Emotional Speech Recognition toward Modulating the Behavior of a Social Robot

Chris LYTRIDIS, Human-Machines Interaction Laboratory (HUMAIN-Lab), Department of Computer and Informatics Engineering, Eastern Macedonia and Thrace Institute of Technology (EMaTTech), Kavala, Greece, lytridic@teiemt.gr
Eleni VROCHIDOU, HUMAIN-Lab, Department of Computer and Informatics Engineering, EMaTTech, Kavala, Greece
○ Vassilis KABURLASOS, HUMAIN-Lab, Department of Computer and Informatics Engineering, EMaTTech, Kavala, Greece

The effectiveness of human-robot interaction depends on the capacity of a robot to modulate its behaviour according to human emotion(s). This work presents preliminary results regarding a Behaviour Modulation System (BeMoSys) implementable on a social robot with a capacity for emotional speech recognition based on signal-processing followed by machine-learning techniques.

***Key Words*:** Emotional speech recognition, Human-robot interaction, Robotic behaviour, Social robots

1. Introduction

A more feasible incorporation of social robots in domains such as education, special treatment and therapy, requires designing autonomous robots that can interact socially with individuals. Autonomy in social robots have only been partly achieved, since it is difficult to achieve robot control that can handle complex, dynamic and unpredictable situations such as those found in typical real-life environments. Interaction with autonomous robots that sense and respond to user emotional state and behaviour is an emerging field of scientific interest. By developing a robot control architecture for specific applications such as therapy, researchers can realise emerging robot behaviours in response to appropriate human stimuli, which is very important for social assistive applications. Such control architectures must include sensing modules that can interpret the emotional state, mood and intentions of humans by observing their behaviour. Sensing modules such as blood pressure, pulse, skin conductance and brain activity may be psychological detectors [1]. However, such sensors are difficult to be applied to children or people with special needs who are sensitive to touch, although they provide more precise information than distance-based estimations of emotional and mental state. Therefore, cameras to observe and analyse gestures [2] and microphones to detect speech emotional content [3], are alternatively employed.

This work proposes a novel behaviour modulation system for social robots based on emotional speech recognition. Human emotions are detected using extracted linguistic features to bias the robot towards appropriate behaviours. The proposed method is divided in three phases; a) emotional speech recognition based on speech signal processing, b) machine learning for emotional state classification, and c) mapping of emotions into appropriate robotic behaviours.

The paper is organized as follows. The proposed methodology is presented in Section 2. Sections 3 and 4 describe the emotional speech recognition process and the utilized machine learning algorithm of the BeMoSys, respectively. Results and possible future directions of this research are presented in Section 4. Finally, Section 5 summarizes the conclusions of the paper.

2. The Proposed Methodology

The proposed methodology involves three stages. The first stage is feature extraction from the voice signal. The second stage is signal classification to various emotional states. Finally, the proposed BeMoSys is formed according to the detected emotional state. The overview of the proposed closed-loop interaction scheme between human and a robot is illustrated in Fig. 1.

While the human and the robot interact, human speech is recorded by the robot. This is converted to a signal in order to extract emotional speech features, which are then fed as inputs in the classification model of the second stage. This model determines the perceived emotional cues and returns this information as an output. Emotional information is then used by a fuzzy inference system to determine the adopted robotic action. Robot modulates its behaviour according to the received emotional information, and as human-robot interaction continues, human behaviour is in turn influenced by the robot's behaviour and the closed-loop interaction cycle continues for the entire engagement session.

3. Emotion Recognition

Emotional speech recognition involves two different aspects; the syntactic-semantic part of the talking (what the person is literally saying), and the paralinguistic information of the speaker (tone, emotional state and gestures). For the emotional speech recognition, Vokaturi software is used [4]. Vokaturi is a novel state-of-the-art software in emotion recognition from the human voice. Introduced in 2016, it is developed in several libraries, in C and Python, so as to be easily integrated in several applications. Moreover, it is validated with existing emotion databases and works in a language-independent manner.

For prosodic detection version OpenVokaturi 2.2b was used, to extract the primary dimensions of pitch, intensity, and spectral slope when detecting emotion. The nine extracted features are compared to two databases of audio recordings to predict five emotions; happiness, sadness, anger, fear, and neutrality. Predictions are returned as values that represent the probability of each emotion

being present. The highest value among all predictions, points out the dominant emotion of the audio recording. The detection of emotion using this technology has been cross validated as 66.5% accurate.

4. The proposed Behaviour Modulation System

The proposed BeMoSys is based on a Fuzzy Inference System (FIS). A FIS essentially defines a nonlinear mapping of the input data vector into a scalar output, utilizing fuzzy rules [5]. The mapping process involves input/output membership functions (MFs), Fuzzy Logic (FL) operators, fuzzy if-then rules, aggregation of output set and defuzzification. In this work, there are five inputs to the FIS which correspond to the five emotions extracted from the speech signal. The knowledge base of the inference system consists of 47 rules, which map the inputs to a single output value referring to the level of estimated happiness.

Based on the value of estimated happiness, one of five pre-defined robotic behaviours is selected. Each robotic behaviour is defined by combining three categories of robot actions, namely facial expressions, sounds and animations. Facial expressions refer to the eye LEDs of the robot that can change colours and blink in different frequencies. Sounds refer to various happy, relaxing audio tracks and music. Animation refers to gestures and postures, involving movements of the robot's head, arms and legs. The action categories and possible individual robot behaviours are summarized in Table 1.

These individual robot actions are combined and form the five pre-set distinct robot behaviours, A to E, summarised in Table 2. Behaviour A is the robot's response to lowest levels of estimated happiness and behaviour E is the robot's response to the highest levels of estimated happiness. These behaviours are formed in such a way that they enable the robot to respond differently to the various affective human states. The main goal is to select appropriate behaviours that would calm and reassure humans and lead them into a more pleasant and engaged interaction [1]. Behaviours from A to E, are gradually driven from being condescending and reassuring, to being cheerful and enthusiastic. Thus, negative emotions detections would activate robot behaviours that tend to gradually improve the emotional state of the human. On the other hand, positive emotions detection would activate robot behaviours that tend to reward and cheer up the human.

5. Results and Discussion

BeMoSys was implemented and tested for different emotional scenarios. More specifically, for the verification of the reliability of the proposed method, 399 speech recordings retrieved from the Berlin Database of Emotional Speech [6] were used; 74 recordings for neutrality, 71 for happiness, 126 for anger, 67 for fear and 61 for sadness. For each recording, the resulting emotional context of the speech signal was then supplied to the FIS, in order to derive the estimated level of happiness. According to this value, the appropriate robot behaviour was selected.

Table 3 shows the occurrence of different behaviours that were selected by the BeMoSys according to each detected emotional state.

It can be seen in Table 3 that the proposed BeMoSys appears to be adaptable and consistent, since according to the detected emotion, it leads towards a specific different behaviour in most of the cases. For detection of negative emotions, the BeMoSys leads more to behaviours A and B, while for detection of positive emotions leads more to behaviours D and E. The mismatches between emotion and selected behaviour that are observed, can be attributed to the inherent inaccuracy of emotion detection of OpenVokaturi library.

A graphical illustration of the experimental results of Table 3 is provided in Fig.2. As it can be observed, for extreme values of the FIS output, the BeMoSys appears very consistent. When the detected emotional state is predominantly anger, sadness or fear, then the system is driven to lower values of happiness level, and therefore triggers the appropriate robot behaviours. On the other hand, clearly happy emotional input, leads to the activation of reinforcing robot behaviours. Finally, it must be noted that neutrality can be interpreted both as a negative or a positive feeling for the proposed system and so it mainly activates behaviours A and E, for the emotional scenarios under study.

The promising preliminary experimental results motivated the authors to implement the BeMoSys on the NAO social robot. Emotion detection is achieved by integrating the emotional speech recognition library and the FIS to the robot's software. In this setup, instead of using prerecorded audio samples, as done previously, NAO's microphones are now used to record human voice. The robot evaluates the voice's emotional content and executes the appropriate behaviours.

6. Conclusions

The effectiveness of human-robot interaction depends on the ability of the social robot to adapt its behaviour according to human emotional states. This paper is the first step towards the development of more intelligent and autonomous social robots based on emotional speech recognition. It presents a behaviour modulation system implemented on the social robot NAO. The robot records the human voice during interaction and extracts the emotional content with the help of a well-known software application. Extracted emotions act as inputs to a FIS in order to map emotions to robotic actions. Preliminary results are encouraging, and future work will focus on applying the proposed system in the areas of special treatment and education.

The potential impact of a robot that can detect the emotional states of a child/adult with special needs (such as autism) and interact with him/her based on its perception offers numerous opportunities. Complex social stimuli, more engaged participants, various interactions and flexibility to unpredictable situations can be introduced as an extension to the present work. By recording emotional changes in a child's voice, one can create personalized training sets, to improve the robot's perception of individual characteristics and adopt different ways of responding according to the needs of different children. A multidimensional system that can combine emotional speech data with facial expressions and gestures, could improve the accuracy of classification significantly. Enrichment of the robotic actions could additionally enhance the diversity in behaviours.

ACKNOWLEDGEMENT

This work has been supported, in part, by the European Commission Horizon 2020 MSCA-RISE Project no. 777720 "Cyber-Physical Systems for PEdagogical Rehabilitation in Special EDucation (CybSPEED)".

REFERENCES

[1] Liu, C., Conn, K., Sarkar, N., and Stone, W., "Online affect detection and robot behavior adaptation for intervention of children with autism," IEEE transactions on robotics, vol. 24, no. 4, pp. 883–896, 2008.

[2] Feil-Seifer, D., and Matarić, M. J., "Automated detection and classification of positive vs. negative robot interactions with children with autism using distance-based features," In Human-Robot Interaction (HRI), 2011 6th ACM/IEEE International Conference on. IEEE, pp. 323–330, 2011.

[3] El Ayadi, M., Kamel, M. S., and Karray, F., "Survey on speech emotion recognition: Features, classification schemes, and databases," Pattern Recognition, vol. 44, no. 3, pp. 572–587, 2011.

[4] Vokaturi Homepage. Available on: https://developers.vokaturi.com [Last Accessed: 05/02/2018]

[5] Kulkarni, A. D., "Computer vision and fuzzy-neural systems," Prentice Hall PTR, 2001.

[6] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., and Weiss, B., "A database of German emotional speech," In Ninth European Conference on Speech Communication and Technology, 2005.

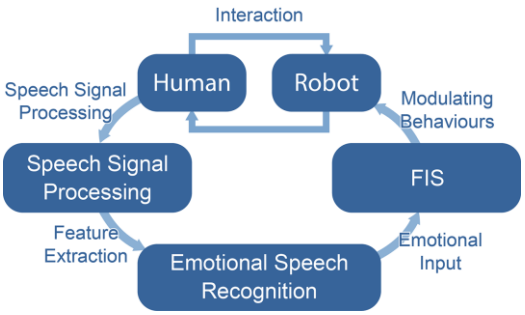

Fig. 1: The proposed framework.

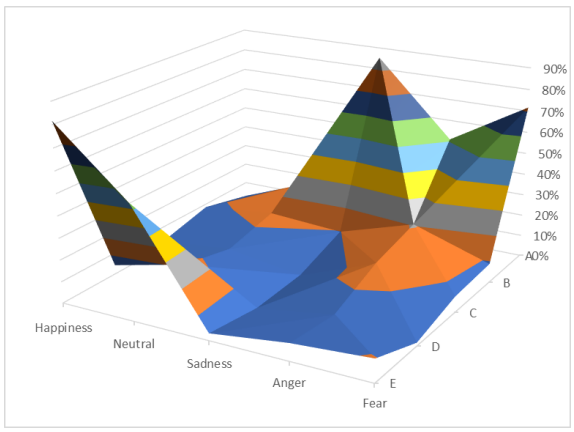

Fig. 2: Graphical distribution of behaviours according to calculated happiness levels.

Table 1: Categories of robot actions and possible individual behaviours.

Action category	**Individual Behaviours**
Facial expressions	Eye LEDs (colours, blinking frequency)
Sounds	Happy sounds Relaxing sounds Music
Gestures/Postures	Head movements Arm movements Leg movements

Table 2: Pre-set robot behaviours.

Behaviour	**Description**
A	Eye LEDs (colour set to white); Relaxing sound; Relaxing animation
B	Eye LEDs (blinking); Relaxing sound; Animation to attract attention
C	Eye LEDs (rapidly blinking); Cheerful sound; Zestful animation
D	Eye LEDs (rapidly alternating colours); Enthusiastic sound; Cheerful animation
E	Eye LEDs (rapidly alternating colours); Upbeat music; Enthusiastic animation

Table 3: Occurrence of different behaviours according to detected emotion.

Behaviour / Emotions	**A**	**B**	**C**	**D**	**E**
Happiness	3%	8%	3%	4%	82%
Sadness	85%	10%	2%	0%	3%
Anger	51%	21%	15%	6%	8%
Fear	72%	21%	15%	6%	8%
Neutrality	72%	9%	7%	1%	53%