

Detecting Depression Severity from Vocal Prosody

Ying Yang, Catherine Fairbairn, and Jeffrey F. Cohn *Associate Member, IEEE*

Abstract—To investigate the relation between vocal prosody and change in depression severity over time, 57 participants from a clinical trial for treatment of depression were evaluated at seven-week intervals using a semi-structured clinical interview for depression severity (Hamilton Rating Scale for Depression: HRSD). All participants met criteria for Major Depressive Disorder at week 1. Using both perceptual judgments by naive listeners and quantitative analyses of vocal timing and fundamental frequency, three hypotheses were tested: 1) Naive listeners can perceive the severity of depression from vocal recordings of depressed participants and interviewers. 2) Quantitative features of vocal prosody in depressed participants reveal change in symptom severity over the course of depression. And 3) Interpersonal effects occur as well; such that vocal prosody in interviewers shows corresponding effects. These hypotheses were strongly supported. Together, participants' and interviewers' vocal prosody accounted for about 60% of variation in depression scores, and detected ordinal range of depression severity (low, mild, and moderate-to-severe) in 69% of cases ($\kappa = 0.53$). These findings suggest that analysis of vocal prosody could be a powerful tool to assist in depression screening and monitoring over the course of depressive disorder and recovery.

Index Terms—Prosody, switching pause, vocal fundamental frequency, depression, interpersonal influence, Hierarchical Linear Modeling (HLM).

1 INTRODUCTION

DIAGNOSIS and assessment of symptom severity in mental health are almost entirely informed by what patients, their families, or caregivers report. Standardized procedures for incorporating nonverbal behavior and vocal prosody, in particular, are lacking. Their absence is especially salient for depression, a mood disorder for which disruption in emotion experience, communication, and self-regulation are key features [3], [12], [14], [17]. Within the past decade, significant progress has been made in linking vocal prosody to emotion [22], [28], [44], [47], turn-taking, reciprocity [15], [41], and a broad range of interpersonal outcomes [26], [38]. There is strong reason to believe that automatic analysis of vocal prosody could provide a powerful tool to assist in detection and assessment of depression over the course of treatment and recovery. Improved measurement and understanding of the relation between depression and vocal prosody could aid early detection and lead to improved interventions. Because depression is one of the most prevalent mental health disorders [30] and a leading cause of disability worldwide [36], the potential contribution of improved measurement is great.

Vocal prosody is a composite of supra-segmental acoustic features of speech (i.e., beyond the lexical, syntactic, and semantic content of the signal). Primary

features are fundamental frequency (F_0), which is perceived as pitch; intensity, which is perceived as loudness; and timing, which is perceived as speech rate, rhythm, and patterning in normal conversation. Related features include jitter and shimmer (cycle-to-cycle variation in frequency and intensity), energy distribution among formants, and cepstral features. Many of these features have been explored with respect to emotion expression [22], [28], [43], [44] and to a lesser extent depression, as noted below. In our research, we focus on timing and F_0 , which have been emphasized in the psychology of emotion and nonverbal behavior.

Unlike most studies that have compared depressed and non-depressed participants with respect to *intrapersonal* behavior (e.g., timing of pauses within a speaking turn) at a single point in time, we focus on both *intra-* and *interpersonal* behavior within a clinical sample over the course depression. We investigate whether vocal prosody varies with severity of depression and identify inter-personal effects of depression (e.g., longer and more variable turn-taking when depression is most severe). We use perceptual judgment studies to investigate whether people can perceive vocal prosody of depression, and quantitative methods to investigate the extent to which features of vocal prosody can reveal change in symptom severity over the course of depressive disorder.

From a psychopathology perspective, one would expect depression to be associated with decreased intensity, irregular timing, and decreased F_0 variability. These features are conceptually related to what is referred to as psychomotor retardation, or slowing, insensitivity to positive and negative stimuli, and the attenuated interest

• Y. Yang is with the Rehabilitation and Neural Engineering Laboratory, University of Pittsburgh.

E-mail: yiy17@pitt.edu

• C. Fairbairn and J.F. Cohn are with the Department of Psychology, University of Pittsburgh.

Email: cef24@pitt.edu and jeffcohn@cs.cmu.edu.

Corresponding author: Jeffrey F. Cohn

in other people that are common in depression.

Two sets of findings are consistent with the hypothesis that prosody reveals depression. One is cross-sectional comparison between persons with and without depression. These studies suggest that vocal prosody strongly covaries with depressive symptoms. If further validated, such studies could support the utility of using vocal prosody to screen for evidence of depression [39]. The other, and more challenging, is longitudinal studies of change in depression over the course of a depressive episode. If successful, this line of research could have significant impact on treatment planning and evaluation of treatment efficacy.

A related issue is the influence of depression on other persons. Because depression occurs in social contexts, it is likely have reciprocal effects on interlocutors. Two early studies found that depressed mothers are slower and more variable in their responses to their infants [4], [48], which may lead to changes in how their infants in turn respond to them. In an analog study, Boker and Cohn [6], [7] found that young adults became more expressive in response to dampened facial and vocal expression of peers in a computer-mediated interaction. In actual depression, interpersonal effects could differ. While the initial reaction to depressed individuals may be attempts to elicit responsiveness, the experience may soon become aversive and prompt efforts to withdraw [13]. We will include evidence for interpersonal influence in our review of the two types of studies.

1.1 Cross-Sectional Studies

Cross-sectional studies compare individuals with and without depression at a single point in time. At least seven cross-sectional studies [4], [11], [21], [33], [37], [46], [48] have compared prosodic features in relation to presence of depression, as assessed using diagnostic interviews or less specific symptom rating scales¹. While these studies vary with respect to which prosodic features they consider, overall they find that prosodic features discriminate between individuals with and without depression. Adults with depression in comparison with non-depressed persons have slower, less consistent timing, lower intensity, and less variable F_0 . With the exception of [11], all involved comparisons of individuals with and without depression at a single point in the disorder. [11] found that change in severity of depression covaried with vocal prosody. Possible interpersonal influence has been neglected.

1.2 Longitudinal Studies

The cross-sectional findings suggest that prosody may be a useful marker of depression. However, the question remains whether the discriminability of prosodic patterns

¹ Symptoms of depression may result from other disorders, diseases, or causes. In part for this reason, self-report measures of depression may correlate only moderately with diagnosis as determined by clinical interview [10]. For diagnosis, it is necessary to rule out other factors [3].

that have been found are specific to depression or are common to the types of people most likely to become depressed. Depression is strongly related to individual differences in neuroticism, introversion, and conscientiousness [31]. These personality characteristics remain relatively stable across the lifespan. Differences in vocal prosody between those with and without depression could be revealing of personality differences rather than time-limited variation in depression. Thus, personality rather than depression per se may account for much or all of the between-group differences in vocal prosody that have been reported previously [42].

To investigate whether vocal prosody varies as individuals recover from depression, longitudinal studies are needed that assess change in depression severity over the course of depressive disorder. The few that exist [1], [16], [32], [34] suggest that vocal timing and F_0 may be responsive to recovery from depression. [32] and [1] found that intra-personal pause duration and speaking rate are closely related to change in depression severity over time. With one exception [34], however, relevant studies have been limited to inpatient samples that are more severely depressed than those found in the community. They also tend to use structured speaking tasks, which leave open the question of whether vocal prosody in depression impacts interlocutors and turn-taking, which is known to influence rapport [27]. We asked whether vocal prosody in clinical interviews varies with change in depression severity and the extent to which it influences the vocal prosody of interviewers, who are not themselves depressed.

1.3 Hypotheses and Study Design

To investigate the relation between change in depression severity and vocal prosody, we recruited participants and interviewers from a clinical trial for treatment of depression. Participants and interviewers were observed from recordings of clinical interviews at seven-week intervals over the course of treatment. Using convergent measures (perceptual judgments and quantitative measures of vocal timing and F_0), we tested three hypotheses.

One, naive listeners will perceive differences in vocal prosody related to depression severity. This hypothesis evaluates whether vocal prosody in depression is perceivable, and thus potentially could influence the vocal prosody of non-depressed people with whom depressed persons communicate. Two, for a given participant, specific features of vocal prosody will co-vary with the change in depression severity. When depression is moderate to severe, F_0 will be lower and less variable and switching pauses longer and less predictable than when depression is remitted (i.e. no longer clinically significant). Switching pause is the time between one speaker's "turn" and that of the other. Three, interpersonal effects will be found in the vocal prosody of interviewers. We investigate whether vocal timing and F_0 variability in depressed participants could be contagious.

The first question was investigated in Study 1. Naive listeners rated the severity of depression from brief segments of low-pass filtered audio recordings of symptom interviews. Filtering rendered speech unintelligible while preserving prosody. In this way, verbal content did not confound ratings of depression. In Study 2, the role of specific prosodic features was investigated in the full data set using quantitative methods.

2 METHODS

The primary data were audio recordings of clinical interviews. As noted above, the audio recordings were analyzed two ways. Study 1 was a perceptual study in which naive listeners rated the severity of depression in low-pass filtered recordings of the interviews. The goal of Study 1 was to determine whether listeners could detect differences in severity from the vocal exchanges of depressed participants and clinical interviewers. Because the capacities of human listeners are limited relative to machine processing, Study 1 used only the first three questions of the interview and a subset of the recordings for which depression score was either low (HRSD score of seven or less) or moderate to severe (HRSD score of 15 or higher). Extreme groups were chosen to maximize variance. Audio from 26 interviews was used. The independent variable was ratings; the dependent variable was two ranges of depression score (HRSD).

Study 2 investigated how prosodic features of the depressed participants and their clinical interviewers may reveal depression severity. Study 2 used the full length of all audio recordings and the full range of depression scores (integer values ranging from 0 to 35). The independent variables were prosodic features (e.g., switching pause mean); the dependent variable was depression score. Thus, the two studies differed in the number and length of audio recordings, types of independent variables, and representation of depression score.

In this section, we describe the depressed participants, the observational and clinical procedures with which severity was ascertained, and procedures specific to Study 1 (perceptual ratings) and Study 2 (specific prosodic features). We refer to participants (or listeners) in the ratings study as “raters;” depressed participants in the clinical interviews as “participants;” and clinical interviewers, who also effectively were participants, as “interviewers.”

2.1 Participants

Fifty-seven depressed participants (34 women, 23 men) were recruited from a clinical trial for treatment of depression. They ranged in age from 19 to 65 years (mean = 39.65) and were Euro- or African-American (46 and 11, respectively). At the time of study intake, all met DSM-IV [3] criteria [18] for Major Depressive Disorder (MDD). MDD is a recurrent disorder, and the participants all had had prior episodes (Range = 1-8, mean =

3.15). Although not a focus of this report, participants were randomized to either anti-depressant treatment with a selective serotonin reuptake inhibitor (SSRI) or Interpersonal Psychotherapy (IPT). Both treatments are empirically validated for treatment of depression [25]. Of the 57 participants, data from 7 could not be included because it was either missing or invalid at the initial (week 1) visit. In two cases, the week 1 visit did not take place; in three others audio was not recorded; and in two, participants were chewing gum, which would have been a potential confound.

2.2 Interview and Observational Procedures

Symptom severity was evaluated on up to four occasions at 1, 7, 13, and 21 weeks by clinical interviewers (11, all female). Interviewers were not assigned to specific participants, and they varied in the number of interviews they conducted. Five interviewers were responsible for the bulk of the interviews. The median number of interviews per interviewer was 17; five conducted six or fewer.

Interviews were conducted using the Hamilton Rating Scale for Depression (HRSD) [23], which is a criterion measure for assessing severity of depression. Interviewers all were expert in the HRSD and reliability was maintained above 0.90. HRSD scores of 15 or higher are generally considered to indicate moderate to severe depression; and scores of 7 or lower to indicate a return to normal [20].

Interviews were recorded using four hardware-synchronized analogue cameras and two unidirectional microphones. Two cameras were positioned approximately 15° to the participant’s left and right to record their shoulders and face. A third camera recorded a full-body view while a fourth recorded the interviewer’s shoulders and face from approximately 15° to their right. Audio was digitized at 48,000 Hz. Findings from the video data will be subject of another report.

Missing data occurred due to missed appointments, attrition, or technical problems. Two participants were transferred to another protocol when they showed evidence of suicidal intent. Technical problems included failure to record audio or video, audio or video artifacts, and insufficient amount of data. To be included for analysis, we required a minimum of 20 speaker turns and 100 seconds of vocalization. Thus, the final sample was 130 sessions from 49 participants.

2.3 Signal Processing

Because audio was recorded in a clinical office setting rather than an anechoic chamber or other laboratory setting, some acoustic noise was unavoidable. To attenuate noise as well as to equalize intensity and remove any overlap between channels (i.e. a speaker’s voice occurring on both channels), Adobe Audition II [38] was used to reduce noise level and equalize intensity. An intermediate level of 40% noise reduction was used to achieve

the desired signal-to-noise ratio without distorting the original signal.

To remove overlap between channels and precisely measure timing, a supervised learning approach was used. Each pair of recordings was transcribed manually using Transcriber software [8] and then force-aligned using CMU Sphinx III [45] post-processed using Praat [5]. Because session recordings exceeded the memory limits of Sphinx, it was necessary to segment recordings prior to forced alignment. While several approaches to segmentation were possible, we segmented recordings at transcription boundaries; that is, whenever a change in speaker occurred. Except for occasional overlapping speech, this approach resulted in speaker-specific segments. This approach may have increased the accuracy of forced alignment because cepstral features extracted each time were based on only a single utterance.

Forced alignment produced a matrix of four columns: speaker (which encoded both individual and simultaneous speech), start time, stop time, and utterance. To assess the reliability of the forced alignment, audio files from 30 sessions were manually aligned and compared with the segmentation yielded by Sphinx. Mean error (s) for onset and offset, respectively, were .097 and .010 for participants and .053 and .011 for interviewers.

2.4 Study 1: Perceptual Ratings

To maximize experimental variance [29], interviews were selected from sessions having HRSD scores of 7 or less (absence of depression) or 15 or higher (moderate to severe). Interviews were randomly sampled with the constraint that no more than one session could be included from any participant. Fifteen sessions had HRSD scores of 7 or less; 11 had scores of 15 to 25. The former were from week 7 or 13, and the latter from week 1.

Audio samples were limited to the first three questions of the HRSD, thus providing relatively “thin slices” of behavior [2]. To eliminate recognizable speech, the recordings were low-pass filtered using an 850Hz threshold. A higher threshold of 1000Hz was considered initially, but some intelligible speech remained. The 850Hz threshold proved sufficient. To convey the back and forth of the interview, audio from the interviewer and the participants was digitally separated using CMU Sphinx [45] and Praat [5] and played over separate speakers positioned approximately 8 feet apart. All audio was played at uniform intensity. Order of presentation was random.

The raters were eight young adults. Six were women, and two were men. Seven were Euro-American and one was Hispanic. All were blind to depression status. They were told that they would listen to a series of audio clips extracted from interviews; that the audio clips had been low-pass filtered so that the speakers’ speech would be unintelligible; and that the interviewer and interviewee voices would be heard through separate audio speakers; the interviewer to their left and the interviewee to their right.

Using a Likert scale, they were asked to rate the severity of interviewee depression from 0 (none) to 6 (most severe). To minimize error and maximize effective reliability [42], ratings were averaged across raters. The intraclass correlation for the depression ratings was $r = 0.94$, $p < 0.001$, which indicates high internal consistency. The participants also rated the extent to which the conversation seemed awkward and the extent to which the conversation seemed comfortable using similar Likert scales. Because all of the ratings were highly correlated (all $r > .85$), only the depression ratings were analyzed to avoid redundancy.

2.5 Study 2: Prosodic Features

2.5.1 Switching Pause Duration

Switching pause (SP), or latency to speak, was defined as the pause duration between the end of one speaker’s utterance and the start of an utterance by the other. Switching pauses were identified from the matrix output of Sphinx. So that back channel utterances would not confound switching pauses, overlapping voiced frames were excluded. Switching-pauses were aggregated to yield mean duration and coefficient of variation (CV) for both participants and interviewers. The CV (σ/μ) is the ratio of standard deviation to the mean. It reflects the variability of switching pauses when the effect of mean differences in duration is removed.

2.5.2 Vocal fundamental frequency (F_0)

For each utterance, vocal fundamental frequency (F_0) was computed automatically using the autocorrelation function in Praat [5] using a window shift of 10 ms. As with switching pause, we computed mean and coefficient of variation of F_0 for both participants and interviewers. Because microphones had not been calibrated for intensity, intensity measures were not considered. Thus, we analyzed prosodic features from two of the three domains of prosody (timing and frequency) for both participants and interviewers.

3 RESULTS

We first present descriptive data with respect to change in depression severity over time. We then present results from Study 1 (perceptual judgments) and Study 2 (specific prosodic features), respectively. Due to the nature of the sampling procedures employed in Study 1 (See Methods above) the depression measure in Study 1 was dichotomous (i.e. low and moderate to severe). We therefore treated HRSD as a binary outcome variable in this study, using regression procedures that account for the non-normal distribution. In contrast, Study 2 sampled the full range of HRSD scores (mean = 12.73, standard deviation = 7.22, range = 0 to 35). Therefore, analytic procedures assuming a normally distributed outcome were employed. In follow-up analyses in Study 2, discriminant analysis was used to detect range of depression severity.

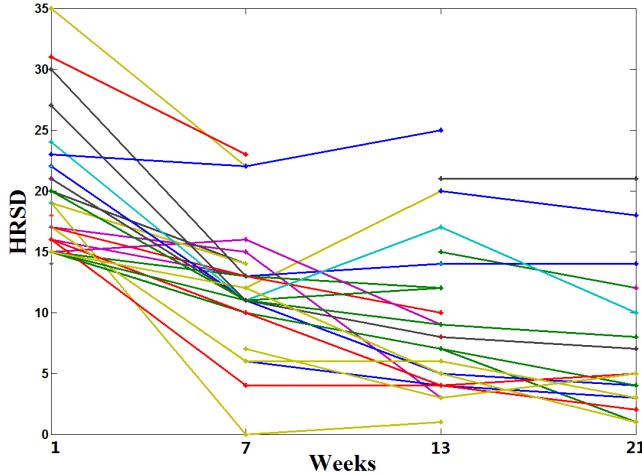


Fig. 1. Change in depression severity (HRSD score) from interviews 1 through 4 at weeks 1, 7, 12, and 21, respectively. Scores of 15 or higher are considered moderate to severe depression. Scores of 7 or lower indicate absence of clinically significant symptoms. Breaks in the individual trajectories indicate missing observations.

TABLE 1
Predicting Depression Severity from
Perceptual Judgments

Predicted	HRSD	
	Low	Moderate to Severe
Low	12	4
Moderate to Severe	3	7
Wald=4.683, df _{1,24} , p=0.03		

3.1 Course of depression

Over the course of the study, depression severity decreased for most participants. At week 7, about 20% remained above clinical threshold for moderate to severe depression, which is an HRSD score of 15 or higher. By week 21, only 10% were still above this threshold. Symptoms remitted ($HRSD \leq 7$) in 55%. In a few cases, severity increased after initially decreasing. Individual trajectories of depression symptoms are shown in Fig. 1.

3.2 Study 1: Perceptual ratings of depression

To evaluate the association between perceptual ratings and depression severity, we used logistic regression. Regression of severity group onto perceived depression predicted 73% of cases ($\text{Wald} = 4.683, p = .03$). The corresponding kappa coefficient, which corrects for agreement due to chance [19], was 0.44, which suggests moderate predictability. Discriminability was higher for low severity than for moderate to severe (Table 1).

3.3 Study 2: Prosodic Features

In the full longitudinal dataset, recordings from 49 participants on multiple occasions (as many as 4 times over 21 weeks) were used, producing two distinct comparisons: between-subject cross-sectional comparisons

and within-subject longitudinal comparisons. Between-subject comparisons evaluate average differences between participants when their severity scores are averaged across time. For instance, do participants with higher or lower averaged severity scores differ on the vocal response measures? Within-subject comparisons evaluate the variability over time of each participant's scores on the vocal response measures. For instance, is change in severity within participants revealed by corresponding changes in vocal response measures? The latter is a key question. It informs whether we can know whether an individual's depression severity has changed by attending to their vocal behavior.

These two sources of variation (between- and within-subject comparisons) in depression severity were accounted for and partialled using Hierarchical Linear Modeling (HLM) [9], [40]. HLM can be considered an extension of multiple regression that enables separation of between- and within-subject effects while remaining robust to missing observations, which are common in longitudinal behavioral research. Support vector regression would not enable inclusion of within-subject effects.

In the HLM model building procedure, both interviewer and participant scores were entered as predictors. By entering both in the model, we were able to isolate the predictive effects of each individual's prosody while controlling for the prosody of the other. Between-subject factors were entered at level two, and within-subject factors were entered at level one. Averaged variables for each subject were entered at level two to isolate between-subject variation. Group mean centering of variables at level one was applied to isolate within-subject variation. Separate models were used for switching pause and F_0 . Because sex was unrelated to depression severity, it was omitted from the models. Table 2 reports descriptive statistics averaged across interviews for each of the measures.

TABLE 2
Descriptive Statistics

	Participant Mean	Participant SD	Interviewer Mean	Interviewer SD
HRSD (Depression Score)	12.73	7.22	—	—
Switching Pause (s)	0.69	0.26	0.68	0.26
Switching Pause CV (s)	1.01	0.24	1.07	0.31
F_0 (Hz)	198.35	35.40	213.43	22.23
F_0 CV (Hz)	0.23	0.09	0.20	0.07

Note. SD = standard deviation, CV = coefficient of variation (SD/mean).

3.3.1 SWITCHING PAUSE

No between-subject differences were found in switching pause mean or CV for either participants or interviewers.

In contrast, within-subject effects for participant switching pause mean and variability (CV) and interviewer variability were highly significant. As depression severity decreased, participant switching pauses became shorter and less variable and interviewer switching

TABLE 3
HLM statistics for prediction of depression severity (HRSD) from switching pause

Effect	Switching Pause Mean					Switching Pause CV				
	Value	s.e.	t	df	p	Value	s.e.	t	df	p
Between-subjects										
Participant	0.006	0.027	0.21	46	N.S.	6.607	4.356	1.53	46	N.S.
Interviewer	-0.071	0.042	-1.46	46	N.S.	-1.834	4.180	-0.440	46	N.S.
Within-subjects										
Participant	0.086	0.027	3.15	77	0.002	8.823	3.297	2.69	77	0.009
Interviewer	0.050	0.041	1.21	77	N.S.	7.489	2.281	3.28	77	0.002

Note: s.e. = standard error, t = t-ratio, df = degrees of freedom, p = probability

TABLE 4
HLM statistics for prediction of depression severity (HRSD) from F_0

Effect	F_0 mean					F_0 CV				
	Value	s.e.	t	df	p	Value	s.e.	t	df	p
Between-subjects										
Participant	-0.024	0.026	-0.93	46	N.S.	3.208	10.064	0.319	46	N.S.
Interviewer	-0.108	0.044	-2.43	46	0.019	24.742	10.613	2.331	46	.024
Within-subjects										
Participant	0.006	0.039	0.15	78	N.S.	-15.388	13.526	-1.138	78	N.S.
Interviewer	-0.061	0.300	-2.04	78	0.045	20.642	8.672	2.380	78	0.020

Note: s.e. = standard error, t = t-ratio, df = degrees of freedom, p = probability

pauses became less variable as well. (Table 3). In order to estimate the combined effect size of switching pause mean and variance, coefficients reaching significance were entered as predictors in the same model. Together, these variables accounted for 32.04% of the variation over time in a subject's depression score. In behavioral science, the variance accounted for is a criterion for how well a model performs.

3.3.2 VOCAL FUNDAMENTAL FREQUENCY

For interviewers but not participants, both between- and within-subject effects were found for F_0 mean and CV (Table 4). Interviewers used lower and more variable F_0 when speaking with participants who were more depressed than they did when speaking with participants who were less depressed. Within-subject differences in interviewer F_0 mirrored these between-subject differences. Together, significant predictors relating to fundamental frequency accounted for 27.51% of the variation between subjects in depression score and 6.30% of the variation over time in a subject's depression score.

3.3.3 DETECTING DEPRESSION SEVERITY FROM PROSODIC FEATURES

To further evaluate the predictive value of prosodic features to detect severity, all of the significant parameters identified in the previous section were entered together into a single HLM. These consisted of five within-subject parameters, participant and interviewer switching pause CV, participant switching pause mean, and interviewer F_0 mean and CV, and two between-subjects parameters, interviewer F_0 mean and CV. The resulting model accounted for 64% of the variation in depression scores. Sixty-six percent of estimated depression scores were within 4.44 points (1 SD) of the actual score; 87% were within 6.66 points (1.5 SD) of the actual score.

For many purposes, only an estimate of severity range is desired. To evaluate such molar predictability, actual scores were divided into three ordinal ranges: low (0 to 7), mild (8 to 14), and moderate-to-severe (15 and above). Using linear discriminant analysis, estimated scores were used to detect these three levels of depressive symptoms. The resulting discriminant function was highly significant (Wilks' lambda = 0.476, $p = .001$). Sixty-nine percent of cases were correctly estimated. Kappa, a measure of agreement which adjusts for chance, was 0.526, which represents moderate agreement. (Table 5). Because the computational model was intended to detect severity in participants seen previously, leave-one-subject-out or k-fold cross-validation was not used. In a clinical context, a goal is to assess severity at each interview, and a participant's baseline is valuable input.

Two additional sets of analyses were pursued. First, we asked how switching pause and F_0 alone would compare to the joint model in detecting range of severity. To answer this question, significant switching pause and F_0 parameters were entered into separate HLMs. The estimated continuous depression scores from each HLM then were entered into separate linear discriminant classifiers to detect range of severity. Detecting range of depression in this way, switching pause parameters resulted in 69.5% accuracy ($\kappa = .554$), which was comparable to that for the joint model (i.e. switching pause plus F_0). When F_0 parameters were used alone, accuracy decreased to 57.8% ($\kappa = .373$). Thus, F_0 parameters were less effective detectors of severity and provided no incremental advantage relative to switching pause parameters alone.

Last, we asked whether accuracy was higher when both participant and interviewer parameters were used relative to when only participant parameters were used.

TABLE 5
Depression Range Predicted from Prosodic Features

Predicted	Actual		
	Low	Mild	Moderate to Severe
Low (HRSD i= 7)	30	7	3
Mild (HRSD 8 to 14)	6	19	15
Moderate to Severe (HRSD ≥ 15)	0	9	39

Kappa=0.526.

Most previous research has focused on single participants to the exclusion of conversational partners. We wanted to evaluate how much prediction power is lost when interpersonal effects are ignored. We found that when interviewer parameters were omitted, accuracy decreased from 69% to 63.3% (kappa decreased from .526 to .471).

4 DISCUSSION

Because most previous research has compared depressed and non-depressed participants, depression effects in previous research have often been confounded by myriad ways in which depressed and non-depressed comparison participants may differ. People who become depressed are far more likely to have high trait neuroticism and low trait extraversion, as but one example. Personality factors such as these have moderate heritability that is non-specific for depression [31]. By restricting our focus to a clinical sample that met criteria for Major Depressive Disorder and by sampling each participant over the course of their depression, we were able to rule out personality and other correlates of depression. The variation in prosody we identified was specific to variation in depression within a clinical sample.

We investigated intra- and interpersonal influence of depression severity on vocal prosody in depressed participants and their interviewers. We first consider the findings for switching pauses. As depression became less severe, participant switching pauses became shorter and less variable. Interviewer switching pauses became less variable in tandem with these changes. The dual effect for participants and interviewers is compelling when one considers that they were statistically independent. Each was highly related to depression severity. Together, they accounted for a third of the variation in depression severity over the course of time. To our knowledge, this is the first demonstration of mutual influence in vocal prosody of depression.

The findings for interpersonal timing (i.e. switching pause) extend previous findings that intra-personal timing (e.g., pauses between utterances) [11] is strongly related to depression severity. Considered together, timing appears to be a robust measure of change over the course of depression. Because timing can be readily measured with relatively low-cost instrumentation, routine measurement of intra- and interpersonal timing in clinical settings would appear feasible. Its adoption

could contribute to significant advances in understanding, monitoring, and treating depression.

Previous work by Mundt [34] found that F_0 became higher and more variable as patients recovered from depression. We found no evidence of this effect. Participant F_0 mean and variability failed to vary with severity. Several factors may account for this failure to replicate the previous findings. The participants studied by Mundt were inpatients, who may have been more severely affected than the outpatients we studied, and Mundt evaluated depression over a shorter time frame. Our findings suggest that F_0 may be a better marker of personality traits than of fluctuating changes in depression severity.

As noted, interviewer F_0 mean and variability showed a strong relationship with severity of depression. Interviewer F_0 and variability accounted for nearly 30% of the variation in depression severity between participants and about six percent of the variation in individual participants over time. As depression became less severe, interviewer F_0 became higher and less variable. Stated differently, interviewers became more expressive when participants were more depressed. This is similar to the findings of Boker and Cohn [6], [7] that participants increase their expressiveness when the expressiveness of their partners is attenuated.

The within-subject effects for both participants and interviewers are remarkable in that the pairing of interviewers and participants was not fixed across sessions. At any one session, they may have been meeting for the first time. The change in interviewer expressiveness, therefore, was most likely driven by something about the participant within interviews. While it is possible that participant nonverbal behavior other than vocal prosody or their answers to the interview may have influenced interviewer prosody, the strong variation we found in participant prosody likely played an important role. Further research will be needed to ferret out these possibilities.

The combination of participant and interviewer vocal timing and F_0 proved a powerful predictor of both numeric depression score and range of depression severity. Together, they accounted for over 60% of variation in depression scores. Sixty percent of estimated scores were within 4.44 points of the actual score; 87% were within 6.66 points. For a nonverbal measure, this is a striking degree of prediction of a language-based measure.

When range of depression severity was considered, the combination of participant and interviewer vocal prosody led to correct classification in 69% of cases. The observed kappa of 0.526 approached the level of agreement acceptable between observers when using the same measurement device. This suggests that moderate to high congruence between verbal and nonverbal behavior occurs over the course of depression. This effect was strongest when both participant and interviewer effects were included. Omitting interviewer effects reduced detection rates. The combination of participant

and interviewer vocal prosody was paramount. An unexpected finding was that F_0 contributed little to severity range prediction beyond the contribution of vocal timing. When interviewer parameters were omitted, classification accuracy was attenuated.

Several mechanisms might be considered with respect to interpersonal influence. While behavioral mimicry [24] or mirroring [35] might have played a role, neither appears to have been sufficient. First, these mechanisms would imply a high correlation between the switching pauses of each person. Yet, switching pauses of participants and interviewers were sufficiently uncorrelated that each independently contributed to variation in participant depression score and range of depression severity. Second, F_0 for participants and interviewers showed very different associations with depression severity. Participant F_0 mean and variability were unrelated to depression severity; while interviewer F_0 mean and variability were strongly related to depression severity. These findings appear more consistent with the hypothesis that very different intentions and goals underlie the vocal prosody of interviewers and participants. Confronted with a more depressed participant, interviewers decreased their F_0 and became more expressive, perhaps in an attempt to elicit more normal mood in the participant. The challenges of coordinating interpersonal timing with a depressed participant may have played a role as well. Time-series modeling and novel experimental paradigms [6], [7] will be needed to pursue these hypotheses.

Clinically, attention to vocal prosody and especially timing could provide a powerful, means of monitoring course of disorder and response to treatment at relatively low computational cost. Because vocal timing may be less susceptible than verbal report or even facial expression to efforts to misrepresent depression, its inclusion in assessment could improve reliability of measurement and enable more fine-tuned interventions. Interpersonal approaches to treatments that emphasize social stressors and skills could benefit from attention to interpersonal timing as well. Vocal timing could inform therapeutic decisions within diagnostic and treatment sessions and contribute to new forms of treatment that emphasize social communication in recovery from depression.

5 CONCLUSION

In summary, we found strong evidence that change in depression severity is revealed by vocal prosody. Listeners naive to depression scores differentiated symptom severity from the voices of participants. Specific prosodic features appeared to carry this information. Four were considered. They were switching pause mean and variability and F_0 mean and variability. Switching pause measures for both participants and interviewers were strongly related to severity. These findings suggest that vocal prosody is a powerful measure of change in severity over the course of depressive disorder. They

encourage use of these measures to screen populations at risk for depression and in considering novel approaches to traditional diagnostic measures in evaluating response to treatment. Further research is needed to investigate vocal features in addition to those we studied. The interpersonal effects of depression we found point to exciting directions for research in coordinated interpersonal timing and mental health.

6 ACKNOWLEDGEMENTS

The authors wish to thank Joan Buttenworth, Wen-Sheng Chu, Fernando De la Torre, Ellen Frank, Jeff Girard, Zakkia Hammal, Mohammad Mahoor, Long Qin, Alex Rudnicki, and Nicole Siverling for their generous assistance and the editors and anonymous reviewers for their constructive suggestions. The work was supported in part by US National Institutes of Health grants R01MH65376 to Ellen Frank and R01MH51435 and R01MH096951 to Jeffrey F. Cohn.

REFERENCES

- [1] M. Alpert, E. R. Pouget, and R. R. Silva. Reflections of depression in acoustic measures of the patient's speech. *Journal of Affective Disorders*, 66(1):59–69, 2001.
- [2] N. Ambady and R. Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, 111(2):256–274, 1992.
- [3] A. P. Association. *Diagnostic and statistical manual of mental disorders*. American Psychiatric Association, Washington, DC, 1994.
- [4] B. A. Bettes. Maternal depression and motherese: Temporal and intonational features. *Child Development*, 59:1089–1096, 1988.
- [5] P. Boersma and D. Weenink. Praat: Doing phonetics by computer, Undated.
- [6] S. M. Boker and J. F. Cohn. Real-time dissociation of facial appearance and dynamics during natural conversation. In C. Curio, H. H. B. Ithoff, and M. A. Giese, editors, *Dynamic faces: Insights from experiments and computation*, pages 239–254. MIT, Cambridge, MA, 2011.
- [7] S. M. Boker, J. F. Cohn, B. J. Theobald, I. Matthews, J. Spies, and T. Brick. Effects of damping head movement and facial expression in dyadic conversation using real-time facial expression tracking and synthesized avatars. *Philosophical Transactions B of the Royal Society*, 364:34853495, 2009.
- [8] K. Boudahmane, M. Manta, F. Antoine, S. Galliano, and C. Barras. Transcriberag, 2011.
- [9] A. S. Bryk and S. W. Raudenbush. Application of hierarchical linear models to assessing change. *Psychological Bulletin*, 101:147–158, 1987.
- [10] S. B. Campbell and J. F. Cohn. Prevalence and correlates of postpartum depression in first-time mothers. *Journal of Abnormal Psychology*, 100(4):594–9, 1991.
- [11] M. Cannizzaro, B. Harel, N. Reilly, P. Chappell, and P. J. Snyder. Voice acoustical measurement of the severity of major depression. *Brain and Cognition*, 56:30–35, 2004.
- [12] J. F. Cohn and S. B. Campbell. Influence of maternal depression on infant affect regulation. In D. Cicchetti and S. L. Toth, editors, *Developmental perspectives on depression*, pages 103–130. University of Rochester Press, Rochester, New York, 1992.
- [13] J. C. Coyne. Toward an interactional theory of depression. *Psychiatry*, 39:28–40, 1976.
- [14] R. Davidson, editor. *Anxiety, depression, and emotion*. Series in Affective Science. Oxford University, New York, 2000.
- [15] S. Duncan. Some signals and rules for taking speaking turns in conversations. *Journal of Personality & Social Psychology*, 23(2):283–292, 1972.
- [16] H. Ellgring and K. R. Scherer. Vocal indicators of mood change in depression. *Journal of Nonverbal Behavior*, 20(2):83–110, 1996.
- [17] R. Elliott, R. Zahn, J. F. W. Deakin, and I. M. Anderson. Affective cognition and its disruption in mood disorders. *Neuropsychopharmacology*, 36:153–182, 2011.

- [18] M. B. First, R. L. Spitzer, M. Gibbon, and J. B. W. Williams. *Structured clinical interview for DSM-IV axis I disorders*. Biometrics Research Department, New York State Psychiatric Institute-Patient Edition, New York, scid-i/p, version 2.0 edition, 1995.
- [19] J. L. Fleiss. *Statistical methods for rates and proportions*. Wiley, New York, 1981.
- [20] J. C. Fournier, R. J. DeRubeis, S. D. Hollon, S. Dimidjian, J. D. Amsterdam, R. C. Shelton, and J. Fawcett. Antidepressant drug effects and depression severity: A patient-level meta-analysis. *Journal of the American Medical Association*, 303(1):47–53, 2010.
- [21] D. J. France. Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Transactions on Biomedical Engineering*, 47(7):829–837, 2000.
- [22] R. W. Frick. Communicating emotion: The role of prosodic features. *Psychological Bulletin*, 97(3):412–429, 1985.
- [23] M. Hamilton. A rating scale for depression. *Journal of Neurology and Neurosurgery*, 23:56–61, 1960.
- [24] E. Hatfield, J. T. Cacioppo, and R. L. Rapson. Primitive emotional contagion. In M. S. Clark, editor, *Emotion and Social Behavior*, volume 14 of *Review of Personality and Social Psychology*, pages 151–177. Sage Publications, Newbury Park, CA, 1992.
- [25] S. D. Hollon, M. E. Thase, and J. C. Markowitz. Treatment and prevention of depression. *Psychological Science in the Public Interest*, 3(2):38–77, 2002.
- [26] J. Jaffe, B. Beebe, S. Feldstein, C. L. Crown, and M. Jasnow. Rhythms of dialogue in early infancy. *Monographs of the Society for Research in Child Development*, 66(2, Serial No. 264), 2001.
- [27] J. Jaffe and S. Feldstein. *Rhythms of dialogue*. Academic Press, New York, NY, 1970.
- [28] P. N. Juslin and P. Laukka. Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, 129:770–814, 2003.
- [29] F. N. Kerlinger. *Foundations of behavioral research: Educational, psychological and sociological inquiry*. Holt, Rinehart and Winston, NY, 1973.
- [30] R. Kessler, W. Chiu, O. Demler, and E. E. Walters. Prevalence, severity, and comorbidity of 12-month dsm-iv disorders in the national comorbidity survey replication. *Archives of General Psychiatry*, 62:617–627, 2005.
- [31] R. Kotov, W. Gamez, F. Schmidt, and D. Watson. Linking big personality traits to anxiety, depressive, and substance use disorders: A meta-analysis. *Psychological Bulletin*, 136(5):768–821, 2010.
- [32] S. Kuny and H. Stassen. Speaking behavior and voice sound characteristics in depressive patients during recovery. *Journal of Psychiatric Research*, 27(3):289–307, 1993.
- [33] E. Moore, M. Clements, J. Peifert, and L. Weisser. Analysis of prosodic variation in speech for clinical depression, 2003.
- [34] J. C. Mundt, P. J. Snyder, M. S. Cannizzaro, K. Chappie, and D. S. Geraltsa. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (ivr) technology. *Journal of Neurolinguistics*, 20:50–64, 2007.
- [35] P. M. Niedenthal. Embodying emotion. *Science*, 316:1002–1005, 2007.
- [36] W. H. Organization. *The global burden of disease: 2004 update*. World Health Organization, Geneva, Switzerland, 2008.
- [37] A. Ozdas, R. G. Shiavi, S. E. Silverman, M. K. Silverman, and D. M. Wilkes. Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk. *IEEE Transactions on Biomedical Engineering*, 51(9):1530–1540, 2004.
- [38] A. Pentland. *Honest signals: How they shape our world*. MIT, Cambridge, MA, 2008.
- [39] A. Pentland. *Kith and kin*, 2010.
- [40] S. W. Raudenbush and A. S. Bryk. *Hierarchical linear models: Applications and data analysis methods*. Sage, Newbury Park, CA, 2nd edition, 2002.
- [41] B. S. Reed. Speech rhythm across turn transitions in cross-cultural talk-in-interaction. *Journal of Pragmatics*, 42(4):1037–1059, 2010.
- [42] R. Rosenthal. Conducting judgment studies. In J. A. Harrigan, R. Rosenthal, and K. R. Scherer, editors, *Handbook of nonverbal behavior research methods in the affective sciences*, pages 199–236. Oxford, NY, 2005.
- [43] K. R. Scherer and G. Ceschi. Criteria for emotion recognition from verbal and nonverbal expression: studying baggage loss in the airport. *Personality and Social Psychology Bulletin*, 26(3):327–339, 2000.
- [44] B. Schuller, A. Batliner, S. Steidl, and D. Seppi. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech and Communication*, 53(9):10621087, 2010.
- [45] Sphinx. Cmu sphinx: Open source toolkit for speech recognition, Undated.
- [46] A. Trevino, T. Quatieri, and N. Malyska. Phonologically-based biomarkers for major depressive disorder. *EURASIP Journal on Advances in Signal Processing*, 42, 2011.
- [47] Z. Zeng, M. Pantic, G. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence*, 31(1):31–58, 2009.
- [48] A. J. Zlochower and J. F. Cohn. Vocal timing in face-to-face interaction of clinically depressed and nondepressed mothers and their 4-month-old infants. *Infant Behavior and Development*, 19:373–376, 1996.



Ying Yang is a Postdoctoral Fellow in the Human Rehabilitation and Neural Engineering Laboratory (hRNEL) at the University of Pittsburgh. Using electrocorticography (ECoG) to record high-resolution cortical activity from the surface of the brain, she studies language processing. She received her PhD from the department of Communication Science and Disorders at the University of Pittsburgh and her bachelors degree in Biological Sciences and Biotechnology from Tsinghua University, China. Her research interest is focused on prosodic parsing for linguistic and nonlinguistic purposes. She is Principal Investigator of two studies on prosody: one is a perceptual study on categorical perception of lexical tones; the other is on electrophysiological responses to English lexical stress.



Catharine Fairbairn is a PhD student in the Department of Psychology at the University of Pittsburgh. She has a Graduate Research Fellowship from the U.S. National Science Foundation (NSF). Her research addresses emotion, self-regulation, and longitudinal modeling of bidirectional influence in social interaction.



Jeffrey F. Cohn is Professor of Psychology at the University of Pittsburgh and Adjunct Professor at the Robotics Institute, Carnegie Mellon University. He received his PhD in psychology from the University of Massachusetts at Amherst. Dr. Cohn has led interdisciplinary and inter-institutional efforts to develop advanced methods of automatic analysis of facial expression and prosody and applied those tools to research in human emotion, interpersonal processes, social development, and psychopathology. He co-developed influential databases, Cohn-Kanade, MultiPIE, and Pain Archive, co-edited two recent special issues of Image and Vision Computing on facial expression analysis, and co-chaired the 8th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2008).