Automated Measurement of Head Movement Synchrony during Dyadic Depression Severity Interviews

Shalini Bhatia[1], Roland Goecke[1], Zakia Hammal[2] and Jeffrey F Cohn[3]
[1] Human-Centred Technology Research Centre, University of Canberra, Canberra, Australia
[2] Robotics Institute, Carnegie Mellon University, Pittsburgh, USA
[3] Department of Psychology, University of Pittsburgh, Pittsburgh, USA

***Abstract*— With few exceptions, most research in automated assessment of depression has considered only the patient's behavior to the exclusion of the therapist's behavior. We investigated the interpersonal coordination (synchrony) of head movement during patient-therapist clinical interviews. Our sample consisted of patients diagnosed with major depressive disorder. They were recorded in clinical interviews (Hamilton Rating Scale for Depression, HRSD) at 7-week intervals over a period of 21 weeks. For each session, patient and therapist 3D head movement was tracked from 2D videos. Head angles in the horizontal (pitch) and vertical (yaw) axes were used to measure head movement. Interpersonal coordination of head movement between patients and therapists was measured using windowed cross-correlation. Patterns of coordination in head movement were investigated using the peak picking algorithm. Changes in head movement coordination over the course of treatment were measured using a hierarchical linear model (HLM). The results indicated a strong effect for patient-therapist head movement synchrony. Within-dyad variability in head movement coordination was found to be higher than between-dyad variability, meaning that differences over time in a dyad were higher as compared to the differences between dyads. Head movement synchrony did not change over the course of treatment with change in depression severity. To the best of our knowledge, this study is the first attempt to analyze the mutual influence of patient-therapist head movement in relation to depression severity.**

I. INTRODUCTION

According to the Diagnostic and Statistical Manual of Mental Disorders (DSM-5), many symptoms of depression are observable [1]. Psychomotor symptoms such as gross motor activity, facial expressiveness, body movements, and speech timing differ between depressed and normal comparison groups [1], [2], [3]. Consequently, an automatic and objective assessment of depression from behavioral signals is of increasing interest to clinical and computer scientists [4], [5]. The last ten years have witnessed major strides in the automated assessment of depression from facial expression [5], [6], voice quality and timing [7], [8], [9], [10], and to a lesser extent body movement [11] and head pose [12], [13].

In nearly all previous efforts, assessment focused on the individual alone, rather than the context in which their behavior unfolds. In some cases, of course, the individual is alone. In AVEC (audio-visual emotion challenge) [4], for instance, patients of varying depression severity were assessed in a human-machine interaction task. Outside of such constrained challenges, however, depression is assessed in a social context by a clinician. This context almost certainly impacts both the intensity and the quality of interpersonal behavior of both the clinician and the patient. In social contexts, behavioral signals are influenced by the dynamics of interaction as well as social intentions (e.g. wanting to appear more or less depressed or suicidal than may be the case) [14]. In this paper, we investigate to what extent this interpersonal behavior, or coordination, varies with change in depression severity.

With a few exceptions, most research in automated detection of depression and depression severity has focused on the depressed patient rather than the interpersonal influence of the social context in which depression is assessed (e.g., [4], [5], [15]). One of these exceptions is the work of Scherer *et al.* [16]. The authors analyzed the interpersonal correlation between the acoustic characteristics of patients and therapists and between depression severity in clinical interviews [16]. They found that the acoustic characteristics of patients did not vary with depression severity, whereas those of therapists varied strongly with depression severity. Accommodation - the tendency of interactants to adapt their communicative behavior to each other - between patients and therapists was inversely related to depression severity. They found that accommodation of voice quality increased when depression remitted. Their findings suggest that therapists modify their acoustic features in response to depressed patients, and depression severity strongly impacts interpersonal accommodation [16].

Another exception to the singular focus on individual patients is the study by Yang *et al.* [17] who investigated the intra- and interpersonal influence of depression severity on vocal prosody in depressed patients and their therapists. They found that for the therapist, but not patient, f0 mean and variability showed a strong relationship with severity of depression. Therapists used lower and more variable f0 when speaking with the patients when the latter were more severely depressed. Intra-personal pause duration and speaking rate also changed with depression severity over time. It was found that switching pause latency for both therapists and patients became shorter and less variable when depressive symptoms decreased. These findings for vocal prosody [16], [17] support an interpersonal perspective that motivates our work.

As a contribution to the previous work on interpersonal effects of depression on vocal characteristics [16], [17], we explored whether similar patterns of interpersonal influence occur for other non-verbal communicative behavior, in partic-

978-1-7281-0089-0/19/$31.00 ©2019 IEEE

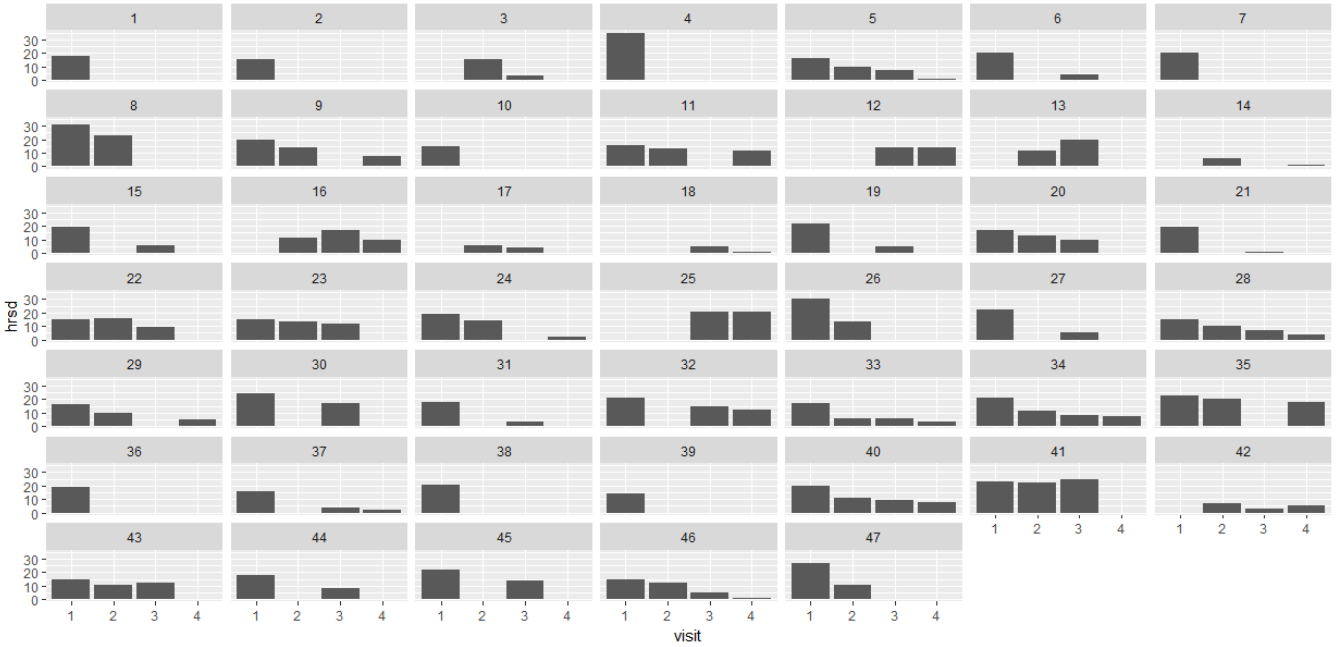

Fig. 1. Overall data - The x-axis represents the visits (1 to 4) and the y-axis represents the HRSD scores (0 to 35). The bar graphs represent the HRSD scores of the patients at each visit. The numbers on top of each group (visit) of bars are the patient IDs. Empty bars indicate missing data.

ular head movement. Head movements regulate turn-taking [18], serve back-channeling functions [19], communicate messages such as agreement or disagreement in interpersonal interaction [20], [21], and vary markedly with depression severity [13]. We investigated the interpersonal coordination between patients' and therapists' head movements over the course of treatment for depression.

Little is known about the coordination of head movement between patients and therapists in clinical assessments of depression severity. We hypothesized that the interpersonal coordination of head movement (measured as head movement synchrony) increases as depression severity decreases. Zface [22] – an automatic, person-independent, generic face tracking approach – was used to track the three degrees of out-of-plane rigid head movements (i.e. pitch and yaw) from synchronized 2D videos of patients and therapists. Windowed cross-correlation between head angles (pitch and yaw) of time series of patient and therapist was used to quantify interpersonal coordination [23]. The peak picking algorithm [23] was then used to analyze the variation in peak correlation at each time instant from the cross-correlation matrix and to measure synchrony. A hierarchical linear model [24], which accounts for both within-dyad and between-dyad variations, then was used to compare head movement synchrony across depression severity scores.

II. Observational Procedures

Fifty-seven depressed patients (34 women, 23 men) were recruited from a clinical trial for treatment of depression. They ranged in age from 19 to 65 years (mean = 39.65yr) and were Euro- or African-American (46 and 11, respectively). At the time of the study, all met DSM-IV criteria for Major Depressive Disorder (MDD) [25]. Data from 49 patients were available for analyses. Missing data occurred due to missed appointments or technical problems. The latter included failure to record audio or video, occurrence of audio or video artifacts and insufficient data. Patient loss was due to change in original diagnosis, severe suicidal ideation and methodological reasons (e.g. missing audio or video). Symptom severity was evaluated on up to four occasions at 1, 7, 13, and 21 weeks post diagnosis and intake by ten clinical therapists (all female). Therapists were not assigned to specific patients. Four therapists were responsible for the bulk of the interviews but the number of interviews per therapist varied. The median number of interviews per therapist was 14.5; four conducted six or fewer.

Structured interviews were conducted using the Hamilton Rating Scale for Depression (HRSD) [26], which is a clinician-rated multiple item questionnaire to rate depression severity and response to treatment. The HRSD rates the severity of depression by probing mood, feelings of guilt, suicidal ideation, insomnia, agitation or retardation, anxiety, weight loss, and somatic symptoms. Each item is scored on a 3- or 5-point Likert type scale, depending on the item, and the total score is compared to the corresponding descriptor, although only 17 items count towards the total score. Therapists were well trained in the HRSD and reliability was maintained above 0.9. Variation in HRSD scores is used as a guide to evaluate recovery by detecting ordinal ranges of depression severity. HRSD scores $\geq$ 15 are generally considered to indicate moderate to severe depression; scores between 8 and 14 indicate mild depression; and scores $\leq$ 7 indicate remission [27]. There are no healthy controls in this dataset.

Interviews were recorded using four hardware synchronized

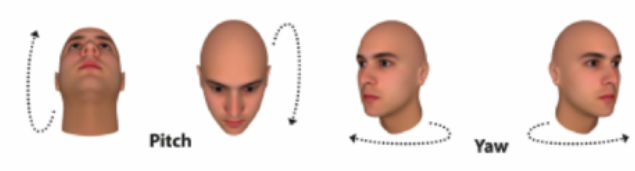

Fig. 2. Orientation of the head: pitch and yaw [28]

analogue cameras (video at 30fps) and two microphones (audio at 48kHz). Two cameras recorded the patient's face and shoulders; these cameras were positioned approximately 15 degrees to the patient's left and right. A third camera recorded a full-body view of the patient. A fourth camera recorded the therapist's shoulders and face from about 15 degrees to their right.

As the data is longitudinal in nature, each patient was seen on at least one or more occasions, represented in terms of $visit$ on the x-axis and evaluated on a continuous HRSD from 0 to 35, represented in terms of $hrsd$ on the y-axis (see Figure 1). The bar graphs represent the HRSD scores of the patients at each visit. The numbers on top of each group of bars are the patient IDs. Empty bars indicate missing data.

III. Methods

Interpersonal dyadic behavior is analyzed as follows: *(i)* automatic detection of head pose of patients and therapists, *(ii)* estimation of the time varying correlations between patients' and therapists' head movement using windowed cross-correlation, *(iii)* measurement of patterns of change in a lead-lag relationship between patients' and therapists' head movement using the peak picking algorithm, *(iv)* measurement of synchrony as the normalized mean of peaks of correlations per dyad and per session, and finally *(v)* analysis of changes in head movement synchrony across the course of treatment using hierarchical linear modeling.

A. Automatic head tracking

Zface [22] – an automatic, person-independent, generic face tracking system – was used to track the 3 degrees of out-of-plane rigid head movement (i.e. pitch (head nods), yaw (head turns), and roll (lateral head inclination)) from 2D videos of patients and therapists. The robustness of the tracker for head pose estimation has been validated in a series of experiments. In the Boston University dataset [29], which uses a magnetic flock-of-bird system to measure pose, mean absolute angular error was 2.66°, 3.93°, and 2.41° for pitch, yaw, and roll, respectively [22]. Angles of the head in the horizontal and vertical directions were selected in this paper to measure head movement coordination across depression severity scores (see Figure 2).

Using Zface, the head was successfully tracked in 98.3% of the patients' and 89.4% of the therapists' video frames. Therapists often looked down at their notes, which may have accounted for the small difference between patients and therapists in the number of tracked frames. Across therapist-patient dyads, the percentage of *simultaneously* tracked frames was 87.74%. For two sessions, the percentage of tracked frames was 0% and 0.02%. After excluding these two sessions, 125 sessions from 48 dyads were available for analysis.

B. Data Selection

We considered simultaneously tracked segments of at least 300 frames (10 seconds) or longer. This choice of minimum segment length is motivated by the observation of our data and the concept of "thin-slicing", which refers to observing a small selection of an interaction, usually less than 5 minutes, and still accurately drawing conclusions about the mutual influence of the interacting partners. It has been previously shown that a longer exposure time of a thin-slice (2-, 5-, and 10-second clips of non-verbal behavior) does not significantly improve the accuracy of judgment [30]. Motivated by Rosenthal *et al.* [30], Hammal *et al.* [20] for instance used a minimum duration of thin slices of 30 seconds and higher to analyze the interpersonal coordination of rigid head motion in intimate couples with a history of interpersonal violence. In another study by Rosenthal *et al.* [31], the tone of voice in which therapists spoke about their alcoholic and/ or drug abusing patients was used to predict the therapists' tone of voice when talking to the same patients. Thin slices of 10 seconds were found to generally capture the bulk of therapists' comments to patients.

We only included sessions for which at least 50% of the video frames were available after segments less than 300 frames long were discarded. Around 90% of the sessions had sufficient data for analysis. This additional constraint reduced the number of available sessions by 12 out of the 125 sessions (i.e. 113 remaining sessions). Out of the 12 sessions that were disregarded, 4 were from remission, 3 from mild and 5 from severe categories. Again, 3 sessions were from visit 4, 5 from visit 2 and 4 from visit 1. Therefore, missing data was unrelated to depression severity or visit.

The final sample consisted of 113 sessions from 47 patients (31 female and 16 male). Fifty were moderately to severely depressed, 32 were mildly depressed, and 31 were remitted. Figure 3 shows the final distribution of the data used for analysis. The mean duration of available data per session was 8.397min (std = 3.865min, median = 7.73min).

C. Windowed Cross Correlation

A common assumption in time series analysis is that signals are stationary. That is, the statistical properties (e.g. mean, standard deviation, auto-correlation and cross-correlation) remain stationary, or stable, over time. Social behavior, however, often is not stationary. Covariation between social partners can vary markedly over time [23]. Windowed cross-correlation (WCC) is a method of time-series analysis appropriate when the relationship between time series varies over time. WCC estimates time varying correlations between signals [32] and produces positive and negative correlation values for each (time, lag) pair of values.

In previous work, WCC has been used to investigate local correlation between infant and mother smiling over time [33] and to investigate dynamic changes in head movement

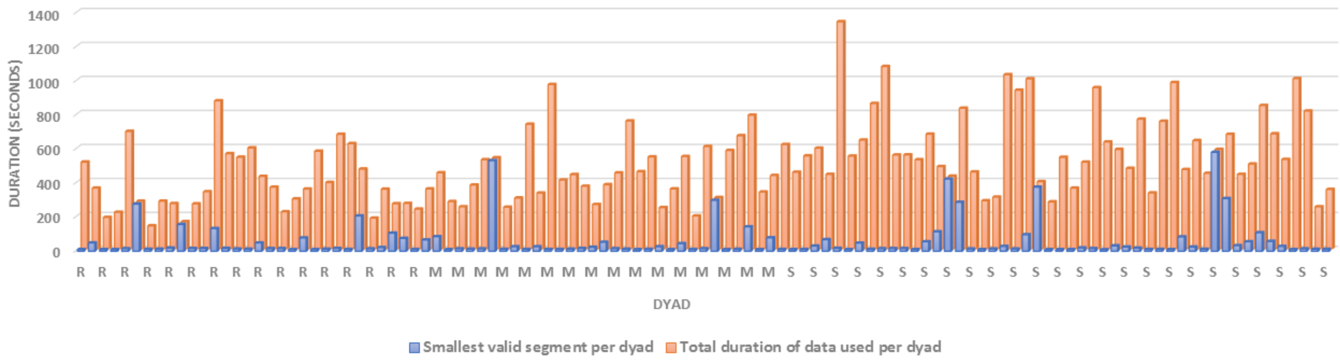

Fig. 3. The bar chart shows the data selection process. The orange bars represent the total duration (in seconds) of valid data used per dyad and the blue bars represent the smallest valid segment (of at least 10 seconds) per dyad. The first 31 sessions (labelled R) are from the category remission, the next 32 sessions (labelled M) are from category mild and the last 50 sessions (labelled S) are from the category severe.

between intimate partners during episodes intended to elicit conflict [20]. Guided by these previous studies, WCC was used to measure time series coordination of head movements between the patients and therapists over the course of treatment for depression. So that missing data would not bias measurements, the WCC for each session were computed for each consecutive valid segment separately and then combined.

In WCC, the signals are split into overlapping segments and a matrix of correlation values C is generated. Columns of C have cross-correlation values up to the maximum lag ($Maxlag$). Correlations are calculated within a window of length W_{max}. Time increases linearly across the columns of C, from left to right and lag increases linearly down the rows, from $-Maxlag$ to $+Maxlag$. The number of rows $N_{Row} = (2 \times Maxlag){+}1$ and the number of columns $N_{Col} = \frac{(N - N_{overlap})}{(W_{max} - N_{overlap})}$, where N is the signal length.

$Maxlag$ is the greatest interval of time separating a behavior from participant X and a behavior from participant Y. A $Maxlag$ of 45 frames (1.5 seconds) was used. A maximum window size (W_{max}) of 90 frames (3 seconds) was chosen in order to preserve the assumption of small change in a lead-lag relationship within the number of samples in the window [23]. Applying this window to a signal with 0% overlap would result in the analysis signal being almost exactly the same. Window overlap of 50% reduces the processing time and does not re-average the same data again. Therefore, window overlap ($Noverlap$) of 45 frames (50% overlap) was chosen for calculating the windowed cross-correlation [34].

A segment from therapist's pitch (resp., yaw) time series is represented as vector X and from patient's pitch (resp., yaw) time series as vector Y (see Section III-A). WCC was calculated on these segments as follows [23]:

$$C(W_x, W_y) = \frac{1}{W_{max}} \sum_{i=1}^{W_{max}} \frac{[W_{x,i} - \mu(W_x)]\,[W_{y,i} - \mu(W_y)]}{SD(W_x) SD(W_y)} \quad (1)$$

For each value of $Maxlag$ from $-Maxlag$ to $+Maxlag$, a pair of windows W_x and W_y were selected from the two data vectors X and Y respectively. $\mu(W_x)$ and $\mu(W_y)$ are the means and $SD(W_x)$ and $SD(W_y)$ are the standard deviations of the windows W_x and W_y, respectively. Figure 4 shows an example of a WCC correlogram for pitch. Yellow patches indicate high positive correlation, blue patches indicate high negative correlation. The area depicting positive lag (i.e. $Lag > 0$), represents the therapist leading the patient and the area depicting negative lag (i.e. $Lag < 0$) represents the patient leading the therapist. $Lag = 0$ indicates that both patient and therapist are moving their heads simultaneously.

D. Peak Picking

The peak picking algorithm [23] is used to analyze the patterns of change in the peak cross–correlation between patients and therapists. The peak picking algorithm obtains for each elapsed time, an estimate of the maximum association between two variables (in our case patients and therapists head movement) with the minimum time lag. The peak is defined as the maximum value of cross-correlation centred in a local region in which values are monotonically decreasing on each side of the peak. To do so, the resulting matrices from the windowed cross–correlation analysis were submitted to the peak picking algorithm to calculate peak correlations nearest a lag of zero and their respective time lags (See Figure 4). Input to the peak picking algorithm is a vector, V of cross-correlations (one column from the matrix C). The algorithm requires the definition of a set of parameters: *(i)* L_{size} (local search region) and *(ii)* P_{span} (degree of smoothing). L_{size} is the size of the local region that defines a peak. It decides how wide a window we want in order to consider the obtained peak as a maximum. L_{size} should be large enough so that spurious local noise is rejected but small enough such that meaningful peaks are not rejected. LOESS smoothing [35] is a non-parametric form of smoothing that uses a sliding-window average. Within each "window", a weighted average is calculated. The span (P_{span} in this case) determines the width of the moving window when smoothing the data. After careful analyses of the data [23], the parameter set for the peak picking algorithm were selected: *(i)* L_{size} (local search region) = 8 frames, *(ii)* P_{span} (degree of smoothing) = 0.1. The peak picking algorithm smooths and interpolates the WCC matrix, so from the output of peak picking, the lag

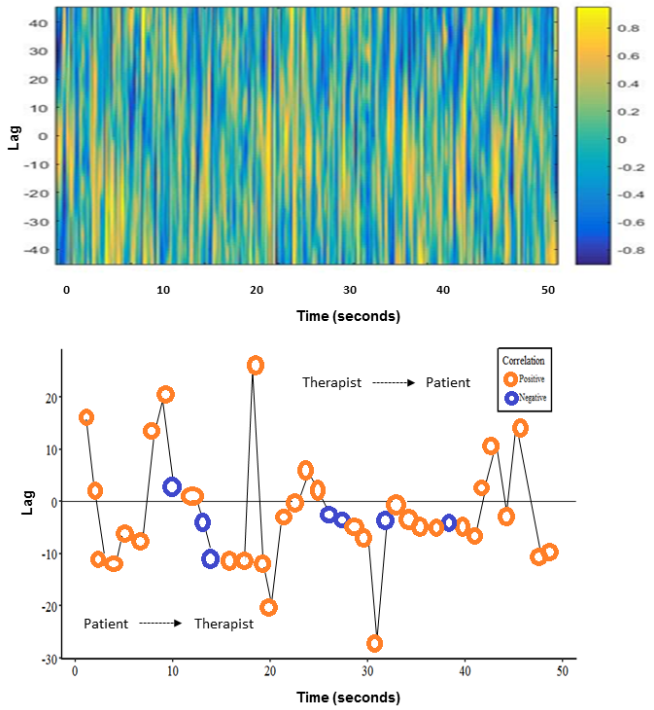

Fig. 4. *Top:* Output of windowed cross correlation shown as a correlogram ($Lag\ vs.\ Time$). *Bottom:* Output of peak picking ($Lag\ vs.\ Time$). Yellow patches in the correlogram indicate high positive correlation and blue patches indicate high negative correlation. Changes between subsequent vertical slices through the graph can be observed. Thus, the lags of the cross-correlational association between the two interactants' head movements change with elapsed time and the peak correlations change between positive and negative lags. This pattern of association is non-stationary [23] and can be observed in the peak picking output where orange dots represent positive correlation peaks and blue dots represent negative correlation peaks. When the lag is negative, the therapist follows the patient (Patient →Therapist); when the lag is positive, the patient follows the therapist (Therapist →Patient).

of the selected peak column was divided by 2 (such that $Maxlag$ ranged from -45 to +45). These are the offsets from zero lag.

The output of the peak picking algorithm is a list of local peaks of cross-correlation with the corresponding time lag. Figure 4 shows an example for pitch. For each graph, the area above the midline of the plot ($Lag > 0$) represents the relative magnitude of correlations for which head movement of therapist predicts head movement of the patient; the corresponding area below the midline ($Lag < 0$) represents the opposite. The midline ($Lag = 0$) indicates that both participants are changing their head movement at the same time. Positive correlation (orange dots) indicate that the direction of head movement of both patient and therapist is the same, whereas negative correlations (blue dots) indicate that the direction of head movement of both patient and therapist is changing out of phase. The visual inspection of the obtained peaks in Figure 4 shows dynamic changes in the direction of the peaks correlation with frequent changes in which partner is leading the other. Peak picking was performed separately on windowed cross correlation matrix segments of each valid segment separately and then combined for each session.

The source code for the windowed cross correlation and peak picking algorithms [23] is available and can be downloaded[1].

E. Synchrony

We used the time series peaks of correlation as defined above to measure head movement synchrony over time for each session. At each time point, synchrony between head movement of patient and therapist is defined as the peak of correlation greater than or equal to 0.5 (medium to large effect sizes [36])[2]. The overall synchrony ($sync$) during each session was measured as the sum of detected peak correlation values greater than or equal to 0.5 and normalized by the duration of the session (i.e. total number of frames). Thus comparisons could be made between sessions over the course of treatment for depression (across different depression severity scores). Synchrony was calculated for each dyad within each session separately and for pitch and yaw respectively.

F. Hierarchical Linear Modeling

Windowed cross correlation and peak picking [23] discussed in sections III-C and III-D were used to quantify interpersonal coordination between patients and therapists and investigate patterns of coordination in head movement (measured using synchrony). Thus, comparisons could be made between estimates of the overall synchrony between patients and therapists over the course of treatment. To do so, we used a variant of hierarchical linear modeling, a mixed effects model. A mixed effects model [37] is a statistical model that contains both fixed effects and random effects predictors [38] and is useful when repeated measurements are made in the context of a longitudinal study in which differences are assessed over time, such as in the dataset used in this study. Variables that vary by group are treated as random effects. A continuous variable is treated as a fixed effect as it would be incorrect to measure the variance across a continuous variable. Also, a categorical (binary) variable with only two values is a fixed effect as it would be incorrect to take two measures and then try to estimate variance [24]. If a variable f is treated as fixed then the model estimates and reports a value for $f1$, $f2$, etc. If a variable r is treated as random, the model estimates values for $r1$, $r2$, etc. to control for them, but only reports the variance of all the effects, rather than the value of each one. In order to investigate differences between patients and variability over time in patient-therapist interpersonal coordination, a mixed effects hierarchical linear model was built for the purpose of this study. Also, because of their advantage in dealing with missing values, hierarchical linear models are preferred over traditional methods such as repeated measures ANOVA.

HLM is a statistical tool for modeling data with a "nested" or interdependent structure [39]. In this study, repeated measurements (i.e. visits) were nested within participants

[1] http://people.virginia.edu/~smb3u/windcross2011

[2] Because the choice of peak threshold could influence the findings, we evaluated thresholds in the range 0.3 to 0.8. The findings were similar for peaks across this range.

(i.e. dyads). Two-level hierarchical linear model was built to represent the nested structure of the data (see Figure 5). Level 1 equation for the hierarchical linear model is as given below. A dyad i at visit j has synchrony defined as follows:

$$sync_{ij} = \beta_0 + \beta_1 * dyad_j + \beta_2 * hrsd_{ij} + \beta_3 * sex_j + e_{ij} \quad (2)$$

$sync$ is the normalized count of positive peak correlation values ≥ 0.5 (as defined in Section III-D). $sync_{ij}$ means the model has more than one variance component, where i is the subscript for level 1 unit, i.e. $dyad$, and j is the subscript for level 2 unit, i.e. $visit$. $hrsd$ is the depression severity of the patient on a continuous HRSD from 0 to 35, sex is the sex of the patient and $visit$ refers to the four occasions (week 1, 7, 13, and 21) up to which symptom severity was evaluated (see Section II). $sync$ is the response variable and predictors include $hrsd$, sex, $dyad$ and $visit$. Based on the description of fixed and random effects given above, $hrsd$ and sex were chosen as fixed effects predictors and $dyad$ and $visit$ as random effects predictors.

β_0 is the intercept which is the estimated $sync$ when $hrsd$ equals 0 and sex is female. The slope β_1 is the effect of $dyad$ on $sync$. The slope β_2 is the effect of $hrsd$ on $sync$. The slope β_3 is the effect of sex on $sync$. e_{ij} is the residual error term that encompasses variability that the predictors can not explain about the response variable.

For each level 1 regression parameter, there is one level 2 regression equation in a two-level hierarchical linear model. Level 2 regression equations for the hierarchical linear model are as given below. Level 1 regression parameters are modeled as response variables in level 2 regression equations:

$$\beta_0 = \gamma_{00} + \gamma_{01} * visit_j + u_{0j} \quad (3)$$

$$\beta_1 = \gamma_{10} + u_{1j} \quad (4)$$

$$\beta_2 = \gamma_{20} + u_{2j} \quad (5)$$

$$\beta_3 = \gamma_{30} + u_{3j} \quad (6)$$

Level 2 regression parameters γ_{00}, γ_{01}, γ_{10}, γ_{20} and γ_{30} are called hyperparameters of the model [24]. u_{0j}, u_{1j}, u_{2j} and u_{3j} are the error terms at level 2.

A dyad-wise analysis could be done (i.e. averaging over visits), which would be disregarding by-visit variation. Performing a visit-wise analysis would disregard by-dyad variation. A hierarchical linear model accounts for both sources of variation in a single model. Between-dyad variation assesses the differences in interpersonal synchrony when the severity scores are averaged across time. Within-dyad variation assesses the variability over time of each interpersonal synchrony score. Even if we measured all of these factors, there could still be other random factors influencing interpersonal synchrony that we can not control, such as personality, age, language, dialect, culture or ethnicity [40]. This model is able to capture the existence of these random factors in the form of residual variance.

IV. Results

Windowed cross correlation and peak picking algorithms [23] were used to quantify interpersonal synchrony between patients and therapists and to investigate patterns of coordination in head movements. Within-dyad and between-dyad analysis was done using a hierarchical linear model. The results of the hierarchical linear model for pitch and yaw synchrony estimates are explained in this section.

Fig. 5. Two level hierarchical linear model

A. Hierarchical Linear Modeling

In order to analyze changes in head movement synchrony between patients and therapists, a hierarchical linear model was built. As the response variable $sync$ was found to be gamma distributed, the function glmer ('family' = Gamma, link = 'inverse') of package lme4 [41] using R (R Core Team, 2017) [42] was used to fit the model. glmer() is used to fit a generalized linear mixed model, which incorporates both fixed effects and random effects parameters in a linear predictor. While lmer() assumes the probability distribution of data as normal, glmer() allows the choice of distribution of data through the argument 'family' [41]. In this model, $hrsd$ and sex were chosen as fixed effects predictors and $visit$ and $dyad$ as random effects predictors. The variable $dyad$ was entered at level 1. The variable $visit$ was entered at level 2. The results of the hierarchical linear model for pitch and yaw synchrony estimates from the fixed and random effects parameters are given in Table I.

1) Fixed and random effects: The estimates for intercept (i.e. $sync$ for pitch and yaw) were gamma transformed. The inverse function was used to transform the values back for ease of interpretation. For both pitch and yaw, the effect of $hrsd$ and sex was found to be statistically not significant (see Table I), which means that they did not contribute to explain the outcome variable $sync$. This finding was in contrast to our hypothesis that synchrony increases as depression severity decreases. The intraclass correlation coefficient (ICC) is a measure of how much the units in a group resemble one another. It can be computed as the ratio of group-level error variance over the total error variance [43]. In case of pitch, ICC(dyad) was 0.81 and ICC(visit) was 0.08, which means that 81% of the total variation in $sync$ is explained by $dyad$ and 8% of the total variation in $sync$ is explained by $visit$.

TABLE I

HIERARCHICAL LINEAR MODEL FOR PITCH AND YAW SYNCHRONY ESTIMATES

		Fixed Effects				Random Effects		
Model outcome	**Model parameter**	**Estimate**	**Standard error**	**Wald statistic**	**p value**	**Variance**	**Standard deviation**	**ICC**
sync (pitch)	Intercept	0.238	0.279	15.070	<2e-16 ***			
	hrsd	0.01	0.015	0.691	0.490			
	sex	0.105	0.259	0.407	0.684			
	dyad					0.213	0.461	0.81
	visit					0.021	0.145	0.08
	residual					0.029	0.170	
Observations = 113	Groups: dyad = 47, visit = 4	AIC = -386.9	BIC = -370.6	logLik = 199.5	Deviance = -398.9		df.resid = 107	
sync (yaw)	Intercept	0.205	0.273	17.910	<2e-16 ***			
	hrsd	0.003	0.015	0.178	0.859			
	sex	0.399	0.268	1.488	0.137			
	dyad					0.232	0.481	0.865
	visit					0.017	0.131	0.064
	residual					0.019	0.138	
Observations = 113	Groups: dyad = 47, visit = 4	AIC = -458.1	BIC = -441.7	logLik = 235.1	Deviance = -470.1		df.resid = 107	

In case of yaw, ICC(dyad) was 0.865 and ICC(visit) was 0.064, which means that 86.5% of the total variation in $sync$ is explained by $dyad$ and 6.4% of the total variation in $sync$ is explained by $visit$ (See Table I). For both pitch and yaw, ICC for $dyad$ was found to be very strong, but ICC for $visit$ was poor, which means that within-dyad variability in $sync$ was found to be much higher than between-dyad variability. This indicates that the differences over time in dyads were higher as compared to the differences between dyads. As $hrsd$ and $visit$ were highly correlated and the effect of $hrsd$ on $sync$ was not significant, it can be explained why the ICC for $visit$ was poor.

2) Goodness-of-fit Criteria, model diagnostics and inference [44]: According to the goodness of fit criteria, the yaw model fits the data better than the pitch model (see Table I). Deviance is defined as a measure of lack of fit between model and data. In general, the smaller the deviance, the better the fit to the data. Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are information-based criteria that assess model fit. Both are based on deviance. When comparing the AIC and BIC values of two models, the model with the smaller value is considered better [44]. The residual plots of both models did not indicate any deviations from a linear form. They showed relatively constant variance across the fitted range. There were no evident patterns/ clusters, hence the assumption of homoscedasticity of residuals was met. Also, the quantile-quantile plots did not raise any significant concerns with the distribution of the weighted residuals. Overall, both pitch and yaw models were good fits.

V. DISCUSSION

Motor mimicry and emotion contagion [45] would suggest that interpersonal effects for depression would mirror intra-individual effects of depression. Contrary to this hypothesis, synchrony failed to vary with depression severity. Several factors may have accounted for this null finding.

One, we measured synchrony in the time domain using cross-correlation and peak picking. Alternatively, one could measure coherence [46] in the frequency domain. Using coherence, we could have considered differences among high frequencies (rapid head movements) and low frequencies (slow head movements). For example, one would expect more coherence/similarity in slow head movements for severe depression as compared to remission. Two, therapists rotated among patients over the course of treatment. 25% of patients saw the same therapist in their sessions. Another 21% of patients saw the same therapist for at least two of their sessions but other therapists for the other sessions. The remainder saw different therapists for each session. The same patient could have seen one to four different therapists. This lack of consistency may have impacted on the display of synchrony over the course of visits. $Three$, when patients were severely depressed, the therapists may have worked harder to achieve or maintain synchrony. When patients became less depressed, they may have taken more of the responsibility for achieving synchrony. Synchrony, thus, became more co-constructed. To test this hypothesis, time series modeling would be needed. And $four$, we only examined head movement synchrony. Synchrony in other nonverbal modalities may be more affected by depression.

VI. CONCLUSIONS

We analyzed interpersonal coordination of patient-therapist head movement during clinical interviews for depression severity assessment. Windowed cross-correlation was used to quantify interpersonal coordination between patients' and therapists' head movement. Peak picking was used to analyze the variation in peak correlation and to measure synchrony. Both within-dyad and between-dyad variation were accounted for in the hierarchical linear model analysis. For both pitch and yaw, within-dyad variation was found to be higher than between-dyad variation. Head movement synchrony did not change over the course of treatment with change in depression severity. Future work will use time-frequency analysis to

further investigate possible changes in high frequencies and low frequencies (rapid vs. slow head movements).

VII. ACKNOWLEDGMENTS

This research was supported in part by the NINR of the NIH under Award Number R21NR016510, the United States NSF award IIS-1721667, and NIH award MH096951. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors would like to thank Anthony Davidson and Julio Romero at University of Canberra for statistical support.

REFERENCES

[1] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*. Washington, DC, USA: APA, 5th edition, 2013.

[2] C. Sobin and H.A. Sackeim. Psychomotor symptoms of depression. *American Journal of Psychiatry*, 154(1):4–17, 1997.

[3] M.P. Caligiuri and J. Ellwanger. Motor and cognitive aspects of motor retardation in depression. *J Affective Disorders*, 57(1):83–93, 2000.

[4] M. Valstar, J. Gratch, S. Schuller, F. Ringeval, L. Lalanne, M. Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic. Avec 2016 – workshop and challenge. In *International Workshop on AVEC*, 2016.

[5] H. Dibeklioglu, Z. Hammal, and J.F. Cohn. Dynamic multimodal measurement of depression severity using deep autoencoding. *IEEE BHI*, 22(2), 2018.

[6] J.M. Girard, J.F. Cohn, M.H. Mahoor, S. Mavadati, and D.P. Rosenwald. Social risk and depression: Evidence from manual and automatic facial expression analysis. In *IEEE FG*, 2013.

[7] S. Scherer, G. Stratou, G. Lucas, M. Mahmoud, J. Boberg, J. Gratch, A. Rizzo, and L-P. Morency. Automatic audiovisual behavior descriptors for psychological disorder analysis. *Image and Vision Computing*, 2014.

[8] N. Cummins, J. Epps, V. Sethu, M. Breakspear, and R. Goecke. Modeling spectral variability for the classification of depressed speech. In *Interspeech*, 2013.

[9] J.F. Cohn, T. Kruez, I. Matthews, Y. Yang, M. Nguyen, M. Padilla, F. Zhou, and F. De la Torre. Detecting depression from facial actions and vocal prosody. In *ACII*, 2009.

[10] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, G. Parker, and M. Breakspear. Characterising depressed speech for classification. In *Interspeech*, 2013.

[11] J. Joshi, A. Dhall, R. Goecke, and J.F. Cohn. Relative body parts movement for automatic depression analysis. In *ACII*, 2013.

[12] A. Kacem, Z. Hammal, M. Daoudi, and J.F. Cohn. Detecting depression severity by interpretable representations of motion dynamics. In *IEEE FGAHI*, 2018.

[13] J.M. Girard, J.F. Cohn, M.H. Mahoor, S. Mavadati, Z. Hammal, and D.P. Rosenwald. Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses. *Image and Vision Computing*, 32(10), 2014.

[14] A. J. Fridlund. Sociality of solitary smiling: Potentiation by an implicit audience. *Journal of Personality and Social Psychology*, 60(2), 1991.

[15] J.M. Girard and J.F. Cohn. Automated audiovisual depression analysis. *Current opinion in psychology*, 4, 2015.

[16] S. Scherer, Z. Hammal, Y. Yang, L-P. Morency, and J.F. Cohn. Dyadic behavior analysis in depression severity assessment interviews. In *ICMI*, November 2014.

[17] Y. Yang, C. Fairbarn, and J.F. Cohn. Detecting depression severity from vocal prosody. *IEEE TAC*, 4(2), 2013.

[18] S. Duncan. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23(2):283–292, 1972.

[19] M.L. Knapp and A. Judith. *Nonverbal communication in human interaction*. Boston, MA, USA: Cengage, 7th edition, 2010.

[20] Z. Hammal, J.F. Cohn, and D.T. George. Interpersonal coordination of head motion in distressed couples. *IEEE TAC*, 5(2), 2014.

[21] Z. Hammal and J.F. Cohn. Intra- and interpersonal functions of head motion in emotion communication. In *RFMIR*, pages 19–22, November 2014.

[22] L.A. Jeni, J.F. Cohn, and T. Kanade. Dense 3d face alignment from 2d videos in real-time. In *IEEE FG*, 2015.

[23] S.M. Boker, J.L. Rotondo, M. Xu, and K. King. Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. *Psychological methods*, 7(3), 2002.

[24] A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2006.

[25] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders: DSM-IV*. Washington, DC, USA: APA, 4th edition, 2000.

[26] M. Hamilton. Development of a rating scale for primary depressive illness. *British Journal of Social and Clinical Psychology*, 6(4), 1967.

[27] 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. *JAMA*, 303(1), 2010.

[28] Arcoverde E.N. Neto, R.M. Duarte, R.M. Barreto, J.P. Magalhaes, Carlos C.M. Bastos, T.I. Ren, and George D.C. Cavalcanti. Enhanced real-time head pose estimation system for mobile device. *Integrated Computer Aided Engineering*, 21(3), 2014.

[29] M.L. Cascia, S. Sclaroff, and V. Athitsos. Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3d models. *IEEE Trans. PAMI*, 22(4), 2000.

[30] N. Ambady and R. Rosenthal. Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology*, 64(3), 1993.

[31] R. Rosenthal, P.D. Blanck, and M. Vannicelli. Speaking to and about patients: Predicting therapists' tone of voice. *Journal of Consulting and Clinical Psychology*, 52(4), 1984.

[32] G. Laurent, M. Wehr, and H. Davidowitz. Temporal representations of odors in an olfactory network. *Journal of Neuroscience*, 16(12):3837–3847, June 1996.

[33] D.S. Messinger, M.H. Mahoor, S-M Chow, and J.F. Cohn. Automated measurement of facial expression in infant–mother interaction: A pilot study. *Infancy*, 14(3):285–305, May 2009.

[34] G. Heinzel, A. Rudiger, and R. Schilling. Spectrum and spectral density estimation by the discrete fourier transform (dft), including a comprehensive list of window functions and some new at-top windows. Technical report, Max-Planck-Institute for Gravitational Physics, 2002.

[35] W.S. Cleveland and S.J. Devlin. Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403), 1988.

[36] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 1988.

[37] M.J. Lindstrom and D.M. Bates. Newton-raphson and em algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83(404), 1988.

[38] N.M. Laird and J.H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4), 1982.

[39] J.M. Girard, J.F. Cohn, L.A. Jeni, M.A. Sayette, and F. De la Torre. Spontaneous facial expression in unscripted social interactions can be measured automatically. *Behavior research methods*, 47(4), 2015.

[40] J.F. Cohn. Beyond group differences: specificity of nonverbal behavior and interpersonal communication to depression severity. In *3rd International Workshop on Audio/Visual Emotion Challenge*, 2013.

[41] D.M. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 2015.

[42] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[43] The Department of Statistics and Data Sciences. Multilevel modeling tutorial using sas, stata, hlm, r, spss, and mplus. Technical report, The University of Texas at Austin, 2015.

[44] Social Science Computing Cooperative. Mixed models: Diagnostics and inference. Technical report, The University of Wisconsin, Madison, 2016.

[45] E. Prochazkovaa and M.E. Kret. Connecting minds and sharing emotions through mimicry: A neurocognitive model of emotional contagion. *Neuroscience and Biobehavioral Reviews*, 80, 2017.

[46] J.M. Gottman and J.T. Ringland. The analysis of dominance and bidirectionality in social development. *Child Development*, 52(2), 1981.