Multimodal Analysis and Estimation of Intimate Self-Disclosure

Mohammad Soleymani*
University of Southern California
Institute for Creative Technologies
Los Angeles, CA, USA
soleymani@ict.usc.edu

Kalin Stefanov*
University of Southern California
Institute for Creative Technologies
Los Angeles, CA, USA
kstefanov@ict.usc.edu

Sin-hwa Kang
University of Southern California
Institute for Creative Technologies
Los Angeles, CA, USA
kang@ict.usc.edu

Jan Ondras
University of Southern California
Institute for Creative Technologies
Los Angeles, CA, USA
jondras@ict.usc.edu

Jonathan Gratch
University of Southern California
Institute for Creative Technologies
Los Angeles, CA, USA
gratch@ict.usc.edu

ABSTRACT

Self-disclosure to others has a proven benefit for one's mental health. It is shown that disclosure to computers can be similarly beneficial for emotional and psychological well-being. In this paper, we analyzed verbal and nonverbal behavior associated with self-disclosure in two datasets containing structured human-human and human-agent interviews from more than 200 participants. Correlation analysis of verbal and nonverbal behavior revealed that linguistic features such as affective and cognitive content in verbal behavior, and nonverbal behavior such as head gestures are associated with intimate self-disclosure. A multimodal deep neural network was developed to automatically estimate the level of intimate self-disclosure from verbal and nonverbal behavior. Between modalities, verbal behavior was the best modality for estimating self-disclosure within-corpora achieving $r = 0.66$. However, the cross-corpus evaluation demonstrated that nonverbal behavior can outperform language modality in cross-corpus evaluation. Such automatic models can be deployed in interactive virtual agents or social robots to evaluate rapport and guide their conversational strategy.

*Soleymani and Stefanov contributed equally to this work.

ICMI '19, October 14–18, 2019, Suzhou, China

ACM ISBN 978-1-4503-6860-5/19/10…$15.00
https://doi.org/10.1145/3340555.3353737

CCS CONCEPTS

• **Computing methodologies** → Artificial intelligence; Information extraction; Neural networks; • **Human-centered computing** → Laboratory experiments.

KEYWORDS

self-disclosure, neural networks, nonverbal behavior, natural language understanding

ACM Reference Format:
Mohammad Soleymani, Kalin Stefanov, Sin-hwa Kang, Jan Ondras, and Jonathan Gratch. 2019. Multimodal Analysis and Estimation of Intimate Self-Disclosure. In *2019 International Conference on Multimodal Interaction (ICMI '19), October 14–18, 2019, Suzhou, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3340555.3353737

1 INTRODUCTION

Self-disclosing personal and intimate information to others is a prerequisite for verbal psychotherapy [7]. Sharing private information with others has demonstrated emotional, relational and psychological benefits [21]. A recent study found similar benefits for self-disclosure to machines [21]. Self-disclosure, if not done in a safe and secure environment, can be also harmful.

Past work have shown that self-disclosure is facilitated by anonymity and rapport [30]. Anonymity, the feeling that one's identity is protected, makes one at ease for revealing intimate information. Rapport, the feeling of "clicking" with another, is enhanced by the nonverbal coordination of positive emotions and mutual attention during a conversation [39]. In a meta-analysis, Weisband and Kiesler [42] found that patients are more likely to self-disclose to computers rather than other humans as computers induce a higher sense of anonymity. Virtual humans are computer-based agents and have the additional ability to facilitate rapport through coordinated nonverbal displays [15]. Research suggests that

these "virtual rapport" behaviors provide benefits above and beyond those provided by standard computer forms [31]. Therefore, virtual humans are well positioned to deliver mental well-being benefits by leveraging both anonymity and rapport. Understanding the nonverbal and verbal behavior associated with self-disclosure can enable virtual agents to evaluate the situation and steer the interaction towards further disclosure. Despite its potential benefits, self-disclosure on social media can be harmful when sensitive personal information are shared publicly [40]. Similar technologies can be used to monitor users' activities and inform them about the potential risks for sharing too much information.

In this work, we aim to better understand verbal and nonverbal behavior associated with self-disclosure, with the goal of its automatic estimation. To this end, we analyzed the verbal and nonverbal behavior of participants from two datasets containing conversations and semi-structure interviews with ratings for the level of self-disclosure [24, 30]. We automatically extracted facial expressions, head gestures, speech prosody features, and verbal content and performed correlation analysis to identify the associated behavior. To estimate self-disclosure, we formulated the problem as regression to estimate self-disclosure from behavior. Multimodal machine learning models performing deep representation learning were developed for this purpose. We performed within-corpus and between-corpus evaluations to demonstrate the generalizability of the proposed pipeline.

The major contribution of our work is as follows. First, we study nonverbal and verbal behavior associated with self-disclosure across two datasets. Second, we propose a machine learning method for automatic estimation of self-disclosure from both verbal and nonverbal behavior in a conversation and evaluate its effectiveness within and across corpora.

2 BACKGROUND

Farber identified benefits and risks of self-disclosure in therapeutic settings [13]. The benefits of self-disclosure in psychotherapy includes experiencing a greater sense of closeness, validation and affirmation, gaining a more cohesive sense of self, expanding one's sense of self, achieving a greater sense of authenticity, and relieving the psychological pressure of painful experiences. The negative consequences of self-disclosure include the risk of being rejected by the recipient of self-disclosure, burdening another with our secret, creating undesired impression about ourselves, experiencing greater vulnerability and experience of shame [13]. Kang *et al.* [25] studied the nonverbal behavior associated with intimate self-disclosure. They found that participants performed more head nods and tilts and paused more when revealing their own intimate information. The authors then developed a conversational agent imitating such behavior which was found by users via an online survey to express higher level of intimate statements. Zhao *et al.* [44] proposed a computational framework for building rapport in dyadic interaction. Similar to [25], they identified self-disclosure from the agent to be an important factor to consider for recognizing rapport. In a later work, Zhao *et al.* [45] focused on the automatic recognition of social conversational strategies including self-disclosure in dyadic interactions. They performed self-disclosure recognition from verbal and nonverbal behavior, including speech prosody and manually annotated smiles, head nods and gaze behavior. They achieved the accuracy of 85% for recognizing self-disclosure from no self-disclosure. There is a growing interest in automatic understanding of human behavior for human-agent and human-robot interaction [19, 29, 44, 45]. A survey by McColl *et al.* [32] identified a large body of work in HRI that utilizes emotion and behavior tracking for improving human-robot communication. Among others, past research in HRI, demonstrated how emotions and engagement can be used to improve user experience [8].

With the exception of [45] that partially used manually annotated nonverbal behavior, past work on automatic recognition of self-disclosure has been limited to language understanding in interactions with spoken dialogue system [35], online patient support group forums [43], and social media [2, 40]. Wang *et al.* [40] developed a language-based self-disclosure assessment for classifying the instances of self-disclosure on Facebook posts. Facebook users were recruited to share and self-rate the level of self-disclosure on their posts. External observers additionally annotated the posts on the level of self-disclosure. The following set of features were extracted to detect disclosure: word count, emotions, social distance, a measure of social relationship between people mentioned and self, social normality, capturing the extent of uniqueness of the information in language and topic features, extracted by latent Dirichlet allocation (LDA). Support vector regression trained on social media posts achieved $r = 0.6$ for self-disclosure detection. The model was then applied to a larger data collected on Facebook and revealed a number of patterns in self-disclosure behavior. Female and older users were more likely to self-disclose.

In [35], users interaction with a dialogue agent on Amazon Alexa was used to train a machine learning model for recognizing self-disclosure. Utterances spoken during the conversation were labeled to identify the instances of voluntary disclosure of personal information. The following features were extracted for disclosure recognition: bag-of-words TF-IDF features, linguistic style such as the length of utterances, frequency of part-of-speech tags, presence of filler words, etc., and Linguistic Inquiry and Word Count (LIWC) features [22], including affect, pronounces, etc. Additionally, chatbot's utterances were also considered to better capture conversation context. A support vector classifier

achieved F1-score of 0.67 for recognizing disclosure considering both users and chatbot's utterances.

3 DATA

We used two datasets with human-human and human-agent quasi-monologue conversations to study self-disclosure and train and evaluate machine learning models [24, 30]. Both datasets are recorded in studies related to mental health applications (anxiety and distress) labeled by external observers for self-disclosure on responses to specific questions.

Distress Analysis Interview Corpus

This dataset is a subset of Distress Analysis Interview Corpus (DAIC) which was collected with SimSensei system that conducted semi-structured interviews with participants [5], originally labeled for [30]. The interview included different phases starting by an introductory rapport building and a series of questions probing potential symptoms for depression and post-traumatic stress disorder (PTSD). The participants behavior was captured by a front facing camera and a wearable microphone. From 239 participants (149 male, 90 female) who have participated in DAIC studies, 727 responses from 102 participants were labeled and analyzed in this paper. DAIC was collected both by the agent being puppeteered by an experimenter (Wizard-of-Oz scenario) and in fully autonomous mode. The data analyzed in this work includes both the Wizard-of-Oz and autonomous conditions. A snapshot of the agent and a participant are shown in Figure 1.

Eight questions from the more intimate and sensitive part of the interview were selected based on their more frequent usage in the interview. The questions were as follows:

1) "Tell me about a situation that you wish you had handled differently."
2) "How close are you to your family?"
3) "Tell me about an event, or something that you wish you could erase from your memory."
4) "Tell me about the hardest decision you've ever had to make."
5) "Tell me about the last time you felt really happy."
6) "What are you most proud of in your life?"
7) "What's something you feel guilty about?"
8) "When was the last time you argued with someone and what was it about?"

External observers read the transcripts of participants' response to these questions and rated the response on the extent they revealed information. Each response was rated from −3 (completely unwilling to disclose) to +3 (completely willing to disclose) on a seven-point scale. The inter-rater reliability between two raters for an initial subset of the data was sufficient ($\alpha = 0.78$) and the rest of the responses were only labeled by one rater [30].

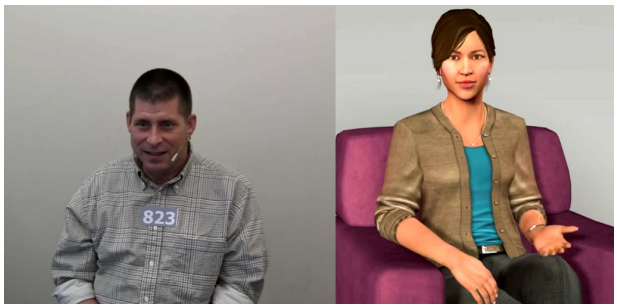

Figure 1: Structured interview with a virtual agent.

Social Anxiety and Self-disclosure Dataset

This dataset was originally collected in the form of computer-mediated one-on-one human-human or agent interview for the purpose of studying the effect of social anxiety on self-disclosure. The interaction scenario was a typical social interaction such as exchanging questions and answers to get to know each other between two individuals performed through a video conferencing system [24] (see Figure 2). Since interactions with virtual agents always happen through a media (*e.g.*, computer screen), the use of computer mediation between two individuals was motivated by creating a situation that is similar to what a human experiences with a virtual agent. In the study [24], interviewers (confederates) asked questions to users without talking about themselves, thus the authors analyzed the intimacy of self-disclosure in interviewees (users)' answers. The interview dataset used in the paper involved three conditions in which avatars differed in visual realism: a raw human video, a degraded human video, and a virtual agent. One hundred and eight participants (54 male, 54 female) participated in the study. The participants were asked to respond to ten questions that were ordered to increasingly ask for more intimate information, as follows:

1) "How old are you?"
2) "What is your hometown?"
3) "What are your favorite things to do in your free time?"
4) "What characteristics of yourself are you most proud of?"
5) "What are some of the things you hate about yourself?"
6) "What do you dislike about your physical appearance?"
7) "What has been the biggest disappointment in your life?"
8) "What have you done in your life that you feel most guilty about?"
9) "What are some of the things that really hurt your feelings?"
10) "What is your most common sexual fantasy?"

The self-disclosure was independently rated by two observers. The observers rated transcribed data of interviewees' answers by first identifying each "disclosure" utterance using Altman and Taylor's three-layer categorization scheme [1]: a peripheral layer, an intermediate layer, and core layer.

According to Altman and Taylor, biographic information, e.g., age, is at the peripheral layer. The examples of each layer included: "I am 30-years old (peripheral layer)," "I like to go shopping (intermediate layer)," and "I feel most guilty about cheating on my girlfriend (core layer)." The observers then rated intimacy levels of verbal self-disclosure using four levels: 0 - no intimacy, 1 - lower intimacy, 2 - intermediate intimacy, and 3 - higher intimacy. After the intimacy levels were judged, inter-coder reliability was measured. To assess inter-coder reliability, the authors performed Krippendorff's alpha for interval data obtained by rating intimacy levels [36]. The results of Krippendorff's alpha showed good inter-coder reliability ($\alpha = 0.84$).

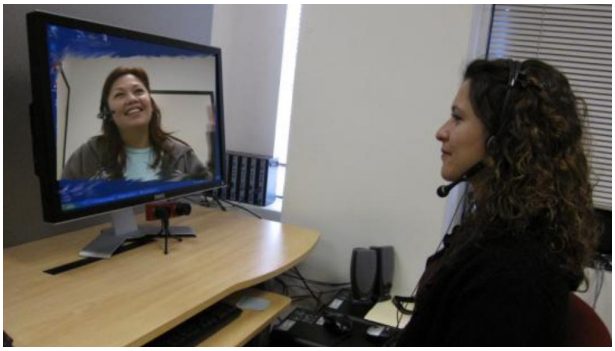

Figure 2: Computer-mediated one-on-one interview.

4 SELF-DISCLOSURE AND BEHAVIOR

Kang *et al.* [25] found that head nods, pauses and head tilts are significantly correlated with intimate self-disclosure. We are similarly interested in understanding the associations between verbal and nonverbal behavior and intimate self-disclosure. To this end, we automatically extracted a number of verbal and nonverbal behavior to measure their correlation with self-disclosure, to validate the results with previous findings [25] and confirm the significance of the captured modalities for automatic recognition of self-disclosure. Higher-level behaviors are not used in our deep learning pipeline.

Automatic Behavior Extraction

To analyze head gestures, we used the dataset labeled in [16] for head nod and head shake detection. It contains excerpts from SEMAINE database collected in interaction with Sensitive Artificial Listener [33]. Each recording is labeled by *other*, *nod* and *shake* classes. First, we extracted the head rotation angles (pitch and yaw) with OpenFace [3], resampled them to 30Hz, and took their first order differences. We then applied a median filter (to remove noise from head motion data [17]) with the empirically selected kernel size of nine samples (300ms). Using the SciPy library [23] we detected maxima and minima in the resulting signal. Each pair of consecutive extreme points (2CEPs) was considered to constitute a head gesture (nod/shake) characterized by its width w and height h. The width of the head gesture was calculated as a duration between a nearest extreme point preceding the 2CEPs and a nearest extreme point following the 2CEPs (if no preceding/following extreme point was found, we used the start/end point of the recording instead). The height of the head gesture was computed as the height difference between 2CEPs. The pair of features (w, h) was extracted for all 2CEPs and averaged per recording, resulting in a feature vector $(\tilde{w}, \tilde{h})$ for each recording.

We treated the nod and shake recognition as two separate binary classification tasks. For nod recognition, we used the features extracted from pitch angle data and extended the negative *other* class with data from the *shake* class. For shake recognition, we used the features extracted from yaw angle data and extended the negative *other* class with data from the *nod* class. These data extensions were possible due to the fact that the three classes of data were mutually exclusive. The resulting feature sets were z-normalized and in each case, the minority class was randomly oversampled using the imbalanced-learn library [28] to balance the number of samples per class. This resulted in 146 and 174 examples per class for nod and shake detection tasks respectively.

Hidden Markov Model (HMM) classifiers have been extensively used for head gesture recognition [16, 26, 38, 41]. Therefore, we first trained a HMM which achieved comparable performance to [16]. However, HMM head gesture recognition performed poorly on our data, and we thus opted for a simple and more generalizable machine learning model, the k-nearest neighbors (kNN) classifier.

We trained two kNN classifiers and performed five-fold cross-validation to find the optimal number k, searching in the range $\{1, 2, \ldots, 24\}$. We found the optimal number of k for nod recognition to be $k_{nod} = 21$ with the mean accuracy (classification rate) of 0.80 ± 0.08 and $k_{shake} = 15$ with the mean accuracy of 0.89 ± 0.04 for shake recognition. Our kNN predictions are also more stable that HMM.

After tracking head pose and facial action units (AUs) with OpenFace [3], we extracted other nonverbal behavior indicators including the standard deviations of head rotation angles (roll, yaw and pitch) and the expression of smile based on AU6 and AU12 intensities. We further used the voice activity detector (VAD) of OpenSMILE [11, 12] to detect the pauses during speech. After thresholding the VAD output by their median value and discarding pauses shorter than 150ms, we calculated pause rate and mean pause duration. For verbal behavior, using LIWC, we extracted a number of linguistic features from the spoken content, such as total function words, total pronouns, articles, verbs and psychological processes such as affective and cognitive processes.

Correlation Analysis

After calculating the correlation coefficients, we visually inspected the scatter plots and removed the correlations that were mainly driven by few samples, after removing the outliers. We only report the correlation coefficients larger than 0.15 with $p < 5 \times 10^{-5}$. The correlation coefficient between behaviors and disclosure levels are given in Table 1.

Table 1: Significant Pearson correlation coefficients between different verbal and nonverbal behavior and level of self-disclosure ($\rho > 0.15$).

Dataset	**Distress**	**Social anxiety**
Behavior	Spearman ρ	Spearman ρ
Verbal		
Word count	0.57	0.17
Articles	-	0.16
Prepositions	-	0.22
Common adjectives	-	0.15
Function words	0.21	0.20
Conjunctions	0.36	0.26
Negation	-0.26	-0.27
Present focus	-0.20	-
Affect	-	0.17
Comparisons	-	0.19
Drives	-	0.21
Tentative	-	0.20
Nonverbal		
Head pose std pitch	0.16	0.17
Head pose std yaw	0.19	0.16
Head nods	0.23	0.17
Pause rate	-	0.16

A number of linguistic characteristics including the numbers of spoken terms, negations, conjunctions and prepositions are associated with disclosure. The longer participants spoke and the more complex their sentences were, they revealed more. Not all verbal behavior were consistently correlated for both datasets. From nonverbal behavior, head movements and head nod counts are correlated with self-disclosure for both datasets. Pause rate was only significantly correlated for the social anxiety dataset. Kang *et al.* [25] have also found head nods and head tilt to be associated with disclosure. Similarly, smile was not associated with self-disclosure. Unlike [25], we found a weak correlation with pauses which might be as a result of using an automated method for extracting pauses.

5 AUTOMATIC ESTIMATION OF SELF-DISCLOSURE

Modalities and Features

For both datasets, three modalities capturing verbal and nonverbal behavior of participates were analyzed. Participants' spoken content was manually transcribed. Nonverbal behavior was captured by front facing cameras and microphones. Videos were used to track facial expression and head pose, and speech prosody was analyzed from audio recorded by head-worn microphones.

Language. To represent the spoken words, we used two different tools for mapping the spoken utterances to a representation (vector), *i.e.*, a data-driven one (BERT) [6] and a dictionary-based tool (LIWC) [22].

Bidirectional Encoder Representations from Transformers (BERT) [6] is a method for learning a language model that can be trained on large amount of data in an unsupervised manner. This pre-trained model is very effective in representing a sequence of terms as a fixed-length representation (vector). BERT architecture is a multi-layer bidirectional Transformer network that encodes the whole sequence at once. BERT representation achieves state-of-the-art results in multiple natural language understanding tasks. In this paper, we used pre-trained BERT for transforming participants' responses for each instance into a 768-dimensional vector.

Linguistic Inquiry and Word Count (LIWC) is a lexical tool that matches the terms in a document with its dictionary and generates scores along different dimensions including linguistic variables such as number of conjunctions and pronouns and affective and cognitive constructs such as "present focus" and "positive emotion" [22]. The terms in each category or selected by experts and is extensively validated on different content. Using LIWC, we extracted 93-dimensional features from the verbal content of each instance.

Speech. We extracted three types of features in order to capture the speech prosody, namely extended Geneva Minimalistic Acoustic Parameter set (GeMAP), MFCC and a deep representation (VGGish). 13 band mel-frequency cepstral coefficients (MFCC) were extracted from audio signals from 25ms audio frames. The MFCC and their first and second order derivatives were extracted using OpenSMILE [11, 12] and generated a $T \times 39$ matrix for each audio sample. GeMAP is a set of acoustic features selected by experts in speech processing psychology of emotion for their potential to index affective content in voice production, their proven performance in literature and their theoretical significance [10]. The extended set of GeMAP or eGeMAP consists of 23 features such as fundamental frequency, loudness and formants.

Deep neural networks trained on large quantities of data have shown to be able to learn powerful representations [6, 18, 20]. VGGish is a deep convolutional neural network trained on audio spectrograms extracted from a large database of videos to recognize an ontology of 632 audio event categories, for example, vehicle noise, music genre, human locomotion [14, 20]. The audio files were first converted to log-mel spectrogram and the resulting images were fed to

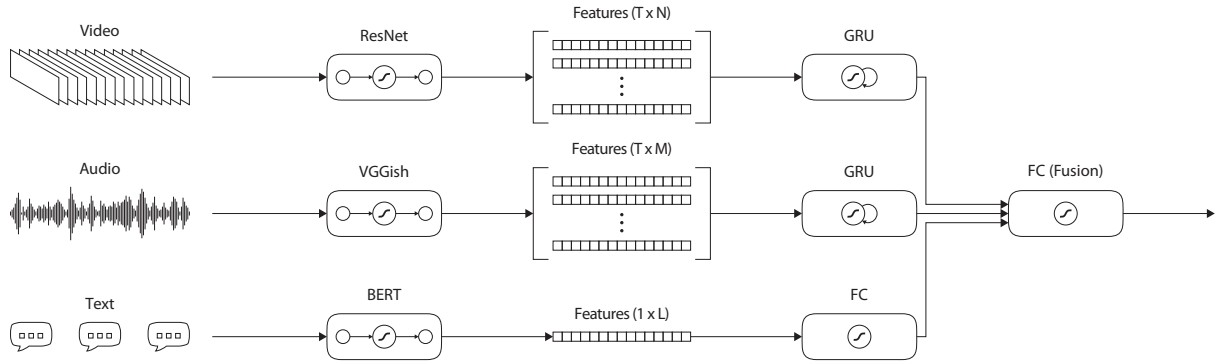

Figure 3: Multimodal method for estimation of self-disclosure. The inputs from different modalities are passed through specific data representation modules, in this case, ResNet for vision, VGGish for speech, and BERT for language. Then the output representations of these modules are passed through instance-based (language) and sequence-based (vision and speech) encoders and mapped to a 128-dimensional vector. Finally, the representations are fused and the level of self-disclosure is estimated.

a modified VGG deep convolutional neural network [37] and trained to recognize Audioset classes. We used the 128-dimensional embedding that can be generated by VGGish after dimensionality reduction with Principal Component Analysis (PCA)[1]. We used the hop size of 33ms, meaning a 128-dimensional vector was extracted for every 33ms of the audio signals. As a result, each audio sequence is represented by a $T \times 128$ matrix, where T is the number of frames.

Vision. Two sets of features capturing head pose and facial expressions were extracted. OpenFace [3] uses computer vision techniques to track head pose and detect facial action units. The Facial Action Coding System (FACS) [9] is a taxonomy of facial movements that describes facial expressions, *e.g.*, lip puller and chin raiser. We used OpenFace to detect the intensity of facial action units, head pose variations and images of aligned faces per frame. We used the intensity of 17 facial action units in addition to the head pose angels as features. Additionally for every frame, we extracted a representation from the penultimate layer of ResNet-50 [18], trained on ImageNet [4], after feeding the network with aligned faces from each frame.

Methods

We formulated the automatic estimation of self-disclosure as a regression problem. For every modality we developed an encoder that maps its input to a 1×128 embedding. Each of these encoders is then followed by a regression module that outputs a continuous value for the level of self-disclosure. Given the nature of self-disclosure, we evaluated the proposed models with Spearman correlation (r).

Language information is encoded with instance-based encoders. These encoders consist of a single fully connected (FC) layer of size 128 that maps the language representations to a 128-dimensional embedding. Since temporal dynamics of human behavior is important in communication we used recurrent layers as sequence-based encoders for the speech and visual modalities. These encoders consist of a single layer gated recurrent unit (GRU) of size 128 that maps video and speech to a 128-dimensional embedding (*i.e.*, only the last state of the GRU is kept). All encoders are then followed by one fully connected (FC) layer of size 64 and a linear layer that outputs a scalar continuous value for the level of self-disclosure.

Additionally, we developed a multimodal model that uses the unimodal encoders. First, we trained all unimodal models and evaluated their performances. Then, we used the trained encoders of the best performing unimodal models (one per modality) in a multimodal model. This model includes three pre-trained encoders followed by one FC layer of size 192 for fusion. Then a final linear layer outputs a scalar value for the level of self-disclosure. We have illustrated the multimodal model in Figure 3. We have also performed late fusion by simply averaging the output of all modalities.

6 EXPERIMENTS

We conducted experiments with both datasets described in Section 3. We present three experiments with the proposed methods and datasets: within-corpus or corpus-dependent, combined dataset and cross-corpus evaluation.

The goal of the within-corpus experiment is to test to what extent the proposed methods can estimate the level of self-disclosure when data from the same participants are considered. For each dataset we used a k-fold ($k = 10$) cross-validation procedure to train and evaluate the models. As described previously, first we evaluated the performance of all unimodal models. Given this information, we used the best performing unimodal within-corpus encoders for fusion, and trained and evaluated a multimodal model.

[1] https://github.com/tensorflow/models/tree/master/research/audioset

The experiment on combined dataset is evaluated similarly to the within-corpus experiment, *i.e.*, through k-fold cross-validation. In this case, however, we combined all data from both datasets. The goal of this experiment is to investigate the effect of the size of the dataset and to evaluate on a more diverse set. The goal of the cross-corpus experiment is to test to what extent the performance of the proposed methods generalizes to unseen data and participants. In this experiment, we trained the models using all data from one of the datasets. Then, we evaluated the models with all the instances from the other dataset. Given the information about the best performing unimodal corpus-independent method, we used the corresponding encoders for fusion, and trained and evaluated a multimodal cross-corpus model.

During training we held out $\sim$ 20% of the data for the validation set. All models were trained with Adam [27] optimizer with default parameters ($\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$) and mean squared error (MSE) loss function. Each model was trained for 100 epochs with dropout rate of 0.1 for all layers. The models achieving the best validation performance (*i.e.*, lowest MSE) were selected for evaluation on the test set. The models are evaluated in terms of Spearman's rank correlation coefficient between the estimations and the labels. All models are implemented in PyTorch [34].

7 RESULTS

We have summarized all evaluation results in Table 2. For within-corpus and combined experiments, the mean and standard deviation of the performance over folds are reported.

The results from the within-corpus experiment are summarized in the first and second column of Table 2. For both datasets the best performance is achieved with representations extracted from language, $r = 0.66$ and $r = 0.58$ by BERT for the Distress and Social anxiety datasets, respectively. The best result for each modality is shown in boldface font. The deep representations learned from large amount of data achieves the best performance in most cases, *i.e.*, BERT for language, ResNet for vision and VGGish for speech. Furthermore, the results show that different fusion strategies of the best performing models (marked in boldface font) fall short of outperforming the best modality (language) for within-corpus evaluations.

The results from the combined within-corpus experiment are summarized in the third column of the table. The obtained results suggest that the methods can scale to more and diverse data without significant decrease in performance. The results from the cross-corpus experiment are summarized in the forth and fifth column in the table. Comparing the result of this experiment with the results from the corpus-dependent experiment, suggests that the methods can generalize to unseen data to certain extent. Features representing audiovisual data (nonverbal behavior) generalized better than language modality. The social anxiety dataset was recorded with lower quality microphones and that might partly explain the failure of MFCC features in cross-corpus evaluation. In cross-corpus evaluation, the multimodal fusion outperforms the unimodal results.

8 DISCUSSION

The result of the within-corpus experiments clearly shows that language is the best-performing modality for estimation of self-disclosure. In all three corpus-dependent experiments, the representations obtained from verbal input yielded the best performance of the proposed methods. On the contrary, the results of the between-corpus experiments are in favor of the representations extracted from audiovisual data or nonverbal behavior. One interpretation of this result might be that the verbal content indicative for self-disclosure changes significantly across different corpora and participants as it is biased to the context of the conversation, *e.g.*, datasets involved different questions from the agent or confederate. Moreover, verbal expressions of self-disclosure require a more conscious effort that might be specific to each participant. On the other hand, the nonverbal behavior, which is less controllable (requires less conscious effort) is somewhat similar across participants. Therefore, one might speculate that nonverbal indicators are more appropriate for estimation of self-disclosure that can generalize beyond the data used for training the models.

The Social anxiety dataset is considerably more challenging than the Distress dataset from data quality point of view. Both image quality and audio quality are considerably worse, due to the poorer quality of the instrumentation, which might explain the persistently worse results on that dataset. Social anxiety dataset is also about 1.5 times bigger than the portion of the Distress dataset that is used in this paper. The size of the datasets might partly explain the differences in cross-corpus performance as the results reported on Distress dataset were trained on a larger training set (Social anxiety). Finally, the annotation procedures, as described in Section 3 are slightly different and might introduce further discrepancies in the cross-corpus experiments. For both datasets, labelling was performed based on observations only, and we do not have any knowledge about the truthfulness of the statements. The effect of potential deception assessed in the statements is considered for the future work.

In addition to nonverbal behavior, deep representations from audiovisual data can also capture features related to participants' appearance and gender. One caveat of using such features is that our machine learning model might learn confounding factors such as gender, appearance or participant-specific features. However, a t-test on disclosure scores revealed no significant differences between male and female participants' disclosure ($p = 0.97$). Our cross-corpus

Table 2: Disclosure estimation results under different conditions. Within-corpus results are trained and tested on each dataset and evaluated with k-fold cross-validation. Combined refers to the case where we cross-validate over both datasets combined. For cross-corpus, we train on one dataset and evaluate on the other one. For the firs three columns, the numbers correspond to mean (standard deviation) over folds.

Evaluation	Within-corpus		Combined	Cross-corpus	
Dataset	Distress	Social anxiety	Both	Eval. on Distress	Eval. on Social anxiety
Features	r [μ (σ)]	r [μ (σ)]	r [μ (σ)]	r	r
Language					
BERT	**0.66 (0.04)**	**0.58 (0.06)**	**0.58 (0.05)**	0.20	**0.40**
LIWC	0.64 (0.06)	0.47 (0.07)	0.54 (0.04)	**0.35**	-0.00
Speech					
VGGish	**0.61 (0.05)**	0.40 (0.09)	**0.49 (0.09)**	**0.60**	**0.39**
eGeMAPS	0.48 (0.08)	0.37 (0.10)	0.42 (0.08)	0.44	0.34
MFCC	0.53 (0.08)	**0.42 (0.10)**	0.40 (0.11)	0.04	-0.12
Vision					
ResNet	**0.62 (0.05)**	**0.39 (0.10)**	0.49 (0.08)	**0.61**	**0.39**
HeadPose	0.61 (0.05)	0.39 (0.10)	**0.50 (0.07)**	0.60	0.39
AU	0.14 (0.16)	0.33 (0.09)	0.43 (0.06)	0.42	0.01
Multimodal					
Late Fusion	0.64 (0.04)	0.53 (0.08)	0.58 (0.06)	**0.62**	**0.42**
NN Fusion	0.64 (0.04)	0.57 (0.06)	0.58 (0.06)	0.39	0.40

results achieved significant results despite being trained and evaluated on separate populations which should rule out most of participant-dependent factors.

9 CONCLUSIONS

Intimate self-disclosure has potential benefits for mental well-being. It can be also harmful when done in an insecure environment. Learning its verbal and nonverbal markers is thus beneficial for designing better interactive social robots or virtual agents. In this paper, we studied and analyzed verbal and nonverbal behavior during intimate self-disclosure. We trained a multimodal deep neural network to estimate the level of self-disclosure which achieved promising performance, in line or superior to the state-of-the-art [40].

Correlation analysis on verbal and nonverbal behavior revealed that the linguistic content of the verbal behavior is associated with self-disclosure. Overall, word count, affective and cognitive processes verbally expressed and sentence constructions were important indicators of intimate self-disclosure. Head gestures such as nods and speech pauses were also associated with self-disclosure.

We trained and evaluated our deep neural network to estimate the level of self-disclosure on two datasets. For all modalities, deep representations learned from large corpora were found to be the best features for recognizing self-disclosure. We performed within-corpus and cross-corpus evaluations to study the performance and robustness of different modalities for this purpose. The verbal channel provided the best features for estimating self-disclosure in within-corpus evaluations, achieving $r = 0.66$. For both datasets, the disclosure labels were given to the transcribed content which introduces a bias toward the verbal content. However, we found nonverbal behavior (audiovisual) to provide more robust performance in cross-corpus evaluation. Multimodal fusion achieved comparable performance to the best modality (language) in within-corpus evaluation and outperformed the best modality in cross-corpus evaluation.

This work was among the first attempts to understand both verbal and nonverbal behavior of self-disclosure. We demonstrated the feasibility of estimating the level of self-disclosure and showed its robustness when evaluated across corpora. Automatic estimation of self-disclosure will enable machines to better sense their users' state and provide an engaging experience. In certain contexts such as social media, similar methods can be used to alert users to avoid potential risks associated with disclosing personal information.

ACKNOWLEDGMENTS

This work was supported in part by the U.S. Army. Any opinion, content or information presented does not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

REFERENCES

[1] I. Altman and D. A. Taylor. 1973. *Social Penetration: The Development of Interpersonal Relationships.* Holt, Rinehart & Winston.

[2] S. Balani and M. De Choudhury. 2015. Detecting and Characterizing Mental Health Related Self-Disclosure in Social Media. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '15.* ACM Press, New York, New York, USA, 1373–1378.

[3] T. Baltrusaitis, P. Robinson, and L. P. Morency. 2016. OpenFace: An open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016.* IEEE, 1–10.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition.* IEEE, 248–255.

[5] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, G. Lucas, S. Marsella, F. Morbini, A. Nazarian, S. Scherer, G. Stratou, A. Suri, D. Traum, R. Wood, Y. Xu, A. Rizzo, and L.-p. Morency. 2014. SimSensei Kiosk : A Virtual Human Interviewer for Healthcare Decision Support. In *International Conference on Autonomous Agents and Multi-Agent Systems.* 1061–1068.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL 201.*

[7] R. Digiuseppe and M. E. Bernard. 2006. REBT assessment and treatment with children. In *Rational Emotive Behavioral Approaches to Childhood Disorders: Theory, Practice and Research*, A.Ellis and M. E.Bernard (Eds.). Springer US, Boston, MA, 85–114.

[8] K. Drejing, S. Thill, and P. Hemeren. 2015. Engagement: A traceable motivational concept in human-robot interaction. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII).* IEEE, 956–961.

[9] P. Ekman and W. Friesen. 1978. *The Facial Action Coding System (FACS).* Consulting Psychologists Press, Stanford University, Palo Alto.

[10] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong. 2016. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing* 7, 2 (apr 2016), 190–202.

[11] F. Eyben, F. Weninger, F. Gross, and B. Schuller. 2013. Recent developments in openSMILE, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia - MM '13.* ACM Press, New York, New York, USA, 835–838.

[12] F. Eyben, M. Wöllmer, and B. Schuller. 2010. OpenSMILE: The Munich Versatile and Fast Open-source Audio Feature Extractor. In *Proceedings of the 18th ACM International Conference on Multimedia (MM '10).* ACM, New York, NY, USA, 1459–1462.

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

[14] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 776–780.

[15] J. Gratch, A. Okhmatovskaia, F. Lamothe, S. Marsella, M. Morales, R. J. van der Werf, and L.-P. Morency. 2006. Virtual Rapport. In *Intelligent Virtual Agents*, J.Gratch, M.Young, R.Aylett, D.Ballin, and P.Olivier (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 14–27.

[16] H. Gunes and M. Pantic. 2010. Dimensional Emotion Recognition from Spontaneous Head Gestures for Interaction with Sensitive Artificial Listener. In *Proceedings Int'l Conf. Intelligent Virtual Agents (IVA'10).* Philadelphia, USA, 371–377.

[17] S. Gurbuz, E. Oztop, and N. Inoue. 2012. Model free head pose estimation using stereovision. *Pattern Recognition* 45, 1 (2012), 33–42.

[18] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 770–778.

[19] J. Hemminghaus and S. Kopp. 2017. Towards Adaptive Social Behavior Generation for Assistive Robots Using Reinforcement Learning. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction - HRI '17.* ACM Press, New York, New York, USA, 332–340.

[20] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, et al. 2017. CNN architectures for large-scale audio classification. In *2017 IEEE international conference on acoustics, speech and signal processing (icassp).* IEEE, 131–135.

[21] A. Ho, J. Hancock, and A. S. Miner. 2018. Psychological, Relational, and Emotional Effects of Self-Disclosure After Conversations With a Chatbot. *Journal of Communication* 68, 4 (aug 2018), 712–733.

[22] A. E. M. Jeffrey H. Kahn, Renée M. Tobin and J. A. Anderson. 2007. Measuring Emotional Expression with the Linguistic Inquiry and Word Count. *JSTOR: The American Journal of Psychology* 120, 2 (2007), 263–286.

[23] E. Jones, T. Oliphant, and P. Peterson. 2014. SciPy: open source scientific tools for Python. (2014). http://www.scipy.org/ [Online; accessed 05/12/2019].

[24] S.-H. Kang and J. Gratch. 2010. Virtual humans elicit socially anxious interactants' verbal self-disclosure. *Computer Animation and Virtual Worlds* 21, 3-4 (2010), 473–482. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/cav.345

[25] S.-H. Kang, J. Gratch, C. Sidner, R. Artstein, L. Huang, and L.-P. Morency. 2012. 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 (AAMAS '12).* International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 63–70.

[26] A. Kapoor and R. W. Picard. 2001. A real-time head nod and shake detector. In *Proceedings of the 2001 workshop on Perceptive user interfaces.* ACM, 1–5.

[27] D. P. Kingma and J. Ba. 2014. Adam: a method for stochastic optimization. *Computing Research Repository* abs/1412.6980 (2014).

[28] G. Lemaître, F. Nogueira, and C. K. Aridas. 2017. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research* 18, 17 (2017), 1–5.

[29] P. Liu, D. F. Glas, T. Kanda, and H. Ishiguro. 2016. Data-Driven HRI: Learning Social Behaviors by Example From Human–Human Interaction. *IEEE Transactions on Robotics* 32, 4 (aug 2016), 988–1008.

[30] G. M. Lucas, J. Gratch, A. King, and L.-P. Morency. 2014. It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior* 37 (2014), 94–100.

[31] G. M. Lucas, A. Rizzo, J. Gratch, S. Scherer, G. Stratou, J. Boberg, and L.-P. Morency. 2017. Reporting Mental Health Symptoms: Breaking Down Barriers to Care with Virtual Human Interviewers. *Frontiers in Robotics and AI* 4, 51 (2017).

[32] D. McColl, A. Hong, N. Hatakeyama, G. Nejat, and B. Benhabib. 2016. A Survey of Autonomous Human Affect Detection Methods for Social Robots Engaged in Natural HRI. *Journal of Intelligent & Robotic Systems* 82, 1 (apr 2016), 101–133.

[33] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder. 2012. The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent. *IEEE Transactions on Affective Computing* 3, 1 (2012), 5–17.

[34] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. 2017. Automatic differentiation in PyTorch. In *Autodiff Workshop, NIPS*.

[35] A. Ravichander and A. W. Black. 2018. An Empirical Study of Self-Disclosure in Spoken Dialogue Systems. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*. Association for Computational Linguistics, Melbourne, Australia, 253–263.

[36] M. Shelley and K. Krippendorff. 1984. Content Analysis: An Introduction to its Methodology. *J. Amer. Statist. Assoc.* (1984).

[37] K. Simonyan and A. Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[38] W. Tan and G. Rong. 2003. A real-time head nod and shake detector using HMMs. *Expert Systems with Applications* 25, 3 (2003), 461–466.

[39] L. Tickle-Degnen and R. Rosenthal. 1990. The Nature of Rapport and Its Nonverbal Correlates. *Psychological Inquiry* 1, 4 (oct 1990), 285–293.

[40] Y.-C. Wang, M. Burke, and R. E. Kraut. 2016. Modeling Self-Disclosure in Social Networking Sites. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing - CSCW '16*. ACM Press, New York, New York, USA, 74–85.

[41] H. Wei, P. Scanlon, Y. Li, D. S. Monaghan, and N. E. O'Connor. 2013. Real-time head nod and shake detection for continuous human affect recognition. In *2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*. IEEE, 1–4.

[42] S. Weisband and S. Kiesler. 1996. Self disclosure on computer forms. In *Proceedings of the SIGCHI conference on Human factors in computing systems common ground - CHI '96*. ACM Press, New York, New York, USA, 3–10.

[43] D. Yang, Z. Yao, and R. Kraut. 2017. Self-Disclosure and Channel Difference in Online Health Support Groups. *Eleventh International AAAI Conference on Web and Social Media* (may 2017).

[44] R. Zhao, A. Papangelis, and J. Cassell. 2014. Towards a Dyadic Computational Model of Rapport Management for Human-Virtual Agent Interaction. In *Intelligent Virtual Agents*, T.Bickmore, S.Marsella, and C.Sidner (Eds.). Springer International Publishing, 514–527.

[45] R. Zhao, T. Sinha, A. Black, and J. Cassell. 2016. 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*. 381–392.