkers of depression in voice acoustics

Both clinician ratings and patient self-reports rely on subjective perceptions and judgments of the severity and/or change in depression symptoms. This is necessary, in part, because no reliable, objective physiologically-based measure of depression severity currently exists. Research into potential biomarkers of central nervous system disorders such as schizophrenia, affective disorder, and Parkinson's disease have explored subtle changes in eye movements and speech characteristics as possible physiologically based indicators of disease progression, severity or treatment efficacy (e.g., Harel, Cannizzaro, & Snyder, 2004; Kathmann, Hochrein, Uwer, & Bondy, 2003).

That depression affects patterns of speech has been recognized for many years (Moses, 1954). Darby and Hollien (1977) found listeners of speech recorded from depressed patients could perceive differences in the pitch, loudness, speaking rate, and articulation before and after treatment. Acoustical signal processing technology available at the time was not adequate to cross-validate the subjective observations, but has been refined for analysis of depression-related speech characteristics over subsequent decades (e.g., Darby, Simmons, & Berger, 1984; Nilsson, 1987; 1988; Scherer & Zei, 1988; Stassen, 1988; Stassen, Bomben, & Gunther, 1991; Kuny & Stassen, 1993). In recent years, advances in the techniques for eliciting and analyzing speech samples have been reported and validated in the literature (e.g., Stassen, Kuny & Hell, 1998; Garcia-Toro, Talavera, Saiz-Ruiz, & Gonzalez, 2000; France, Shiavi, Silverman, Silverman, & Wilkes, 2000; Wuyts, et al., 2000; Parsa & Jamieson, 2001; Alpert, Pouget & Silva, 2001; Brenznitz, 2001; Alpert, Shaw, Pouget, & Lim, 2002). The measurement of acoustic properties of speech related to depression and other central nervous system diseases provides a natural intersection in applied research between objective physical manifestations and subjective clinical observations.

1.5 Telephony, mental health, and voice acoustics

While advances were being made in applied analysis of depression-related speech, telecommunication processes were also being transformed. Development of dual-tone modulated frequency (DTMF) technology led to the commonplace computer-automation of touch-tone telephone systems known as interactive voice response (IVR). Interactive voice response systems for assessing and diagnosing psychopathology have been used in research and practice settings for over a decade (e.g., Kobak et al 1997; Piette, 2000; Corkrey & Parkinson, 2002), and are now widely used in randomized clinical trial research to obtain primary outcome data in FDA registration trials (e.g., Krystal et al., 2003).

Clinicians are trained to attend closely to both 'what' patients say, as well as 'how' they say it. Vocal inflection and emotive expression may be under less conscious control than the voluntary selection of constrained response options among limited sets of choices. The current study was designed to elicit, record, and analyze speech samples from depressed patients receiving treatment to investigate voice acoustic patterns associated with depression severity and treatment response. The speech elicitation and analysis methods were derived from prior research.

Cannizzaro, Harel, Reilly, Chappell, and Snyder (2004) analyzed the speech of depressed patients from videotapes of clinical interviews. Results converged with previous findings that motor timing parameters reflecting speech production (e.g., speaking rate and pause time) and frequency modulation (e.g., pitch variability) were significantly correlated with depression severity. A subsequent pilot study found that such speech measures could be adequately collected over the telephone using IVR technology and were acoustically consistent with speech samples concurrently recorded in an acoustics laboratory and analyzed for both timing and frequency related measures (Cannizzaro, Reilly, Mundt, & Snyder, 2005).

The present study was conducted to replicate and extend prior findings using a naturalistic, six-week observational study design. It included a larger sample of depressed patients than previous studies, and collected speech data longitudinally to assess characteristics associated with depression severity and changes related to treatment response. Validated depression severity measures were obtained weekly along with speech samples for acoustic analyses. It was expected that many of the patients would respond to treatment during the course of the study, but that others would not or would respond idiosyncratically. The analyses reported examine the association between the accepted clinical depression severity measures and the vocal acoustic parameters extracted from the recorded speech samples.

2. Methods

2.1 Participants

Thirty-five patients (20 women and 15 men, Mean age = 41.8 years) were recruited to participate. Study participants were predominately Caucasian (88.6%), with four participants of African American, Bi-racial, Greek, or Hispanic descent. Study inclusion criteria required study participants to be 18 years or older, recently started on pharmacotherapy and/or psychotherapy treatment for depression, referred by a treating physician, and access to a touch-tone telephone. Written informed consent was obtained in accordance with required federal and institutional guidelines, including oversight of human subjects protection in research by the Allendale Institutional Review Board (Allendale, NJ). Study eligibility or participation was not influenced the depression treatments prescribed by the referring physicians.

Given the exploratory nature of assessing depression longitudinally using vocal acoustic measures, no severity criterion was required for study participation; sixty-six percent of the patients had Hamilton Depression Rating Scale scores over 17, a common clinical trial inclusion criterion. Since the treatment provided each patient was individually determined, self-reported side-effects of the treatments were not collected. Thirty-four of the 35 patients entering the study continued through the final follow-up visit.

2.2 Apparatus

An automated IVR telephone interface was developed by Healthcare Technology Systems (Madison, WI) to collect data from study participants. The system was developed using VoiceXML and was accessed through a commercial application service provider (Voxeo, Inc.; Orlando, FL). Call flow branching was restricted to touch-tone inputs. Speech samples recorded by the IVR system were saved as single channel, 8-bit .wav files, sampled at 8 kHz and were sent to the Voice Acoustics Laboratory at Pfizer Global Research & Development – Groton Laboratories (Groton, CT) for analysis and extraction of the voice acoustic metrics. Acoustic analyses were performed using freely available PRAAT software version 4.2.17 (Boersma & Weenik, 2003) and commercial speech and voice analysis software programs on the Computerized Speech Laboratory (Kay Elemetrics, 2001).

2.3 Procedures

Patients who were referred to the study by their treating physicians and interested in study participation contacted the study coordinator by telephone. Details of the study procedures were explained and an appointment was scheduled to complete informed consent procedures and baseline assessment. Consenting participants were assigned an identification number and created their own personal pass code for accessing the IVR system to ensure data integrity.

Study participants were asked to call a toll-free telephone number using any touch-tone telephone over the following six weeks. The schedule of assessments is presented below. Study participants were compensated proportionally for their compliance with the requested reporting procedures. In total, participants could earn up to \$370.00 for complete compliance with all data collection procedures. The IVR system reminded callers of the compensation earned to date after the each call, a procedure designed to maximize call compliance (Mundt, Searles, Perrine, & Walter, 1997).

2.3.1 Weekly IVR HAMD and QIDS assessment of depression severity—Patients also completed a HAMD and Quick Inventory of Depressive Symptomatology (QIDS) by IVR each week. The IVR HAMD and QIDS assessments use touch-tone inputs to obtain caller responses to standard queries in a structured clinical interview format. These scales are established, validated depression severity measures and currently used in clinical trial research (Mundt et al 1998; Kobak, Mundt, Greist, Katzelnick, & Jefferson, 2000; Rush et al 2005).

Memory Enhanced Retrospective Evaluation of Treatment (MERET[®]) is an assessment technique that prompts patients with any disorder to record personal descriptions of difficulties regarding emotional, physical, and functional problems at baseline. Subsequently, the baseline recordings are played back to the patients and they are asked to rate clinical change since then on a 7-point where 1 = Very Much Better; 2 = Much Better; 3 = A Little Better; 4 = Unchanged; 5 = A Little Worse; 6 = Much Worse; and 7 = Very Much Worse (Mundt, DeBrot, Moore, & Greist, 2005). In the current study, after listening to the baseline recordings and rating change, the patients were asked to record statements about experiential changes since the baseline recordings. These recordings provided the “free speech” samples of extemporaneous speech data in the voice acoustic analyses described below. The MERET prompts and additional structured tasks for eliciting speech samples over the telephone are described below in section 2.3.4.

2.3.2 Bi-weekly clinician HAMD assessment of depression severity—During office visits at baseline, and at weeks 2, 4, and 6, the patients completed face-to-face HAMD assessments with an experienced clinical trial rater. The order of clinician or IVR assessments during the office visits was counterbalanced across subjects and alternated from one office visit to the next.

2.3.3 Vocal acoustic measures obtained weekly—Each week, several tasks elicited and recorded speech samples from study participants, including the extemporaneously produced MERET recordings in response to the following queries: (1) “Describe how you’ve been feeling emotionally during the last week. Record your experiences as well as any comments other people made about your emotions.” (2) “Describe how you’ve been feeling physically during the past week. If you can, give examples of how your physical feelings have been affecting your life.” (3) “Please describe how your emotional and physical feelings have affected your general ability to function in the past week. Consider

such things as your ability to work, manage your home, get along with others, and participate in leisure activities.”

Additional tasks asked callers to count from 1 to 20 and to recite the alphabet. They were also asked to pronounce the vowels /a/, /i/, /u/, and /ae/ for 5 seconds. The subjects also rapidly repeated the syllables /pa ta ka/ for five seconds to provide a measure of diadochokineic rate. Standardized instructions and examples were provided aurally prior to recording speech samples. Recording time of speech samples was measured and durations less than five seconds were repeated. Finally, the subjects were asked to read *The Grandfather Passage*, a standard 132-word paragraph commonly used in the assessment of communication disorders.

Telephone frequency response is generally reported to range from 300 Hz to 3000 Hz acting essentially as a band pass filter rejecting higher and lower frequencies (Kent, & Read, 2002). Pilot testing of the IVR system found accurate frequency portrayal of telephone transmissions between 250 Hz to about 3200 Hz with a steep frequency response roll off outside these parameters. The accurate measurement of frequencies beyond these boundaries, for example F_0 below 300 Hz, is dependent on the algorithms used in the analysis. The periodicity-to-pitch autocorrelation function in Praat was used because it is not dependent on the actual presence of a fundamental frequency to make the measurement. Autocorrelation uses the repeat length of the complex waveform to determine the fundamental period (P. Boersma, personal communication). The lowest common frequency in a complex periodic wave will generate a repeat pattern at that frequency. Since the voice fundamental frequency is generated by the periodic vibration of the vocal folds and each harmonic is a simple multiple of that fundamental, the lowest common repeat pattern is at the same rate as the fundamental frequency of the voice. Previous testing found that a wide range of frequency excursions around the fundamental frequency could be accurately assessed in telephone recordings (Cannizzaro, Reilly, Mundt, & Snyder, 2005).

Pitch variability about F_0 , F_1 , and F_2 was measured by the coefficient of variation (COV), the ratio of the standard deviation about the mean frequency divided by the mean frequency for each fundamental. The COV about the F_0 was also obtained for continuous speech samples (extemporaneous MERET responses and reading the *Grandfather Passage*). Voice samples obtained from the counting, alphabet recitation, reading, and extemporaneous speech tasks provided the following voice acoustic measures: (1) total recording duration, (2) vocalization time, (3) pauses while speaking (number, mean length, and standard deviation of pause length), (4) percent time pausing, (5) vocalization/pause ratio, and (5) speaking rate (syllables per second). Diadochokinesis measures included: (1) mean and variability of syllable durations, (2) mean and variability of vocal intensities, and (3) syllable rate.

3. Results

3.1 Depression severity measures

Three clinical measures of depression severity were obtained: (1) clinician-rated HAMDS, (2) IVR HAMDS, and (3) IVR QIDS. The FDA has accepted all three measures as valid outcomes for randomized antidepressant clinical trials. Figure 1 shows the mean scores for all three measures across the six weeks of the study.

All three measures show similar strong trends of diminishing depression severity over time, as expected. The correlation between the HAMD scores obtained from paired clinician and IVR assessments was .90, ranging from .84 to .94 within the study weeks when both were obtained. Correlations between IVR QIDS scores and the clinician and IVR HAMDS scores

were .82 and .84, respectively. Within-week correlations between the QIDS and HAMD scores ranged from .73 to .90.

IVR HAMD scores were used below for analysis of the vocal acoustic measures. The high correlation between clinician and IVR HAMDs indicates both are essentially interchangeable, and comparable correlations with the independent QIDS assessment support the construct validity of both. The IVR HAMDs allows analysis of vocal acoustic measures obtained at all study weeks, including those obtained between office visits at weeks 1, 3, and 5.

3.2 Association between HAMD scores and vocal acoustics

The internal consistencies of the vocal acoustic measures obtained multiple times within each call and the overall correlation of each measure with depression severity are shown in Table 1. Acoustic measures from the diadochokinesis task are not presented. Internal consistency of measures for a single task cannot be computed, and none were significantly correlated with depression severity. The COVs about F_0 or F_1 associations with depression severity were not statistically significant, but the correlation with the F_2 COV was. The poor reliability of this measure across the vowel production tasks and analyses examining the source of the recordings (home versus office, see Section 3.4) suggests spurious Type I error. The reliability of total recording durations, total pause time, variability of pause durations, ratio of vocalization to pause, and speaking rates within speech samples were more reliable and significantly correlated with the HAMD scores. These data suggest that greater depression severity results in longer total recording lengths, but that the increased length is due to longer and more variable pauses during speech production. The result is a lower vocalization/pause ratio and slower speaking rates.

The total pause time during automatic speech production tasks (counting, reciting the alphabet, passage reading) was better correlated with depression severity ($r = .27, p < .001$) than pause time during the free speech tasks ($r = .21, p < .01$). This pattern reversed for the pause variability and vocalization/pause measures which correlated with depression severity .34 ($p < .001$) and -.22 ($p = .001$) during the free speech tasks and .30 ($p < .001$) and -.18 ($p < .01$) during the automatic speech production tasks.

3.3 Association between depression and vocal acoustic measures change over time

The data in Table 1 reflect overall associations between depression severity and voice acoustic measures, collapsed across study participants and study weeks. They do not directly assess sensitivity to within-subject change in depression severity over time. For each subject, the HAMD score obtained each week was subtracted from the baseline HAMD. The resulting change score was divided by the baseline HAMD to produce a percent drop in depression severity from baseline to each follow-up week. At each week, subjects were classified as a treatment “responder” if the drop was 50% or greater. The mean HAMD change from baseline and number of subjects classified as responders at each follow-up week are shown in Table 2. While 23 (66%) of the 35 subjects indicated a treatment response at one of more of the follow-up weeks, in many instances the response was not sustained through the final week.

To examine whether subjects who responded to treatment also manifested changes in their voice acoustics measures, the 13 subjects reflecting a treatment response in their HAMD scores at week 6 were compared to the 19 subjects who did not. Between-group t-tests of depression severity and vocal acoustic measures did not find any significant differences at baseline (all p -values $> .05$). Changes in each vocal acoustic measure were computed by subtracting the value obtained from the week 6 speech samples from the baseline value.

Statistically significant between-group differences were found in five of the vocal acoustic change measures, shown in Table 3.

Treatment responders increased pitch variability about F_0 more than nonresponders; the F_0 COV increase among responders is significantly different from 0 ($p = .01$), whereas the small decrease in pitch variability among nonresponders is not ($p = .31$). Total recording sample durations decreased from baseline to a greater degree among treatment responders than nonresponders. The shorter recording lengths reflect treatment responders making fewer pauses while speaking, with less overall pause time. Consistent with the F_0 COV data, the reductions in total recording length, total pause time, and number of pauses from baseline were significantly different from 0 ($p < .01$ for all three VA measures) for treatment responders; only the reduction in the total pause time was significant ($p = .03$) among the nonresponders. By contrast, the changes in total vocalization time from baseline (data not shown) were not statistically different between the groups, $t(30) = -1.89$, $p = .07$, nor was the change significantly different than 0 for either group. The significantly reduced number of pauses and total pause time, while producing equivalent vocalizations, is reflected in significantly faster speaking rates in the treatment responders. The mean increase of 0.75 syllables per second among responders is significantly greater than 0 ($p < .001$), whereas the 0.18 syllables per second increase among the nonresponders is not ($p = .06$).

3.4 Influence of standardized versus nonstandardized telephones

This study also allowed examination of the influence of using different telephones for collection of vocal acoustic measures. During office visits (weeks 0, 2, 4, and 6) all subjects used a standard commercial multi-line desktop telephone (Meridian Model 2616); “off-week” assessments (weeks 1, 3, and 5) permitted subjects to use any telephone they choose (home, cell, or public). Comparisons of the HAMD and QIDS depression severity measures obtained during office and off-week calls did not indicate significant differences with respect to assessment variability (Levene’s test for equality of variance) or means (independent group t-tests).

This was not the case for vocal acoustic measures. Signal amplitudes during the diadochokinesis tasks, total recording lengths, vocalization times, and pause measures were significantly more variable when obtained during the off-week calls than from office calls. Additionally, t-tests of speech samples obtained during off-week calls found significantly lower and more variable signal intensities, less over all vocalization time, and more pauses in the recordings obtained during off-weeks than in those obtained using the standardized telephone during office calls. Differences in the variability and acoustical properties associated with the use of different telephones affected data quality and subsequent results.

The Week 6 treatment responder/nonresponder comparisons described above (Table 3) were repeated using Week 5 vocal acoustic change measures. Only the change in COV about F_0 separated treatment responders from nonresponders, $t(27) = 2.23$, $p = .04$. Responders at Week 5 had greater COV about F_0 and nonresponders less variation, but neither group’s change was significantly different than their baseline value. None of the changes in the other vocal acoustic measures obtained at Week 5 separated responders from nonresponders.

4 Discussion

This study builds upon a growing body of literature demonstrating structured IVR clinical interviews produce reliable and valid measures of depression severity (e.g., Kobak et al., 2000; Rush et al., 2005). The results also contribute to the emerging research relating vocal acoustics metrics and depression (e.g., Alpert et al., 2002), and are consistent with previous

results regarding the effect of depression on speaking rate, speech pauses, and pitch variability (Cannizzaro et al., 2004). The data obtained and analyzed in this study also clearly establish the feasibility of automating procedures for speech collection using telephone transmission lines to obtain reliable and valid voice acoustic measures related to depression severity and clinical response to treatment. Methodological constraints on task implementation were also identified.

Not surprisingly, the telephone device used by study subjects to transmit speech data influenced the variability and sensitivity of the acoustical measures obtained. It is not surprising that the quality and reliability of the voice acoustics data were influenced by the variety of telephones used to provide speech data from outside the office. There was no evidence that the depression measures obtained from touch-tone responses to clinical questions were influenced by the calling location. Clearly, future research on voice acoustic measures obtained over the telephone must choose measures and metrics carefully that can be reliably captured by the telephone system utilized.

The study results obtained are consistent with previous research (Cannizzaro et al., 2004; 2005). Measures of pitch variability and speech production were significantly influenced by depression severity and clinical response to treatment. The results also suggest additional avenues of inquiry that should be addressed by future research.

This study found the COV about F_2 to correlate significantly with overall depression severity, but the measure had little reliability across speech samples recorded within the same call. The COV about F_0 showed better internal consistency across speech samples obtained within a given telephone call, but did not correlate with overall depression severity. Within-subject change in the F_0 COV over time, however, was correlated with clinical change in response to treatment. More research is needed to disentangle pitch variability in speech, automated data extraction methods, telephone instrumentation and transmission bandwidths, and their relationship to depression.

The relationships between depression and time-based measures of speech production, such as total speaking time and the interplay between vocalization and pausing, are more definitive and readily interpreted. Data from this study indicate that more depressed patients take longer to express themselves. They don't vocalize more; rather they speak with greater hesitancy, producing more cumulative and variable pauses. Consequently, voice acoustic measures such as the percent pause time, vocalization/pause ratio, and speaking rates reflect depression severity. Not surprisingly, these measures also dominate the acoustic measures most sensitive to clinical change. The significantly reduced recording duration of treatment responders was due to fewer pauses, less total pause time, and faster speaking rates than observed in nonresponders. The change in the total amount of vocalizations produced before and after treatment did not differ between treatment responders and nonresponders. The responders were more efficient with their verbal communication, completing the speech tasks required with less hesitancy.

Interestingly, these data suggest that total pause time observed during completion of "automatic" speech tasks such as counting, reciting the alphabet, or reading may correlate better with depression severity than total pause time observed during production of extemporaneous "free" speech. Conversely, measures of pause variability and the vocalization/pause ratio observed during free speech may correlate better with depression severity than these measures obtained during automatic speech production. Such patterns, if replicated in future research, would be interesting and important both theoretically and clinically. Summative measures of the total number and cumulative duration of pauses while speaking may reflect a general level of psychomotor activation/retardation. During

performance of highly automated speech tasks with minimal cognitive demand, the effects of depression on psychomotor retardation may be most evident in these measures and tasks. Generation of extemporaneous speech requires greater cognitive demand, requiring selection and preparation of the verbal response as well as articulatory execution. Greater pause variability and the relationship between vocalization and pausing behaviors may reflect the effects of depression on cognitive slowing, showing better association in these measures when completing free speech tasks. Such speculation requires further research and replication, but is consistent with the data obtained in this study and patterns previously observed in persons with schizophrenia (Cannizzaro, Cohen, Reppard, & Snyder, 2005).

This study has several limitations that caution against over interpretation. First, the naturalistic observational study design limits generalization of these results to those that might be expected in a placebo-controlled clinical trial. Second, while larger than previous studies investigating telephone methods for obtaining vocal acoustic measures, and the first to include longitudinal assessments, the sample was limited to 35 patients and only six weeks of follow-up. Third, this study involved a homogeneous sample of primarily Caucasian patients recruited from a regional HMO. Whether or not these results would replicate in a longer study with a larger and more diverse patient population remains to be investigated.

Such limitations notwithstanding, this pilot study clearly establishes the feasibility of voice acoustic measures as potential depression severity biomarkers and indicators of treatment response. It also establishes the feasibility of using computer-automated speech solicitation techniques to obtain meaningful and analyzable data over standardized telephone interfaces. Voice acoustic measures associated with pitch variability and motoric articulation were significantly related to cross-sectional differences in depression severity between patients, as well as within-patient clinical change between treatment responders and nonresponders. These results replicate and extend prior research and identify needs for further research.

Acknowledgments

The assistance and cooperation of physicians at Dean Medical Systems, Drs. Peter Clagnaz, Leslie Taylor, David Hahn, Steven Singer, and Evan Weiden who referred patients to participate in this study is gratefully acknowledged. The administrative assistance of Sharon Bailey at the Dean Foundation for Health Research and Education was also appreciated. We also thank Nicole Reilly and Dr. Ania Wisniecki for technical assistance in extracting voice acoustics measures at the Pfizer Speech Acoustics Laboratory. Clinical consultation with Dr. John Greist and IVR programming skills of Ben Barth at Healthcare Technology Systems were essential to the success of this project. MERET[®] is a registered trademark of Healthcare Technology Systems. Finally, the outstanding clinical ratings services of Dr. Kenneth Kobak and Tammy Zinser are gratefully acknowledged. Dr. Peter J. Snyder is currently at the University of Connecticut (Storrs, CT), and Dr. Michael S. Cannizzaro, currently at the University of Vermont (Burlington, VT). Support for this research was provided by the National Institute on Mental Health (R43MH68950 JC Mundt, PI) and by Pfizer, Inc.

References

- Alpert M, Pouget ER, Silva RR. Reflections of depression in acoustic measures of the patient's speech. *Journal of Affective Disorders* 2001;66:59–69. [PubMed: 11532533]
- Alpert M, Shaw RJ, Pouget ER, Lim KO. A comparison of clinical ratings with vocal acoustic measures of flat affect and alogia. *Journal of Psychiatric Research* 2002;36:347–353. [PubMed: 12127603]
- Bech P, Grosby H, Husum B, Rafaelsen L. Generalized anxiety or depression measured by the Hamilton Anxiety Scale and the Melancholia Scale in patients before and after cardiac surgery. *Psychopathology* 1984;17:253–63. [PubMed: 6537541]
- Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J. An inventory for measuring depression. *Archives of General Psychiatry* 1961;4:561–571. [PubMed: 13688369]

- Boersma, P.; Weenik, D. Praat (Version 4.1.9) [computer software]. Amsterdam: Institute of Phonetic Sciences; 2003.
- Boersma, P. Ppersonal communication. Nov 12. 2003
- Breznitz Z. Verbal indicators of depression. *The Journal of General Psychology* 2001;119:351–363. [PubMed: 1491239]
- Byrom B, Mundt JC. The value of computer-administered self-report data in central nervous system clinical trials. *Current Opinions in Drug Discovery Development* 2005;8:374–383.
- Cannizzaro M, Cohen H, Reppard F, Snyder PJ. Bradyphrenia and bradykinesia both contribute to altered speech schizophrenia: A quantitative acoustic study. *Cognitive and Behavioral Neurology* 2005;18:206–210. [PubMed: 16340393]
- Cannizzaro M, Harel B, Reilly N, Chappell P, Snyder PJ. Voice acoustical measurement of the severity of major depression. *Brain and Cognition* 2004;56:30–35. [PubMed: 15380873]
- Cannizzaro M, Reilly N, Mundt JC, Snyder PJ. Remote capture of human voice acoustical data by telephone: A methods study. *Clinical Linguistics & Phonetics* 2005;19:649–658. [PubMed: 16147408]
- Corkrey R, Parkinson L. Interactive voice response: review of studies 1989-2000. *Behavior Instrument, Methods, & Computers* 2002;34:342–53.
- Darby JK, Hollien H. Vocal and speech patterns of depressive patients. *Folia Phoniatica* 1977;29:279–291. [PubMed: 604242]
- Darby JK, Simmons N, Berger PA. Speech and voice parameters of depression: A pilot study. *Journal of Communication Disorders* 1984;17:75–85. [PubMed: 6725627]
- DeBrot, DJ.; Demitrack, MA.; Landin, R.; Kobak, KA.; Greist, JH.; Potter, WZ. A comparison between interactive voice response system- administered HAM-D and clinician-administered HAM-D in patients with major depressive disorder. New Clinical Drug Evaluation Unit, 39th Annual Meeting; Boca Raton, FL. 1999.
- DeBrot, DJ.; Manner, DH.; Morrow, RJ.; Padich, RA.; Peng, X.; Mundt, JC.; Greist, JH. Daily and Weekly Assessment of Depression Severity in a Duloxetine Clinical Trial. National Institute of Mental Health, New Clinical Drug Evaluation Unit, 45th Annual Meeting; Boca Raton, FL. 2005.
- Demitrack MA, Faries D, Herrera JM, DeBrot DJ, Potter WZ. The problem of measurement error in multisite clinical trials. *Psychopharmacology Bulletin* 1998;34:19–24. [PubMed: 9564194]
- Edwards BC, Lambert MJ, Moran PW, McCully T, Smith KC, Ellingson AG. A meta-analytic comparison for the Beck Depression Inventory and the Hamilton Rating Scale or Depression as measures of treatment outcome. *British Journal of Clinical Psychology* 1984;23:93–99. [PubMed: 6722384]
- Faries D, Herrera J, Rayamajhi J, DeBrot D, Demitrack M, Potter WZ. The responsiveness of the Hamilton Depression Rating Scale. *Journal of Psychiatric Research* 2000;34:3–10. [PubMed: 10696827]
- France DJ, Shiavi RG, Silverman S, Silverman M, Wilkes DM. Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Transactions of Biomedical Engineering* 2000;47:829–837.
- Garcia-toro M, Talavera JA, Saiz-Ruiz J, Gonzalez A. Prosody impairment in depression measured through acoustic analysis. *The Journal of Nervous and Mental Disease* 2000;188:824–829. [PubMed: 11191583]
- Gibbons RD, Clark DC, Kupfer DJ. Exactly what does the Hamilton Depression Rating Scale measure? *Journal of Psychiatric Research* 1993;27:259–273. [PubMed: 8295158]
- Greenberg PE, Stiglin LE, Finkelstein SD, Berndt ER. Depression: A neglected major illness. *Journal of Clinical Psychiatry* 1993;54:419–424. [PubMed: 8270584]
- Hamilton M. A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry* 1960;23:56–62.
- Harel B, Cannizzaro M, Snyder PJ. Variability in fundamental frequency during speech in prodromal and incipient Parkinson's disease: a longitudinal case study. *Brain and Cognition* 2004;56:24–29. [PubMed: 15380872]

- Kathmann N, Hochrein A, Uwer R, Bondy B. Deficits in gain of smooth pursuit eye movements in schizophrenia and affective disorder patients and their unaffected relatives. *American Journal of Psychiatry* 2003;160:696–702. [PubMed: 12668358]
- Kay Elemetrics Corporation. CSL Model 4400 [computer software]. Lincoln Park, NJ: 2001.
- Kent, RD.; Read, C. Acoustic analysis of speech. 2. Canada: Singular; 2002.
- Kessler RC, McGonagle KA, Zhao S, Nelson CB, Hughes M, Eshleman S, Wittchen H, Kendler KS. Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States. *Archives of General Psychiatry* 1994;51:8–19. [PubMed: 8279933]
- Khan A, Leventhal RM, Khan SR, Brown WA. Severity of depression and response to antidepressants and placebo: An analysis of the Food and Drug Administration database. *Journal of Clinical Psychopharmacology* 2002;22:40–45. [PubMed: 11799341]
- Kobak KA, Greist JH, Jefferson JW, Katzelnick DJ, Mundt JC. New technologies to improve clinical trials. *Journal of Clinical Psychopharmacology* 2001;21:255–256. [PubMed: 11386486]
- Kobak KA, Mundt JC, Greist JH, Katzelnick DJ, Jefferson JW. Computer assessment of depression: Automating the Hamilton Depression Rating Scale. *Drug Information Journal* 2000;34:145–156.
- Kobak KA, Taylor LvH, Dottl SL, Greist JH, Jefferson JW, Burroughs D, Mantle JM, Katzelnick DJ, Norton R, Henk HJ, Serlin RC. A computer-administered telephone interview to identify mental disorders. *JAMA* 1997;278:905–910. [PubMed: 9302242]
- Krystal AD, Walsh JK, Laska E, Caron J, Amato DA, Wessel TC, Roth T. Sustained efficacy of eszopiclone over 6 months of nightly treatment: Results of a randomized, double-blind, placebo-controlled study in adults with chronic insomnia. *Sleep* 2003;26:793–799. [PubMed: 14655910]
- Kuny S, Stassen HH. Speaking behavior and voice sound characteristics in depressive patients during recovery. *Journal of Psychiatric Research* 1993;3:289–307. [PubMed: 8295161]
- Lambert MJ, Hatch DR, Kingston MD, Edwards BC. Zung, Beck, and Hamilton rating scales as measures of treatment outcome: A meta-analytic comparison. *Journal of Consulting and Clinical Psychology* 1986;54:54–59. [PubMed: 3958302]
- Mann JJ. A current perspective on suicide and attempted suicide. *Annals of Internal Medicine* 2002;136:302–311. [PubMed: 11848728]
- Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. *British Journal of Psychiatry* 1979;134:382–389. [PubMed: 444788]
- Moses, PJ. *The Voice of Neurosis*. New York, NY: Grune and Stratton; 1954.
- Mundt, JC.; DeBrot, DJ.; Moore, HK.; Greist, JH. Memory Enhanced Retrospective Evaluation of Treatment (MERET): Anchoring Patients' Perceptions of Clinical Change in the past. National Institute of Mental Health, New Clinical Drug Evaluation Unit, 45th Annual Meeting; Boca Raton, FL. 2005.
- Mundt JC, Kobak KA, Taylor LvH, Mantle JM, Jefferson JW, Katzelnick DJ, Greist JH. Administration of the Hamilton Depression Rating Scale using interactive voice response technology. *M D Computing* 1998;15:31–39. [PubMed: 9458661]
- Mundt, JC.; Moore, HK.; DeBrot, DJ.; Greist, JH. Recency Effects in Standard Depression Measures Using Daily Telephone Assessment Ratings. National Institute of Mental Health, New Clinical Drug Evaluation Unit, 45th Annual Meeting; Boca Raton, FL. 2005.
- Mundt JC, Searles JS, Perrine MW, Walter D. Conducting longitudinal studies of behavior using interactive voice response technology. *The International Journal of Speech Technology* 1997;2:21–31.
- Nilsson A. Acoustic analysis of speech variables during depression and after improvement. *Acta Psychiatrica Scandinavica* 1987;76:235–245. [PubMed: 3673650]
- Nilsson A. Speech characteristics as indicators of depressive illness. *Acta Psychiatrica Scandinavica* 1988;77:253–263. [PubMed: 3394527]
- Parsa V, Jamieson DG. Acoustic discrimination of pathological voice: Sustained vowels versus continuous speech. *Journal of Speech, Language, and Hearing Research* 2001;44:327–339.
- Petkova E, Quitkin FM, McGrath PJ, Stewart JW, Klein DF. A method to quantify rater bias in antidepressant trials. *Neuropsychopharmacology* 2000;22:559–565. [PubMed: 10788756]

- Piette JD. Interactive voice response systems in the diagnosis and management of chronic disease. *American Journal of Managed Care* 2000;6:817–827. [PubMed: 11067378]
- Rayamajhi, JN.; Lu, Y.; DeBrot, DJ.; Demitrack, MA.; Greist, JH. A comparison between interactive voice response system and clinician administration of the Hamilton Depression Rating Scale. Poster II-41. New Clinical Drug Evaluation Unit, 42nd Annual Meeting; Boca Raton, FL. 2002.
- Robinson DS, Rickels K. Guest editorial: Concerns about clinical drug trials. *Journal of Clinical Psychopharmacology* 2000;20:593–596. [PubMed: 11106129]
- Rush AJ, Bernstein IH, Trivedi MH, Carmody TJ, Wisniewski S, Mundt JC, Shores-Wilson K, Biggs MM, Woo A, Nierenberg AA, Fava M. An evaluation of The Quick Inventory of Depressive Symptomatology and The Hamilton Rating Scale for Depression: A sequenced treatment alternatives to relieve depression trial report. *Biological Psychiatry*. 2005
- Rush AJ, Gullion CM, Basco MR, Jarrett RB, Trivedi MH. The Inventory of Depressive Symptomatology (IDS): Psychometric properties. *Psychological Medicine* 1996;26:477–286. [PubMed: 8733206]
- Rush, AJ.; Ryan, ND. Current and emerging therapeutics for depression. In: Davis, KL.; Charney, D.; Coyle, JT.; Nemeroff, C., editors. *Neuropsychopharmacology The Fifth Generation of Progress*. Philadelphia, PA: Lippincott Williams & Wilkins; 2002. p. 1081-1095.
- Rush AJ, Trivedi MH, Ibrahim HM, Carmody TJ, Arnow B, Klein DN, Markowitz JC, Ninan PT, Kornstein S, Manber R, Thase ME, Kocsis JH, Keller MB. The 16-Item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. *Biol Psychiatry* 2003;54:573–583. [PubMed: 12946886]
- Santor DA, Coyne JC. Examining symptom expression as a function of symptom severity: Item performance on the Hamilton Rating Scale for Depression. *Psychological Assessment* 2001;13:127–139. [PubMed: 11281034]
- Scherer KR, Zei B. Vocal indicators of affective disorders. *Psychotherapy and Psychosomatics* 1988;49:179–186. [PubMed: 3070621]
- Stassen HH. Modeling affect in terms of speech parameters. *Psychopathology* 1988;21:83–88. [PubMed: 3249797]
- Stassen HH, Bomben G, Gunther E. Speech characteristics in depression. *Psychopathology* 1991;24:88–105. [PubMed: 1886971]
- Stassen HH, Kuny S, Hell D. The speech analysis approach to determining onset of improvement under antidepressants. *European Neuropsychopharmacology* 1998;8:303–310. [PubMed: 9928921]
- Uhlenhuth EH, Matuzas W, Warner TD, Thompson PM. Methodological issues in psychopharmacological research growing placebo response rates: The problem in recent therapeutic trials? *Psychopharmacology Bulletin* 1997;33:31–39. [PubMed: 9133749]
- Walsh BT, Seidman SN, Sysko R, Gould M. Placebo response in studies of major depression. *JAMA* 2002;287:1840–1847. [PubMed: 11939870]
- Williams JB. A structured interview guide for the Hamilton Depression rating scale. *Archives of General Psychiatry* 1988;45:742–747. [PubMed: 3395203]
- Wuyts F, De Bodt MS, Molenberghs G, Remacle M, Heylen L, Millet B, Van Lierde K, Raes J, Van de Heyning PH. The dysphonia severity index: An objective measure of vocal quality based on a multiparameter approach. *Journal of Speech, Language, and Hearing Research* 2000;43:796–809.
- Zigmond AS, Snaith RP. The Hospital Anxiety and Depression Scale. *Acta Psychiatrica Scandinavica* 1983;67:361–370. [PubMed: 6880820]
- Zimmerman M, Posternak MA, Chelminski I. Is it time to replace the Hamilton Depression Rating Scale as the primary outcome measure in treatment studies? *Journal of Clinical Psychopharmacology* 2005;25:105–10.
- Zung WWK. A self-rating depression scale. *Archives of General Psychiatry* 1965;12:63–70. [PubMed: 14221692]

Figure 1. Mean depression scores for each week of the study. Clinician-rated HAM-Ds were obtained during office visits at weeks 0, 2, 4, and 6. IVR measures were obtained every week.

Table 1

Reliabilities of voice acoustic measures and correlations with Hamilton Rating Scale for Depression (HAMD) severity.

VA measure	Reliability ^b	HAMD Association ^c	<i>p</i>
F ₀ COV ^a	.61	-.01	<i>ns</i>
F ₁ COV ^a	.24	-.03	<i>ns</i>
F ₂ COV ^a	-.02	-.17	< .05
Total recording length of speech sample	.57	.20	< .01
Total vocalization time	.66	.07	<i>ns</i>
Total pause time	.71	.29	< .001
Number of pauses	.71	.10	<i>ns</i>
Pause variability (St. Dev.)	.64	.38	< .001
Percent pause time	.04	.18	< .01
Vocalization/pause ratio	.48	-.22	= .001
Speaking rate (syllables/sec)	.65	-.23	< .001

^a Coefficient of Variation

^b Cronbach's alpha, assessed across speech tasks within each telephone call

^c Pearson's correlation

Table 2

Mean (SD) HAM-D change from baseline and treatment response (50% or greater reduction from baseline) by study week.

	Study Week					
	1	2	3	4	5	6
N ^a	33	34	30	34	29	32
HAM-D change	-1.7 (4.9)	-2.9 (5.0)	-4.3 (6.5)	-5.5 (7.0)	-6.3 (6.7)	-6.2 (8.7)
Responders	3	5	9	13	13	13

^a Differences in N reflect failures to complete IVR calls and/or uninterpretable speech recordings.

Table 3

Significant vocal acoustic change from baseline measures between treatment responders and non-responders at Week 6.

Change from Baseline ^a	Responders (13)		NonResponders (19)		t-test	p.
	Mean	SD	Mean	SD		
F ₀ COV	-0.007	0.008	0.003	0.013	2.42	.02
Total recording length	19.60	19.97	4.36	15.02	-2.45	.02
Total pause time	13.32	6.87	6.15	11.22	-2.05	.05
Number of pauses	32.85	34.38	5.95	33.30	-2.22	.03
Speaking rate	-0.75	0.56	-0.18	0.38	3.43	.002

^aBaseline – Week 6