Designing Social Robots for Early Detection of Mental Heath Conditions

by

Amrita Krishnaraj

A thesis submitted to The Johns Hopkins University
in conformity with the requirements for the degree of
Master of Science in Engineering

Baltimore, Maryland
May, 2019

Abstract

Globally, mental health is a growing socio-economic burden and leads to negative ramifications including mortality and poor quality of life. Successful early detection of mental illness will make a significant, positive economic and societal impact. Social robots show potential to be integrated as tools for psychological therapy and early detection. This thesis seeks to design and develop social robots for early detection of mental illness.

I explore how multi-modal inputs can be used to infer user's mental state and to direct appropriate robot behaviour. I have employed an iterative design process for the design of robot morphology, personality, and behaviour. Design 1 is a social robot with 6 DOF and exhibits non-verbal behaviours. In this design, I explore audio, video, and haptic inputs to detect user's emotional state. Design 2 is an interactive device that aims to collect audio data for the detection of early signs of depression. In this design, acoustic features are explored for depression detection, and the device uses audio and LEDs to communicate it's internal state. Finally, I have conducted a pilot experiment to investigate how the users interact with the robot. This thesis informs the design of future robots that aim to support early detection of mental illnesses.

Reader

Dr. Chien-Ming Huang
John C. Malone Assistant Professor
Department of Computer Science
Johns Hopkins Whiting School of Engineering

Acknowledgments

I would like to thank all the professors with whom I have had the pleasure of interacting throughout my time at JHU, particularly those in Laboratory for Computational Science and Robotics. In particular, I am grateful for the guidance provided by Dr. Chien-Ming Huang throughout my graduate study and research. I appreciate the countless hours he has spent with me answering all my questions. I have learnt a lot working with him.

I would also like to extend my appreciation to my lab mates for supporting me throughout the research. In particular, I would like to thank Erica Huang for her help with the sketches and the character design of the robot, Brandon Lax for his work on the audio recording framework for Melo, and Candy for her contribution to digitization of depression scale.

Finally, I would like to thank all my friends and family, especially my parents and my brother who have always been a source of inspiration and support.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Mental health is a growing concern in both the developed and the developing countries. Around 1-in-6 people globally (15-20%) have one or more mental illnesses (Ritchie and Roser, 2018). Globally, this means over one billion people in 2016 experienced mental illness. In the United States, approximately 1 in 5 adults (46.6 million) experienced mental illness in 2017 (National Institute of Mental Health, 2017) and over one-third (37%) of students aged 14-21 suffer from a mental health condition (Education, 2013). Financial burden associated with mental illness is substantial and costs America approximately $193.2 billion per year (Insel, 2008). Mental illness includes many different conditions, such as Autism Spectrum Disorder, Schizophrenia, Dementia, Depression Disorder, and Anxiety Disorder, that vary in degree of severity and affect wide demographic populations. Individuals living with serious mental illness face an increased risk of having chronic medical conditions (Colton and Manderscheid, 2006). Recent research has reported that adults with serious mental illness die on average 10 years earlier than others with similar treatable medical conditions (Walker, McGee, and Druss, 2015). Mental

illness also leads to other complications including suicide (Isometsä, 2001), dropout from school (Education, 2013), violence towards other, indulgence in antisocial activities, and smoking (Lasser et al., 2000). Research has also shown prolonged hospitalization and delayed recovery due to negative psychological consequences throughout recovery.

Despite being critical to overall well-being and physical health, diagnoses and treatment of mental illnesses remain low. It is identified that only 40% of the affected population receive treatment (National Institute of Mental Health, 2019). Successful early identification of mental health conditions will make a significant, positive economic and societal impact. Emerging research indicates that intervening early can interrupt the negative course of some mental illnesses and may, in some cases, lessen long-term disability (American Mental Health Councillors Association, 2011).

With the advent of technology, researchers have explored a variety of technologies including mobile and computer technologies (Callan et al., 2017) and robots (Robinson, MacDonald, and Broadbent, 2014) for use in mental health care. The use of robot technology in mental health care is nascent, but represents a potentially useful tool in the professional's toolbox. Robots afford appealing characteristics such as embodiment, tangibility and interactivity that are conductive for psychological therapy (Deng, Mutlu, and Mataric, 2019). These features make them better suited for therapy compared to mobile and web-based interventions. Such robots that provide assistance to human users through social, rather than physical interaction are called Socially Assistive Robots or SARs (Matarić and Scassellati, 2016). Researchers have

investigated robots as mental health care therapy tools for Autism (Scassellati, Admoni, and Matarić, 2012), Dementia (Mordoch et al., 2013), Alzheimer's, Depression (Chen, Jones, and Moyle, 2018) and Distress (Trost et al., 2019). These studies often report increased engagement, increased levels of attention (Ricks and Colton, 2010) and novel social behaviors such as joint attention (Scassellati et al., 2018), spontaneous imitation when robots are part of the interaction (Scassellati, 2007), increased communication with other humans and improved sleep patterns (Tapus, Maja, and Scassellatti, 2007). Almost all of the SARs developed and used thus far focused on the intervention for mental illnesses. Little prior research has explored the use of SARs for early detection of mental illness.

The central aim of this thesis is to explore the design space of social robot for early detection of mental illnesses. To this end, I have undertaken iterative design of the robot morphology, features and behaviour inspired by (Arsand and Demiris, 2008). My design explored multi-modal inputs and interactive nonverbal behaviors. The first prototype is a social robot with 6 DOF capable of proactive and reactive non-verbal behaviours. In this prototype, I have explored audio, video, and haptic input channels to help the robot understand its environment and interpret the user's emotional state. To understand how the user will interact with the robot, I conducted a preliminary study guided by prior research on psychological therapy. Further, I have designed the second prototype, a device for collecting audio information for early detection of depression. In this design, I focus on the use of acoustic features for early detection of depression, and LEDs are used to communicate internal state of

the device. I hope that findings of this research would inform future research in improving people's quality of life in addition to providing companionship.

The rest of this thesis is organized as follows. Chapter 2 provides background knowledge on SARs, current state of the art in health care robotics and outlines the barriers that are central to the SARs in mental health care. Chapter 3 presents the two prototypes I developed for detection of mental states. Design 1 explores multi-modal information, robot embodiment and personality with an acute focus on haptics, a novel approach to detect emotions. Design 2 has been developed for collecting audio data for detecting depression and explores an audio based interactive design. In Chapter 4, I discuss findings found in and lessons learned from the study. I also discuss limitations of this research that motivates future research. Finally, this thesis is concluded with it's contributions in Chapter 5.

Chapter 2

Background

2.1 Socially Assistive Robots

There is no formal definition of assistive robotics (Bemelmans et al., 2012), but Feil-Seifer and Matarić (Feil-Seifer and Matarić, 2005) describe socially assistive robots as the meeting point of assistive robots and socially interactive robots, and further stated that this kind of robots have the purpose of aiding humans by emphasizing the importance of social interactions. Assistive robots can be broadly categorized into two groups. First, rehabilitation robots; these robots focus on physical assistive technology features and are principally not communicative, such as smart wheelchairs (Gomi and Griffith, 1998) and exoskeletons (Kazerooni, 2005). Second, assistive social robots are subgrouped into service and companion robots. Service robots are used to support basic tasks of independent living, such as eating and bathing; mobility and navigation, or monitoring (e.g., Graf, Hans, and Schraft, 2004). Companion robots aim to enhance the health and psychological well-being of human users. Fong et al. (Fong, Nourbakhsh, and Dautenhahn, 2003) emphasized the critical role

of social interaction and used the term ”socially interactive robots.” Any robot developed to have the ability to interact and possibly able to communicate with users falls into the category of socially interactive robot.

2.2 Socially Assistive Robots in Mental Health Care

Compelling opportunities for social robots in the context of health care include their ability to educate, to enhance people's communication and social connection with others, to collect data to augment clinician’s understanding of patient 's mental condition, to help in cognitive and behavioural therapy, and to assist with adherence to care regimen through social support.

One of the promising applications of SARs is use in Autism therapy (Scassellati, Admoni, and Matarić, 2012). Studies report positive effects of robot presence on attention and engagement in therapy-like scenarios (Dautenhahn et al., 2009), spontaneous joint attention behaviours (Kozima, Nakagawa, and Yasuda, 2007), and sharing and turn taking which are usually difficult for children with autism. SAR systems for autism have had success as social mediators and embodiments that elicit social interactions between people (Scassellati et al., 2018).

Another area of health care where social robots have been predominantly used is for Dementia therapy. A variety of robots including PARO (Wada et al., 2008), AIBO (Kramer, Friedmann, and Bernstein, 2009), Keepon and Nao (Shamsuddin et al., 2012) have been used in intervention for dementia. Results of the studies indicated enjoyment, acceptance of the robot, increased socially interactive behavior, reduced sadness scores (Moyle et al., 2017), improved

stress recovery, maximize the user's task performance in the cognitive game (Tapus, Tapus, and Mataric, 2009), increased social network density (Wada and Shibata, 2007), and improved sleep patterns.

Social robots have also been studied for intervention of depression and stress. Study by Jøranson et al. shows significant reduction of depression and agitation among participants from baseline to followup (Jøranson et al., 2015). Similarly, a control trial for 10 weeks reported decrease in depression and blood pressure and an increase in cognitive activity (Thodberg et al., 2016) among participants in robot-assisted group activity. Robot therapy has shown to improve distress among children getting flu vaccinations (Beran et al., 2013). Robots have played an instrumental role in developing emotional security and reducing anxiety levels among adopted children (Trujillo, 2010).

Thus, the ability to capture physiologic data in an disencumbering way using robots has great potential for improving health assessment, diagnosis, and treatment. The promise of SARs for use in mental health care has been uncovered and developments are in progress.

2.3 Barriers for SAR in Mental Health Care

While there are exciting advances in health care robotics, it is important to carefully consider some of the challenges inherent in health care robotics. One of the major challenges is that user's perspective are often excluded from the robot design process, which leads to unusable and unsuitable technology. Robots that have a high degree of freedom and sophisticated features require a high level of cognitive function to control (Tsui et al., 2011). However,

many people needing such robots often have co-morbidities, which can make complex interaction an exhausting process. Thus the challenge is developing functionally simple and transparent robots.

The next underlined challenge is creating organisational therapy sessions and environments in which the robots and human can interact effectively (Lee et al., 2017). Considering that the robots in mental health care interact with mentally unstable people, the properties of interactions will vary drastically. Hence the challenge is developing robots that are intuitive yet have the computational ability to respond in unexpected situations. Another factor is the special knowledge required to operate the robot that pose a challenge for caregivers and clinicians (Lluch, 2011). This combined with the lack of evidence-based clinical effectiveness leads to clinical resistance and ignorance of the technology. Further, acceptability is an important barrier for health care robots. When a user uses a robot in public, they are immediately calling attention to their disability, disorder, or illness. Hence the robot morphology and behaviour must not embarrass the users. The stigma around technology, fear of leak of classified health information, and loss of privacy are huge setbacks from both users and clinicians perspective.

Lack of research on safe cognitive levels of human-robot interaction also contributes to the barriers of use of SARs in mental heath care. When used in mental health care, most users will not have the sufficient cognitive ability for limited dependence thus resulting in a closed relationship between the robot and user and avoidance of real interactions with others outside. Lastly, the cost associated with buying a robot also is a contributing barrier. There is

no guidance available on what robot to buy and the different robot features required for effective therapy of mental conditions preventing people from investing in the technology.

Chapter 3

Designing Social Robots for Mental Health Care

In this chapter, I summarize the required robot features based on the current state of the art of robots used in mental health care and the barriers discussed in Chapter 2. I then describe the two robots designed and developed for investigating the use of social robots for early detection of mental conditions. Design 1 explores multi-modal inputs and non-verbal interactions of social robot for detection of user's emotional state. Design 2 focuses on the collection of data for using acoustic features for detection of depression. In addition to the designs, I explore a pilot experiment to understand how the user's interact with the robot.

3.1 Robot Requirements

A key difference between conventional and social robots is that the way in which a human perceives a robot establishes expectations that guide his interaction with it (Fong, Nourbakhsh, and Dautenhahn, 2003). The appearance of

a robot is important because it helps establish social expectations. Physical appearance guides interaction. A relative familiarity of a robot's morphology can have profound effects on its accessibility, desirability, and expressiveness. In addition, since most of the target population might lack physical strength, the design must be relatively light in weight. Hygiene is a predominant concern in hospital and care environments. The robot must be easily sterilize-able.

A robot's morphology must match its intended function (DiSalvo et al., 2002). Since peer interaction is important in psychological therapy, the robot must project an amount of ”humanness” so that the user will feel comfortable in socially engaging the robot. At the same time, however, a robot's design needs to reflect an amount of ”robotness”. This is needed so that the user does not develop false expectations of the robot's capabilities.

A therapy robot must proficiently perceive and interpret human behavior. This includes detecting and recognizing gestures, monitoring and classifying activity, detecting intent and social cues. A social therapy robot must send signals to the human in order to provide feedback of its internal state and allow human to interact in a facile, and transparent manner. Emotions play a significant role in human behavior, communication and interaction. Artificial emotions helps facilitate believable human-robot interaction (Cañamero and Fredslund, 2001). Artificial emotion can also provide feedback to the user, such as indicating the robot's internal state and intentions (Fong, Nourbakhsh, and Dautenhahn, 2003; Breazeal, 1998). To achieve this, the robot must manifest believable behavior through the use of natural cues (gaze, gestures, etc.) and it must follow social convention and norms.

In order to be successfully employed in therapy sessions, the robots need to exhibit a certain degree of adaptability and flexibility to pro-actively encourage social interaction. This can be achieved by developing deep models of human interaction. These robots would be used by doctors, nurses, therapists, caregivers, and volunteers. Hence, it is important that the robots are designed in such a manner that anyone can operate them and that no specialized knowledge is required to do so.

With these design requirements in mind, I have developed two prototypes that are discussed in Sections 3.2 and 3.3.

3.2 Robot Design 1: DOT

Design 1 focuses on developing a social robot that perceives and interprets human emotional state and provides artificial emotion support. First, I describe the iterative design process that guided the robot design (Section 3.2.1). I then present the hardware and software of the robot that I developed (Sections 3.2.2 and 3.2.4). The developed robot is called DOT and is capable of emotional support and empathetic interactions. Haptics has been explored for detecting emotional state and generating responses. The haptic guesture processing framework and visual examples are included in Section 3.2.3. Finally, a pilot experimental study is presented to examine the effect of robot on human users (Section 3.2.5).

3.2.1 Design Iterations

The design process started with an iterative development of the robot's morphology and nonverbal behavior (NVB). An iterative process involving sketching, 3D modeling, and rapid prototyping, inspired by the movement-centric design approach introduced in (Hoffman and Ju, 2014) was followed to design the appearance of the robot. A story-boading approach was followed for designing the robot's character and behaviours.

Sketching - The designing started with freehand sketches exploring widely varying shapes based on animated and Disney-inspired designs (Sten and Walsh, 2006). Since children usually get attracted to warm squishy animals, one of the initial designs was modelled after a squirrel. Figure 3.1 shows CAD model and 3.2 shows the 3D printed prototype of the design.

Figure 3.1: CAD model of an early design variation: the squirrel

Figure 3.2: 3D printed prototype of an early design variation: the squirrel

However, a quick survey with eight participants selected through convenience sampling informed that a few people had inherent repulsion to some animals. Another key finding was that people had lower trust and expectation of robots designed after animals.

Thus, the subsequent robot models were designed to have unique identity and personality and not bear similarities to any animals or existing characters. Our methodology led us to a robot design inspired by simple geometric shapes. Some key morphology and interaction ideas related to our design goals emerged at the sketching stage: (a) The robot structure must consist of two independent geometric structures, relatively fastened to each other in order to express a rich set of NVBs efficiently. (b) The robot must have high degree of freedom for the head and eyes to exhibit social cues. (c) The robot structure must support easy change of external appearance without affecting the robot internal design. This is motivated by two factors: ease of identity change depending on the user and easy sterilisation in hospital environment. (d) The robot must be capable of table and lap interactions and hence have a robust and solid base to sit on. Figure 3.3 was one of the early designs considered for the robot but was later disregarded due to structural instability.

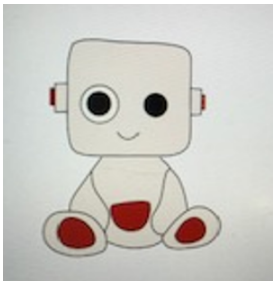

Figure 3.3: An example design from the iterative design process followed for the design of robot

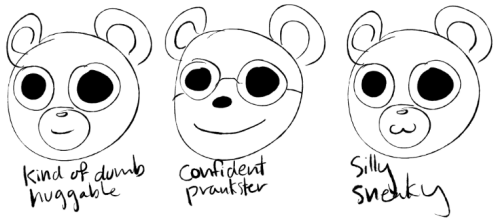

Figure 3.4: An example sketch from the iterative character design approach (Credits: Erica Huang)

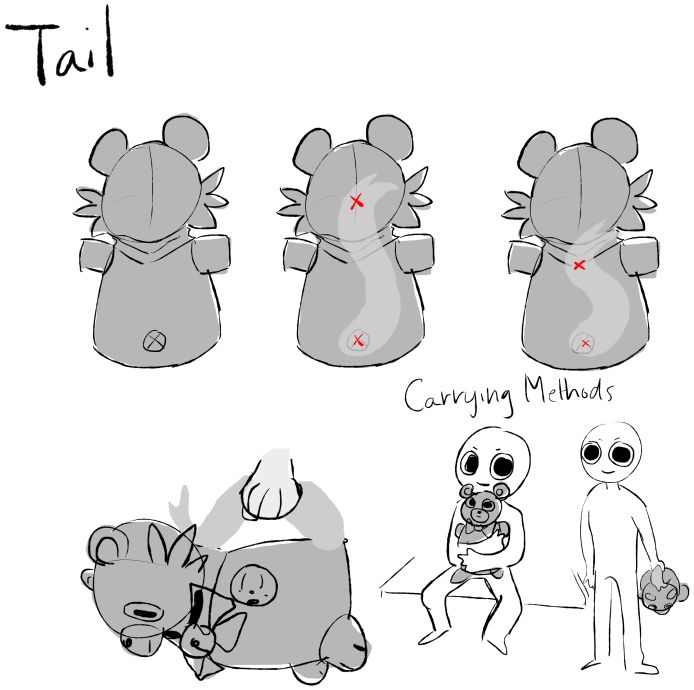

Figure 3.5: An example design from the iterative design process followed where the tail of robot and handling mechanism were explored (Credits: Erica Huang)

The robot's external appearance (artificial fur covering) was also designed through an interactive process. Different colours and accessories like the bow-tie, spectacles, eye aspect ratio, nose, mouth and tail were explored to design the robot's personality. Figure 3.4 shows the different facial features explored and the personality perceived due to appearance. Similarly, Figure

Figure 3.6: An example sketch from the iterative design approach followed that shows an interaction situation simulated (Credits: Erica Huang)

3.5 shows the different lengths of the tails explored and the simulated robot handling mechanism in presence of a tail.

The robot's movements and behavior were developed in parallel to robot's morphology. In order to understand when social relationships are needed in human-robot interaction or when the perception of such relationships need to be changed, social relations were modeled. Various scenarios and robot behaviours were designed using free hand sketching before implementation on

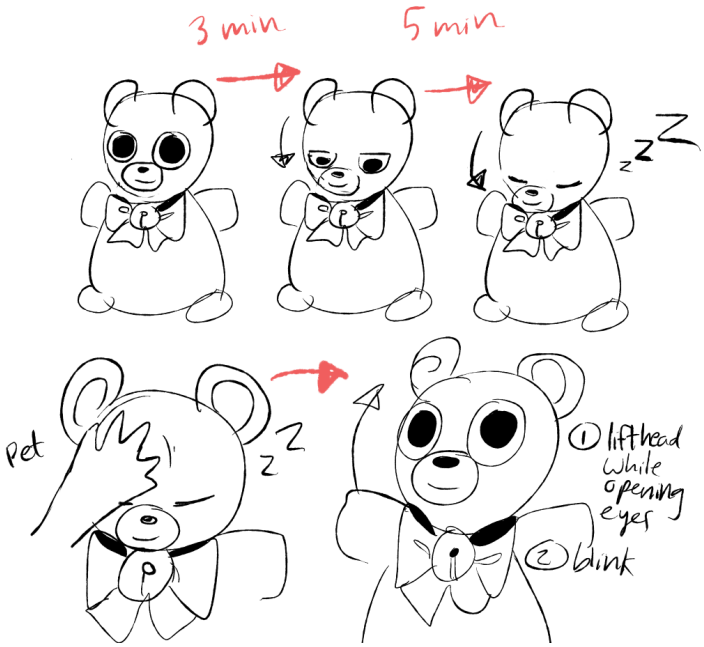

Figure 3.7: An example sketch from the iterative design process that shows the robot behaviour design for idle state (Credits: Erica Huang)

the actual robot platform. Figure 3.6 shows a developed interaction between the robot and human where the person is displaying affection by hugging the robot. Figure 3.7 shows the sequence of behaviours the robot exhibits during idle state.

3.2.2 Robot Hardware

This sections describes the developed hardware of the robot. The target population is older adults and children and the robot has been developed for lap, handheld and table interactions. The maximum height of the robot is 1 feet which allows it to be carried around and manipulated easily. The physical dimensions of the robot are 17x15x30 cms and weighs about 2.36 lbs. The robot has 6 DOF, eye-lids open and closing mechanism (2 DOF), eyeballs pan

and tilt mechanism (2 DOF) and neck rotation similar to human head (2 DOF). The entire robot is covered with artificial fur to encourage the users to make physical contact with the robot.

The robot skeleton is composed of two main parts: the Head and the Thorax, both connected by a neck. The thorax houses the actuation mechanism for the neck movements and the controllers of the robot. The head houses the actuation mechanism and mechanics for the eyeballs and eye-lids motion. The robot design in adapted from the open source robot Maki [1]. The original design was modified to accommodate speakers, improve fluidity between moving parts and to reduce the weight of the robot. In addition, the shape of the thorax was redesigned for better robot handling and aesthetics. All parts of the robot were 3D printed. The skeleton of the robot can be seen in figure 3.8 and figure 3.9 shows the robot in its artificial fur covering.

A newly-developed fabric tactile sensor is inserted between the hard inner skeleton and the fur to facilitate haptic iterations. The tactile sensor array covers the head and thorax regions with about 320 contact points. DOT is equipped with the four primary senses; sight(camera), auditory(microphone), balance(Inertial Measurement Unit or IMU) and the above-stated tactile sense.

The robot is equipped with a two layer control architecture. The high-level controller (Raspberri pi 3) supports speech and video processing, determining robot's internal state and generating high-level response to environment stimulus. There are low-level controllers: an Arduino Uno that processes the input

[1] https://www.kickstarter.com/projects/391398742/maki-a-3d-printable-humanoid-robot/posts

Figure 3.8: Picture of 3D printed robot prototype

Figure 3.9: A picture of the robot prototype with its artificial fur covering

from tactile sensors and IMU. An Arbotix-m controller that generates joint-level response gestures and commands the motors to generate movement. Both the controller layers interact via serial communication. Individual joints are actuated using the Dynamixel motors. Thus the robot totally consists 6 motors connected in series and commanded by the Arbotix-m controller. The robot controller programs, haptic and emotion recognition are available in github repository [2].

3.2.3 Haptic and Posture gestures

The differentiating feature of DOT from other social robots is its ability to recognise haptic cues. The input from tactile sensors is used to infer the gestures mentioned in table 3.1. Further, posture analysis is performed from the data received from the IMU to inform the robot about its surrounding

[2] https://github.com/akrishn9/Robot-for-emotional-support

and its position relative to the user. This gesture information helps the robot understand better about the environment and the user.

Gestures	Identification Pattern
Stroke	move fingers over the creature
Contact	touch any part of the robot
Hug	contact multiple sensor locations
Hold	make contact with both hands on robot
Rub	move hands over the stomach repeatedly
Pat	repeated taping gesture over the head/body
Squeeze	high contact force on robot
Poke	single point high pressure activation
Lift	lift the robot from rest position
Toss	throw the device off its standing position
Rock	move robot repeatedly front and back/side to side

Table 3.1: Summary of the haptic and posture gestures that the robot can interpret

Posture of the robot is achieved by using the 6 DOF IMU. The gyroscope is used to measure the robot's angular velocity and the accelerometer is employed to measure the external specific force acting on the robot. Since the accelerometer is a good indicator of orientation in static conditions and gyroscope is a good indicator of tilt in dynamic conditions, the accelerometer

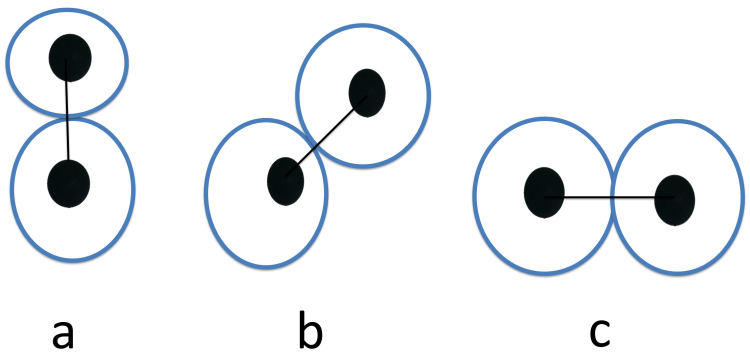

Figure 3.10: The robot is capable of detecting various postures by processing the IMU data: (a) stand position; (b) tilted position; (c) fallen pose which calls for help

signals are passed through a low-pass filter and the gyroscope signals through a high-pass filter and combined to obtain the final orientation (yaw, pitch, roll) of the robot. These values are compared with a base value to detect the posture. Figure 3.10 shows the different postures of the robot that can be detected using IMU data. These postures when repeated in sequence as described in Table 3.1 will be interpreted as gestures.

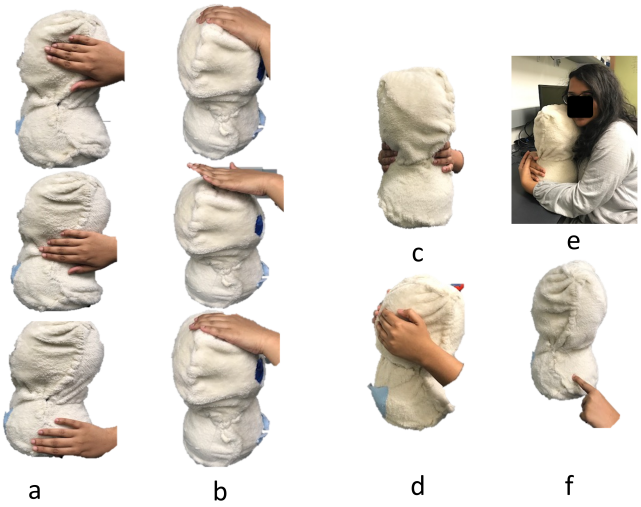

Figure 3.11: The robot is capable of detecting haptic gestures: (a) Stroke gesture sequence; (b) Pat gesture sequence; (c) Hold gesture; (d) Squeeze gesture; (e) Hug gesture; and (f) Poke gesture

A fabric tactile sensor, formed by sandwiching a resistive fabric between two conductive fabrics resulting in a matrix of mxn contact points, is used for obtaining tactile information. Due to contact, the contact pressure increases and the pressure at each point is derived by measuring the potential difference across it. This is performed by activating each column n with a digital high signal (5V TTL) while deactivating other rows with a digital low. This construction allows for about 320 contact points on the robot's body. The contact

regions in combination with previous contact history are used to classify the tactile input into one of the haptic gestures.

The different haptic gestures and visuals of the sequence of actions to determine the gestures can be found in Figure 3.11. The gesture recognition is performed by the low-level controller and transmitted to the high-level controller which then triggers appropriate responses.

3.2.4 Robot Software

This section describes the software framework that generates robot behaviours based on multi-modal inputs. DOT has two layers to generate its proactive behavior: a behavior-planning layer and a behavior-generation layer. Depending on its internal states DOT generates behavior. However, the internal state of the robot is influenced by the users mood and emotions. The behaviour-planning layer takes input from the face tracking and emotion frameworks and generates robot 's internal state. This layer then decides a particular response from a pool of predefined responses and sends basic behavioral patterns to the behavior-generation layer. The behavior-generation layer generates control references for each actuator to perform the determined behavior. The behavior-generation layer adjusts parameters of priority of behaviors based on the internal states. This creates lifelike behavior that the user will be able to interpret. Figure 3.12 shows the behavior generation system of DOT.

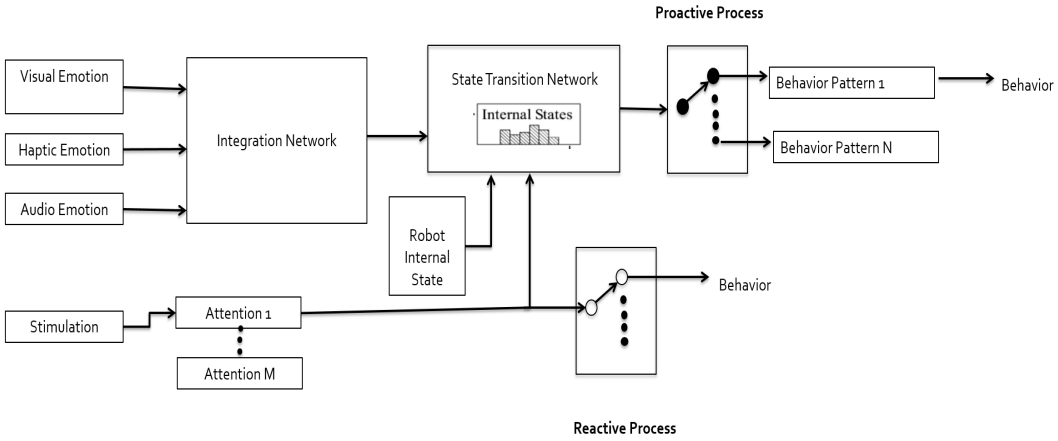

Figure 3.12: The software architecture of the robot for emotion recognition and response generation

3.2.4.1 Face Tracking Module

The face tracking module is designed to track the user across a room. This is accomplished by moving the head of the robot to maintain the person's face at the centre of image plane. The maximum head tilts are limited to 140deg along z axis and 170deg along x axis. Initially the face is detected using single short multibox detector and visual servoing is used to accomplish the tracking.

- *Face Detection Module*

  The Single Short MultiBox detector(SSD) based on (Liu et al., 2016) is used for real-time detecting of faces. This architecture achieves high accuracy in general object detection task together with real-time run-time performance (59FPS). The initial part of the network is based on YOLO base architecture(truncated before classification layers) unlike as proposed in the paper. Convolutional feature layers which decrease in size progressively and allow predictions of detection at multiple scales are added to the truncated base

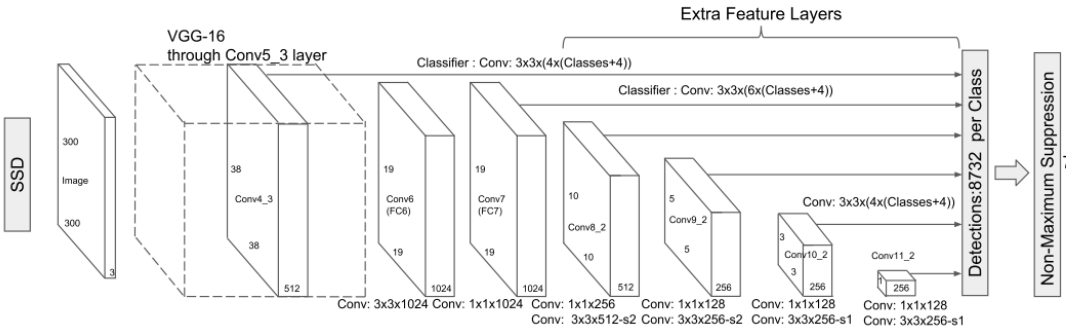

Figure 3.13: Single Short Detector Network for face detection (Liu et al., 2016)

network as shown in Figure 3.13. A set of default bounding boxes are associated with each feature map cell, for multiple feature maps at the top of the network. Specifically, for each box out of k at a given location, c class scores and the four offsets relative to the original default box shape are computed. This results in a total of $(c+4)kmn$ outputs for a mxn feature map.

The network was trained on both FER2013 (Carrier and Courville, 2013) and IMDB (Rothe, Timofte, and Gool, 2015) data sets. In case of training SSD, the ground truth information needs to be assigned to specific outputs in the fixed set of detector outputs. To begin with, each ground truth box is matched to the default box with the best Jaccard overlap (higher than 0.6 threshold). The overall objective loss function as measured using Eqn. 3.1 is a weighted sum of the localization loss (loc) and the confidence loss (conf):

$$L(x,c,l,g) = \frac{1}{N}(L_{con}f(x,c) + \alpha L_{loc}f(x,l,g)) \tag{3.1}$$

where N is the number of matched default boxes, and the localization loss is the Smooth L1 loss between the predicted box (l) and the ground truth

box(g) parameters.

To enable specific feature maps to be responsive to particular scales of the objects, the default boxes were tilted. The scale of the default boxes for each feature map is computed according to the measure in Eqn. 3.2:

$$s_k = s_{min} + \frac{s_{max} - s_{min}}{m - 1}(k - 1), k \in [1, m] \tag{3.2}$$

where s_{min} and s_{max} values were empirically chosen to be 0.3 and 0.85 respectively. A diverse set of predictions, covering various input object sizes and shapes is obtained by combining predictions for all default boxes. After the matching step, most of the default boxes are negatives. In order to avoid imbalance, the default boxes are sorted using the highest confidence loss so that the ratio between the negatives and positives is around 2:1. Data augmentation was not performed.

- *Tracking Module*

  The purpose of tracking is to keep the face at the center of the image plane by controlling the head position. The camera(robot head) remains static until the a person is detected, the person's location is simply determined by difference between two successive images. However, because of noise in the images, a threshold difference between two successive averaged images is considered. The center of gravity of the detected face gives the initial position to be zero.

  The desired position s* of any measured value s = (x,y) is the centre of image (i.e. s* = (0,0)) and s is view as the error vector. The aim then is to minimize

the error vector by controlling the head movements. Since in this case, only the translation along the x and z axis is considered, a linear rotation at the neck is made to minimize the error vector. The linear rotation is performed to maintain the face at the centre of image using the control law in Eqn. 3.3.

$$\dot{s} = [t_x, t_z]^T + \frac{\partial s}{\partial t} \tag{3.3}$$

3.2.4.2 Emotion Recognition Module

The emotion recognition framework is implemented to take the sensor(camera, microphone, and tactile) inputs and interpret the emotional state of the robot. I first obtain the audio signal into segments and then the DNN computes the emotion state distribution for each segment. The frames of video obtained from the camera is fed through the mini-Xception network to obtain the confusion matrix. In order to synchronize the audio and video frames, a time stamp is attached to each video frame and audio segment and the average emotion of all the frames corresponding to an audio segment is used for prediction of emotion. Finally the two probabilities are combined to obtain an emotional state.

- *Video-based Emotion Recognition Module*

  The deep network used for emotion classification is a fully-convolutional neural network that contains four residual depth-wise separable convolutions where each convolution is followed by a batch normalization operation and a ReLU activation function as described in Arriaga, Valdenegro-Toro, and Plöger, 2017. The last layer applies a global average pooling

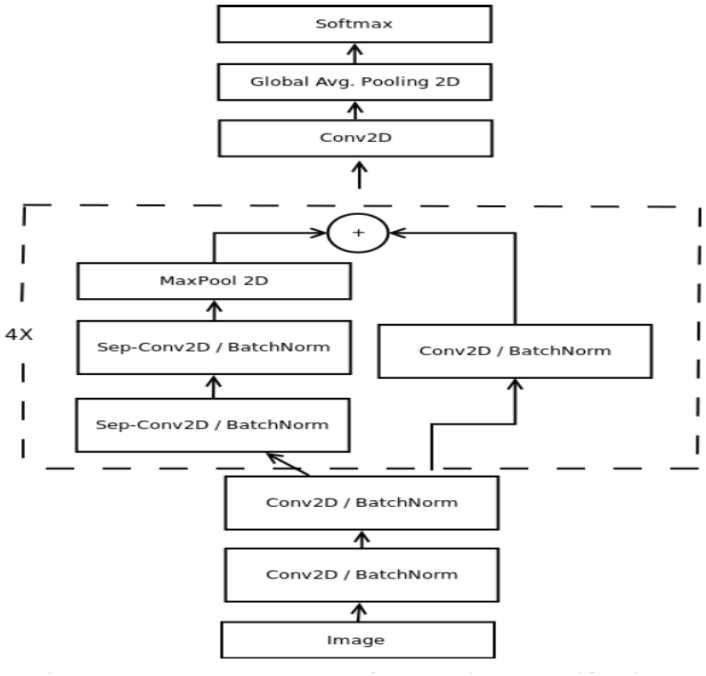

Figure 3.14: Network architecture for emotion and gender classification (Arriaga, Valdenegro-Toro, and Plöger, 2017)

and a soft-max activation function to produce a prediction. The different emotions classified are happy, amusement, surprise , sad, angry, fear and neutral.

The benefit of adding the residual modules is that they modify the desired mapping between two subsequent layers, so that the learned features become the difference of the original feature map and the desired features. The Depth-wise separable convolutions included in the deepnet help reduce the computation and are composed of two different layers: depth-wise and point-wise convolutions. These layers first applying a DxD filter on every M input channels and then apply $N1x1xM$ convolution filters to combine the M input channels into N output channels thereby separating the

spatial cross-correlations from the channel cross-correlations. Further, applying 1x1xM convolutions combines each value in the feature map without considering their spatial relation within a channel.

This architecture is called the mini-Xception and is shown in Figure 3.14. It has approximately 60,000 parameters; which corresponds to a reduction of 80x parameters compared to the original CNN. Emotion classification is trained using FER-2013 emotion dataset. The emotion classification module achieves 82.6% accuracy on the testing data of the FER-2013 dataset. The top 3 accurate emotion classes are happy, surprise and neutral. This is reasonable because there are very obvious facial cues when these emotions are shown on peoples face while emotions like fear and sad can be more subtle and harder to tell.

- *Audio-based Emotion Recognition Module*

  The entire audio track is broken into segments and the size of the segment level feature is set to 25 frames. The input signal is sampled at 16 kHz and converted into frames using a 25-ms window sliding at 10-ms each time. So the total length of a segment is 10ms x 25 + (25-10)ms= 265ms. The emotion recognition form the audio input is achieved by using a DNN network (Han, Yu, and Tashev, 2014). The segment-level DNN unlike as proposed in paper was identified to have 770-unit input layer corresponding to the dimensions of the feature vector. The DNN contains three hidden layers and each hidden layer has 256 rectified linear hidden units. Mini-batch gradient descend method is used to learn the weights in DNN and the objective function is cross-entropy.

The different features used include Mel-frequency cepstral coefficients(MFCC) features, spectral roll-off, pitch-based features and their delta feature across time frames. The Spectral Rolloff is the point where it is in the 85^{th} percentile of the power spectral distribution. A function W(ω) is said to be in the 85^{th} percentile if satisfies Eqn. 3.4.

$$|W(\omega)| < \frac{M}{w_{n+1}} forall\omega > \omega_0 \tag{3.4}$$

where ω_0 is the reading at any time.

MFCC are coefficients that represent sound as a short-term power spectrum. The process for generating them is as follows:

1. Segment the audio signal into short frames.
2. For each frame calculate the Discrete Fourier Transform and periodogram-based power spectral estimate of the power spectrum.
3. Apply the mel filter bank to the power spectra and sum the total energy in each filter.
4. Take the log of all filter bank energies.
5. Take the Discrete Cosine Transform of the log filter bank energies and coefficients 2-13.

The pitch-based features include pitch period and the harmonics-to-noise ratio(HNR). The segment-level feature vectors are formed by stacking features in the neighboring frames according to Eqn. 3.5.

$$x(m) = [z(m-w), ..., z(m), ..., z(m+w)] \tag{3.5}$$

where w is the window size on each side.

- *Combined Emotion Recognition Module*

  The output from both the audio and video based networks is used to predict the final emotion. Song et al., 2004 had used a tripled hidden Markov model(THMM) for emotion recognition. For combining the audio and video based emotion, I have used a modified version of the THMM which uses a combination of two HMMs, one for each data stream. It can be used to compute the joint likelihood of the two sequences. The choice of the initial parameters is critical as the maximum likelihood estimation of the parameters only converges to a local optimum. Hence, a method for the initialization of the maximum likelihood training that uses Viterbi algorithm modified from the algorithm described in the paper is used to derive the double HMM. The combined(audio and video) accuracy of emotion classification module is 92.3%.

3.2.5 Pilot Experiment

To study the effectiveness of the robot an experiment was conducted. During the experiment, artificial emotions were simulated and the interaction between the robot and user under different emotions was observed. Two participants, both female (M=23 yrs) who had neither previously interacted with DOT nor familiar with the research work were recruited for the study through convenience sampling.

The study took place in a controlled lab environment which was set like a home-theatre with a comfortable chair and side table. The schematic of the top

view of room setting is as shown in Figure 3.15. Prior to the study, the participants were introduced to DOT and various features of DOT were explained. During the study, artificial emotions were simulated in the participants. This was achieved by the participant watching a video for 22 mins which was created using the ravdness (Eerola and Vuoskoski, 2011) and International Affect Picture System (Lang, Bradley, and Cuthbert, 1997) data-sets. This protocol has been successfully used in prior research for generating emotions. The different emotions triggered were happiness, fear, sadness, anger, amusement, disgust and calmness. The participants were allowed to interact with the robot without any restrictions. After the study, participants were interviewed in addition to filling a questionnaire.

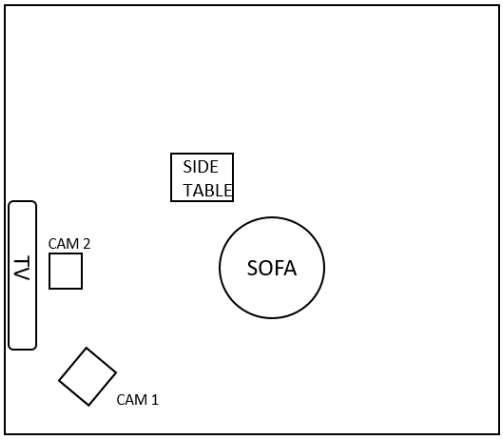

Figure 3.15: Top view of the setting used for pilot study

The questionnaire was a 7-point Likert scale measuring the acceptance of the robot. Regarding the relationship with DOT, people were interviewed along the following questions: i) Did you speak to and touch the robot? ii) Was it comfortable holding the robot? iii) How often do you play with the

robot? iv) What do you call the robot? v) What is the robot to you? In order to objectively investigate the interaction of the participants with DOT, the activities of the participants during the study was recorded.

3.2.6 Preliminary Observations

- *Interactions between the participants and DOT:*

  The participants were excited to meet DOT and greeted it like a friend or a new person during the introduction. The participants interacted with DOT willingly from the beginning, speaking to it, stroking and hugging it. During the study, though they watched a video, the participants continuously held the robot on their lap and kept stroking or patting. During the study, one of the participants felt protective over DOT and covered its eye and ears during simulated fear emotion. The interactions between the participant and DOT during study can be seen in Figures 3.16 and 3.17. During interview, while referring to the robot, both the participants personified the robot. It was also noticed that both the participants associated male gender to the robot and referred to the robot as ”him” or ”he”.

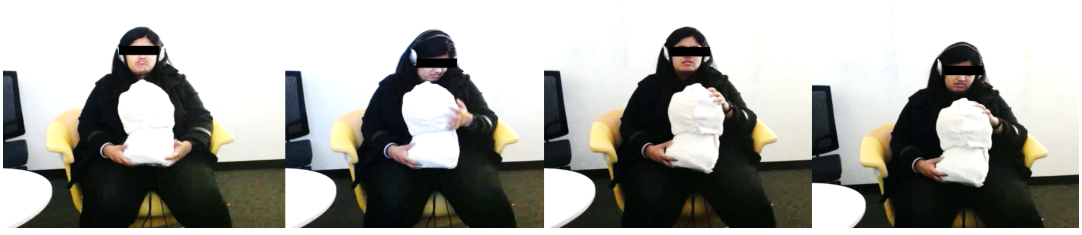

Figure 3.16: Sequence of interaction between the participant and robot during pilot study

Figure 3.17: Interaction between the participant and user during pilot during which the participant protects the robot

- *Results of Video analysis:*

  The recorded video was analysed to validate the effectiveness of the experimental video in simulating emotions and to understand the interaction between the participants and DOT. The result showed that the emotions felt by the participants was in line with the ones attempted to simulated through the short film. Analysis of the video showed that the participants continuously interacted with DOT. It was also observed that both participants held the robot facing away from them and towards the TV for most parts of the experiment. Further, it was observed that the users turned the robot to face them at points when they were talk to the robot or checking on the robot. Apart from this, I was not able to detect common interactions or behaviours between the robot and the user.

- *Results of Questionnaire:*

  The results of the questionnaire showed that the participants positively rated the robot behaviour and appearance. Both participants answered that they wanted the robot to exhibit nonverbal behaviour. Further, the participants rated strongly positive for question if they would like to interact with the robot in the future.

3.2.7 Discussion

The experimental study provided insight on the interactions between the robot and the user and, also provided feedback on the capabilities of the developed system. As mentioned in the results, it was observed that the participants held the robot facing away from them and in the direction they were focused. Due to this, the camera located in the robot 's eye was not able to track the user or detect emotions. The behaviour that both participants exhibited while interacting with the robot suggests that they were considering the robot as a person (maybe a child or pet) who needed protection.

Even though the emotion recognition module performed well on the data-sets, the performance of the module in the wild was limited. In order to improve the accuracy of the deepnets, it is mandatory to train the network on actual demographic population. The use of haptics for the detection of emotional state is a new and open challenge. Due to the limited sampling frequency of the low-level controller, the haptic module was unable to detect contacts that were established for very small duration (typically less than 2sec). As a result, the gestures were mis-classified. Another challenge is the absence of baseline for haptic data. Contact pressure varies widely depending on the person and this contributed to wrong classification of gestures. In order to train the haptic emotion recognition module, meaningful data needs to be collected.

The developed robot prototype focused on nonverbal behaviours to display social cues and internal state of the robot. Although studies show that nonverbal behaviour is irrepressibly impactful (DePaulo, 1992), the intention

to produce a particular nonverbal expression for self-presentational purposes cannot always be successfully interpreted as that emotion by the observer. Hence to make more transparent robots, verbal behaviours need to be explored.

The focus of this robot prototype is to interpret the emotional state of the user and direct robot responses. Studies have shown that the mental illnesses have an effect on the emotional state of the person (Mayo Clinic, 2015). Thus, the developed prototype is an initial design exploration towards the development of robots for detection of mental conditions.

3.3 Robot Design 2: Melo

Mental illnesses include many different conditions and developing technology for mental heath care is a rich problem space. To design robot for a specific mental condition, it was decided that the robot would be designed to detect early signs of depression. Thus, prototype 2 was designed to collect audio data to predict early stage depression. Hence in this section I first introduce the problem of depression to motivate the research.

According to the World Health organisation, about 322 million people around the world suffer from depression (Geneva:WHO, 2017). However, it is estimated that only 3% of the affected population receive treatment for depression (Olfson and Pincus, 1996). Earlier, treatment of depression was almost exclusively medication or cognitive therapy or a combination of both (McLean, 1981). With the advent of technology, mobile and computer based therapy programs have been explored to improve the care. Though these

technologies have shown some positive results, there is still a need for early detection of depression (not based on self reported questionnaires).

Symptoms of Depression include loss of pleasure, suicidal intentions, feeling of guilt and insomnia (Palagini et al., 2013). Physical symptoms include loss of energy and fatigue. All of these symptoms bear reflections in the speech of a person. The aim of this prototype is to leverage audio acoustics for early detection of depression.

The developed prototype is an interactive device that will leverage an intensive longitudinal research methodology called Experience Sampling to collect audio data to train a network for early detection of depression. Experience Sampling also known as Ecological Momentary Assessment is fine-grained way of measuring the mental state of a patient in context and over time, via questions posed to the patient (Mikus et al., 2018). In the subsequent sections, I summarise the design iterations undertaken and the final prototype. I then describe the software framework implemented and propose an experiment for data collection. In this design, only the acoustic features of speech have been explored for the detection of depression. The actual speech content will not be analysed for detection. The various acoustic features that could be explored are also detailed. Finally I talk about the digitisation of the Geriatric Depression Scale developed for user study.

3.3.1 Design Iterations

The design process started with an iterative development of the device's structure and interaction cues. An iterative process involving sketching, 3D

modeling, and rapid prototyping was followed to design the morphology of the robot.

Figure 3.18: An example design from the iterative design process followed for the design of interactive device (Credits: Erica Huang)

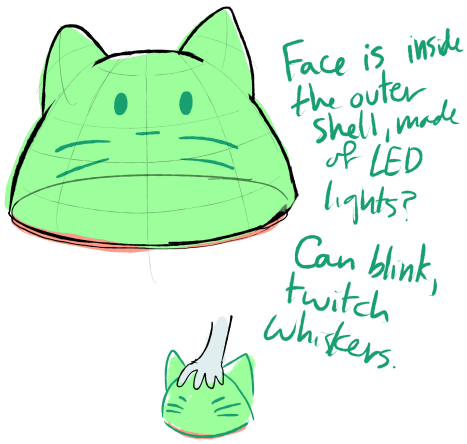

Figure 3.19: An example design from the iterative design process followed for the design of interactive device (Credits: Erica Huang)

Firstly, the design needs to be warm and friendly as it is designed for

interactions. Hence shapes that do not blend into the background were chosen. An early design developed to stand out of the background is shown in Figure 3.18. Further, designs that are overly complicated need to be avoided in order to facilitate easy assembly of the device. Since the component houses electronics a sturdy base was required. The above mentioned criteria lead to two more design aspects: (1) Designs must be easy to grip and not slip easily. (2) A wide flat base to make sure it does not fall over. (3) Air pockets to allow high quality audio input. The design in Figure 3.19 was developed based on a cat model to invite people to interact.

3.3.2 Device Hardware

The final design is a rectangular doom shaped structure consisting of 2 individual parts. The upper, a dome structure and the base, are coupled using screws. The outed structure was 3D printed. The device consists of a Raspberri pi3 and is called the ”Melo”. It is equipped with speakers and a USB microphone for audio data collection. The Raspberri pi receives audio readings from mic attached to it. The CAD model of the individual parts is shown in Figure 3.20 and Figure 3.21 shows the 3D printed final prototype. The final prototype and previous CAD models have been made open source and can be found in github [3].

The raspberri pi is power by cables connected to the socket. In order to display the device 's internal state, LEDs are actuated in certain patterns. This design uses acoustics and LEDs to display proactive behaviour and is inspired

[3] https://github.com/akrishn9/Melo-Design

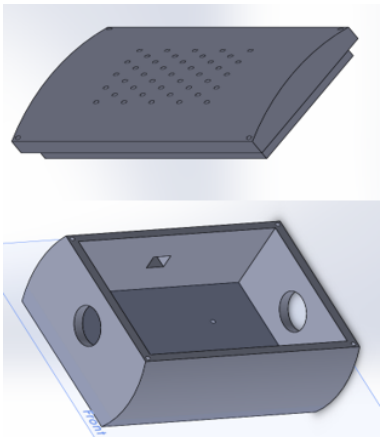

Figure 3.20: 3D CAD models of the two individual parts in Melo

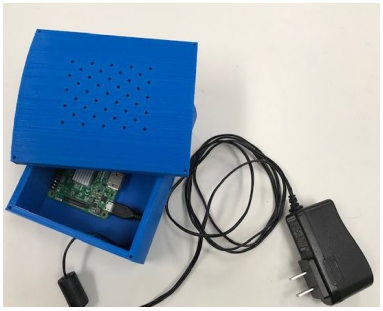

Figure 3.21: A picture of the device prototype that was 3D printed and assembled

by "Amazon Echo".

3.3.3 Device Software

The software module is designed to interact with the user multiple times a day and collect the responses. It consists of 3 major modules: Audio recording module for voice activity detection and audio recording, Depression detection framework that will detect the users depression from the recorded audio and finally the interaction module which will trigger questions to collect information. The project files can be found at site [4]. The time of interaction and the questions will be randomly selected each day.

3.3.3.1 Audio Recording Framework

The audio module is designed to record and pre-process the audio input. Since the device on all the time, storing the complete audio information will

[4] https://github.com/akrishn9/Social-robot-for-depression-detection

consume a large amount of space. In order to avoid this, the voice activity is detected and only audio frames that have speech signals are recorded and processed. This is achieved by converting the stereo to mono audio lines and moving the windows of 20ms along the audio data. The ratio between energy of speech band and total energy for window is computed and ratios above a threshold (0.6) is labelled as speech and recorded in 4sec frames. Prior to recording, median filter with length of 0.5s is applied to smooth detected speech regions.

3.3.3.2 Features for depression detection

In order to avoid cultural barriers, our framework uses only the audio acoustics for detection of depression. The proposed framework in the modeling and classification of the depressed speech is illustrated in Figure 3.22.

Grouping of acoustic features into categories and subcategories that are closely related to the human speech production model is proposed as in Moore II et al., 2008 and Low et al., 2011. In the study, the acoustic features

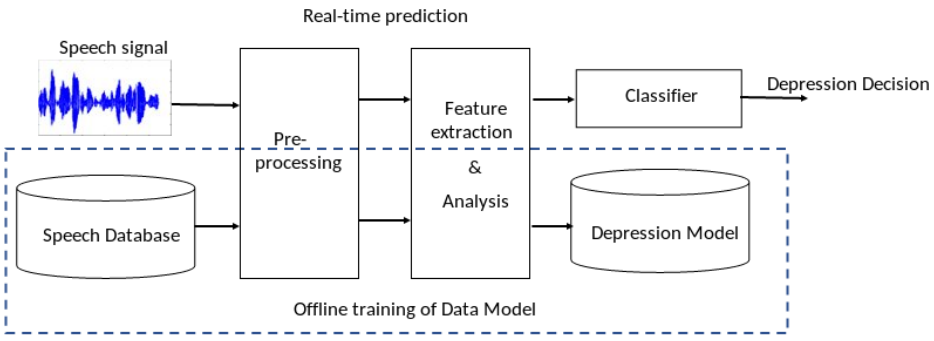

Figure 3.22: Network for modelling and classification of depression

are grouped into five main categories: TEO-based, cepstral (C), prosodic (P), spectral (S), and glottal (G) features. The acoustic features are briefly discussed in the following paragraphs.

1) TEO-based features: TEO-based features have shown good performances in stress recognition (Zhou, Hansen, and Kaiser, 2001). During stress, fast air flow causes vortices located near the false vocal folds, which provide additional excitation signals other than the usual speech pitch. Teager (Teager, 1980) proposed an energy operator called the TEO to model the time-varying vortex. Several TEO-based features have been proposed in the literature. TEO-critical-band-based autocorrelation envelope (TEO-CB-Auto-Env) feature, based on the method discussed in (Zhou, Hansen, and Kaiser, 2001) is proposed.

2) Cepstral feature: Similar to the emotion recognition module, the MFCC can be effectively used in speech content characterization.

3) Prosodic features: The Prosodic features are described below

- Fundamental frequency: The auto-correlation method was chosen for the full dataset to compute the fundamental frequency. The values of F0 can be determined on a frame-by-frame basis by finding the maximum values of the auto-correlation function.

- Log energy: To determine the changes in speaking behavior in response to factors relating to depression the LogE of the speech time waveform can be calculated as in (J. R. Deller, 1999).

- Jitter: Jitter refers to the cycle-to-cycle fluctuations in pitch. It is obtained

by measuring the fundamental frequency (F0) of each cycle of vibration, subtracting it from the previous F0 values, and dividing it by the average of F0.

- Shimmer: Shimmer is calculated in similar fashion to jitters. However, the period-to-period variability of the signal peak-to-peak amplitude is calculated instead.

- Formants (FMTS) and formant bandwidths (FBWS): A 13^{th}-order LP filter is proposed to calculate the formant frequencies. Only values of the first three formants (FMT 1-FMT 3) and formant bandwidths (FBW 1-FBW 3) below its Nyquist frequency needs to be taken.

4) Spectral features: Below are some of the spectral features used for depression detection.

- Spectral centroid: Center of a signal's spectrum power distribution is the Spectral centroid. It is the calculated as the weighted mean of frequencies present in the signal as in Eqn. 3.6.

$$SC = \frac{\sum_{n=1}^{M} f(n)X(n)}{\sum_{n=1}^{M} X(n)} \tag{3.6}$$

  where $X(n)$ represents the magnitude of frequency bin number n, $f(n)$ represents the center frequency bin, and M is the total number of frequency bins.

- Spectral flux: Spectral flux is the measure of how quickly the power spectrum of a signal is changing and it can be calculated by relative comparison of the power spectrum for one frame against the previous.

- Spectral roll-off: Spectral roll-off is the point, where the frequency that is below some percentile.

- Power spectral density: Using a 4096-point fast Fourier transform with a 5-ms nonoverlapping hamming window size, the one-sided PSD can be computed.The PSD have been effectively used to discriminate between speech of control and depressed adults (France et al., 2000).

5) Glottal features: The glottal pulse and shape have been documented to play an important role in the analysis of speech in clinical depression (Moore II et al., 2008). Quantitative analysis of the glottal flow pulses can be performed in the time and frequency domains. It should be noted that glottal waveform extraction is still a matter of study and accurate representations are still difficult to determine and verify. The glottal flow can be divided into the opening phase (OP), closing phase (CP), and closed phased (C). Once these instances are acquired, several timing and frequency parameters can be easily calculated and used for training.

3.3.4 Proposed experiment

The experiment will closely follow the protocol used by Silk et al. for studying the emotional dynamics in depressed youth (Silk et al., 2011). Participants will be requested to complete a preliminary questionnaire to gather information such as average hours spent at home, details for installation of device and demographics. The participants will also be asked to answer questionnaires to understand their depression, and cognitive levels which will serve as baseline. Big-Five personality test will also be included to explore the personality of

the individual (Goldberg, 1992). Before the experiment, the devices will be installed at the participants homes.

A four week field study will follow during which the device will be triggered multiples times a day and information will be collected. Once a week, the participants will be asked to answer the Geriatric Depression Scale(GDS) (Yesavage et al., 1982), a well used scale in research for the measurement of depression. The data will be labelled based on the results of the scale and used for training the network

3.3.5 Digitization of Scale

Digitization is the process of conversion of an item in printed text, manuscript, image or sound, film and video recording from into digital form (Devi and A.V. Murthy, 2005). For the experiments, a digitized version of the GDS scale will be used for depression level measurement. The need arose cause, in case of digitised scales, the data can be collected and monitored remotely without having to trouble the participant or the researcher. Another reason was the ease of data access over distributed research team. This way, we will be able to collect data from participants regularly without visiting their homes. Most of our material is of a sensitive nature, including many personal information and digitization would allow us to restrict access to this fragile resources.

The GDS was digitised using the PyQt software. A interactive 5 window navigable GUI was created to collect mood data and personal information. Two windows for the GDS questions, 1 for the personal information and an introductory and a debriefing window. Link to the Big-Five personality test

Figure 3.23: An example screen from the developed interactive digital version of the Depression Scale

was included in the personal information section. The GUI is designed in such a manner that it does not allow the user to navigate to the next screen without answering all question in a given screen. An interactive dialog box shows up if the user tries to change his initial choice for a particular question. This information was then logged to the database. Figure 3.23 shows a screen from the GUI. The tick marks are displayed as a marker to help user's identify the unanswered sections of the questionnaire.

Chapter 4

General Discussion

In this chapter, I provide discussions on the lessons I learned from the research undertaken (Section 4.1), limitations of this research (Section 4.2), and directions for future work (Section 4.3).

4.1 Lessons Learned

Robotics in mental heath care is a rich problem space. As mentioned earlier, the illnesses vary in condition and severity and affect wide demographics. The symptoms that an individual shows to any particular mental condition is not the same as another individual. For example, Cauffman et al., 2007 show the difference in symptoms based on gender. Similarly, Weissman et al., 1977 have identified difference in symptoms of depression across demographics. It is therefore important to approach a specific and defined problem while designing robots for mental health care.

Detection of symptoms is a multi-faceted problem. The robot must be

aware of its surroundings and the user's intentions, be able to recognise behavioural changes, must be able to maintain a long-term relationship with the user and finally collect useful data for interpretation of mental conditions. To facilitate this, the robot must be quipped with muti-sensor input and be capable of processing this information in real time. The different inputs explored in this research are audio, visual, haptic and orientation(IMU). Though multi-modal inputs increase the detection compared to individual sensor inputs, there is still scope for improvements. One avenue is better data for training the detection networks. Most of the data sets available currently are not designed for a particular demographic or any particular mental condition. Thus networks trained using the data have low prediction accuracy. More specific data will help in better detection of mental conditions. Another area of potential improvement is modelling the proactive behaviour of the robot to fully explore the potential of it's features (Huang and Mutlu, 2014; Huang and Mutlu, 2012). The robot behaviors could be modelled based on a clinician to improve the detection. However, the robot must be autonomous and intelligent to respond to uncertain situations which can be common when dealing with people with mental conditions.

4.2 Limitations

Limitations of this research are discussed to provide directions for future research. Firstly, the research approach employed in this thesis does not involve the patients to inform the design of robot. This research explores an iterative design approach and multi-modal sensing methods to infer the

emotional state of the person. However, it imposes limitations on how effective these features are when interaction with the users. Another limitation is the sensors itself. The tactile sensor input was not always reliable and in certain condition led to misjudgment of haptic gestures. This was because, the controller took minimum time to loop through all the 320 contact points and contacts for very short time were not recorded during this interval.

In addition to the design limitation, only a pilot study was conducted in this research. Complete, rigorous HRI evaluations are required to study the evident-based effectiveness of robot in therapy. Additionally, observations conducted in this research involved one-time short-period interactions. Such short exposure to the robot and the experimental manipulations might yield results different from those obtained through long-term interactions. To realize natural, effective human-robot interaction, future research is necessary to study long-term deployment of robots in natural human environments and explore how those robots might be integrated into human daily activities and therapy sessions.

Lastly, contextual factors, such as cultural background, personal disposition, and gender, were not explicitly taken into account in developing the deep networks for emotion recognition and the robot behaviour. Prior research has shown how these contextual factors might influence ways in which people perceive and express behaviors. For example, Triandis describes how cultural factors shape the private, public and collective behaviour of a person (Triandis and Charalambos, 1994). Robot trust, likability and engagement and response are influenced by the culutural achground (Li, Rau, and Li, 2010). Future

research is needed to explore how robots might take these factors into account when engaging human users.

4.3 Future Directions

To ultimately enable robots therapy and to enable robots detect mental illness, we need a better understanding of how people with mental conditions live in their natural environments and how robots might be integrated. To gain such understanding, field deployment of robots in hospitals and individual homes is necessary. Such inclusion of field studies bridges the gap between controlled laboratories and real-world environments. Further, such experiments will increase the trust and acceptance of therapy robot amidst both clinicians and users.

Throughout the discussion, it should be clear that enabling effective human-robot interactions and detection of mental illnesses is an interdisciplinary problem, requiring applications of various techniques, methods, and theories. This research therefore also motivates future work in related fields, modelling robot behaviour to individuals and based on situations, data to identify symptoms, affecting computing and social signal processing. Advances in social signal processing will allow robots to better understand and utilize social signals displayed by human for early detection of mental conditions. Similarly, a better interpretation of affect and natural human behaviour will facilitate building rapport between the person and the robot. Moreover, data collected from the actual population will lead to creating better detection models resulting in better and accurate early detection of illnesses.

Chapter 5

Conclusion

This thesis seeks to explore the use of social robots for early detection of mental illnesses. To this end, I drew on the mechanisms of artificial emotional support, natural social interaction, empathy and structures intervention proposed in psychological science to develop mental therapy robot. My approach followed an iterative design process to develop a highly suited robot morphology and behaviour. In the initial part of research, to enable consistent detection of the user emotional state and help the robot better understand its environment, I have explored multi-modal inputs. To this end, deep networks for face recognition, emotion recognition using audio and video have been implemented. To explore the effectiveness of touch, I have proposed a haptic based emotion recognition and implemented the gesture recognition and reactive responses. In investigating how to provide artificial emotional support and structured response, a proactive nonverbal behaviour set has been developed in response to the user's actions. Further, I have studied the effectiveness of the robot through a pilot experiment.

A second prototype has been developed for the detection of mental conditions. In this prototype, I have explored the rich set of acoustics speech features for early detection of depression. The prototype developed is for collecting acoustic data for early detection of mental illnesses. Similar to the initial process, I have explored designs iteratively and have developed a device that is capable of communicating it's internal state through audio and LEDs. Finally, I have proposed an user study experiment for collecting data.

Overall, I present a series of designs to motivate the future design of social robots for early detection of mental conditions. This thesis informs future research on the design of robots and motivates the integration of social robots for early detection of mental illnesses.

References

Ritchie, Hannah and Max Roser (2018). *Mental Health*. URL: https://ourworldindata.org/mental-health.

National Institute of Mental Health (2017). *Any Mental Illness (AMI) Among Adults*. URL: http://www.nimh.nih.gov/health/statistics/prevalence/any-mental-illness-ami-among-adults.shtml.

Education, U.S. Department of (2013). *35th Annual Report to Congress on the Implementation of the Individuals with Disabilities Education Act*. URL: http://www2.ed.gov/about/reports/annual/osep/2013/parts-b-c/35th-idea-arc.pdf.

Insel, Thomas R (2008). *Assessing the economic costs of serious mental illness*.

Colton, Craig W and Ronald W Manderscheid (2006). "PEER REVIEWED: Congruencies in increased mortality rates, years of potential life lost, and causes of death among public mental health clients in eight states". In: *Preventing chronic disease* 3.2.

Walker, Elizabeth Reisinger, Robin E McGee, and Benjamin G Druss (2015). "Mortality in mental disorders and global disease burden implications: a systematic review and meta-analysis". In: *JAMA psychiatry* 72.4, pp. 334–341.

Isometsä, ET (2001). "Psychological autopsy studies–a review". In: *European psychiatry* 16.7, pp. 379–385.

Lasser, Karen, J Wesley Boyd, Steffie Woolhandler, David U Himmelstein, Danny McCormick, and David H Bor (2000). "Smoking and mental illness: a population-based prevalence study". In: *Jama* 284.20, pp. 2606–2610.

National Institute of Mental Health (2019). *Any Mental Illness (AMI) Among Adults*. URL: https://www.nimh.nih.gov/health/statistics/mental-illness.shtml.

American Mental Health Councillors Association (2011). *Need for Early Mental Health Screening*. URL: http://www.amhca.org/HigherLogic/System/

DownloadDocumentFile.ashx?DocumentFileKey=2ca60afe-8be0-af27-2ad9-7100b61ad636&forceDialog=0.

Callan, Judith A, Jesse Wright, Greg J Siegle, Robert H Howland, and Britney B Kepler (2017). "Use of computer and mobile technologies in the treatment of depression". In: *Archives of psychiatric nursing* 31.3, pp. 311–318.

Robinson, Hayley, Bruce MacDonald, and Elizabeth Broadbent (2014). "The role of healthcare robots for older people at home: A review". In: *International Journal of Social Robotics* 6.4, pp. 575–591.

Deng, Eric, Bilge Mutlu, Maja J Mataric, et al. (2019). "Embodiment in Socially Interactive Robots". In: *Foundations and Trends® in Robotics* 7.4, pp. 251–356.

Matarić, Maja J and Brian Scassellati (2016). "Socially assistive robotics". In: *Springer Handbook of Robotics*. Springer, pp. 1973–1994.

Scassellati, Brian, Henny Admoni, and Maja Matarić (2012). "Robots for use in autism research". In: *Annual review of biomedical engineering* 14, pp. 275–294.

Mordoch, Elaine, Angela Osterreicher, Lorna Guse, Kerstin Roger, and Genevieve Thompson (2013). "Use of social commitment robots in the care of elderly people with dementia: A literature review". In: *Maturitas* 74.1, pp. 14–20.

Chen, Shu-Chuan, Cindy Jones, and Wendy Moyle (2018). "Social robots for depression in older adults: a systematic review". In: *Journal of Nursing Scholarship* 50.6, pp. 612–622.

Trost, Margaret J, Adam R Ford, Lynn Kysh, Jeffrey I Gold, and Maja Mataric (2019). "Socially Assistive Robots for Helping Pediatric Distress and Pain: A Review of Current Evidence and Recommendations for Future Research and Practice". In: *The Clinical Journal of Pain* 35.5, pp. 451–458.

Ricks, Daniel J and Mark B Colton (2010). "Trends and considerations in robot-assisted autism therapy". In: *2010 IEEE international conference on robotics and automation*. IEEE, pp. 4354–4359.

Scassellati, Brian, Laura Boccanfuso, Chien-Ming Huang, Marilena Mademtzi, Meiying Qin, Nicole Salomons, Pamela Ventola, and Frederick Shic (2018). "Improving social skills in children with ASD using a long-term, in-home social robot". In: *Science Robotics* 3.21, eaat7544.

Scassellati, Brian (2007). "How social robots will help us to diagnose, treat, and understand autism". In: *Robotics research*. Springer, pp. 552–563.

Tapus, Adriana, Mataric Maja, and Brian Scassellatti (2007). "The grand challenges in socially assistive robotics". In: *IEEE Robotics and Automation Magazine* 14.1, N–A.

Arsand, Eirik and George Demiris (2008). "User-centered methods for designing patient-centric self-help tools". In: *Informatics for health and social care* 33.3, pp. 158–169.

Bemelmans, Roger, Gert Jan Gelderblom, Pieter Jonker, and Luc De Witte (2012). "Socially assistive robots in elderly care: A systematic review into effects and effectiveness". In: *Journal of the American Medical Directors Association* 13.2, pp. 114–120.

Feil-Seifer, David and Maja J Matarić (2005). "Defining socially assistive robotics". In: *9th International Conference on Rehabilitation Robotics, 2005. ICORR 2005.* IEEE, pp. 465–468.

Gomi, Takashi and Ann Griffith (1998). "Developing intelligent wheelchairs for the handicapped". In: *Assistive Technology and Artificial Intelligence*. Springer, pp. 150–178.

Kazerooni, H (2005). "Exoskeletons for human power augmentation". In: *2005 IEEE/RSJ International conference on intelligent Robots and Systems*. IEEE, pp. 3459–3464.

Graf, Birgit, Matthias Hans, and Rolf D Schraft (2004). "Care-O-bot IIâĂŤDevelopment of a next generation robotic home assistant". In: *Autonomous robots* 16.2, pp. 193–205.

Fong, Terrence, Illah Nourbakhsh, and Kerstin Dautenhahn (2003). "A survey of socially interactive robots". In: *Robotics and autonomous systems* 42.3-4, pp. 143–166.

Dautenhahn, Kerstin, Chrystopher L Nehaniv, Michael L Walters, Ben Robins, Hatice Kose-Bagci, N Assif Mirza, and Mike Blow (2009). "KASPAR–a minimally expressive humanoid robot for human–robot interaction research". In: *Applied Bionics and Biomechanics* 6.3-4, pp. 369–397.

Kozima, Hideki, Cocoro Nakagawa, and Yuriko Yasuda (2007). "Children–robot interaction: a pilot study in autism therapy". In: *Progress in brain research* 164, pp. 385–400.

Wada, Kazuyoshi, Takanori Shibata, Toshimitsu Musha, and Shin Kimura (2008). "Robot therapy for elders affected by dementia". In: *IEEE Engineering in medicine and biology magazine* 27.4, pp. 53–60.

Kramer, Stephen C, Erika Friedmann, and Penny L Bernstein (2009). "Comparison of the effect of human interaction, animal-assisted therapy, and AIBO-assisted therapy on long-term care residents with dementia". In: *Anthrozoös* 22.1, pp. 43–57.

Shamsuddin, Syamimi, Hanafiah Yussof, Luthffi Ismail, Fazah Akhtar Hanapiah, Salina Mohamed, Hanizah Ali Piah, and Nur Ismarrubie Zahari (2012).

"Initial response of autistic children in human-robot interaction therapy with humanoid robot NAO". In: *2012 IEEE 8th International Colloquium on Signal Processing and its Applications*. IEEE, pp. 188–193.

Moyle, Wendy, Cindy J Jones, Jenny E Murfield, Lukman Thalib, Elizabeth RA Beattie, David KH Shum, Siobhan T O'Dwyer, M Cindy Mervin, and Brian M Draper (2017). "Use of a robotic seal as a therapeutic tool to improve dementia symptoms: A cluster-randomized controlled trial". In: *Journal of the American Medical Directors Association* 18.9, pp. 766–773.

Tapus, Adriana, Cristian Tapus, and Maja J Mataric (2009). "The use of socially assistive robots in the design of intelligent cognitive therapies for people with dementia". In: *2009 IEEE international conference on rehabilitation robotics*. IEEE, pp. 924–929.

Wada, Kazuyoshi and Takanori Shibata (2007). "Social effects of robot therapy in a care house-change of social network of the residents for two months". In: *Proceedings 2007 IEEE International Conference on Robotics and Automation*. IEEE, pp. 1250–1255.

Jøranson, Nina, Ingeborg Pedersen, Anne Marie Mork Rokstad, and Camilla Ihlebæk (2015). "Effects on symptoms of agitation and depression in persons with dementia participating in robot-assisted activity: a cluster-randomized controlled trial". In: *Journal of the American Medical Directors Association* 16.10, pp. 867–873.

Thodberg, Karen, Lisbeth Uhrskov Sørensen, Janne Winther Christensen, Pia Haun Poulsen, Birthe Houbak, Vibeke Damgaard, Ingrid Keseler, David Edwards, and Poul B Videbech (2016). "Therapeutic effects of dog visits in nursing homes for the elderly". In: *Psychogeriatrics* 16.5, pp. 289–297.

Beran, Tanya N, Alex Ramirez-Serrano, Otto G Vanderkooi, and Susan Kuhn (2013). "Reducing children's pain and distress towards flu vaccinations: A novel and effective application of humanoid robotics". In: *Vaccine* 31.25, pp. 2772–2777.

Trujillo, Kate (2010). "Developing Emotional Security Among Children Who Have Been Adopted". In:

Tsui, Katherine M, Dae-Jin Kim, Aman Behal, David Kontak, and Holly A Yanco (2011). ""I want that": Human-in-the-loop control of a wheelchair-mounted robotic arm". In: *Applied Bionics and Biomechanics* 8.1, pp. 127–147.

Lee, Hee Rin, Selma Šabanović, Wan-Ling Chang, David Hakken, Shinichi Nagata, Jen Piatt, and Casey Bennett (2017). "Steps toward participatory design of social robots: mutual learning with older adults with depression".

In: *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*. IEEE, pp. 244–253.

Lluch, Maria (2011). "Healthcare professionals'organisational barriers to health information technologies-A literature review". In: *International journal of medical informatics* 80.12, pp. 849–862.

DiSalvo, Carl F, Francine Gemperle, Jodi Forlizzi, and Sara Kiesler (2002). "All robots are not created equal: the design and perception of humanoid robot heads". In: *Proceedings of the 4th conference on Designing interactive systems: processes, practices, methods, and techniques*. ACM, pp. 321–326.

Cañamero, Lola and Jakob Fredslund (2001). "I show you how I like you-can you read it in my face?[robotics]". In: *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and humans* 31.5, pp. 454–459.

Breazeal, Cynthia et al. (1998). "A motivational system for regulating human-robot interaction". In: *Aaai/iaai*, pp. 54–61.

Hoffman, Guy and Wendy Ju (2014). "Designing robots with movement in mind". In: *Journal of Human-Robot Interaction* 3.1, pp. 91–122.

Sten, Ekman, Sanderson Susan Walsh, et al. (2006). *Design-inspired innovation*. World Scientific.

Liu, Wei, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg (2016). "Ssd: Single shot multibox detector". In: *European conference on computer vision*. Springer, pp. 21–37.

Carrier, Pierre-Luc and Aaron Courville (2013). *Challenges in Representation Learning: Facial Expression Recognition Challenge*.

Rothe, Rasmus, Radu Timofte, and Luc Van Gool (2015). "DEX: Deep EXpectation of apparent age from a single image". In: *IEEE International Conference on Computer Vision Workshops (ICCVW)*.

Arriaga, Octavio, Matias Valdenegro-Toro, and Paul Plöger (2017). "Real-time convolutional neural networks for emotion and gender classification". In: *arXiv preprint arXiv:1710.07557*.

Han, Kun, Dong Yu, and Ivan Tashev (2014). "Speech emotion recognition using deep neural network and extreme learning machine". In: *Fifteenth annual conference of the international speech communication association*.

Song, Mingli, Jiajun Bu, Chun Chen, and Nan Li (2004). "Audio-visual based emotion recognition-a new approach". In: *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.* Vol. 2. IEEE, pp. II–II.

Eerola, Tuomas and Jonna K Vuoskoski (2011). "A comparison of the discrete and dimensional models of emotion in music". In: *Psychology of Music* 39.1, pp. 18–49.

Lang, Peter J, Margaret M Bradley, and Bruce N Cuthbert (1997). "International affective picture system (IAPS): Technical manual and affective ratings". In: *NIMH Center for the Study of Emotion and Attention* 1, pp. 39–58.

DePaulo, Bella M (1992). "Nonverbal behavior and self-presentation." In: *Psychological bulletin* 111.2, p. 203.

Mayo Clinic (2015). *Mental illness*. URL: https://www.mayoclinic.org/diseases-conditions/mental-illness/symptoms-causes/syc-20374968.

Geneva:WHO (2017). *Depression and Other Common Mental Disorders:Global Health Extimates*.

Olfson, Mark and Harold Alan Pincus (1996). "Outpatient mental health care in nonhospital settings: distribution of patients across provider groups". In: *The American journal of psychiatry* 153.10, p. 1353.

McLean, PD (1981). "Matching treatment to patient characteristics in an outpatient setting". In: *Behavior therapy for depression*, pp. 197–206.

Palagini, Laura, Chiara Baglioni, Antonio Ciapparelli, Angelo Gemignani, and Dieter Riemann (2013). "REM sleep dysregulation in depression: state of the art". In: *Sleep medicine reviews* 17.5, pp. 377–390.

Mikus, Adam, Mark Hoogendoorn, Artur Rocha, Joao Gama, Jeroen Ruwaard, and Heleen Riper (2018). "Predicting short term mood developments among depressed patients using adherence and ecological momentary assessment data". In: *Internet interventions* 12, pp. 105–110.

Moore II, Elliot, Mark A Clements, John W Peifer, and Lydia Weisser (2008). "Critical analysis of the impact of glottal features in the classification of clinical depression in speech". In: *IEEE transactions on biomedical engineering* 55.1, pp. 96–107.

Low, Lu-Shih Alex, Namunu C Maddage, Margaret Lech, Lisa B Sheeber, and Nicholas B Allen (2011). "Detection of clinical depression in adolescentsâĂŹ speech during family interactions". In: *IEEE Transactions on Biomedical Engineering* 58.3, pp. 574–586.

Zhou, Guojun, John HL Hansen, and James F Kaiser (2001). "Nonlinear feature based classification of speech under stress". In: *IEEE Transactions on speech and audio processing* 9.3, pp. 201–216.

Teager, H (1980). "Some observations on oral air flow during phonation". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28.5, pp. 599–601.

J. R. Deller J. G. Proakis, J. H. Hansen (1999). *Discrete Time Proc. Speech Signals*.

France, Daniel Joseph, Richard G Shiavi, Stephen Silverman, Marilyn Silverman, and M Wilkes (2000). "Acoustical properties of speech as indicators of depression and suicidal risk". In: *IEEE transactions on Biomedical Engineering* 47.7, pp. 829–837.

Silk, Jennifer S, Erika E Forbes, Diana J Whalen, Jennifer L Jakubcak, Wesley K Thompson, Neal D Ryan, David A Axelson, Boris Birmaher, and Ronald E Dahl (2011). "Daily emotional dynamics in depressed youth: A cell phone ecological momentary assessment study". In: *Journal of experimental child psychology* 110.2, pp. 241–257.

Goldberg, Lewis R (1992). "The development of markers for the Big-Five factor structure." In: *Psychological assessment* 4.1, p. 26.

Yesavage, Jerome A, Terence L Brink, Terence L Rose, Owen Lum, Virginia Huang, Michael Adey, and Von Otto Leirer (1982). "Development and validation of a geriatric depression screening scale: a preliminary report". In: *Journal of psychiatric research* 17.1, pp. 37–49.

Devi, Satyabati and T A.V. Murthy (2005). *The need for digitization.*

Cauffman, Elizabeth, Frances J Lexcen, Asha Goldweber, Elizabeth P Shulman, and Thomas Grisso (2007). "Gender differences in mental health symptoms among delinquent and community youth". In: *Youth Violence and Juvenile Justice* 5.3, pp. 287–307.

Weissman, Myrna M, Diane Sholomskas, Margaret Pottenger, Brigitte A Prusoff, and Ben Z Locke (1977). "Assessing depressive symptoms in five psychiatric populations: a validation study". In: *American journal of epidemiology* 106.3, pp. 203–214.

Huang, Chien-Ming and Bilge Mutlu (2014). "Learning-based modeling of multimodal behaviors for humanlike robots". In: *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. ACM, pp. 57–64.

Huang, Chien-Ming and Bilge Mutlu (2012). "Robot behavior toolkit: generating effective social behaviors for robots". In: *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, pp. 25–32.

Triandis and Harry Charalambos (1994). *Culture and social behavior*. McGraw-Hill New York.

Li, Dingjun, PL Patrick Rau, and Ye Li (2010). "A cross-cultural study: Effect of robot appearance and task". In: *International Journal of Social Robotics* 2.2, pp. 175–186.