Attention Level Approximation of a Conversation in Human-Robot Vocal Interaction Using Prosodic Features of Speech

Sajila D.Wickramaratne and A. G. Buddhika P. Jayasekara
Robotics and Control Laboratory
Department of Electrical Engineering
University of Moratuwa
Moratuwa 10400, Sri Lanka
sajila@elect.mrt.ac.lk and buddhika@elect.mrt.ac.lk

***Abstract*—With the widespread use of robots and integration of robots into the daily human life, the communication abilities of the robots are highlighted in order to ensure smooth interaction. Domestic robots with conversation capabilities are getting an increasing demand due to the fact that modern lifestyles are driving elderly community into isolation. One of the requirements of a conversation system is to ensure that all the people involved are interested in carrying out the conversation. Humans use various physical parameters in order to determine the interest level of the speaker to ensure the conversation is interesting for both parties. The interest level estimation can be used by the robot as a form of feedback to evaluate the effectiveness of the conversation system. In order to evaluate the attention level of the user a system using the prosodic features of speech is proposed. Prosodic features can be associated with the auditory and acoustic measures. These features are important in indicating emotions and attitudes of the speaker. By using the proposed system the robot will be able to determine the interest level using the responses given by the user to the robot by analyzing the prosodic parameters of the vocal response.**

***Keywords*—human-robot interactions, prosody, speech, affective states, interaction decisions**

I. Introduction

The rapid growth of the aging population in the recent years has posed a significant challenge to elderly care sector [1]. With the professionals involved in elderly care not increasing at the rate of population growth there will be a shortage of caretakers. Further the large number of elderly population will face isolation due to inadequate caretakers. Isolation among elders can lead to deterioration of mental health which can result in depression. Intelligent domestic service robots can be used in elderly care as assistants to care taking professionals or to support the elders for independent living in their own home environment [2]. Both physical and cognitive support for the elderly can be given using domestic service robots with socially communicative capabilities [3].

The intelligent robots that interact with the humans should possess the ability to communicate in a more human-like manner to facilitate smooth bidirectional communication. The service robots are considered as autonomous agents which are designed to carry out complex and unstructured tasks while following the social norms of humans [4]. In recent times several researches have been carried out to integrate a vocal interaction system for domestic robots to enhance human-robot interaction by increasing human likeness. One of the characteristics which is expected from the conversation system includes affective interaction [5]. Affective interaction ensures that the conversation system is aware of the emotions of the other party in a conversation. One of the important aspects of a conversation is to make sure that all the parties are equally interested in the conversation and can also be used as a form of feedback to self evaluate the performance of the vocal interaction system.

Emotions play a prominent role in the decision making process of humans. If the robots need to coexist with humans it is important for the robots to understand the emotions of the human and adapt the behavior accordingly. One of the guidelines that studies suggest to be met by future designers of social robots to be suitable for long term interaction regarding affect is for the robots being able to show affective qualities and empathy [6]. This guideline recommends that the robot should be able to understand the user's affective state and react accordingly. In order to improve affective interaction the robot should have a methodology to determine the affective state of the user. Apart from vocal interaction the behavior of the robot when performing other household tasks can also be planned or modified according to the emotional state of the human.

Emotions by nature are defined as complex, fuzzy and indeterminate construct. Emotional Classification is required to distinguish one emotional state from another. Emotions can be described either as discrete or with multiple dimensions. The discrete emotional theory focuses on the the set of six basic emotions which were proposed by Ekman [7]. The significance of this set of emotions is that they are recognizable across different human cultures and easily distinguishable by individuals facial expressions. Dimensional theories use one or more dimensions to represent an emotional state,unlike the discrete theory it can be used to represent a wide array of

978-1-5090-6491-5/17/$31.00 ©2017 IEEE

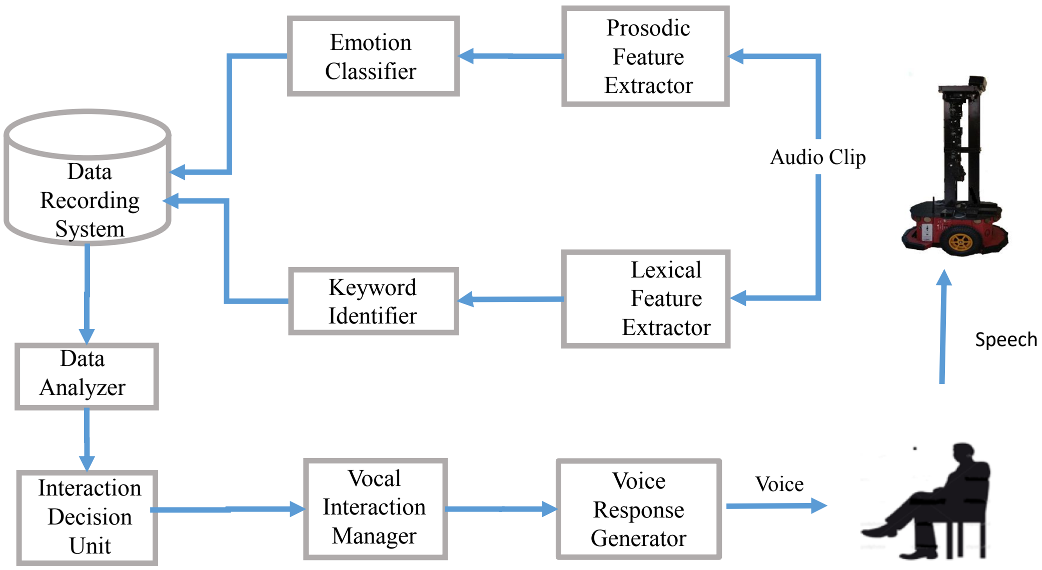


Fig. 1. System Overview

emotions [8]–[10].

Researchers have developed techniques to determine the affective state using facial expressions, voice, body language, physiological activity and brain imaging. Emotions can also be represented as a fusion of several physiological and behavioral responses, hence can use a multi modal approach to identify them [11].

Speech is one of the most important channels of communication and is relatively easier to process than other input modalities. Hence extracting affective information from audio data has been one of the first methods used. Audio based affect recognition has used a wide array of features including prosody, lexical cues and Mel Frequency Cepstral Coefficents [12]. Even though literature shows a range of features which can be used for affect recognition the optimal set of features to determine the affective state is not conclusive. Emotional classification are based on methods such as fuzzy rules,decision trees,state vector machines and Hidden Markov Models [13]–[15]. The emotion classification is done in both speaker dependent mode and speaker-independent mode.

The systems that are mentioned above have evaluated the instantaneous emotional state. For a human-robot conversation system merely recognizing the emotional state is not enough to improve its performance and adapt to the user for long term interaction. This paper presents a model to recognize the affective state and use the information to estimate the enthusiasm of the user to converse based on the user's vocal response. The level of enthusiasm will be used to determine whether the vocal interaction should proceed or not. In Section II the system overview is presented. Section III illustrates how the extracted prosodic parameters are analyzed to understand the user's enthusiasm about the conversation. The experiment results are discussed in the Section IV. The final conclusion presented in the Section V.

II. System Overview

The functional overview of the proposed system is given in Fig. 1. This system is used to evaluate the user engagement level of a human-robot vocal interaction system. Both temporary and persistent emotional states are analyzed by the system. The final objective of the system is to change the topic of the conversation if the user engagement is at a low value persistently. The system takes the audio waveform file of the users utterance as the input for the system. The system then feeds this audio waveform separately into the prosodic feature and the lexical feature extractor. The extracted prosodic and lexical features are the passed to the emotional classifier.

The emotional classifier will use a neural network based system to determine the emotional state. The lexical feature analyzer will also determine the key phrases of the users utterance. The emotional state and the key phrases are then stored in the data recording system. The data is then passed to the analyzer which will compare the received data to historical data and decide whether the emotional state is temporary or persistent.

III. Approximation of Attention Level

A. Rules for the Dialog System

The characteristics of the human-robot dialog system is based on natural human-human dialog systems. The human-human conversation is an activity done jointly with two or more interlocutors. The conversation is built up by a number of consecutive turns each of which include one or more utterances. Turn is a joint activity between the speaker and

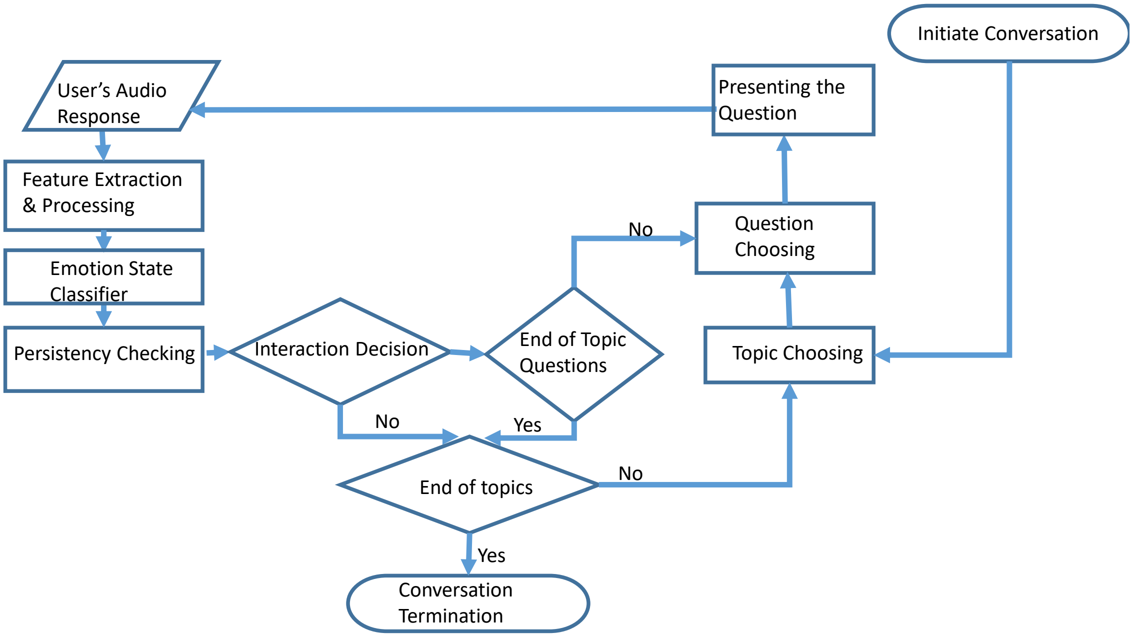


Fig. 2. Functional flow of the system implementation

the listener. When the speaker is presenting an utterance, the listener reasons towards the speakers intended meaning using the specific details presented and prior experience.

The dialog system that is been used for this experiment is made specifically for two party conversation. Therefore the simple turn taking rule applies, in which the human and the robot takes turns one after another. The amount of silent time between the turns is kept as minimum as possible. In case the dialog system takes excessive amount of time to produce a response the human operator interferes with the operation and propose a response. This is done in order prevent the user from misinterpreting the silence as a significant silence which can be interpreted as refusal to respond.

The dialog system has a set of predefined questions belonging to a set of topics chosen prior to the experiment after discussing with the participants. The questions are presented in an order until the system detects that the participant has lost the interest in the conversation. Then the system can either change the topic or abort the conversation depending the level of enthusiasm.

B. Layered Structure in Operation

The system operates with several layers starting from the bottom most layer consisting with raw acoustic data to the upper most layer where interaction decisions are made as shown in Fig. 3. The data collection begins as soon as the robot's turn ends in the conversation. The acquisition of audio data ends when the turn of the speaker ends and is indicated by the human operator. At Layer 0 consists of the raw data collected from the audio system.The raw data is then preprocessed in order to reduce the disturbances. Preprocessing is done before parameter extraction to improve the quality of the audio data by reducing noise.

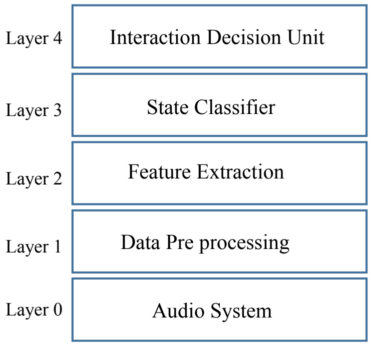


Fig. 3. Layered Structure in evaluation of extracted parameters

Prosodic parameter extraction layer uses the preprocessed data from layer 1. Layer 3 produces a single feature vector by concatenating all the extracted values for the parameters. The emotion classifier is fed with the feature vector. The emotion classifier will then determine the instantaneous emotional state. The emotion classifier uses a neural network based approach. The Interaction Decision Unit (IDU) will finally determine how the conversation should proceed or aborted. IDU uses a rule base to determine the user's interest to converse. If the IDU decides to continue the conversation the dialog system will receive the indication and it will proceed with the next question of the set. Fig. 2 illustrates the functional flow of the vocal interaction system when the system is implemented.

C. Extraction of Prosodic Features

Prosody is associated with larger units of speech or sequence of speech sounds. Prosody can be used to understand

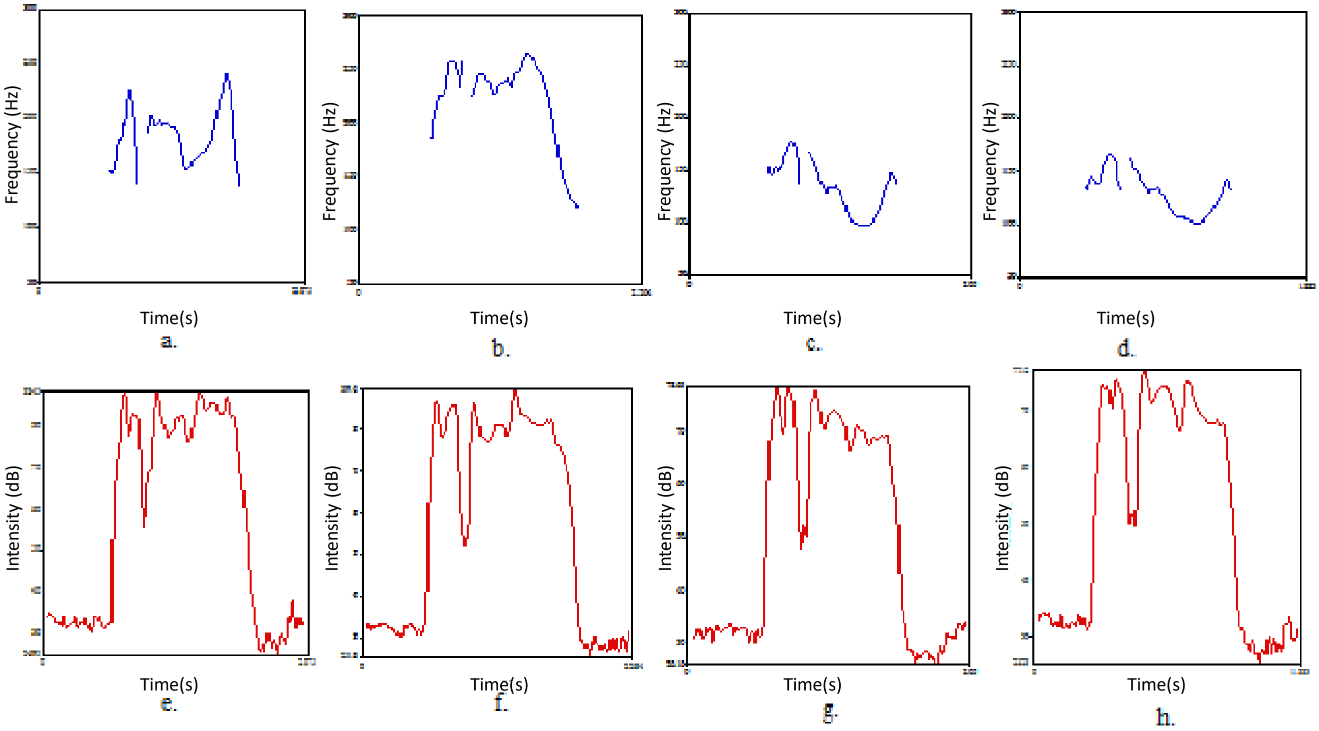


Fig. 4. Pitch contours(upper row) and Intensity Contours(lower row) for the four affective states when the user presents the same utterance a) & e) Angry b) & f) Happy c) & g) Bored d) & h) Calm

the features related to the speaker such as emotional state or related to the utterance such as the form. Prosody is important in recognizing emotions that can be linguistically difficult to distinguish.

Prosodic parameters can be analyzed by using either auditory or acoustic measures. The auditory measures are associated with the impressions produced by the listener's mind and can be measured using auditory scales such as mel scale. The acoustic measures can be obtained through the physical properties of the sound wave. In this model only acoustic parameters are considered which are given below .

- Maximum Pitch (Hz)
- Minimum Pitch (Hz)
- Mean Pitch (Hz)
- Speech Rate (syllables per minute)
- Amplitude Root Mean Square (Pa)
- Mean Power (dB)

The prosodic parameters are extracted using the Praat software for each sound wave file [16]. The extracted parameters are then concatenated to form a single feature matrix.

D. Classification of Affective State

The classification of affective states is done using a neural net. The affective states that the neural net is trained to identify are happy, angry, calm and bored. These affective states are located in the four quadrants of the Circumplex Model of Affect [10]. The categorized audio clips from standard data bases were used for the initial training of the neural net. The extracted feature matrix and the expected affective state were given as the input and the output of the neural network. The classification of affective state is performed for every turn of the user in the conversation.

The advantage of using a prosodic feature based method over a lexical feature based method is that the users can use the same utterance in different affective states. Hence affective classification cannot be done purely based on lexical features. As Fig. 4 shows the pitch and intensity contours of 4 audio samples that were obtained during the experiment. These represents four affective state even though the utterance is same.

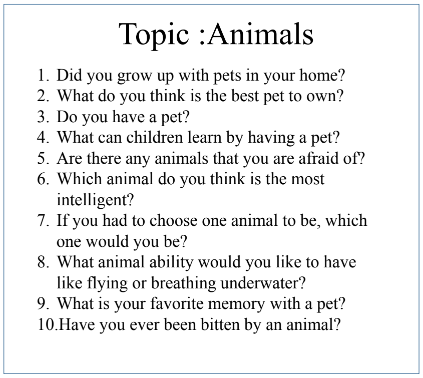
E. Interaction Decision Determination

The affective state for the respective turn of the conversation is stored in a database. It is used to determine whether the affective states are persistent or temporary. If the affective state is seen in three of the last four turns the affective state is considered persistent or else it is considered as a temporary state. The rules that are used in the conversation decision unit are given in Table I. These rules are applied until all the system has covered all the three topics and come to the end of the conversation.

Happy and calm states are considered as positive states which shows that the user is interested the conversation. Angry and Bored states are considered negative states. Hence those states indicate lack of interest in the conversation. It is important to note that the system immediately changes the

TABLE I
RULES SYSTEM FOR THE CONVERSATION DECISION SYSTEM

Affective State	Persistent or Temporary	Action
Happy	Persistent	Continue the conversation
Happy	Temporary	Continue the conversation
Calm	Persistent	Continue the conversation
Calm	Temporary	Continue the conversation
Bored	Persistent	Topic Change/Termination
Bored	Temporary	Continue the conversation
Angry	Temporary	Topic Change/Termination



Topic :Animals

1. Did you grow up with pets in your home?
2. What do you think is the best pet to own?
3. Do you have a pet?
4. What can children learn by having a pet?
5. Are there any animals that you are afraid of?
6. Which animal do you think is the most intelligent?
7. If you had to choose one animal to be, which one would you be?
8. What animal ability would you like to have like flying or breathing underwater?
9. What is your favorite memory with a pet?
10. Have you ever been bitten by an animal?

Fig. 5. Question set used for the topic animals

topic when angry state is detected and hence a persistent angry state does not exist. Although bored state is considered negative, behavior modification does not need to happen unless the state is persistent.

IV. RESULTS AND DISCUSSION

A. Research Platform

The concept has been implemented on MIRob platform with a Microsoft Kinect sensor attached [17]. The required navigation maps are created with Mapper3 Basic software. The navigation maps are used to keep the set social distance required for the conversation. The audio clips were captured using the array of microphones in the Kinect. Further a voice synthesis unit is used to produce the vocal response from robot. The experiment was carried out in a simulated domestic environment. Fig. 6 illustrates a participant engaged in vocal interaction with the robot during the experiment.

B. Experiment

The robot and the participant were kept at a distance of 1m apart facing each other. Prior to the experiment each of the participants were given the opportunity to pick 2 topics from a given pool of topics. Each of these topics have a set of 10 questions associated with it. Fig. 5 illustrates a question set that was used in the experiment. Once the topics are chosen they are given as inputs along with the vocal interaction system of the robot. The system will choose another random topic from the rest.

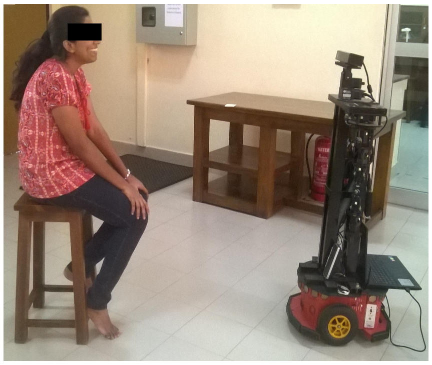

[redacted]

Fig. 6. A participant and MIRob platform during the experiment

The system will initiate the conversation with a friendly greeting. The robot will start randomly with one of the 3 selected topics. The robot and the participant will take turns one after the other until the conversation decision system indicates the user has lost interest of the conversation. When the user has lost the interest the system will switch to a new topic. This process happens until the all the 3 topics are used in the conversation.

C. Results

The experiment was conducted with the participation of 20 individuals in the age range of 18-58 (Mean-30.8 and Standard Deviation-11.47). Before the experiment 12 audio samples were obtained from each of the participant with 3 each displaying the 4 affective states mentioned in Section D. The sample audio clips used to train the emotion classification neural network were obtained from the participants as well as the audio clips obtained from the participants and Surrey Audio Visual Expressed Emotion Database [18].

Throughout the conversation the user had to converse about 3 topics. Within the conversation there will be two topic changes and finally the termination of the conversation. After the conversation each of the participant was asked whether they are satisfied about when the changes and termination happened in the conversation.

Table II presents a sample of the results obtained during the experiment of 6 participants. A combined total of 60 topic changes and terminations happened through the experiment. The average number of turns a participant had was 6 per topic. Further the average number of participant turns for the topic chosen by the system was 4.65 turns per topic. Of all the changes and terminations 70.01% of them was considered satisfactory by the participants. A total of 16 participants of the total considered the overall system to be satisfactory.

The reasons for unsatisfactory performances as given by the participants are as follows

- Asking more questions than anticipated from the random topic chosen by the system.
- System not being able to identify some of the sarcastic answers given by the participants to convey displeasure.
- Changing the topic even when the participant is still interested in the conversation

One of the reasons which might have caused unsatisfactory performance is due to the fact anger and happy states having

TABLE II
RESULTS OF THE EXPERIMENT

User No.	No of turns before change /termination			User Reaction			Overall System Satisfaction
	Topic 1	Topic 2	Topic 3	Topic change 1	Topic change 2	Termination	
1	6	4	8	satisfied	unsatisfied	satisfied	satisfied
2	5	5	7	satisfied	satisfied	unsatisfied	satisfied
3	4	6	6	satisfied	satisfied	satisfied	satisfied
4	6	8	5	satisfied	unsatisfied	satisfied	unsatisfied
5	9	5	7	satisfied	satisfied	satisfied	satisfied
6	7	5	6	unsatisfied	unsatisfied	unsatisfied	unsatisfied

similar acoustic properties. Although these two states can be easily distinguished through visual means such as facial emotion recognition, in acoustic measures the high excitement and high annoyance states both have nearly similar values for prosodic parameters. Hence it is highly likely that these two states being misclassified which can produce two completely different results in the system.

Further the use of a static value of 3 out of the last 4 turns being the same affective state considered as a persistent affective state should change according to the individual. The participants suggested that this value does not provide a satisfactory performance when the conversation is about an unfamiliar topic such as the random topic generated by the system. The user had to carry on the conversation for at least 4 turns with a negative affective state in order to change the topic which participants suggested as a lengthy conversation than expected on a topic not favorable for conversation.

V. CONCLUSION

A method was introduced to estimate the interest level of a person who is engaged in a conversation with a human-robot vocal interaction system. The system extracts the prosodic features of the vocal response of the user. The affective state is then determined using a trained neural network. Before further proceeding with the conversation the system should decide whether the user is interested in the conversation. The decision is based on the fact whether there is a persistent negative affective state. A persistent low attention level will result in either change of the conversation topic or termination of the conversation.

The system proposed can be used in a domestic service robot to improve vocal interaction capabilities as well as affective qualities. The ability of the system to adapt over the time will enhance the overall performance by presenting a more personalized behavior. As of future improvements for this system the addition of semantic information of the speech can be considered to improve the overall accuracy of determination of the interest level and affective state of the user. Even further by using a multi modal approach by including a visual channel, the system will be able to determine the attention level of the user when both listening and speaking.

ACKNOWLEDGMENT

This Research is supported by University of Moratuwa Senate Research Grant No: SRC/CAP/2017/03.

REFERENCES

[1] "World population ageing 2015," Population Division, Department of Economic and Social Affairs, United Nations, ST/ESA/SER.A/390, 2015.

[2] D. O. Johnson, R. H. Cuijpers, J. F. Juola, E. Torta, M. Simonov, A. Frisiello, M. Bazzani, W. Yan, C. Weber, S. Wermter *et al.*, "Socially assistive robots: a comprehensive approach to extending independent living," *International journal of social robotics*, vol. 6, no. 2, pp. 195–211, 2014.

[3] E. Broadbent, R. Stafford, and B. MacDonald, "Acceptance of healthcare robots for the older population: review and future directions," *International Journal of Social Robotics*, vol. 1, no. 4, p. 319, 2009.

[4] D. Feil-Seifer and M. J. Matarić, "Socially assistive robotics," *IEEE Robotics & Automation Magazine*, vol. 18, no. 1, pp. 24–31, 2011.

[5] N. Mavridis, "A review of verbal and non-verbal human–robot interactive communication," *Robotics and Autonomous Systems*, vol. 63, pp. 22–35, 2015.

[6] I. Leite, C. Martinho, and A. Paiva, "Social robots for long-term interaction: a survey," *International Journal of Social Robotics*, vol. 5, no. 2, pp. 291–308, 2013.

[7] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[8] R. Plutchik, "The nature of emotions human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice," *American scientist*, vol. 89, no. 4, pp. 344–350, 2001.

[9] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of research in Personality*, vol. 11, no. 3, pp. 273–294, 1977.

[10] J. A. Russell, "Core affect and the psychological construction of emotion." *Psychological review*, vol. 110, no. 1, p. 145, 2003.

[11] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Transactions on affective computing*, vol. 1, no. 1, pp. 18–37, 2010.

[12] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[13] A. Austermann, N. Esau, L. Kleinjohann, and B. Kleinjohann, "Prosody based emotion recognition for mexi," in *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2005.(IROS 2005).* IEEE, 2005, pp. 1138–1144.

[14] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog." in *INTERSPEECH*. Citeseer, 2002.

[15] L. Devillers and L. Vidrascu, "Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs." in *Interspeech*, 2006.

[16] P. P. G. Boersma *et al.*, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, 2002.

[17] M. A. V. J. Muthugala and A. G. B. P. Jayasekara, "Mirob: An intelligent service robot that learns from interactive discussions while handling uncertain information in user instructions," in *Moratuwa Engineering Research Conference (MERCon), 2016*. IEEE, 2016, pp. 397–402.

[18] S. Haq and P. J. Jackson, "Multimodal emotion recognition," *Machine audition: principles, algorithms and systems*, pp. 398–423, 2010.