A Multimodal Adaptive Dialogue Manager for Depressive and Anxiety Disorder Screening: A Wizard-of-Oz Experiment

Konstantinos Tsiakas
HERACLEIA Lab
Computer Science and Engineering Department
University of Texas, Arlington
konstantinos.tsiakas@mavs.uta.edu

Lynette Watts
Psychology Department
University of Texas, Arlington
lynette.watts@mavs.uta.edu

Cyril Lutterodt
HERACLEIA Lab
Computer Science and Engineering Department
University of Texas, Arlington
cyril.lutterodt@mavs.uta.edu

Theodoros Giannakopoulos
National Center for Scientific Research DEMOKRITOS, Athens, Greece
tyiannak@gmail.com

Alexandros Papangelis
Articulab, Computer Science Department,
Carnegie Mellon University, Pittsburgh
apapa@andrew.cmu.edu

Robert Gatchel
Psychology Department
University of Texas, Arlington
gatchel@uta.edu

Vangelis Karkaletsis
National Center for Scientific Research DEMOKRITOS, Athens, Greece
vangelis@iit.demokrtitos.gr

Fillia Makedon
HERACLEIA Lab
Computer Science and Engineering Department
University of Texas, Arlington
makedon@uta.edu

ABSTRACT
In this paper, we present an Adaptive Multimodal Dialogue System for Depressive and Anxiety Disorders Screening (DADS). The system interacts with the user through verbal and non-verbal communication to elicit the information needed to make referrals and recommendations for depressive and anxiety disorders while encouraging the user and keeping them calm. We designed the problem using interconnected Markov Decision Processes using sub-goals to deal with the large state space. We present the problem formulation and the experimental procedure for the training data collection and the system training following the methodology of Wizard-of-Oz experiments.

Categories and Subject Descriptors
I.2.6. [**Artificial Intelligence**]: Learning

Keywords
Multimodal Adaptive Dialogue Systems, Markov Decision Processes, Reinforcement Learning, Mental Health Monitoring, Wizard-of-Oz

1. INTRODUCTION
A traumatic event, such as abuse, combat, an assault, an accident or a natural disaster, may have a long-lasting negative effect on an individual. With the increase of soldiers in combat since 2001, the interest in Post-Traumatic Stress Disorder (PTSD) has increased. "Epidemiologic surveys indicate that the vast majority of individuals with PTSD meet criteria for at least one other psychiatric disorder.... The most common comorbid diagnoses are depressive disorders, substance use disorders, and other anxiety disorders." [1] Moreover, the World Health Organization reports that mental and behavioral disorders were the number one category contributing to U.S. YLDs (years living with disability) in 2010. At 27.1%, this is more than diabetes (8.4%), chronic respiratory diseases (7.9%), and cardiovascular diseases (5.2%) combined. Within the category, Major Depressive Disorder was number one contributing 30.66%, followed by All Anxiety Disorders (18.76%), Drug Use Disorders (13.03%), and Alcohol Use Disorders (8.40%). The National Institute of Mental Health estimates total direct and indirect costs of serious mental illness exceeds $300 billion in the U.S. annually based on 2002 data. More than 60% is the indirect cost of lost earnings from lost productivity. Healthcare expenditures account for about 30% with disability benefits accounting for less than 10% [2]. Given the prevalence and comorbidity of depressive and anxiety disorders with the associated personal and societal costs, there is need for self-screening tools to provide referrals for relevant treatment resources.

2. RELATED WORK
Although many works have proposed multimodal interaction with the user, most of the systems are rule-based or plan-based and they use speech as the primary modality. Moreover, a few works have proposed stochastic dialogue policy optimization in the health domain. In [3], they proposed an Adaptive Dialogue System able to have a conversation in natural language with PTSD-suffering users, and guide the way that allows eliciting information about their disorder and progress of their treatment. They consider speech as the only input, while they continuously monitor the user's emotional state through keywords in order to adapt the dialogue in such way that the system encourages the user and keeps them calm. A similar system that uses multimodal input and output, SimCoach, was designed to provide support and health care information about PTSD following the Information State Update (ISU) approach [4]. Although, ISU and plan-based approaches seem to be effective for this kind of systems, they have a number of general limitations concerning the design and

PETRA '15, July 01-03, 2015, Corfu, Greece
 ISBN 978-1-4503-3452-5/15/07…$15.00
DOI: http://dx.doi.org/10.1145/2769493.2769572

implementation. These approaches require a manual specification for the update or inference rules. Moreover, the system behavior remains static during the interaction without taking into consideration the current user's personal preferences and needs. In [5], they model an Alcohol Brief Intervention system as interconnected Markov Decision Processes (MDP), following a model-based approach.

In the current paper, we present our ongoing work on a multimodal adaptive dialogue system used as a self-assessment tool for depressive and anxiety disorders. The system follows a questionnaire-form dialogue to determine which recommendation to provide the user after the interaction. Moreover, audiovisual data, such as speech and facial expressions, are taken into consideration to estimate the user's emotional state and prevent unwanted emotional states during the interaction by encouraging the user when needed. We focus on the dialogue manager of the system, which is responsible for the decision making of the system. We model the interaction splitting the dialogue into sub dialogues, represented by interconnected MDP in order to screen for the often comorbid disorders of PTSD, General Anxiety Disorder (GAD), Depression, and Substance Use Disorders.

In Section 3, we present the system architecture showing the different levels of screening and how we formulate the problem using interconnected MDP to represent each part of the dialogue. In Section 4, we show the experimental procedure for the data collection and the system training following the methodology of the Wizard-of-Oz (WoZ) experiments, and describe the experimental setup and the modalities used for the experiments. Finally, in Section 5, we present the future work which includes the system training and evaluation.

3. SYSTEM DESCRIPTION

The proposed system interacts with the user using verbal and non-verbal communication to elicit the required information for making appropriate recommendations for each user. During the interaction, the system perceives multimodal input, such as speech (textual and audio information) and facial expressions. The system keeps track of the questionnaire-based score based on the level of screening, as explained in the next section. Moreover, audiovisual emotion recognition is used to estimate the user's emotional state and encourage the user when needed.

3.1 Modeling the Dialogue as MDP

As mentioned before, we formulated the interaction using MDP. An MDP is described by a tuple $\langle S, A, T, R \rangle$ where:

- ***S*** is a finite set of states
- ***A*** is a finite set of actions
- ***T*** is the transition model where *T(s, a, s')* denotes the probability of moving from state *s* to state *s'* by performing action *a*.
- ***R(s, a)*** is a reward function that gives a numerical reward of going to state *s* performing action *a*.

The state space includes the audio and visual information for estimating the user's emotional state and the questionnaire-based scores for each level of screening. At each state, the system can ask a question that is disorder-related (anxiety, depression, substance use, etc.) or encourage the user, based on the user emotional state. The goal of the system is to maximize the average cumulative discounted reward during each interaction. In our system, we have divided the main dialogue into different sub-dialogues that represent a different level of screening. Each sub-dialogue is formulated as a separate MDP with a specific tuple of attributes.

3.2 Levels of Screening

Multi-level screening was chosen to keep the number of questions to a minimum by progressively asking more detailed questions only when indicated as necessary. The American Psychiatric Association (APA) offers all five of the assessments used as "emerging measures" for research and clinical use in the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* (DSM-5) [6]. These adult self-rated measures are intended for an initial interview and to monitor progress. They were developed "to enhance clinical decision-making and not as the sole basis for making a clinical diagnosis." Scores are used to select disorder-specific resources that the user may find helpful for the recommendations. All of the assessments use a 5-point Likert scale, where "1" is never and "5" is always, in order to indicate how much the user has been bothered by the problem during the specific time period.

3.2.1 Level 1: Cross-Cutting Symptom Screening

Initial screening is based on the APA's *DSM-5 Self-Rated Level 1 Cross-Cutting Symptom Measure—Adult*. Up to a total of ten general examples of depressive symptoms, anxiety symptoms, and substance use are presented. A score greater than "1" (never), for any depressive or anxiety symptom, triggers level 2 intermediate screening for the respective disorder. A score greater than "1" (never) for substance use triggers specific recommendations with no further screening for these.

3.2.2 Level 2: Disorder-Specific Screening

Intermediate depression screening is based on the PROMIS Health Organization's (PHO) *LEVEL 2—Depression—Adult (PROMIS Emotional Distress—Depression— Short Form)*. Eight specific examples of depressive symptoms are presented. Based on this score, recommendations are made with no further screening for depressive disorders.

Intermediate anxiety screening is based on the PHO's *LEVEL 2—Anxiety—Adult (PROMIS Emotional Distress—Anxiety— Short Form)*. Seven specific examples of anxiety symptoms are presented. Based on this score, either the severity screening is presented or recommendations are made with no further screening for anxiety disorders.

3.2.3 Level 3: Anxiety Severity Screening

Anxiety severity screening is based on the APA's *Severity Measure for Generalized Anxiety Disorder—Adult* and *Severity of Posttraumatic Stress Symptoms—Adult (National Stressful Events Survey PTSD Short Scale [NSESSS])*. Ten specific examples of GAD symptoms are presented followed by ten specific examples of PTSD symptoms. Based on the scores, recommendations are made with no further screening.

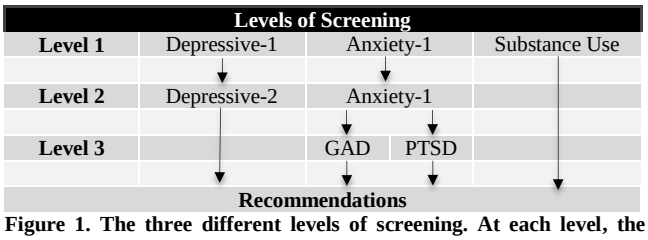

Figure 1. The three different levels of screening. At each level, the system collects the required information needed to calculate a level score. Based on this score, it moves to the appropriate next level. At the end, the system delivers a recommendation and online resources. Each level is represented by one or more MDPs.

3.3 Problem Formulation

Each level of screening is represented in the dialogue as a separate Markov Decision Process with its own state and action space and their specific goals, as described in Section 3.1. In this way, the dialogue branches according to user's input on each screening level, resulting to a reduced state-action space.

We follow a model-based approach for the system implementation and we apply reinforcement learning for the system training. Since the system receives multimodal input, it needs to learn a model of the transitions between different states and actions during the interaction. In the next section, we present the WoZ methodology for our system.

4. WIZARD OF OZ

Model-based approaches require a model that simulates the dynamics of the interaction in order to compute an approximate value of taking an action in a particular state. In this work, we follow the Wizard-of-Oz methodology, in order to collect data to learn the transition model and apply reinforcement learning to train the decision-making system.

WoZ studies are performed in order to simulate the human computer interaction for system evaluation, data collection, and design improvement. A 'wizard' is a hidden human operator that simulates some aspects of the system, where the subjects are led to believe that they interact with a real system. In our case, we follow a semi-manual approach, since the system automatically processes the multimodal input and the wizard decides the next system action.

For our WoZ experiment, the system perceives the multimodal input combined with the user responses on the questionnaire in order to formulate the current state. Based on the current state, the wizard performs the decision making and selects an appropriate action. In this way, we record the interaction data in the form of state-action-state in order to estimate the transition model by applying Maximum Likelihood Estimation (MLE) given the relative frequency of occurrence of each transition.

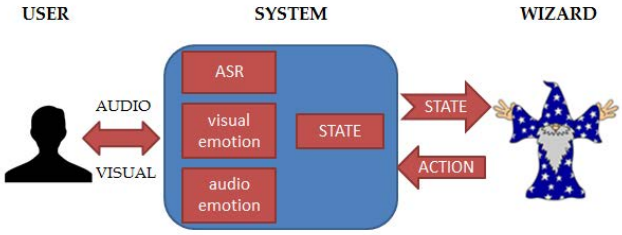

Figure 2. The WoZ experiment architecture. The system interacts with the user and estimates the user's current state based on the multimodal input and the user's responses. Based on this state, the wizard performs the decision making and the interaction continues with the next estimated state.

4.1 Experimental Description

In order to conduct the WoZ experiments, we have implemented a prototype version of the system that collects multimodal data and formulates the current state of the current MDP. Indicatively, the state space of the first level MDP includes the audio emotion, the visual emotion, the scores of the questionnaires for each disorder category, and the goal variable used for the transition to the next MDP. The system formulates the state automatically (ASR, audiovisual emotion detection) and the wizard (user), taking into consideration the estimated state, decides which should be the next action. In contrast with other WoZ approaches, we do not formulate the state manually based on what the wizard hears or sees, but automatically using each modality's recognizer.

4.1.1 Experimental Setup

For the system implementation and the experimental setup we use the Robot Operating System (ROS), which was designed for robots and human computer interaction systems. The ROS framework is a graph architecture where processes are nodes, which publish or subscribe to topics (types of messages) produced by other nodes. Through ROS, we implemented the system to set up the WoZ experiment. The hardware for the experiment is an Asus Xtion Pro sensor for the visual input and an ATR2100-USB microphone for the audio input. The system was implemented in Python. For the face detection, we use the opencv2 library and for speech recognition, we use the Pocketsphinx package for Python. We modified the recognizer in order to capture the user's voice when the recognizer detects speech.

Each modality is captured and processed using different ROS nodes. Each node publishes specific messages for each modality. Then, the state node subscribes to these topics to estimate and publish the current state. The Dialog Manager node subscribes to the state topic and prompts the wizard to select the action. Then, the action is used to play the corresponding pre-recorded question that the user will answer. During the interaction, we keep track of the recorded state-action-state sequences to estimate the transition model for the system training.

In order to make the system more natural and appealing, we created a therapist female avatar in Gazebo. Gazebo is a simulation environment for robots. We use it as the visual simulator to give visual feedback to the user. The avatar was rigged and then made into a URDF model. The output from the dialogue manager is sent to the avatar in Gazebo, which then simulates the text as speech for the user to receive.

4.1.2 Audio Emotion Recognition

Besides extracting information regarding events and language content, a substantial research effort of several audio characterization methodologies and focused on recognizing the affective content of the input signal i.e., the emotions that underlie the audio (and /or visual) information [7, 8].

The most widely-used approach to affective audio content recognition is to apply well-known classifiers (e.g. HMMs, SVMs, etc.) for classifying signals into an *a-priori* known number of predefined distinct categories of emotions (e.g. fear, anger, etc.). A drawback of these techniques is that the emotions of the audio content cannot always be easily classified in distinct categories as the level of categorical taxonomy of emotion is subjective. An alternative way to analyze emotion is the dimensional approach, according to which affective content is represented using specific dimensions that stem from psychophysiology. The most widely adopted dimensional model for affective characterization is that of Valence and Arousal [9, 10].

In this work, we first extract a wide range of audio features in a short-term basis, both from the time and frequency domain: signal energy, entropy of energy, zero crossing rate, spectral centroid, spectral flux, Mel Frequency Cepstral Coefficients, Chroma-based features, etc. [11]. The total number of short-term features is 34, (i.e., each short-term frame is represented by a 34-dimensional feature vector). As a second step, two mid-term statistics are extracted per speech segment, namely the average value and the standard deviation. This mid-term statistic extraction process results in 2x34 = 68 feature statistics per speech segment. In order

to estimate the Valence – Arousal values for a speech segment (represented by 68 feature statistics as explained above), we have adopted the Support Vector Machine regression technique [12, 13]. In particular, one SVM regression model is trained for each dimension (Valence and Arousal).

An annotated dataset of 90 speech segments, recorded in the context of the dialog system has been compiled in order to train the regression models. We collected data from 7 participants reading scripted dialogue sentences with duration of 2-3 seconds. The script provides participants a symptom profile and the 5-point Likert score for each symptom's severity. In addition, a cross-validation procedure has been carried out in order to compute the Mean Square Error (MSE). This experimental procedure indicated that the MSE for the Arousal dimension is 0.18, while for the Valence dimension the error is equal to 0.26. Note that the baseline MSE (i.e., the MSE when the estimation is always equal to the average estimated value of the training dataset) is 0.25 for Arousal and 0.43 for Valence, meaning that the estimation of Valence is a more difficult task (which is rather obvious).

4.1.3 Visual Emotion Recognition

For the facial expression classification, we used a Python wrapper for Indico [14], which uses an implemented predictive model for facial expression regression. Given a face image, it returns a likelihood score for each of the six basic emotions (Angry, Sad, Neutral, Surprise, Fear and Happy). In future implementation, we plan to train our own facial expression classifier, using Local Binary Patterns (LBP) and Histogram of Oriented Gradients (HOG) features for the feature extraction and SVM for the classification.

5. REMARKS AND FUTURE WORK

After each interaction, we give a questionnaire to each participant to rate the interaction with the system and describe their experience. We will use this input in order to get a better insight about user behavior and user preferences. Moreover, upon agreement with each user, we collect the audiovisual interaction data in order to develop a language model for the system and a more efficient facial expression classifier. The next step is to train the algorithm using Reinforcement Learning and conduct a second round of WoZ experiments with psychology experts to evaluate the decision making performance and efficiency. In a future implementation, we plan to modify the dialogue form using natural language interaction instead of a questionnaire-based dialogue. Moreover, we plan to make the avatar an affective agent able to simulate emotions for a more natural and human-alike interaction.

6. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grants No. NSF-CNS 1035913, NSF-IIS 1409897, NSF-CNS 1338118. Moreover, this paper is partially supported by the project "Robots in assisted living environments: Unobtrusive, efficient, reliable and modular solutions for independent ageing – RADIO", which has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 643892. For more details, please see http://www.radio-project.eu.

7. REFERENCES

[1] Brady, K. T., Killeen, T. K., Brewerton, T., & Lucerini, S. (2000). Comorbidity of psychiatric disorders and posttraumatic stress disorder. *Journal Of Clinical Psychiatry*, *61*(Suppl7), 22-32.

[2] Statistics. *National Institute of Mental Health*. Retrieved March 2015 from http://www.nimh.nih.gov/health/statistics/index.shtml

[3] Papangelis, Alexandros, Fillia Makedon, and Robert Gatchel. "Assessing and Monitoring Post-Traumatic Stress Disorder through Natural Interaction with an Adaptive Dialogue System." Journal of Applied Biobehavioral Research 19.3 (2014): 192-215.

[4] Rizzo, Albert, et al. "SimCoach: An intelligent virtual human system for providing healthcare information and support." The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC). Vol. 2011. No. 1. National Training Systems Association, 2011.

[5] Yasavur, Ugan, Christine Lisetti, and Napthali Rishe. "Modeling brief alcohol intervention dialogue with MDPs for delivery by ECAs." Intelligent Virtual Agents. Springer Berlin Heidelberg, 2013.

[6] Online Assessment Measures. *American Psychiatric Association.* Retrieved November 2014 from http://www.psychiatry.org/practice/dsm/dsm5/online-assessment-measures

[7] Cowie, Roddy, et al. "Emotion recognition in human-computer interaction." Signal Processing Magazine, IEEE 18.1 (2001): 32-80.

[8] Wang, Yongjin, and Ling Guan. "Recognizing human emotional state from audiovisual signals*." Multimedia, IEEE Transactions on 10.5 (2008): 936-946.

[9] Hanjalic, Alan, and Li-Qun Xu. "Affective video content representation and modeling." Multimedia, IEEE Transactions on 7.1 (2005): 143-154.

[10] Giannakopoulos, Theodoros, Aggelos Pikrakis, and Sergios Theodoridis. "A dimensional approach to emotion recognition of speech from movies." Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on. IEEE, 2009.

[11] Theodoros Giannakopoulos, "Python Audio Analysis Library", GitHub repository: https://github.com/tyiannak/pyAudioAnalysis

[12] Vapnik, Vladimir, Steven E. Golowich, and Alex Smola. "Support vector method for function approximation, regression estimation, and signal processing." Advances in neural information processing systems (1997): 281-287.

[13] https://indico.io/

[14] Basak, Debasish, Srimanta Pal, and Dipak Chandra Patranabis. "Support vector regression." Neural Information Processing-Letters and Reviews 11.10 (2007): 203-224.