Beyond the Words: Analysis and Detection of Self-Disclosure Behavior during Robot Positive Psychology Interaction

Beyond the Words: Analysis and Detection of Self-Disclosure Behavior during Robot Positive Psychology Interaction

Sharifa Alghowinem[1,2§], Sooyeon Jeong[1§], Kika Arias[1], Rosalind Picard[1], Cynthia Breazeal[1], and Hae Won Park[1]

[1] MIT Media Lab, Cambridge, MA, USA

[2] Computer and Information Sciences College at Prince Sultan University, Riyadh, Saudi Arabia.

***Abstract*— Self-disclosure is an important part of mental health treatment process. As interactive technologies are becoming more widely available, many AI agents for mental health prompt their users to self-disclose as part of the intervention activities. However, most existing works focus on linguistic features to classify self-disclosure behavior, and do not utilize other multi-modal behavioral cues. We present analyses of people's non-verbal cues (vocal acoustic features, head orientation and body gestures/movements) exhibited during self-disclosure tasks based on the human-robot interaction data collected in our previous work. Results from the classification experiments suggest that prosody, head pose, and body postures can be independently used to detect self-disclosure behavior with high accuracy (up to 81%). Moreover, positive emotions, high engagement, self-soothing and positive attitudes behavioral cues were found to be positively correlated to self-disclosure. Insights from our work can help build a self-disclosure detection model that can be used in real time during multi-modal interactions between humans and AI agents.**

I. Introduction

Mental wellbeing is an important issue given its prevalence in all ages, where depression and anxiety are considered as the leading cause of disability worldwide [28]. Especially with the COVID-19 pandemic, the prevalence of mental health problems has tripled according to a late study, given the high exposure to stressors [12]. Digital mental health technologies attempt to reduce the effect of mental health and its prevalence by increasing accessibility to interventions and resources to those who need it. These technologies, such as social media [4], [25], [24], chatbots [23], [35], and voiced agents [31], have become mediums for self-disclosure, and can act as a tool for self-expression, while mitigating the potential risk of disclosing to another person (e.g. vulnerability, embarrassment, shame, lose of privacy [15]).

Social robots have also been studied as a tool to provide mental health interventions. In our recent study, we deployed a social robot that offers seven positive psychology exercises to improve people's psycholgical wellbeing in their homes [20]. Several of the intervention sessions offered opportunities for people to disclose about themselves, their thoughts, and feelings [20]. However, the robot had limited capabilities of encouraging, understanding, and properly addressing self-disclosure events. This limitation led some of the study participants to feel that the robot was not really "listening" to them during the intervention activities and frustrated. Based on these feedback, we believe that equipping the robot with automatic self-disclosure detection from verbal and non-verbal behaviors will improve engagement and rapport with the robot by enabling the robot to elicit and respond to users' self-disclosure with empathy and attentiveness.

Self-disclosure is defined as sharing of personal and intimate information about oneself. It has a great impact on improving one's mental health (e.g. release their stress, depression, and anxiety [24]). The act of self-disclosing has been shown to benefit people even when information was not disclosed to another person, e.g. expressive writing [29].

Most existing works on self-disclosure focus on analyzing linguistic features (e.g. interaction with text-based chatbots), even in the context of human-voiced agent interactions [31]. Unlike text-based chatbots, voiced agents and social robots rely on multi-modal interactions, often including user's voiced speech. This could be advantageous if utilized properly, such as augmenting self-disclosure analysis with prosody and nonverbal behaviors for a better interaction policy. Although emotional and self-disclosure speech often contain prolonged pauses [22], current voiced agents treat a long pause as an indicator for end of user utterance [31]. This is problematic because it could cause a voice-based agent to interrupt users while they are self-disclosing. One solution to avoid agent interruptions is to complement the linguistic self-disclosure analysis with available modalities (e.g. prosody and nonverbal behavior). However, research in this solution is scarce.

Therefore, in this work, we investigate the behavioral cues that indicate self-disclosure beyond the spoken language. Using our video recordings of the interaction sessions between the participants and a robotic positive psychology coach, we extract acoustic and visual behaviors during self-disclosure events. We model different modalities of behaviors (i.e. vocal, head, and body expressions) and provide analyses and interpretation of these behaviors through feature selection, and machine learning algorithms. The main contributions of this research are as follows:

- This paper explores the broadest array of high-level temporal behavioral features to date associated with self-disclosure from speech prosody, head pose, and body gestures.
- This paper is also the first to undertake a detailed and comprehensive approach to provide interpretation and insights into these behaviors to support modeling transparency.
- To the best of our knowledge, we are the first to evaluate this approach in a human-robot interaction context.

§ Equal Contribution

978-1-6654-3176-7/21/$31.00 ©2021 IEEE

- We are also the first to analyze body gestures in correlation to self-disclosure events.

Given that not all modalities can be available during the robot interaction (e.g. missing audio or video segments, the user is not speaking or is out of the camera(s) view), we investigate individual modalities to detect self-disclosure behavior. Moreover, our focus here is to analyze and extract behaviors to create a light-weight real-time model to be implemented on the robot with minimum computational resources.

II. Background and Related Work

Social agents have been widely utilized to support the user to improve their wellbeing and mental health outcomes. With the increase of such utilization, the actual impact and efficacy on user's wellbeing are evaluated based on their adequacy, as reviewed in [35]. Willingness to self-disclose is considered one of the measures to assess the rapport between users and the agent, as well as the social support provided by the agent [35]. For example, the study in [8] showed that participants felt more rapport and willingness for self-disclosure with a virtual agent than face-to-face interaction. Contrary to human clinicians, social agents can provide perceived anonymity [25] and non-judgment [33], which reduce the stress and lower emotional barriers to self-disclosing [17]. These characteristics make them a suitable medium to share personal and sensitive information compare to sharing with other people. These unique characteristics create an opportunity for the users to retain the benefits of expressing themselves (e.g. reduce psychological distress [24]), while avoiding its risks (e.g. vulnerability and shame [15]).

Analyzing self-disclosure from linguistic attributions attracted a wide range of research fields (e.g. psychology, linguistics). Social media provided a platform for users to disclose information about one-self as studied in [25], where they categorized self-disclosure in such platforms based on topics (e.g. Tastes and Interests, Interpersonal Relations and Self-Concept). The results of their study showed the factors affecting willingness to self-disclose in social media platforms, which included the level of topic intimacy, user anonymity, and social ties (i.e. personally knowing the people in the platform). [31] defined self-disclosure as "sharing extraneous information voluntarily as opposed to a direct response to a question" and annotated people's dialogues with voiced conversational agents. Their automatic classification of self-disclosure extracted conversational and linguistic features, and resulted up to 91.7% accuracy in detecting self-disclosure events. They also found that people were more like to reciprocate if the agent exhibited a self-disclosure behavior, and the dialogues that contain self-disclosure are longer in duration. However, their conversational agents rely on Automatic Speech Recognition (ASR), and inaccurate ASR results led to failures in the agent's responses.

A text-based chatbot developed by [23] explored strategies to encourage and elicit users' self-disclosure. The agent's self-disclosure statements were categorized as informational, thoughts, and feeling, and were further rated as one of the three levels (no/low/high self-disclosure). They found that people who interacted with the chatbot that shared low or high self-disclosure were more liked to reciprocate the behavior as previously suggested by [31]. Moreover, virtual conversational agents that express intimacy behaviors (i.e. honesty, positivity, and mutual comprehension) during tourist counseling elicited positive emotions in users and enhanced the user experience [30].

We believe that for stronger human-agent rapport and relationship would enhance the effectiveness in the mental health interventions the agent provides. Previous findings suggest that an artificial agent could achieve better rapport with users by engaging in intimate and engaging self-disclosure interactions. However, this is a difficult task due to the complexity of the linguistic understanding, the errors in ASR, and the effect of mistakenly recognizing end-of-speech signal (i.e. variations in pauses in emotional or self-disclosing speech). Instead of solely depending on single modality, we argue that complementing self-disclosure detection with other available modalities (e.g. vocal and visual behaviors) could enhance the self-disclosure detection.

To the best of our knowledge, only a few studies investigated behaviors beyond the language attributes in recognizing self-disclosure (i.e. analyzing prosody and nonverbal behaviors). One of the first studies manually annotated the nonverbal behavior of participants during virtual consular interaction, including eye gaze, head movement, smiling, and pauses [22]. They analyzed the association of these behaviors with the level of intimate self-disclosure (low, medium, and high), where the results showed that head tilts and nodes behaviors are the strongest in distinguishing self-disclosure. Another study analyzed verbal and nonverbal behaviors that are correlated to conversational strategies including self-disclosure, shared experiences, etc. [36]. The nonverbal behaviors, namely eye gaze, smiles, and head nodes, were manually annotated, while linguistic and prosody features were extracted using available tools. The finding showed that nods and mutual gaze are expressed from the self-disclosing person, while the listener behaviors showed nod and avert their gaze during self-disclosing events.

Automatic behavior extraction from linguistic, acoustic, and nonverbal modalities for the level of self-disclosing estimation was investigated recently in [34], which is, to the best of our knowledge, the only one that fully automate the behavior extraction and estimation of self-disclosure. The dialog interactions with a tele-operated virtual agent were rated for self-disclosure in 7 levels ranging from -3 to +3 (low-high disclosure), where the self-disclosing estimation was treated as a regression problem. They compared the performance of modeling self-disclosure using handcrafted features (e.g. head nods, shake, and orientation, smiles, pauses) with deep representations (extracted from pre-trained models) from the three channels (i.e. text, audio and video) individually and when fused. On average, the model performance was not significantly different between handcrafted features and deep representations in each modality, while handcrafted features were slightly better with small datasets. Moreover, the performance of the multimodal fusion was comparable to signal modalities.

Extending on prior work, we extract an extensive array of acoustic and nonverbal behaviors from head and body gestures. We focused our investigation to handcrafted features modeled with traditional discriminative models (multi-layer perceptron and SVMs) for the following reasons: (1) our dataset is relatively small for deep modeling, (2) according to [34], deep features and deep models were not significantly higher in accuracy than traditional ones, (3) we aim to provide a transparent and objective

Fig. 1: A portable station that integrates a social robot, a tablet, and a Raspberry Pi with a wide-angle USB camera.

interpretation of self-disclosure behavior, and (4) we wish to create a lightweight model suitable for robot implementation with minimum computational resources to work on real-time.

III. Methodology

A. Robotic Positive Psychology Coach

In our previous study [20], we deployed a social robot in participants' homes intending to improve their mental health through positive psychology sessions.

A portable station was used for this purpose that integrated Jibo robot[1] with a tablet, which provides an interface for extra resources, as well as recording the sessions through its camera, and also integrated with a Raspberry Pi for recording the sessions through a wide-angle USB web camera (see Fig. 1).

The commercial robot is equipped with useful skills (e.g. weather report) that participants interact with. To extend the skills in the robot, we implemented a Wellness skill to coach the participants for the positive psychology sessions. To activate the Wellness skill, the robot asks the participant if they would like to start, or by participant's request. Once the Wellness skill is activated, audio and video recording of the session starts from both the tablet camera and the USB camera.

B. Human-Robot Interaction Study

We recruited thirty five undergraduate students living in on-campus dormitories at *jeong2020robotic* to participate in our study (age M=18.94, SD=1.43; 27 female, 7 male, and 1 other). Once the student signed a consent form, a robot station was delivered to the him/her dormitory room and a set of questionnaires were administered to measure their personality traits (Mini-IPIP) [9], psychological wellbeing (Ryff's Psychological Wellbeing Scale [21]), mood (Brief Mood Introspection Scale [26]) and readiness to change for better wellbeing (Readiness Ruler [18]).

Participants were instructed on the overall study procedure, as well as how to use the robot's wake word ("Hey, Jibo"), how to start/stop the robot's positive psychology skill, and how to make the robot go to sleep or turn around for their privacy. We asked our participants to use the robot's *wellness* skill daily at any time during the day they found suitable.

Seven *wellness* sessions were 5-10 minute each and were designed based on evidence-based positive psychology interventions [32]. They included contents about (1) the introduction to positive psychology, (2) character strengths (CS), (3) using your signature strength in a new way (SS), (4) practicing the three good things exercise (TGT), (5) writing a gratitude letter (GL), (6) savoring (S), and (7) overall reviews. Each session was video/audio recorded through the Android tablet camera and the external USB camera on the station.

During each session, the robot first greeted the participants and asked them about how their day was going – small talk task (ST). Then, the robot introduced the intervention content for the day. All of the intervention sessions (except for the introduction session on the first day) included at least one questions/prompts for participants to share their reflections and feelings based on the intervention procedures. In some sessions, they were asked to verbally express their reflections on the previous session. In total, the robot asked fourteen open-ended questions. Seven of them were small talk (ST) tasks, and the other seven were intervention specific questions, e.g. "can you tell me three things you are grateful for today and why they went well?" for the three good things (TGT) exercise, "how did [gratitude letter recipient] react and how did it make you feel?" for the gratitude letter (GL) exercise, etc. The robot used rule-based parsers to classify the participants' intent.

At the end of the study, participants filled out surveys on their psychological wellbeing, mood, readiness to change behavior and working alliance with the robot, and participated in a semi-structured interview for more open-ended feedback on their experience living with the robot. For further details about the protocol, please refer to our paper [20].

C. Self-Disclosure Annotation

The audio recordings were manually transcribed, where we manually annotate for self-disclosure statements. However, due to some technical issues (e.g. lost communication between devices), out of the 245 sessions, the audio recordings of 13 sessions were not available. Therefore, the total number of audio-recorded sessions included in this study is 232 sessions. Given the unique interaction and context in our robot coach study, we used a customized self-disclosure definition adopted from [2] rather than utilizing the self-disclosure definitions used in previous studies (described in Section II). We define self-disclosure and non-self-disclosure in this context as:

Self-disclosure: The participant reveals information about themselves to the agent. The information can be descriptive or evaluative and can include thoughts, feelings, aspirations, goals, failures, successes, fears, and dreams, as well as one's likes, dislikes, and favorites. Typical conversations you might share with a friend. An example statement: "Well, I value close relationships with people so...for example, yesterday I went to Waltham with my friend to spend a few hours with him.".

Non-Self-disclosure : The participant may still reveal some amount of information about themselves. This information is typically inconsequential and something that any person would be comfortable sharing with a complete stranger. An example statement: "Um, I'm on a basketball team so I regularly work with a group of people and contribute to the team environment and team success.".

Even though this self-disclosure definition might not bring the full psychological benefit covered in Section II, it allowed us to categorize utterances that reveal personal and intimate information

[1]https://www.jibo.com/

TABLE I: Number of Samples in each Modality Used for Analysis and Classification

Modality	**Acoustic** (mic)	**Head Pose** (Tablet's cam)	**Body Gestures** (Webcam)
Self-disclosure	56	52	53
Non Self-disclosure	196	179	189
Total	252	231	242

* Sample sizes differ between modalities due to missing recording from the recording device caused by communication failure.

that can potentially deepen the rapport and relationship [3] from utterances that only share superficial and factual information. We annotated the open-ended responses, described above, since these responses have high chances of participants sharing their reflections and feelings. Responses to yes/no or confirmation questions were excluded from the annotation task. This process results in a total of 420 annotated statements, where only 95 statements are identified as self-disclosure. From these annotated statements, we further excluded very short utterances (e.g. "I will read a book") to ensure a fair comparison between self-disclosure and non-self-disclosure, since most self-disclosure statements tend to be longer. The final count of statements used in this analysis is 252, of which 56 are self-disclosure utterances (see Table I).

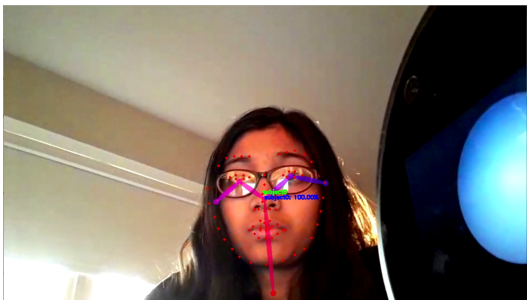
D. Behavioral Feature Extraction

We manually synchronized the audio and video signals from the Tablet and Raspberry Pi recordings as a pre-processing step. The aligned signals (audio, video, and text transcripts) were then segmented for the annotated self-disclosure to extract the vocal and nonverbal behaviors.

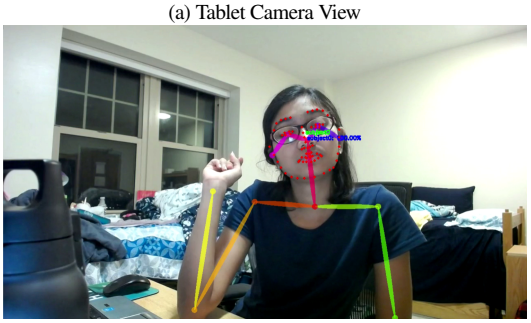
For the **acoustic signal**, we segmented the session audio signal using the transcription timestamps to extract the participant's clean speech (i.e. not contaminated with the robot's speech or other speakers). To account for the different duration between audio segments, the speech utterances are summarized through extracting high-level functional (statistical) acoustic features using OpenSmile [14]. Research in detection emotions from speech showed a high performance using the minimalistic feature set carefully selected features [13]. We extract 88 parameters, that includes Frequency parameters (Pitch, Jitter, and Formants), Energy parameters (Shimmer, loudness, and Harmonic to Noise Ratio (HNR)), and Spectral parameters (including Mel-Frequency Cepstral Coefficients(MFCC).

As mentioned earlier, we video-record the sessions using two cameras; the tablet camera and the wide-angle USB camera. We used these videos to extract nonverbal behaviors from head pose and body gestures. Given the different views of the cameras (Fig. 2 illustrate an example of the cameras' views), we used the wide-angle camera for analyzing body gestures, and the tablet camera for analyzing head pose.

To extract the **head orientation**, we first extracted facial landmarks in an estimated 3D space [6], as shown in Fig. 2a. At every frame, head orientation (yaw, pitch, and roll) is calculated by solving the Direct Linear Transform (DLT) followed by Levenberg-Marquardt optimization from the 3D facial landmarks. Similar to the speech signal, we summarize the head orientation through statistical functions to account for different duration of segments,



(a) Tablet Camera View



(b) Wide Angle USB Webcam View

Fig. 2: An example of detected upper body joints and facial landmarks from both Tablet and USB cameras' Views.

which is performed at the utterance level. For this summary, we calculate 10 functional features from each of the 3 low-level head orientation features and their derivatives (i.e. speed and acceleration), which are: minimum (min), maximum (max), range, average, standard deviations (std), variance (var), skewness, kurtosis, peaks, and valleys. This process produced a total of 90 functional features (10 functional x 3 signal and derivatives x 3 low-level head pose). The number of samples of head pose was less than those of audio samples due to missing or corrupted video recording from the tablet, and in a few instances out-of-view participant. The final number of samples with head features used in this work is 231, where 52 are classified as self-disclosure (see Table I).

Furthermore, we analyze **body movement and gestures** through extracting body joints using the wide-angle view since it is suitable for capturing the upper body gestures (see Fig. 2b). To locate the body joints, OpenPose [7] was utilized, which locates 25 body joints with an estimated 3D space using bottom-up joint detection through multi-stage CNN-based network architecture. From these joints, body orientation (yaw, pitch, and roll) as well as touching behaviors, such as touching the face, the other hand, or the upper body are extracted from both hands. Following the same process as with head pose features, body gestures are summarized by calculating the 10 functional features from each body behavior and its first and second derivatives. The length of the body gesture feature vector is 240 for each sample (10 functional x 3 signal and derivatives x 8 low-level body gestures). Worth noting that the videos recorded from the USB camera had some missing sessions, and therefore, the number of samples is not always aligned to the recordings from the tablet and the audio recordings. The sample size used for the body gestures is 242, of which 53 are identified as self-disclosure (see Table I).

E. Feature Selection for Interpretation

The feature spaces for the extracted features from acoustic (88), head pose (90), and body gestures (240) modalities are relatively large (in comparisons to the sample size), which could hinder the explainability of specific behaviors associated with self-disclosure. A recent study proposed a **feature selection framework (FSF)**, where it aggregates several feature selection methods to systematically identify the most representative features of an independent variable [1]. We utilized the FSF in this work to facilitate the interpretation of behaviors associated to identify self-disclosure and to support modeling transparency. For a binary classification problem, the FSF employs statistical-based methods (e.g. t-test for single-variate analysis, Correlation-based Feature Selection (CFS) for multivariate analysis), machine-learning-based methods (e.g. genetic algorithms), and data structure-based methods (e.g. random forest), and then aggregate their results to narrow the feature space to the most meaningful set. Through this process, the framework not only analyze the features independently but also analyzing the relationship between the them (e.g. removing redundant features, finding a combination of features that correlate together, etc.).

Using the FSF, we selected the top 10% of the behavioral features for each modality that are highly correlated to self-disclosure, through a 10-fold cross-validation process and with two runs to measure the stability of the selected features from the framework. We run the FSF on each channel's feature space to interpret each modality behaviors separately. The selected features are then analyzed for interpretation, as well as used for classification to be compared when using the full feature space. For interpretation purposes, we conduct a Point-Biserial Correlation analysis to get the direction of the correlation of the selected feature with the self-disclosure classes. Moreover, identifying a minimal feature space without compromising the accuracy of detecting self-disclosure behavior helps in creating a lightweight real-time model with minimum computational resources.

F. Classification

To equip the robot with the ability to identify behaviors associated with self-disclosure (or the lack of), we run several experiments for modeling the self-disclosure. Such capability will allow the robot to adjust its interaction to encourage self-disclosure and respond properly to self-disclosure events. To this end, we implemented a modeling pipeline to select the best model and parameters for this task. Given the nature of the extracted behavioral features, we believe that discriminative models such as multilayer perceptron (MLP) and Support Vector Machines (SVMs) are suitable for the task, especially given our small dataset and our goal of creating a lightweight model. Therefore, we model the feature space for each modality using MLP with one hidden layer and SVMs with linear and radial kernels for comparisons.

Since the self-disclosure sample size is small compared to the non-self-disclosure samples, we compare different methods of data augmentation and sampling. This is performed to mitigate the model bias toward the overrepresented class. We compare oversampling (randomly duplicated samples from self-disclosure class) and undersampling (randomly removed samples from non-self-disclosure class), where both classes have the same number of samples for training. Moreover, to mitigate the small sample size, we used 10-fold cross-validation, where each fold the samples are randomly split into a stratified 75%/25% training set and testing set respectively. The testing set is then normalized (using min-max) based on the training set values to avoid contamination.

For the models' hyperparameter tuning, a grid search on the training set is utilized using further 10-fold cross-validation, which split the training set into sub-training and validation sets. For MLP Model parameters, we used 150 epochs or until convergence (when loss improves by less than 1e-4 per iteration), and 144 for batch size, which was optimized using stochastic gradient descent with a learning rate of 0.1, and log-loss for loss function. Worth noting, that none of the MLP experiments reached the 150 training epoch. For SVM parameters, we used linear and radial basis function (RBF) kernels, where their parameters (cost and cost& gamma, respectively) were optimized by the grid search described above.

The model performance is measured by the average of the balanced accuracy (BAcc.) across the 10-folds on the testing sets. We use the balanced accuracy rather than accuracy to account for the imbalanced samples from the self-disclosure classes. Furthermore, we also measure the model performance using Matthews Correlation Coefficient (MCC) score, which is considered a reliable result for classification performance. MCC produces high scores only if all elements in the confusion matrix had good results, accounting for imbalanced samples between classes [37]. With MCC, the closer the score is to 0, the closer the model to random-level, while the closer it is to 1, the more perfect the classification results are. MCC can also be negative, which indicates a high confusion in model performance.

Finally, to compare the model results with that of a random chance, we randomly shuffled the original labels while keeping the feature vectors as is. The results reported on all models (including the random shuffle models) are the best-performing combinations of models and sampling methods. Furthermore, we performed a two-tailed T-test for two samples to analyze the significant differences between models' performances, assuming equal variances and $p = 0.05$.

IV. Results and Discussion

A. Self-disclosure Behavior Classification Results

Table II shows the classification results of the best-performing classifiers and its sampling method of each modality using full features, selected features and randomly shuffled labels. To ensure a fair reporting of the model performance given the imbalanced sample sizes between the self-disclosure classes, we measure and compare model performances using balanced accuracy (BAcc.) and MCC. Comparing modeling from full feature space and the features selected based on the FSF, showed performances significantly higher than that acquired by chance (modeling with randomly shuffled labels) for each of the behavioral signals. This indicates that the extracted behavioral features are representative to the task of recognizing self-disclosure behavior. Interestingly, the performance when using the top 10% features from the FSF in all modalities outperforms that when using full feature space. Even though the performance differences between the full features and the selected features are not significant, it confirms that selecting meaningful features for modeling can reduce the computational resources without compromising the model accuracy [1]. The

TABLE II: Results of Best-performing Classifier in Terms of Balanced Accuracy and MCC of Detecting Self-disclosure Behavior from Different Modalities Comparing Different Feature Spaces with Randomly Shuffled Labels

Modality	**Prosody**			**Head Pose**			**Body Gestures**		
Features (#)	All (88)	FS (9)	Random (88)	All (90)	FS (8)	Random (90)	All (240)	FS (16)	Random (240)
Sampling	oversample	oversample	undersample	undersample	undersample	none	undersample	none	undersample
Best model	SVM (linear)	SVM (linear)	MLP	SVM (radial)	SVM (radial)	SVM (radial)	MLP	SVM (radial)	MLP
avg. BAcc.	0.706	**0.716**	0.551	0.770	**0.805**	0.504	0.776	**0.777**	0.527
std. BAcc.	0.055	0.046	0.056	0.053	0.055	0.012	0.047	0.090	0.069
avg. MCC	**0.376**	0.371	0.086	0.495	**0.600**	0.025	0.483	**0.597**	0.045

best performing models using the FSF features for each of the modalities were linear kernel SVM for prosody features and radial kernel SVM for head and body behaviors. This implies that the selected features are easily separable between classes and that they hold good discriminative power in identifying self-disclosure.

Comparing modalities using the FSF features, head pose followed by the body gestures significantly outperformed the prosody features ($p = 0.0007$ and $p = 0.04$ respectively). However, there are no statistically significant differences between the performance of head pose and body gestures modalities. Interestingly, body behaviors showed a high performance in recognizing self-disclosure, which could not be possible without the wide-angle camera (see Fig. 2b). Worth noting that in some instances when the participant was out of the tablet camera view (e.g. due to station location), the wide-angle camera was able to capture the participant. This could justify adding such cameras in designing the setting for human-robot interaction, where analyzing body gestures could enhance the interaction. Nonetheless, since the signals might not be available during the entire session (e.g. user is not speaking or out of camera view, or missing/corrupted signal), it is valuable to be able to utilize the available signals for detecting self-disclosure behavior from different modalities.

In summary, the results showed that the selected behavioral features are suitable to the task, therefore, they accurately identified self-disclosure behavior even with simple classification algorithms. Moreover, the similarity in the accuracy results from the multimodal approach used in this work emphasizes the need for redundant/complementary information from different channels. That is, given the complexity of recording environments in real-world interactions, and the possibility of missing data from different sensory devices, designing recording setting and analysis protocols should take into account the inclusion of different modalities for reliable modeling and enhanced user interaction experience.

B. Self-disclosure Behavior Interpretation

As mentioned in Section III-E, a feature selection framework was utilized in this work to serve as (1) reducing the feature space by selecting the most meaningful features, (2) reduce the computational resources to create a lightweight model for self-disclosure, and (3) provide interpretation of the prosody and nonverbal behaviors associated with self-disclosure. Using the framework, we reduce the full feature space to the top 10% features that are systematically aggregated using a variety of feature selection methods (e.g. statistical and AI methods).

For **speech prosody** modality, 9 features out of the extracted 88 features were selected. To analyze the direction of the correlation of these features with the self-disclosure classes, we conducted a Point-Biserial Correlation analysis. The results show a positive correlation in the variation of pitch, and a fast power spectrum change, for self-disclosing speech, which is expected for speech that contains positive emotions [5]. Besides recognizing vowel frequencies, the third formant of speech also indicates a high-quality natural speech [19]. In our analysis, the third formant showed a positive correlation with expressing self-disclosure, indicating spontaneous and natural speech in these segments. On the other hand, non-self-disclosure speech showed high jitter values indicating roughness, breathiness, or hesitation in the voice [27].

From the **head pose** behavioral features, only 8 features were selected out of the 90 features. In line with previous work, 4 of these features indicated maximum angle, faster movement, and more frequent head tilting in self-disclosure behavior [22], [36]. Moreover, a positive correlation was found between the frequency of looking to the left (toward the robot) and the maximum angle of looking up, which could indicate positive engagement during the self-disclosure event. On the other hand, looking down and slow head pitch movement was correlated with non-self-disclosure segments, which could indicate disengagement or distractions (e.g. playing with a phone).

Finally, **body gestures** selected features were 16 out of 240. Even though we set the final selected features for the top 10% (i.e. 24 features), the framework eliminated "weakly voted" features during the aggregation phase to ensure a strong and representative final feature set. Interestingly, 15 of these selected features were related to face and hand touching, where fast, frequent and short duration of touching behaviors were positively correlated with self-disclosure behavior, which could indicate a self-soothing behavior [11]. Moreover, a slow movement of the torso leaning forward was positively correlated with self-disclosure behavior, indicating positive attitudes toward the disclosing content [16].

C. Other factors Impacting Self-disclosure

We further investigated factors that can contribute to people's tendency to self-disclosure by analyzing participants' personality traits and rapport with the robot. We first calculated the frequency of self-disclosure behaviors for each participant by dividing the number of total self-disclosure response by the total number of open-ended prompts available in the study (14 prompts in total). Pearson's correlation coefficients were calculated between this self-disclosure frequency metric and participants' self-reported personality traits (Mini-IPIP) and working alliance with the robot (WAI-SR).

TABLE III: Percentage of Self-disclosure Statements in each Session/Part of Interaction

Description	Percentage of Occurrences
Responses w/ self disclosure	22.6%
responses Jibo interrupted	3.6%
interruptions with self disclosure	0.48%
interruptions with no self disclosure	3.10%
ST Disclosure	2.4%
CS Disclosure	58.8%
SS Tomorrow Disclosure	78.8%
TGT Disclosure	45.2%
GL Disclosure	0.0%
GL Post Disclosure	52.6%
Savoring Plan Disclosure	26.5%
Savoring Post Disclosure	37.9%

The total WAI score (mean of three WAI sub-component scores) showed marginally positive correlation with participant's self-disclosure frequency ($p=0.059$,$r=0.322$). However, participants' agreement about the intervention tasks (WAI-SR Task) was significantly correlated with self-disclosure ($p=0.043$,$r=0.343$). In addition, self-disclosure was positively correlated with two of the personality traits: openness ($p = 0.026, r = 0.374$) and agreeableness ($p=0.008$,$r=0.440$). These results are in line with the findings by [10].

Types of interaction and questions could also influence people's willingness to self-disclose. Thus, we measured the ratio of self-disclosure responses per each open-ended prompts (Table III). Overall participants' self-disclosure during the intervention was limited. This could be due to the robot presenting itself as a coach and not disclosing about itself during the interactions. Studies on self-disclosure reciprocity suggests that human interlocutors will be more likely to self-disclose if an artificial agent exhibit self-disclosure behavior [31]. In our follow-up study, we plan to study the effect of robot's self-disclosure on the people's engagement and rapport with the robot.

Another factor that could have contributed to the low self-disclosure frequency is the robot's unintended interruption during participants' open-ended response. During the post-study interview, several participants noted that the robot interrupted them while they were still responding and moved on to the next part of the interaction. This type of behavior is commonly found with voiced agents [31] because most voice user interfaces takes a pause as an end-of-speech signal. Such interruption, when repeated, could have refrained participants from self-disclosing in the later interaction sessions, which was reflected in participants' behavior during the Small Talk (ST) task with only 2% of self-disclosure. After realizing the robot tends to interrupt during long responses, some participants gave only short and simple answers, e.g. "Good", "It's going okay", "It was horrible.", etc.

No participant self-disclosed during the Gratitude Letter (GL) session. This was not surprising given that the question asked was "Who would you write the gratitude letter to?" and could be answered in one or two word answer. However, the follow-up prompt (GL Post) showed a higher rate of self-disclosure since the participants were asked to reflect on their experience after sending the gratitude letter. Similarly, Character Strength (CS) and Signature Strength (SS) sessions were more likely to elicit self-disclosure, 58.8% and 78.8% respectively. These sessions prompted participants to share their personal goals and plans to utilize one of their strengths. It is worth noting that some participants did not complete the tasks given as a "homework", and did not have anything to share in the follow-up session (GL-post and Savoring-Post).

These results suggest that giving users the space to talk, showing interest in what they are talking about with follow-up questions could potentially increase people's willingness to self-disclose. We hope to implement these strategies in our future work and positively impact their wellbeing and rapport with the robot.

V. Conclusions

This paper explores how non-verbal behaviors can be used to detect whether people are self-disclosing or not during positive psychology intervention interactions with a social robot. Unlike most existing works that only used linguistic features, we analyzed both acoustic and visual features during people's self-disclosure behaviors. Our classification experiments and feature selection method showed that non-verbal cues alone without linguistic features can detect self-disclosure behaviors with high accuracy (72% to 81%). This finding is significant because it shows that an AI agent can leverage multiple modalities to infer human interlocutors' self-disclosing intention. Human-agent interactions in the real world are messy, and often times one or more input modalities might not be available, e.g. soft speaking, user not in being in the camera view, etc. Based on our results, we plan to build a classifier that can classify people's self-disclosure behavior in real time during multimodal human-machine interactions. Moreover, we found that spontaneous and positive emotions related features from speech, high engagement behaviors from head pose, and self-soothing and positive attitudes cues from body gestures are positively correlated to self-disclosure. Investigating other factors contributing to people's self-disclosure behavior, personality traits (openness and agreeableness) were found to have significant impact on the tendency to self-disclose. Qualitative feedback from the participants suggest that the failure to correctly detect end-of-speech led the robot to unintentionally interrupt people's response and negatively affected the willingness to self-disclose. These suggest that a robust self-disclosure detection using both verbal and non-verbal cues could improve AI agents' ability to build rapport and therapeutic alliance with their users, and potentially result enhance the outcomes of mental health intervention they offer. Besides investigating implementing multimodal self-disclosure detection in future

studies, we plan to study the effect of robot's self-disclosure on the people's self-disclosure, engagement and rapport with the robot.

ACKNOWLEDGMENTS

This work was supported by the Information and Communication Technology (ICT) R&D program of the Ministry of Science and Institute for Information and Communication Technology Promotion of Republic of Korea under grant 2017-0-00162, "Development of Human-Care Robot Technology for Aging Society,".

REFERENCES

[1] S. Alghowinem, T. Gedeon, R. Goecke, J. Cohn, and G. Parker. Interpretation of depression detection models via feature selection methods. *IEEE Transactions on Affective Computing*, (01):1–1, nov 2020.

[2] I. Altman and D. Taylor. Da (1973): Social penetration: The development of interpersonal relationships. *Holt, New York*.

[3] I. Altman and D. A. Taylor. *Social penetration: The development of interpersonal relationships*. Holt, Rinehart & Winston, 1973.

[4] A. Barak and O. Gluck-Ofri. Degree and reciprocity of self-disclosure in online forums. *CyberPsychology & Behavior*, 10(3):407–417, 2007. PMID: 17594265.

[5] C. Breitenstein, D. V. Lancker, and I. Daum. The contribution of speech rate and pitch variation to the perception of vocal emotions in a german and an american sample. *Cognition & Emotion*, 15(1):57–79, 2001.

[6] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1021–1030, 2017.

[7] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[8] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, et al. Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1061–1068, 2014.

[9] M. B. Donnellan, F. L. Oswald, B. M. Baird, and R. E. Lucas. The mini-ipip scales: tiny-yet-effective measures of the big five factors of personality. *Psychological assessment*, 18(2):192, 2006.

[10] J. A. Driesenaar, P. A. De Smet, R. van Hulten, and S. van Dulmen. The relationship between patients' big five personality traits and medication adherence: A systematic review. *Beliefs and Adherence regarding Inhaled Corticosteroids*, 2018.

[11] P. Ekman and W. V. Friesen. Nonverbal leakage and clues to deception. *Psychiatry*, 32(1):88–106, 1969.

[12] C. K. Ettman, S. M. Abdalla, G. H. Cohen, L. Sampson, P. M. Vivier, and S. Galea. Prevalence of Depression Symptoms in US Adults Before and During the COVID-19 Pandemic. *JAMA Network Open*, 3(9):e2019686–e2019686, 09 2020.

[13] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202, 2015.

[14] F. Eyben, M. Wöllmer, and B. Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462, 2010.

[15] B. A. Farber. *Self-disclosure in psychotherapy*. Guilford Press, 2006.

[16] C. K. Goman. *The nonverbal advantage: Secrets and science of body language at work*. ReadHowYouWant. com, 2009.

[17] J. Hart, J. Gratch, and S. Marsella. How virtual reality training can win friends and influence people. In *Fundamental Issues in Defense Training and Simulation*, pages 235–249. CRC Press, 2017.

[18] M. Hesse. The readiness ruler as a measure of readiness to change poly-drug use in drug abusers. *Harm reduction journal*, 3(1):3, 2006.

[19] J. Holmes. The influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer. *IEEE transactions on Audio and Electroacoustics*, 21(3):298–305, 1973.

[20] S. Jeong, S. Alghowinem, L. Aymerich-Franch, K. Arias, A. Lapedriza, R. Picard, H. W. Park, and C. Breazeal. A robotic positive psychology coach to improve college students' wellbeing. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 187–194. IEEE, 2020.

[21] E. Kállay and C. Rus. Psychometric properties of the 44-item version of ryff's psychological well-being scale. *European Journal of Psychological Assessment*, 2014.

[22] S.-H. Kang, J. Gratch, C. Sidner, R. Artstein, L. Huang, and L.-P. Morency. Towards building a virtual counselor: modeling nonverbal behavior during intimate self-disclosure. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 63–70, 2012.

[23] Y.-C. Lee, N. Yamashita, Y. Huang, and W. Fu. "i hear you, i feel you": Encouraging deep self-disclosure through a chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–12, New York, NY, USA, 2020. Association for Computing Machinery.

[24] M. Luo and J. T. Hancock. Self-disclosure and social media: motivations, mechanisms and psychological well-being. *Current Opinion in Psychology*, 31:110–115, 2020. Privacy and Disclosure, Online and in Social Interactions.

[25] X. Ma, J. Hancock, and M. Naaman. Anonymity, intimacy and self-disclosure in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 3857–3869, New York, NY, USA, 2016. Association for Computing Machinery.

[26] J. D. Mayer and Y. N. Gaschke. The brief mood introspection scale (bmis). 1988.

[27] A. Nunes, R. L. Coimbra, and A. Teixeira. Voice quality of european portuguese emotional speech. In T. A. S. Pardo, A. Branco, A. Klautau, R. Vieira, and V. L. S. de Lima, editors, *Computational Processing of the Portuguese Language*, pages 142–151, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[28] W. H. Organization et al. Depression and other common mental disorders: global health estimates. Technical report, World Health Organization, 2017.

[29] D. Páez, C. Velasco, and J. L. González. Expressive writing and the role of alexythimia as a dispositional deficit in self-disclosure and psychological health. *Journal of Personality and Social Psychology*, 77(3):630, 1999.

[30] D. Potdevin, C. Clavel, and N. Sabouret. A virtual tourist counselor expressing intimacy behaviors: A new perspective to create emotion in visitors and offer them a better user experience? *International Journal of Human-Computer Studies*, 150:102612, 2021.

[31] A. Ravichander and A. W. Black. An empirical study of self-disclosure in spoken dialogue systems. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 253–263, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[32] M. E. Seligman, T. A. Steen, N. Park, and C. Peterson. Positive psychology progress: empirical validation of interventions. *American psychologist*, 60(5):410, 2005.

[33] M. Skjuve and P. B. Brandtzæg. Chatbots as a new user interface for providing health information to young people. *Youth and news in a digital media environment–Nordic-Baltic perspectives*, 2018.

[34] M. Soleymani, K. Stefanov, S.-H. Kang, J. Ondras, and J. Gratch. Multimodal analysis and estimation of intimate self-disclosure. In *2019 International Conference on Multimodal Interaction*, ICMI '19, page 59–68, New York, NY, USA, 2019. Association for Computing Machinery.

[35] M. M. C. van Wezel, E. A. J. Croes, and M. L. Antheunis. "i'm here for you": Can social chatbots truly support their users? a literature review. In A. Følstad, T. Araujo, S. Papadopoulos, E. L.-C. Law, E. Luger, M. Goodwin, and P. B. Brandtzaeg, editors, *Chatbot Research and Design*, pages 96–113, Cham, 2021. Springer International Publishing.

[36] R. Zhao, T. Sinha, A. Black, and J. Cassell. Automatic recognition of conversational strategies in the service of a socially-aware dialog system. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 381–392, Los Angeles, Sept. 2016. Association for Computational Linguistics.

[37] H. Zou, I. Jakovlić, R. Chen, D. Zhang, J. Zhang, W.-X. Li, and G.-T. Wang. The complete mitochondrial genome of parasitic nematode camallanus cotti: extreme discontinuity in the rate of mitogenomic architecture evolution within the chromadorea class. *BMC genomics*, 18(1):1–17, 2017.