Available online at www.sciencedirect.com
SCIENCE DIRECT®
Brain and Cognition 56 (2004) 30–35
Brain and Cognition
www.elsevier.com/locate/b&c

Voice acoustical measurement of the severity of major depression

Michael Cannizzaro,[a,b] Brian Harel,[a,c] Nicole Reilly,[a,c] Phillip Chappell,[d] and Peter J. Snyder[a,c,e,*]

[a] *Voice Acoustics Laboratory, Clinical Technology, Pfizer Global Research & Development, Groton, CT, USA*
[b] *Department of Communication Sciences, University of Connecticut, Storrs, CT, USA*
[c] *Department of Psychology, University of Connecticut, Storrs, CT, USA*
[d] *CNS Clinical Development, Pfizer Global Research & Development, New London, CT, USA*
[e] *Département de Psychologie, Centre de Neuroscience de la Cognition, Université de Québec à Montréal, Montréal, Que., Canada*

Accepted 11 May 2004
Available online 9 June 2004

Abstract

A number of empirical studies have documented the relationship between quantifiable and objective acoustical measures of voice and speech, and clinical subjective ratings of severity of Major Depression. To further explore this relationship, speech samples were extracted from videotape recordings of structured interviews made during the administration of the 17-item Hamilton Depression Rating Scale (HDRS; Hamilton, 1960). Pilot data were obtained from seven subjects (five males, two females) from videotapes that have been used to train expert raters on the administration and scoring of the HDRS. Several speech samples were isolated for each subject and processed to obtain the acoustic measurements. Acoustic measures were selected on the basis that they were correlated with HDRS ratings of symptom severity as seen under ideal voice recording conditions in previous studies. Our findings corroborate earlier reports that *speaking rate* is well correlated (negatively) with HDRS scores, with a strong correlation and nearly significant trend seen for the measure of *pitch variability*. A moderate pairwise correlation between *percent pause time* and HDRS score was also revealed, although this relationship was not statistically significant. The results from this cross-sectional study further demonstrate the ability of voice and speech signal analyses to objectively track severity of depression. In the present case, it is suggested that this relationship is robust enough to be found despite the less than ideal recording conditions and equipment used during the original videotape recording. Voice acoustical analyses may provide a powerful compliment to the standard clinical interview for depression. Use of such measures increases the range of techniques that are available to explore the neurobiological substrates of Major Depression, its treatment, and the dynamic interplay of the systems that govern the motor, cognitive, and emotional aspects of speech production.

1. Introduction

A growing body of research has documented the relationship between subjective estimates of the severity of Major Depression and observed qualitative and/or quantitative changes in speech production. For example, Stasen, Kuny, and Hell (1998) found a strong correlation between change in clinical ratings of symptom severity, and several key voice acoustic measures. This finding was consistent in more than 74% of 43 patients admitted to an inpatient service for pharmacologic treatment of severe Major Depressive Disorder (MDD). In particular, Stasen et al. (1998) reported that fundamental frequency amplitude and mean pause duration were among the voice acoustic variables that most closely track with improvement in symptom severity as measured by the 17-item Hamilton Depression Rating Scale (HDRS; Hamilton, 1960) over the first few weeks of treatment. The former acoustic measure relates to the muscular and respiratory effort exerted to control phonatory intensity and rate, whereas the latter measure relates to motor timing and the execution of speech movements. These findings corroborate earlier reports of increased speech pause time and total speech time

* Corresponding author.
E-mail address: peter_j_snyder@groton.pfizer.com (P.J. Snyder).

0278-2626/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.bandc.2004.05.003

being associated with higher scores on the HDRS and other scales of psychomotor speed and subject self-rating of mood (Hardy, Jouvent, & Widloecher, 1984; Teasdale, Fogarty, & Williams, 1980). In a longitudinal study by Ellgring and Scherer (1996), 11 female and 5 male inpatients were repeatedly examined over the course of treatment for MDD. Significant correlations were found between measures of speech rate, pause duration, and minimum fundamental frequency, and HDRS ratings for their female subjects (similar but non-significant trends were found for their male subjects, most likely due to a substantially smaller sample of males who participated in this study).

In all of the aforementioned studies, patients were treated with a wide variety of antidepressant medications, and regardless of the specific pharmacologic mechanisms of action, it appears that several speech acoustical measures consistently track with subtle changes in symptom severity. Moreover, these measures seem to be sensitive to early symptomatic improvement as well as the degree of response to drug intervention (Stasen et al., 1998). The observed changes seen in these voice acoustical measures quite likely reflect changes in the modulation of both serotonergic and dopaminergic neurotransmitter systems in response to treatment of Major Depression Disorder.

At the root of speech production lies an exceedingly complex and dynamic set of interactions between a number of neuromuscular systems, which subsequently effect the motor execution and production of the speech signal. Neuroradiological and clinical observations have identified differential contributions of the motor cortex, the supplementary motor area, the basal ganglia and the cerebellum within the speech production system (Wildgruber, Ackerman, & Grodd, 2001). While a great number of cortical and sub-cortical mechanisms are known to contribute to speech production, it is hypothesized that the functional contributions of basal ganglia structures play a crucial role in mediating the voice acoustical changes that have been linked to depressive symptomatology. Specifically, it is reasonable to suppose that the control of motor speech movements by basal ganglia structures (cf. Wildgruber et al., 2001) would be directly effected by changes in the modulation of dopaminergic tone that has been repeatedly associated with MDD.

Specifically, numerous studies have shown an inverse relationship between severity of depression and the abundance of homovanillic acid (HVA) and other principal metabolites of dopamine in persons diagnosed with depression (Engstrom, Alling, Blennow, Regnell, & Traskman-Bendz, 1999; Goodnick, Dominguez, DeVane, & Bowden, 1998; Lambert, Johansson, Agren, & Friberg, 2000; Santagostino et al., 1998; Swann, Katz, Bowden, Berman, & Stokes, 1999). As depression spectrum disorders (i.e., Bipolar illness, Dysthymia) are successfully treated by either tricyclic or serotonin reuptake inhibitor drugs, monoaminergic tone in basal ganglia structures and the nucleus acumbens appears to normalize (Pallis, Thermos, & Spyraki, 2001). Moreover, increasing severity of symptoms in Bipolar Depression appear to be well correlated with both deficits in performance on measures of psychomotor ability (e.g., speed, visual tracking, and dexterity) as well as with diminished dopaminergic turnover, as measured by levels of HVA in cerebrospinal fluid (Swann et al., 1999). Hence, it is plausible to assume that the relationship between improvement in psychomotor aspects of speech production are in part due to the underlying changes in dopaminergic tone associated with motor control. By objectively measuring the speech acoustic signal, we are quantifying the observed output of the neurological and physiological subsystems as they coordinate to create speech.

Previous literature supports the conclusion that two general aspects of speech and voice, specifically motor timing and fundamental frequency characteristics, are closely related to mood states as measured by both subject self-report and clinician-administered scales (Stasen et al., 1998). The purpose of this small, cross-sectional study was to: (1) replicate the principal results reported by Stasen et al. (1998); (2) determine whether such a replication might be possible under less-than-ideal recording conditions; and (3) evaluate the within-subject variability of these measures. This study differs from prior published work, in that we attempted to derive the relevant voice acoustical metrics from samples of free discourse (spontaneous speech during a structured interview), as opposed to the standard methods of eliciting responses to of a contextual speech and voice exercises.

2. Methods

A cross sectional design with five men and two women was used in this study. All participants were video and audio recorded during the structured interview assessment of the HDRS 21 question rating scale for depression. These videotapes were made as instructional training tapes for clinicians, and all subjects were interviewed by "master raters" (two of these tapes contained structured interviews administered by the author of the HDRS rating scale, Prof. M. Hamilton). These seven tapes have each been scored by hundreds of trained raters over the past several years, and the consensus HDRS scores for each case are provided below in Table 1A (*Note*: a higher score denotes increased symptom severity with an absolute minimum score = 0, maximum = 66). While 19 of the 21 areas assessed by the HDRS are based upon the content of the patient's speech (e.g., actual answers to the questions), two items

Table 1A
Consensus scores for HDRS and the observed scores for three acoustic measures

Subject	HDRS score	Speaking rate	% Pause time	Pitch variation
1	16	6.994	0.22	0.584275
2	22	3.444	0.26	0.273511
3	23	3.399	0.55	0.05639
4	23	3.367	0.24	0.10876
5	25	2.443	0.38	0.241992
6	26	3.106	0.28	0.146784
7	30	2.448	0.49	0.140525

are based upon clinical observation of the patient regarding psychomotor retardation and agitation. At most, three additional points may be added to a patient's overall depression rating based on the clinicians perception of slowness of thought or speech. This aspect of a patient's speech/thought is based on subjective clinical rating and not objective quantification of actual speaking rate, it is not felt that this measure will confound the results of this study. Because we were unable to control the placement of the microphone during the taping of these sessions, intensity measurements were not thought to be reliable and therefore were not investigated. All recordings were made in standard examination rooms, without the benefit of sound baffling or high-quality recording equipment. Measures of speech prosody and speech timing previously found to be sensitive to depression severity and treatment response (Stasen et al., 1998) were made using the acoustic signal extracted from the videotape.

2.1. Signal capture and manipulation

For each subject, the audio portion of tape was examined until the first 10 s of the participant's uninterrupted speech was located. Five of the seven tapes contained additional segments of uninterrupted speech, and for these cases up to three 10-s segments per subject were chosen to evaluate within-subject variability of measurement. These additional segments represented successively, the next available portions of the patients' speech lasting over 10 s. Average intra-subject variability was very small for all three measures, with the smallest variability of measurement of 0.005% seen for percent pause time. The average intra-subject variability for the measure of speaking rate was 0.003 and 0.269% for pitch variability. Because these estimates of measurement variability were so small, in cases for which more than one sample was obtained the average scores across samples, for each subject, were used.

Speech samples were captured and digitized using the Kay Elemetrics Computerized Speech Laboratory 4400 (CSL; Kay Elemetrics Corporation, 2001). Visual inspections of the spectrographic representations of the patients' speech segments were made to ensure integrity of the videotaped audio recordings. All segments were judged to be suitable for analysis based upon the visible harmonic patterns representing the fundamental frequency and successive formants. Signals contained the bandwidths of interest for speech measures (signal between 0 and greater than 11 KHz normalized spectrographic window) for representing prosodic contours of fundamental frequency. The middle five seconds of each 10-s speech sample was identified by spectrographic and waveform analysis to locate the onset of acoustical energy signaling the beginning of a word. The endpoint of the five second epoch samples was determined by placing the second vertical cursor approximately five seconds later in the time, at the junction of words, within the CSL amplitude waveform window (see Fig. 1). These middle five second clips represent uninterrupted free speech discourse of the participant without the complicating factors of sentence initial or sentence final acoustic variability. Once these five second epoch samples had been isolated, the three acoustic measures (see below) were made on each sample.

2.2. Acoustic variables

Speaking rate was calculated by dividing the number of syllables spoken by the length of the sample measured

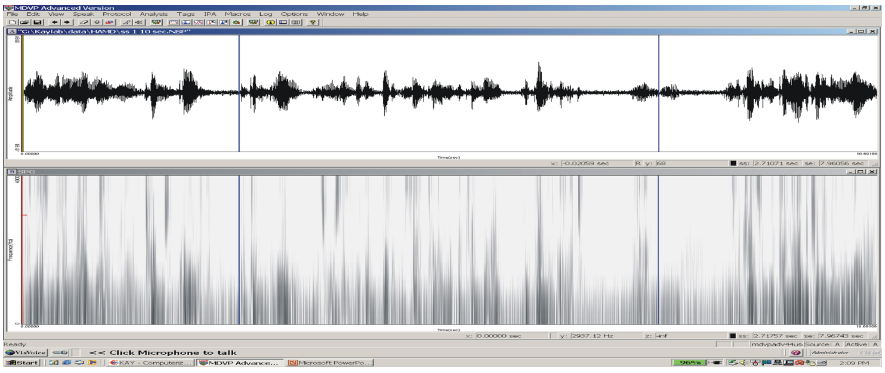

Fig. 1. Signal Capture and Standardization Window (Kay Elemetrics 4400 CSL Software Package). The period between the two vertical lines represents a five second interval of uninterrupted free speech.

in seconds. Using the spectrographic representation of the speech signal, vertical cursors are placed at the initiation of acoustic energy representing the onset of a word and again at the end of acoustic energy approximately five seconds later signaling the offset of a word. The time between the vertical cursors is the sample time recorded in seconds. The number of syllables spoken is counted using *The American Heritage College Dictionary* (Third edition, 2000) as a guide to syllable breakdown of each utterance based upon patterns of American English. This measure is thought to represent the interaction of cognitive, affective, and motoric influences on speech, and it is widely considered to be a useful measure of speech motor control (Robb, Maclagan, & Chen, 2004).

Percent pause time was measured by removing all silent pauses over 250 ms in length, and dividing by the total sample time. This number is then subtracted from 1 and multiplied by 100 to obtain a *percent* pause time score. Pauses of over 250 ms are thought to reflect cognitive and emotive aspects of paralinguistic speech patterns.

Pitch variation was extracted from the CSL pitch contour analysis for each sample. The pitch and standard deviation of fundamental frequency are reported in Hertz (Hz). To derive a standardized measure of pitch variation that is comparable between persons with markedly different fundamental frequencies (i.e., males vs. females), a coefficient of variation of fundamental frequency was calculated. This number is derived by dividing the standard deviation of the fundamental frequency by the average fundamental frequency for each sample. The listener perceives this measured aspect of speech as lying on a continuum of variation from monotonic (relatively little pitch variation) to highly intonated (relatively large pitch variation) speech patterns.

3. Results

Data for all three of the dependant variables, described above, as well as the consensus HDRS ratings for each subject, are provided in (Table 1A). For the seven subjects, scores on the HDRS range from 16 (least symptom severity) to 30, the most severely depressed subject. All participants in this study were classified as having moderate-to-severe depression (a score $\geqslant$17; Hamilton, 1960) except for one male who was classified as having mild to moderate depression (HDRS Score = 16).

The Pearson product-moment correlation coefficients and significant probability of the relationships between the HDRS and the three acoustic measures are provided in Table 1B. An alpha level of .05 was specified for all pairwise comparisons for indication of a significant linear relationship between dependant measures. The relationship between speaking rate and HDRS was the strongest with scores demonstrating a significant negative correlation ($r = -.089$, $p = 0.0076$). That is, as HDRS scores increase (indicating greater symptom severity), speaking rates significantly decreased. Pitch variation also demonstrated a large negative correlation with HDRS scores although this relationship did not reach statistical significance ($r = -0.74$, $p = .0581$) with this small sample size. Although percent pause time correlations are moderately correlated to HDRS scores ($r = .55$), the high probability of making a Type I error suggests that this relationship should be interpreted cautiously. Relationships with r-values within the range of 0.3 to 0.5 are considered to be moderate (Cohen, 1988).

4. Discussion

Two of the three speech acoustical variables demonstrated a large correlation with consensus HDRS rating scores for this small cross-sectional sample of seven subjects. Speaking rate, in particular, was significantly correlated with symptom severity. While this finding has been shown consistently in depression, other CNS disease related to dopaminergic tone, mainly Parkinson's disease, has not shown this clear pattern. Parkinson's disease has been associated with increased, decreased, and typical speech rates, depending on the investigation (cf. Flint, Black, Campbell-Taylor, Gailey, & Leviton, 1992). A plausible explanation for this disparity is likely due to the severity of the disease state and the trade off of between reduced range of articulator movement in Parkinson's hypo-kinetic dysarthria and movement rate. Persons with Parkinson's disease have been shown to increase speaking rate as a function of reduced articulator movement in turn sacrificing articulation clarity (Goberman & Coelho, 2002). This compensation

Table 1B
Pairwise correlations between dependent variables

Variable	By variable	Correlation	Signif. prob
Speaking rate	HDRS score	−0.8876	0.0076
Percent pause	HDRS score	0.5512	0.1997
Pitch variation (coefficient of variation)	HDRS score	−0.7382	0.0581

produces smaller movements and faster speech in direct opposition to rigid muscle tone. Conversely, a person with depression does not need to compensate for hypokinetic muscle tone and will continue to articulate clearly although at a reduced rate. Reduced rates of speech have been demonstrated in the later and more severe stages of Parkinson's disease when articulatory compensation is ineffective in maintaining a more typical speech rate (Goberman & Coelho, 2002). A strong correlational trend was also evident for pitch variation. This trend was not, however, evident for the measure of percent pause time.

One potential reason for the less robust association between symptom severity and the percent pause time measure may lie in the fact that there are basically two ways in which one can alter (slow) the time required to complete any particular utterance. The first involves the insertion of longer or more frequent pauses ($\geqslant$250 ms) increasing the time it takes to complete an utterance. In our sample this does not appear to be the case, at least within this small series of seven subjects with MDD. Alternatively, and as the results indicate in the current study, the increased speaking time may result from a decrease in the rate of speech sound production. This is thought to occur as a result of the effect of the disorder on the neural systems that guide psychomotor speed. In this case, the expected acoustic results would be decreased speaking rate without increased pause time or articulation rate (Robb et al., 2004).

These findings are in direct agreement with findings from other studies that have reported psychomotor slowing in depressed individuals, in which slowed motor performance has been found to correlate with a decrease in dopaminergic tone (as measured, for instance, by its principal metabolites in the CSF) and an increase in depression severity (Swann et al., 1999). This decrease in motor speed and agility can lead to the repeatedly observed decrease in pitch variation as symptom severity increases. Pitch variation and range is governed by the laryngeal musculature, which would also be effected by psychomotor retardation, resulting in reduced vocal variability or monotonous speech prosody. This finding is also strongly indicated by these results.

5. Conclusion

The results of this study lead to several conclusions. First, several quantitative measurement of speech acoustics are well correlated with the careful subjective measurement of mood state and symptom severity in MDD. These measures are likely related to changes in the neurobiology of the disease specific to dopaminergic tone, reflected in the acoustic output. Secondly, this procedure can be easily accomplished during a standard clinical interview for depression by making a recording of the structured conversation. This does not require added effort on the part of the examiner or increased task time or task effort on the part of the examinee. Finally, although best-case scenario dictates that a strictly controlled recording environment is the ideal, less than perfect conditions can yield useful information regarding these objective measures of depression severity.

References

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New Jersey: Lawrence Erlbaum.

Ellgring, H., & Scherer, K. R. (1996). Vocal indicators of mood change in depression. *Journal of Nonverbal Behavior, 20*, 83–110.

Engstrom, G., Alling, C., Blennow, K., Regnell, G., & Traskman-Bendz, L. (1999). Reduced cerebrospinal HVA concentrations and HVA/5-HIAA ratios in suicide attempters. Monoamine metabolites in 120 suicide attempters and 47 controls. *European Neuropsychopharmacology, 9*, 399–405.

Flint, A. J., Black, S. E., Campbell-Taylor, I., Gailey, G. F., & Leviton, C. (1992). Acoustic analysis in the differentiation of Parkinsons's disease, and major depression. *Journal of Psychlinguistic Research, 21*, 383–399.

Goberman, A. M., & Coelho, C. A. (2002). Acoustic analysis of Parkinsonian speech I: Speech characteristics and L-Dopa therapy. *NeuroRehabilitation, 17*, 237–246.

Goodnick, P. J., Dominguez, R. A., DeVane, C. L., & Bowden, C. L. (1998). Bupropion slow-release response in depression: Diagnosis and biochemistry. *Biological Psychiatry, 44*, 629–632.

Hamilton, H. (1960). HDRS: A rating scale for depression. *Journal of Neurosurgery and Psychiatry, 23*, 56–62.

Hardy, P., Jouvent, R., & Widloecher, D. (1984). Speech pause time and the Retardation Rating Scale for Depression (ERD): Towards a reciprocal validation. *Journal of Affective Disorders, 6*, 123–127.

Kay Elemetrics Corporation, 2001. Software instruction manual: Multi-Speech: model 3700 CSL models 4100, 4300B, and 4400. Version 2.4.

Lambert, G., Johansson, M., Agren, H., & Friberg, P. (2000). Reduced brain norepinephrine and dopamine release in treatment of refractory depressive illness: Evidence in support of the catecholamine hypothesis of mood disorders. *Archives of General Psychiatry, 57*, 787–793.

Pallis, E., Thermos, K., & Spyraki, C. (2001). Chronic desipraimine treatment selectively potentiates somatostatin-induced dopamine release in the nucleus accumbens. *European Journal of Neuroscience, 14*, 763–767.

Robb, M., Maclagan, M., & Chen, Y. (2004). Speaking rates of American and New Zealand Varieties of English. *Clinical Linguistics and Phonetics, 18*, 1–15.

Santagostino, G., Cucchi, M. L., Frattini, P., Zebri, F., Di Paolo, E., Preda, S., & Corona, G. L. (1998). The influence of alprazolam on the monoaminergic neurotransmitter systems in dysthymic patients. Relationship to clinical response. *Pharmacopsychiatry, 31*, 131–136.

Stasen, H. H., Kuny, S., & Hell, D. (1998). The speech analysis approach to determining onset of improvement under antidepressants. *European Neuropsychopharmacology, 8*, 303–310.

Swann, A. C., Katz, M. M., Bowden, C. L., Berman, N. G., & Stokes, P. E. (1999). Psychomotor performance and monoamine function in bipolar and unipolar affective disorders. *Biological Psychology, 45*, 979–988.

Teasdale, J. D., Fogarty, S. J., & Williams, J. M. G. (1980). Speech rate as a measure of short term variation in depression. *British Journal of Social and Clinical Psychology, 19*, 271–278.

Wildgruber, D., Ackerman, H., & Grodd, W. (2001). Differential contributions of the motor cortex, basal ganglia, and cerebellum to speech motor control: Effects of syllable repetition rate evaluated by fMRI. *Neuroimage, 13*, 101–109.