Verbal indicators of psychological distress in interactive dialogue with a virtual human

David DeVault, Kallirroi Georgila, Ron Artstein, Fabrizio Morbini, David Traum, Stefan Scherer, Albert (Skip) Rizzo, Louis-Philippe Morency
University of Southern California, Institute for Creative Technologies
Playa Vista, CA
devault@ict.usc.edu

Abstract

We explore the presence of indicators of psychological distress in the linguistic behavior of subjects in a corpus of semi-structured virtual human interviews. At the level of aggregate dialogue-level features, we identify several significant differences between subjects with depression and PTSD when compared to non-distressed subjects. At a more fine-grained level, we show that significant differences can also be found among features that represent subject behavior during specific moments in the dialogues. Finally, we present statistical classification results that suggest the potential for automatic assessment of psychological distress in individual interactions with a virtual human dialogue system.

1 Introduction

One of the first steps toward dealing with psychological disorders such as depression and PTSD is diagnosing the problem. However, there is often a shortage of trained health care professionals, or of access to those professionals, especially for certain segments of the population such as military personnel and veterans (Johnson et al., 2007). One possible partial remedy is to use virtual human characters to do a preliminary triage screening, so that mental healthcare providers can focus their attention on those who are most likely to need help. The virtual human would engage an individual in an interview and analyze some of their behavioral characteristics. In addition to serving a triage function, this automated interview could produce valuable information to help the healthcare provider make their expert diagnosis.

In this paper, we investigate whether features in the linguistic behavior of participants in a conversation with a virtual human could be used for recognizing psychological distress. We focus specifically on indicators of depression and post-traumatic stress disorder (PTSD) in the verbal behavior of participants in a Wizard-of-Oz corpus.

The results and analysis presented here are part of a broader effort to create an automated, interactive virtual human dialogue system that can detect indicators of psychological distress in the multimodal communicative behavior of its users. Realizing this vision requires a careful and strategic design of the virtual human's dialogue behavior, and in concert with the system's behavior, the identification of robust "indicator" features in the verbal and nonverbal responses of human interviewees. These indicators should be specific behavior patterns that are empirically correlated with specific psychological disorders, and that can inform a triage screening process or facilitate the diagnosis or treatment performed by a clinician.

In this paper, we report on several kinds of such indicators we have observed in a corpus of 43 Wizard-of-Oz interactions collected with our prototype virtual human, Ellie, pictured in Figure 1. We begin in Section 2 with a brief discussion of background and related work on the communicative behavior associated with psychological distress. In Section 3, we describe our Wizard-of-Oz data set. Section 4 presents an analysis of indicator features we have explored in this data set, identifying several significant differences between subjects with depression and PTSD when compared to non-distressed subjects. In Section 5 we present statistical classification results that suggest the potential for automatic assessment of psychological distress based on individual interactions with a virtual human dialogue system. We conclude in Section 6.

2 Background and Related Work

There has been a range of psychological and clinical research that has identified differences in the

Proceedings of the SIGDIAL 2013 Conference, pages 193–202,
Metz, France, 22-24 August 2013.

Figure 1: Ellie.

communicative behavior of patients with specific psychological disorders such as depression. In this section, we briefly summarize some closely related work.

Most work has observed the behavior of patients in human-human interactions, such as clinical interviews and doctor-patient interactions. PTSD is generally less well studied than depression.

Examples of the kinds of differences that have been observed in non-verbal behavior include differences in rates of mutual gaze and other gaze patterns, downward angling of the head, mouth movements, frowns, amount of gesturing, fidgeting, emotional expressivity, and voice quality; see Scherer et al. (2013) for a recent review.

In terms of verbal behavior, our exploration of features here is guided by several previous observations in the literature. Cohn and colleagues have identified increased speaker-switch durations and decreased variability of vocal fundamental frequency as indicators of depression, and have explored the use of these features for classification (Cohn et al., 2009). That work studied these features in human-human clinical interviews, rather than in virtual human interactions as reported here. In clinical studies, acute depression has been associated with decreased speech, slow speech, delays in delivery, and long silent pauses (Hall et al., 1995). Aggregate differences in lexical frequencies have also been observed. For example, in written essays, Rude et al. (2004) observed that depressed participants used more negatively valenced words and used the first-person pronoun "I" more frequently than never-depressed individuals. Heeman et al. (2010) observed differences in children with autism in how long they pause before speaking and in their use of fillers, acknowledgments, and discourse markers. Some of these features are similar to those studied here, but looked at children communicating with clinicians rather than a virtual human dialogue system.

Recent work on machine classification has demonstrated the ability to discriminate between schizophrenic patients and healthy controls based on transcriptions of spoken narratives (Hong et al., 2012), and to predict patient adherence to medical treatment from word-level features of dialogue transcripts (Howes et al., 2012). Automatic speech recognition and word alignment has also been shown to give good results in scoring narrative recall tests for identification of cognitive impairment (Prud'hommeaux and Roark, 2011; Lehr et al., 2012).

3 Data Set

In this section, we introduce the Wizard-of-Oz data set that forms the basis for this paper. In this virtual human dialogue system, the character Ellie depicted in Figure 1 carries out a semi-structured interview with a single user. The system was designed after a careful analysis of a set of face-to-face interviews in the same domain. The face-to-face interviews make up the large human-human Distress Assessment Interview Corpus (DAIC) that is described in Scherer et al. (2013). Drawing on observations of interviewer behavior in the face-to-face dialogues, Ellie was designed to serve as an interviewer who is also a good listener, providing empathetic responses, backchannels, and continuation prompts to elicit more extended replies to specific questions. The data set used in this paper is the result of a set of 43 Wizard-of-Oz interactions where the virtual human interacts verbally and nonverbally in a semi-structured manner with a participant. Excerpts from the transcripts of two interactions in this Wizard-of-Oz data set are provided in the appendix in Figure 5.[1]

3.1 Procedure

The participants were recruited via Craigslist and were recorded at the USC Institute for Creative

[1] A sample demonstration video of an interaction between the virtual agent and a human actor can be seen here: http://www.youtube.com/watch?v=ejczMs6b1Q4

Technologies. In total 64 participants interacted with the virtual human. All participants who met requirements (i.e. age greater than 18, and adequate eyesight) were accepted. In this paper, we focus on a subset of 43 of these participants who were told that they would be interacting with an automated system. (The other participants, which we exclude from our analysis, were aware that they were interacting with a human-controlled system.) The mean age of the 43 participants in our data set was 36.6 years, with 23 males and 20 females.

We adhered to the following procedure for data collection: After a short explanation of the study and giving consent, participants completed a series of questionnaires. These questionnaires included the PTSD Checklist-Civilian version (PCL-C) and the Patient Health Questionnaire, depression module (PHQ-9) (Scherer et al., 2013) along with other questions. Then participants engage in an interview with the virtual human, Ellie. After the dialogue concludes, participants are then debriefed (i.e. the wizard control is revealed), paid $25 to $35, and escorted out.

The interaction between the participants and Ellie was designed as follows: Ellie explains the purpose of the interaction and that she will ask a series of questions. She then tries to build rapport with the participant in the beginning of the interaction with a series of casual questions about Los Angeles. Then the main interview begins, including a range of questions such as:

> *What would you say are some of your best qualities?*
>
> *What are some things that usually put you in a good mood?*
>
> *Do you have disturbing thoughts?*
>
> *What are some things that make you really mad?*
>
> *How old were you when you enlisted?*
>
> *What did you study at school?*

Ellie's behavior was controlled by two human "wizards" in a separate room, who used a graphical user interface to select Ellie's nonverbal behavior (e.g. head nods, smiles, back-channels) and verbal utterances (including the interview questions, verbal back-channels, and empathy responses). This Wizard-of-Oz setup allows us to prove the utility of the protocol and collect training data for the eventual fully automatic interaction. The speech for each question was pre-recorded using an amateur voice actress (who was also one of the wizards). The virtual human's performance of these utterances is animated using the SmartBody animation system (Thiebaux et al., 2008).

3.2 Condition Assessment

The PHQ-9 and PCL-C scales provide researchers with guidelines on how to assess the participants' conditions based on the responses. Among the 43 participants, 13 scored above 10 on the PHQ-9, which corresponds to moderate depression and above (Kroenke et al., 2001). We consider these 13 participants as positive for depression in this study. 20 participants scored positive for PTSD, following the PCL-C classification. The two positive conditions overlap strongly, as the evaluated measurements PHQ-9 and PCL-C correlate strongly (Pearson's r > 0.8, as reported in Scherer et al. (2013)).

4 Feature Analysis

4.1 Transcription and timing of speech

We have a set $D = \{d_1, ..., d_{43}\}$ of 43 dialogues. The user utterances in each dialogue were transcribed using ELAN (Wittenburg et al., 2006), with start and end timestamps for each utterance.[2] At each pause of 300ms or longer in the user's speech, a new transcription segment was started. The resulting speech segments were subsequently reviewed and corrected for accuracy.

For each dialogue $d_i \in D$, this process resulted in a sequence of user speech segments. We represent each segment as a tuple $\langle s, e, t \rangle$, where s and e are the starting and ending timestamps in seconds, and t is the manual text transcription of the corresponding audio segment. The system speech segments, including their starting and ending timestamps and verbatim transcripts of system utterances, were recovered from the system log files.

To explore aggregate statistical features based on user turn-taking behavior in the dialogues, we employ a simple approach to identifying turns within the dialogues. First, all user and system speech segments are sorted in increasing order of

[2]ELAN is a tool that supports annotation of video and audio, from the Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands. It is available at `http://tla.mpi.nl/tools/tla-tools/elan/`.

Segment level features
(a) mean speaking rate of each user segment
(b) mean onset time of first segment in each user turn
(c) mean onset time of non-first segments in user turns
(d) mean length of user segments
(e) mean minimum valence in user segments
(f) mean mean valence in user segments
(g) mean maximum valence in user segments
(h) mean number of filled pauses in user segments
(i) mean filled pause rate in user segments
Dialogue level features
(j) total number of user segments
(k) total length of all user segments

Figure 2: List of context-independent features.

their starting timestamps. All consecutive segments with the same speaker are then designated as constituting a single turn. While this simple scheme does not provide a detailed treatment of relevant phenomena such as overlapping speech, backchannels, and the interactive process of negotiating the turn in dialogue (Yang and Heeman, 2010), it provides a conceptually simple model for the definition of features for aggregate statistical analysis.

4.2 Context-independent feature analysis

We begin by analyzing a set of shallow features which we describe as *context-independent*, as they apply to user speech segments independently of what the system has recently said. Most of these are features that apply to many or all user speech segments. We describe our context-independent features in Section 4.2.1, and present our results for these features in Section 4.2.2.

4.2.1 Context-independent features

We summarize our context-independent features in Figure 2.

Speaking rate and onset times Based on previous clinical observations related to slowed speech and increased onset time for depressed individuals (Section 2), we defined features for speaking rate and onset time of user speech segments.

We quantify the speaking rate of a user speech segment $\langle s, e, t\rangle$, where $t = \langle w_1, ..., w_N\rangle$, as $N/(e - s)$. Feature (a) is the mean value of this feature across all user speech segments within each dialogue.

Onset time is calculated using the notion of user turns. For each user turn, we extracted the first user speech segment in the turn $f_{\mathrm{u}} = \langle s_{\mathrm{u}}, e_{\mathrm{u}}, t_{\mathrm{u}}\rangle$, and the most recent system speech segment $l_{\mathrm{s}} = \langle s_{\mathrm{s}}, e_{\mathrm{s}}, t_{\mathrm{s}}\rangle$. We define the onset time of such a first user segment as $s_{\mathrm{u}} - e_{\mathrm{s}}$, and for each dialogue, feature (b) is the intra-dialogue mean of these onset times.

In order to also quantify pause length between user speech segments within a turn, we define feature (c), a similar feature that measures the mean onset time between non-first user speech segments within a user turn in relation to the preceding user speech segment.

Length of user segments As one way to quantify the amount of speech, feature (d) reports the mean length of all user speech segments within a dialogue (measured in words).

Valence features for user speech Features (e)-(g) are meant to explore the idea that distressed users might use more negative or less positive vocabulary than non-distressed subjects. As an exploratory approach to this topic, we used SentiWordNet 3.0 (Baccianella and Sebastiani, 2010), a lexical sentiment dictionary, to assign valence to individual words spoken by users in our study. The dictionary contains approximately 117,000 entries. In general, each word w may appear in multiple entries, corresponding to different parts of speech and word senses. To assign a single valence score $v(w)$ to each word in the dictionary, in our features we compute the average score across all parts of speech and word senses:

$$v(w) = \frac{\sum_{e \in E(w)} \mathrm{PosScore}_{\mathrm{e}}(w) - \mathrm{NegScore}_{e}(w)}{|E(w)|}$$

where $E(w)$ is the set of entries for the word w, $\mathrm{PosScore}_{\mathrm{e}}(w)$ is the positive score for w in entry e, and $\mathrm{NegScore}_{e}(w)$ is the negative score for w in entry e. This is similar to the "averaging across senses" method described in Taboada et al. (2011).

We use several different measures of the valence of each speech segment with transcript $t = \langle w_1, ..., w_n\rangle$. We compute the min, mean, and max valence of each transcript:

$$\begin{aligned} \text{minimum valence of } t &= \min_{w_i \in t} v(w_i) \\ \text{mean valence of } t &= \tfrac{1}{n} \textstyle\sum_{w_i \in t} v(w_i) \\ \text{maximum valence of } t &= \max_{w_i \in t} v(w_i) \end{aligned}$$

Features (e)-(f) then are intra-dialogue mean

values for these three segment-level valence measures.

Filled pauses Another feature that we explored is the presence of filled pauses in user speech segments. To do so, we counted the number of times any of the tokens *uh*, *um*, *uhh*, *umm*, *mm*, or *mmm* appeared in each speech segment. For each dialogue, feature (h) is the mean number of these tokens per user speech segment. In order to account for the varying length of speech segments, we also normalize the raw token counts in each segment by dividing them by the length of the segment, to produce a *filled pause rate* for the segment. Feature (i) is the mean value of the filled pause rate for all speech segments in the dialogue.

Dialogue level features We also included two dialogue level measures of how "talkative" the user is. Feature (j) is the total number of user speech segments throughout the dialogue. Feature (k) is the total length (in words) of all speech segments throughout the dialogue.

Standard deviation features For the classification experiments reported in Section 5, we also included a standard deviation variant of each of the features (a)-(i) in Figure 2. These variants are defined as the intra-dialogue standard deviation of the underlying value, rather than the mean. We discuss examples of standard deviation features further in Section 5.

4.2.2 Results for context-independent features

We summarize the observed significant effects in our Wizard-of-Oz corpus in Table 1.

Onset time We report our findings for individuals with and without depression and PTSD for feature (b) in Table 1 and in Figure 3. The units are seconds. While an increase in onset time for individuals with depression has previously been observed in human-human interaction (Cohn et al., 2009; Hall et al., 1995), here we show that this effect transfers to interactions between individuals with depression and virtual humans. We find that mean onset time is significantly increased for individuals with depression in their interactions with our virtual human Ellie ($p = 0.018$, Wilcoxon rank sum test).

Additionally, while to our knowledge onset time for individuals with PTSD has not been reported, we also found a significant increase in onset time for individuals with PTSD ($p = 0.019$, Wilcoxon rank sum test).

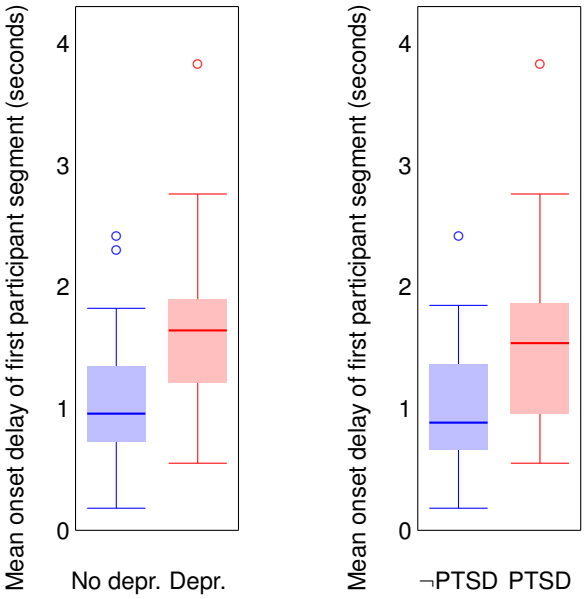

Figure 3: Onset time.

Filled pauses We report our findings for individuals with and without depression and PTSD under feature (h) in Table 1 and in Figure 4. We observed a significant reduction in this feature for both individuals with depression ($p = 0.012$, Wilcoxon rank sum test) and PTSD ($p = 0.014$, Wilcoxon rank sum test). We believe this may be related to the trend we observed toward shorter speech segments from distressed individuals (though this trend did not reach significance). There is a positive correlation, $\rho = 0.55$ ($p = 0.0001$), between mean segment length (d) and mean number of filled pauses in segments (h).

Other features We did not observe significant differences in the values of the other context-independent features in Figure 2.

4.3 Context-dependent features

Our data set allows us to zoom in and look at specific contextual moments in the dialogues, and observe how users respond to specific Ellie questions. As an example, one of Ellie's utterances, which has system ID happy_lasttime, is:

> happy_lasttime = *Tell me about the last time you felt really happy.*

In our data set of 43 dialogues, this question was asked in 42 dialogues, including 12 users positive for depression and 19 users positive for PTSD.

Feature	Depression (13 yes, 30 no)	PTSD (20 yes, 23 no)
(b) *mean onset time of first segment in each user turn*	Depr.: 1.72 (0.89) ↑ No Depr.: 1.08 (0.56) $p =$ 0.018	PTSD.: 1.56 (0.80) ↑ No PTSD.: 1.03 (0.57) $p =$ 0.019
(h) *mean number of filled pauses in user segments*	Depr.: 0.32 (0.19) ↓ No Depr.: 0.48 (0.23) $p =$ 0.012	PTSD: 0.36 (0.24) ↓ No PTSD: 0.49 (0.21) $p =$ 0.014

Table 1: Results for context-independent features. For each feature and condition, we provide the mean (standard deviation) for individuals with and without the condition. P-values for individual Wilcoxon rank sum tests are provided. An up arrow (↑) indicates a significant trend toward increased feature values for positive individuals. A down arrow (↓) indicates a significant trend toward decreased feature values for positive individuals.

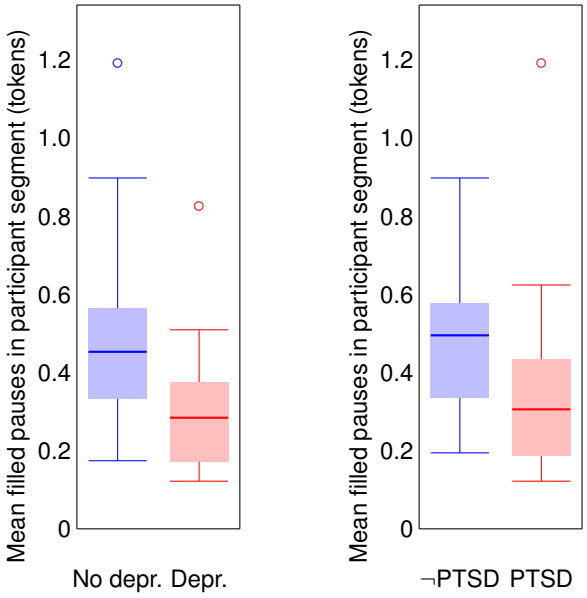

Figure 4: Number of filled pauses per speech segment.

This question is one of 95 *topic setting utterances* in Ellie's repertoire. (Ellie has additional utterances that serve as continuation prompts, backchannels, and empathy responses, which can be used as a topic is discussed.)

To define context-dependent features, we associate with each user segment the most recent of Ellie's topic setting utterances that has occurred in the dialogue. We then focus our analysis on those user segments and turns that follow specific topic setting utterances. In Table 2, we present some examples of our findings for context-dependent features for happy_lasttime.[3]

The contextual feature labeled (g′) in Table 2 is the mean of the maximum valence feature across all segments for which happy_lasttime is the most recent topic setting system utterance. We provide a full example of this feature calculation in Figure 5 in the appendix.

As the figure shows, we find that users with both PTSD and depression show a sharp reduction in the mean maximum valence in their speech segments that respond to this question. This suggests that in these virtual human interactions, this question plays a useful role in eliciting differential responses from subjects with these psychological disorders. We observed three additional questions which showed a significant difference in the mean maximum valence feature. One example is the question, *How would your best friend describe you?*.

With feature (b′) in Table 2, we find an increased onset time in responses to this question for subjects with depression.[4] Feature (d′) shows that subjects with PTSD exhibit shorter speech segments in their responses to this question.

We observed a range of findings of this sort for various combinations of Ellie's topic setting utterances and specific context-dependent features. In future work, we would like to study the optimal combinations of context-dependent questions that yield the most information about the user's distress status.

[3]While we provide significance test results here at the $p < 0.05$ level, it should be noted that because of the large number of context-dependent features that may be defined in a small corpus such as ours, we report these merely as observations in our corpus. We do not claim that these results transfer beyond this data set.

[4]In comparing Table 2 with Table 1, this question seems to induce a higher mean onset time for distressed users than the average system utterance does. This does not seem to be the case for non-distressed users.

Feature	Depression (12 yes, 30 no)		PTSD (19 yes, 23 no)	
(g′) *mean maximum valence in user segments following happy_lasttime*	↓	Depr.: 0.15 (0.07) No Depr.: 0.26 (0.12) $p =$ 0.003	↓	PTSD: 0.16 (0.08) No PTSD: 0.28 (0.11) $p =$ 0.0003
(b′) *mean onset time of first segments in user turns following happy_lasttime*	↑	Depr.: 2.64 (2.70) No Depr.: 0.94 (1.80) $p =$ 0.030	n.s.	PTSD: 2.18 (2.48) No PTSD: 0.80 (1.76) $p =$ 0.077
(d′) *mean length of user segments following happy_lasttime*	n.s.	Depr.: 5.95 (1.80) No Depr.: 10.03 (6.99) $p =$ 0.077	↓	PTSD: 6.82 (5.12) No PTSD: 10.55 (6.68) $p =$ 0.012

Table 2: Example results for context-dependent features. For each feature and condition, we provide the mean (standard deviation) for individuals with and without the condition. P-values for individual Wilcoxon rank sum tests are provided. An up arrow (↑) indicates a significant trend toward increased feature values for positive individuals. A down arrow (↓) indicates a significant trend toward decreased feature values for positive individuals.

5 Classification Results

In this section, we present some suggestive classification results for our data set. We construct three binary classifiers that use the kinds of features described in Section 4 to predict the presence of three conditions: PTSD, depression, and distress. For the third condition, we define distress to be present if and only if PTSD, depression, or both are present. Such a notion of distress that collapses distinctions between disorders may be the most appropriate type of classification for a potential application in which distressed users of any type are prioritized for access to health care professionals (who will make a more informed assessment of specific conditions).

For each individual dialogue, each of the three classifiers emits a single binary label. We train and evaluate the classifiers in a 10-fold cross-validation using Weka (Hall et al., 2009).

While our data set of 43 dialogues is perhaps of a typical size for a study of a research prototype dialogue system, it remains very small from a machine learning perspective. We report here two kinds of results that help provide perspective on the prospects for classification of these conditions. The first kind looks at classification based on all the context-independent features described in Section 4.2.1. The second looks at classification based on individual features from this set.

In the first set of experiments, we trained a Naïve Bayes classifier for each condition using all the context-independent features. We present our results in Table 3, comparing each classifier to a baseline that always predicts the majority class (i.e. no condition for PTSD, no condition for depression, and with condition for distress).

We note first that the trained classifiers all outperform the baseline in terms of weighted F-score, weighted precision, weighted recall, and accuracy. The accuracy improvement over baseline is substantial for PTSD (20.9% absolute improvement) and distress (23.2% absolute improvement). The accuracy improvement over baseline is more modest for depression (2.3% absolute). We believe one factor in the relative difficulty of classifying depression more accurately is the relatively small number of depressed participants in our study (13).

While it has been shown in prior work (Cohn et al., 2009) that depression can be classified above baseline performance using features observed in clinical human-human interactions, here we have shown that classification above baseline performance is possible in interactions between human participants and a virtual human dialogue system. Further, we have shown classification results for PTSD and distress as well as depression.

We tried incorporating context-dependent features, and also unigram features, but found that neither improved performance. We believe our data set is too small for effective training with these very large extended feature sets.

Disorder	Model	Weighted F-score	Weighted Precision	Weighted Recall	Accuracy
PTSD	Naïve Bayes	0.738	0.754	0.744	74.4%
	Majority Baseline	0.373	0.286	0.535	53.5%
Depression	Naïve Bayes	0.721	0.721	0.721	72.1%
	Majority Baseline	0.574	0.487	0.698	69.8%
Distress	Naïve Bayes	0.743	0.750	0.744	74.4%
	Majority Baseline	0.347	0.262	0.512	51.2%

Table 3: Classification results.

In our second set of experiments, we sought to gain understanding of which features were providing the greatest value to classification performance. We therefore retrained Naïve Bayes classifiers using only one feature at a time. We summarize here some of the highest performing features.

For depression, we found that the feature *standard deviation in onset time of first segment in each user turn* yielded very strong performance by itself. In our corpus, we observed that depressed individuals show a greater standard deviation in the onset time of their responses to Ellie's questions ($p = 0.024$, Wilcoxon rank sum test). The value of this feature in classification complements the clinical finding that depressed individuals show greater onset times in their responses to interview questions (Cohn et al., 2009).

For distress, we found that the feature *mean maximum valence in user segments* was the most valuable. We discussed findings for a context-dependent version of this feature in Section 4.3. This finding for distress can be related to previous observations that individuals with depression use more negatively valenced words (Rude et al., 2004).

For PTSD, we found that the feature *mean number of filled pauses in user segments* was among the most informative.

Based on our observation of classification performance using individual features, we believe there remains much room for improvement in feature selection and training. A larger data set would enable feature selection approaches that use held out data, and would likely result in both increased performance and deeper insights into the most valuable combination of features for classification.

6 Conclusion

In this paper, we have explored the presence of indicators of psychological distress in the linguistic behavior of subjects in a corpus of semi-structured virtual human interviews. In our data set, we have identified several significant differences between subjects with depression and PTSD when compared to non-distressed subjects. Drawing on these features, we have presented statistical classification results that suggest the potential for automatic assessment of psychological distress in individual interactions with a virtual human dialogue system.

Acknowledgments

This work is supported by DARPA under contract (W911NF-04-D-0005) and the U.S. Army Research, Development, and Engineering Command. The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

Andrea Esuli Stefano Baccianella and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Jeffery F. Cohn, Tomas Simon Kruez, Iain Matthews, Ying Yang, Minh Hoai Nguyen, Margara Tejera Padilla, Feng Zhou, and Fernando De la Torre. 2009. Detecting depression from facial actions and vocal prosody. In *Affective Computing and Intelligent Interaction (ACII)*, September.

Judith A. Hall, Jinni A. Harrigan, and Robert Rosenthal. 1995. Nonverbal behavior in clinician-patient interaction. *Applied and Preventive Psychology*, 4(1):21 – 37.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November.

Peter A Heeman, Rebecca Lunsford, Ethan Selfridge, Lois Black, and Jan Van Santen. 2010. Autism and

interactional aspects of dialogue. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 249–252. Association for Computational Linguistics.

Kai Hong, Christian G. Kohler, Mary E. March, Amber A. Parker, and Ani Nenkova. 2012. Lexical differences in autobiographical narratives from schizophrenic patients and healthy controls. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 37–47, Jeju Island, Korea, July. Association for Computational Linguistics.

Christine Howes, Matthew Purver, Rose McCabe, Patrick G. T. Healey, and Mary Lavelle. 2012. Predicting adherence to treatment for schizophrenia from dialogue transcripts. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 79–83, Seoul, South Korea, July. Association for Computational Linguistics.

Shannon J Johnson, Michelle D Sherman, Jeanne S Hoffman, Larry C James, Patti L Johnson, John E Lochman, Thomas N Magee, David Riggs, Jessica Henderson Daniel, Ronald S Palomares, et al. 2007. *The psychological needs of US military service members and their families: A preliminary report*. American Psychological Association Presidential Task Force on Military Deployment Services for Youth, Families and Service Members.

Kurt Kroenke, Robert L. Spitzer, and Janet B. W. Williams. 2001. The phq-9. *Journal of General Internal Medicine*, 16(9):606–613.

Maider Lehr, Emily Prud'hommeaux, Izhak Shafran, and Brian Roark. 2012. Fully automated neuropsychological assessment for detecting mild cognitive impairment. In *Interspeech 2012: 13th Annual Conference of the International Speech Communication Association*, Portland, Oregon, September.

Emily Prud'hommeaux and Brian Roark. 2011. Extraction of narrative recall patterns for neuropsychological assessment. In *Interspeech 2011: 12th Annual Conference of the International Speech Communication Association*, pages 3021–3024, Florence, Italy, August.

Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133.

S. Scherer, G. Stratou, M. Mahmoud, J. Boberg, J. Gratch, A. Rizzo, and L.-P. Morency. 2013. Automatic behavior descriptors for psychological disorder analysis. In *IEEE Conference on Automatic Face and Gesture Recognition*.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307, June.

Marcus Thiebaux, Stacy Marsella, Andrew N. Marshall, and Marcelo Kallmann. 2008. Smartbody: behavior realization for embodied conversational agents. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems - Volume 1*, AAMAS '08, pages 151–158, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. Elan: a professional framework for multimodality research. In *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation*.

Fan Yang and Peter A. Heeman. 2010. Initiative conflicts in task-oriented dialogue. *Computer Speech & Language*, 24(2):175 – 189.

Appendix A. Wizard-of-Oz Dialogue Excerpts

Example user with PTSD and depression

	max valence	transcript
Ellie		(happy_lasttime) tell me about the last time you felt really happy
User	0.014	when i play with my dog
Ellie		when was the last time that happened?
User	0.000	this morning
Ellie		aw
Ellie		tell me more about that
User	0.312	i don't know i just
User	0.010	whenever i get
User	0.312	super bored or i just need some time to
User	0.312	for myself i just play with the dog
Ellie		that's good
User	0.000	yeah
		0.1371 = mean maximum valence in user segments following happy_lasttime
Ellie		(ideal_weekendC) tell me how you spend your ideal weekend

Example non-distressed user

	max valence	transcript
Ellie		(happy_lasttime) tell me about the last time you felt really happy
User	0.562	um the last time i felt really happy was
User	0.000	hm
User	0.000	today
Ellie		tell me more about that
User	0.688	uh just from the moment i woke up it was a beautiful sunny day
User	-0.062	i
User	0.565	went to see some friends we had a good time went to school
User	0.565	had some good grades on some papers um wrote a good essay
User	0.292	feel pretty accomplished and
User	0.312	i feel like my day is just
User	0.565	a good day
Ellie		that's so good to hear
		0.3487 = mean maximum valence in user segments following happy_lasttime
Ellie		(BF_describe) how would your best friend describe you?

Figure 5: Examples of maximum valence feature.