PYHRE: PSYCHOACOUSTIC MODEL FOR HUMAN-ROBOT EMOTIONAL INTEGRATION

Fredy H. Martínez S.[1], Faiber Robayo Betancourt[2] and Holman Montiel A.[1]
[1]Facultad Tecnológica, Universidad Distrital Francisco José de Caldas, Bogotá D.C., Colombia
[2]Departamento de Ingeniería Electrónica, Facultad de Ingeniería, Universidad Surcolombiana, Neiva, Huila, Colombia
E-Mail: fhmartinezs@udistrital.edu.co

ABSTRACT

There are still many design problems for service robots. Among the most important problems is human-robot integration, a problem that has many edges, both at the level of physical interaction and at the emotional level. The research group is evaluating different algorithms on its robotic platform ARMOS TurtleBot. Among these algorithms, it recently developed a scheme for the identification of a person's emotions from identifiable facial characteristics in a person's face. Under laboratory conditions, the scheme reached a 92% success rate, however, in low-light conditions or when the person had the face partially covered this rate decreased considerably. Consequently, we propose the design of an alternative loop as a support to increase the success rate by estimating the emotional state of the person from the voice. For this purpose, we train a convolutional neural network with spectral images built from audio characteristics of the same 7 emotions used in the first algorithm. The model achieved a 69% hit rate, and together with our face algorithm raised the total performance of the system to 96.5%.

Keywords: convolutional neural network, emotions, human-robot interaction, image processing, real-time, service robotics, voice intonation, word cadence.

Manuscript Received 14 July 2020; Revised 30 September 2023; Published 10 October 2023

1. INTRODUCTION

The importance of service robotics for the future care of people is undeniable [1]. Our society is increasingly inclined to social models in which elderly people and children spend a lot of time alone in their homes without the care of others, people who mostly spend most of their time in work activities to provide their homes with daily sustenance [2]. This social model is evident in all societies, regardless of their level of development or political structure. These applications of the service robots can be mixed with tasks such as support in training and learning processes, physical health trainers, and health monitoring [3, 4]. However, robots as artificial systems must have a high level of integration with humans to successfully perform all these tasks [5, 6].

The processes of human-robot interaction (HRI) and integration do not only refer to the coordination of contents and processes of the artificial system, but they also relate to the degree of empathy that the robot can express towards the human being [7, 8, and 9]. This is to say, a great part of the success of this interaction is related to the capacity of the robot to perceive or infer the emotional state of the person under its care, and from there to coordinate its responses in coherence with that emotional state [10]. This is fundamental in the development of the robot's behaviors and at the same time in the quality of the interaction of the people with the machine, which would promote better development of the task of care by the robot [11].

The emotional state of an individual can be identified through certain characteristics that can be detected by another person [12]. The most important characteristics can be detected in facial expressions. Individuals are socially conditioned to express their emotional state through facial postures, and it has been shown that even isolated populations identify basic emotions such as sadness, anger, surprise, disgust, joy, and fear on their faces. However, despite being a very important means of expression, it is not the only way through which human beings express emotions, people also use the voice (through words and intonation), and other non-verbal means such as sweating and blushing [13].

The traditional strategy used to identify expressions has been the visual identification of specific characteristics in a person's face [14, 15, and 16]. As a basic strategy, it has been very useful to identify the emotion of a person in a fast way, tracing the facial expressions. However, human beings express emotions in very varied and complex ways and with different intensities from individual to individual. In particular, a person not only expresses emotions in his face, but these are accompanied by body movements (body language), words, voice intonation, and word cadence. Many of these characteristics cannot be communicated just by seeing the person's face, or just by hearing it on the phone. As a general rule, it is necessary to identify most of these parameters to establish a person's emotional level.

As far as artificial systems for voice processing are concerned, today we have very high-performance systems capable of recognizing almost in real-time phonetic sounds such as words and phrases, even identifying with a high degree of precision one person from another. These systems, however, are isolated from the visual component and do not deal with relating their results to the facial characteristics of the person during the conversation. The text identified by itself does not possess the emotional information of the person that is present in the audio and therefore fails in emotional identification tasks unless the person explicitly expresses his or her emotions in words [17].

A robot can autonomously sense many parameters around it [18, 19]. The identification of emotions on a person's face has been widely researched using digital cameras image processing, and even deep neural networks [20]. Image processing provides a simple and standardized way of identifying features and classifying them by learning these characteristics [21, 22]. Other signals such as the person's voice can be analyzed using the same strategies as a parallel strategy for identifying emotions [23]. For this strategy, it is assumed that parameters such as voice intonation and word cadence can be represented as images and therefore processed in the same way [24].

It should be noted that in addition to the social conditions of each individual, there are great differences between the emotions expressed by adults and those expressed by children [25]. Many of the systems documented to date consider their application in human-robot interaction with adults. However, service robots, in addition to the elderly, must primarily serve a child population, and the models developed for adults can hardly be adapted to this new task. For example, young children often use tantrums to express their frustration or to get something they want, and this behavior is reflected in the tone and cadence of their words. This is a virtually undetectable feature of an adult's voice, which requires particular modeling and adjustment.

In this paper, we show our proposal of a trainable model for the identification of human emotions that allows us to increase the level of human-robot integration for robotic platforms conceived for service robotics applications. The model seeks to integrate with other strategies of the research group oriented to visual identification of expressions, which is why it was designed with requirements of high performance, low computational cost, and real-time operation.

2. PROBLEM FORMULATION

The research group is developing a robotic platform for applications in the care of small children and the elderly in homes and special care centers. The platform developed is called ARMOS TurtleBot (Figure-1) and has different algorithms for autonomous navigation in unknown environments, identification, and tracking of people in real-time, estimation of the distance to obstacles, and an algorithm for identifying human emotions from the analysis of facial characteristics.

Figure-1. TurtleBot ARMOS mobile platform.

This last algorithm works very well in laboratory conditions (success rate of 92%), but in real operation, it has several flaws, the most important is its inability to identify the emotion of the user when the face is not in front of the camera, and the confusing characteristics of the emotions when there is poor ambient light or the user has glasses or other elements in the face (such as a mask, a feature not initially contemplated, but forced inclusion due to the social effect produced by the COVID-19). In addition, many of the users we tested do not express their emotions clearly on their faces, which also confuses the algorithm.

In principle, we propose the development of a parallel scheme of emotion identification built from the user's vocal characteristics that allows us to increase the performance of our current image algorithm. Consequently, it is required a reliable algorithm, of low computational cost, and the capacity to allow the incorporation of new characteristics from the operational needs of the robot. In addition, the emotions to be identified initially must be the same as characterized by our facial algorithm.

Our problem can be modeled from the functional conditions of the robot. The robot is restricted in motion to free space E defined as a subset of the navigation environment $W \subset \boldsymbol{R}^2$. W is an open set in the plane containing E and an O set of regions inaccessible to the robot called obstacles. For design purposes, W corresponds to the environment found in indoor spaces where people interact, and therefore obstacles can be fixed or mobile (furniture, people, pets, etc.). The obstacles in O are finite in quantity and are detectable by the robot. In any case, E corresponds to the open set of W without the obstacles and represents an area in the plane that allows the mobility of the robot.

The robot does not know W, but it is equipped with different sensors that allow it to know the environment locally. From the observation history, the robot builds an information space that it uses to make information feedback and define its movement in correspondence with a movement policy. We define the information space S from the information mapping developed along the observations in time that is (equation 1):

$$o: [0, t] \rightarrow S \tag{1}$$

This information space is interpreted to produce the information required for the robot's decision-making. According to this nomenclature, the objective of the investigation corresponds to the definition of a filter that can be applied to the vocal signals detected by the robot, processed as images, and coming from the user of the robot that allows identifying emotional states automatically and continuously. The emotional states to be identified are: neutral, happy, sad, angry, fearful, disgusted, and surprised.

3. MATERIALS AND METHODS

This research aims to build a trainable model capable of detecting emotional states in a person, based on the identifiable characteristics of their voice. This model will then become part of a whole human-robot integration system of our robotic platform.

To train the model we use a dataset The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [26]. This dataset consists of 7356 files corresponding to 24 professional actors (12 men and 12 women) vocalizing two lexical statements with a neutral American accent. These recordings contain audio files (phrases and songs) and video, but we use only the audio files and only the phrases. This last subset contains 1440 files (60 trials per actor x 24 actors) in **.wav* format. The voices characterize eight emotional states, one more than those sought in our previous model: neutral, calm, happy, sad, angry, fearful, surprised, and disgusted. However, for our prototype, the first class (neutral) does not exist so this category is not used. This provides the equivalent category system database to our face analysis model.

Each of the files in the database has a unique filename consisting of a numerical identifier. The structure of this identifier denotes the emotional stimulus performed by the actor, and the actor himself. The seven blocks of the numerical identifier are shaped like this:

- Modality (01 = full-AV, 02 = video-only, 03 = audio-only).
- Vocal channel (01 = speech, 02 = song).
- Emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised).
- Emotional intensity (01 = normal, 02 = strong). There is no strong intensity for the 'neutral' emotion.
- Statement (01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door").
- Repetition (01 = 1st repetition, 02 = 2nd repetition).
- Actor (01 to 24. Odd-numbered actors are male, even-numbered actors are female).

Since we are only taking audio with phrases, all files selected for model training have the initial blocks 03-01. The third block of the name identifies the seven tags of our classifier (01 is not used), i.e.:

- 1 - calm
- 2 - happy
- 3 - sad
- 4 - Angry
- 5 - fearful
- 6 - disgust
- 7 - surprised

The other features add more information to the model so they are not considered when organizing the dataset for training and validation. Each class was balanced with 192 files, for a total of 1344 in the seven categories.

To build the model we selected the Dense Convolutional Network (Figure-2). This topology was selected because of its high efficiency in terms of parameters concerning other topologies, and its high classification capacity. This network is designed to classify images in a three-matrix color body, which is why it is necessary to pre-process the audio files into a graphic format that represents the characteristics to be identified (those that allow a human being to characterize the sound) with this structure.

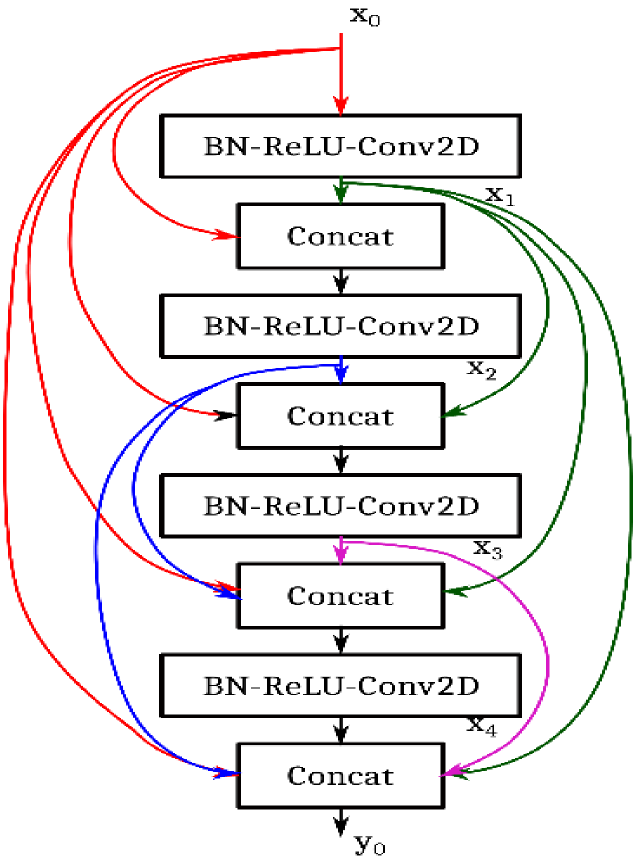


Figure-2. Dense Convolutional Network (DenseNet) architecture.

To convert the audio files to images we choose the MEL scale spectrogram format. The MEL scale is a perceptual musical scale of spaced tones from observers, comparing a tone of 1 kHz above the hearing threshold with a tone of 1000 mels. This scale is based on human perception, so the resulting image includes features related to the particular characteristics of the subject (Figure-3). In the conversion of the audio to images (one figure for each audio file, saved in the folder corresponding to its class label) a length of the FFT window of 1024, and several samples between successive frames of 100 were used.

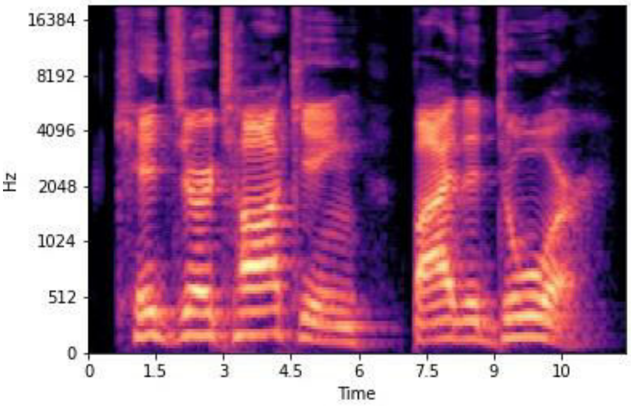


Figure-3. Spectrogram of one of the audios used in the training dataset.

For the DenseNet training, the dataset images were randomly mixed in the data list to improve performance. The images were scaled to 256*256 pixels (the initial image size was 393*258 pixels), this facilitates the design and operation of the network (in fact increases its performance) without loss of information in the images. The matrices corresponding to each color in the images were normalized from 0 to 255 to the range of 0 to 1, values of operation of the neurons.

The dataset was separated into two groups, training, and testing. The separation was done randomly, and 80% of the data was assigned to training, and the remaining 20% to model testing and validation.

4. RESULTS AND DISCUSSIONS

Convolutional neural networks are built with architectures that functionally duplicate the neurons in the primary visual cortex of the brain, which is why they are excellent at image classification and segmentation. They are built with two-dimensional matrices to process information coming from digital images, and in the last years, architectures with very high performance have been proposed.

In particular, we used a DenseNet network of 121 layers deep, these layers were organized in a 5+(6+12+24+16)*2=121 structure, where 5 are (conv, pooling) + 3 transition layers + classification layer. Multiplication by 2 is done because each dense block has two layers (1x1 conv and 3x3 conv). The number of nodes in the input layer is determined by the shape of the images. We scaled all input images to 255*255 pixels and processed them in RGB, so the shape of the input layer was 255*255*3. The number of nodes in the final output layer is the number of possible class labels, in this case, the output layer had seven nodes. With this structure, the neural network required 6,961,031 trainable parameters and 83,648 non-trainable parameters.

As loss function, we use categorical cross-entropy, for optimizer we use stochastic gradient descent, and as metrics to evaluate the performance of the neural model, we calculate the accuracy, recall, f1-score, and support (Figures 4 and 5). Also, to evaluate the model performance, we generate the confusion matrix and ROC curve of the trained network (Figures 6, 7, and 8). The training was carried out during 30 epochs, and it was verified that there was no over-adjustment. We used the Google Colab GPU, and the training took 16 minutes.

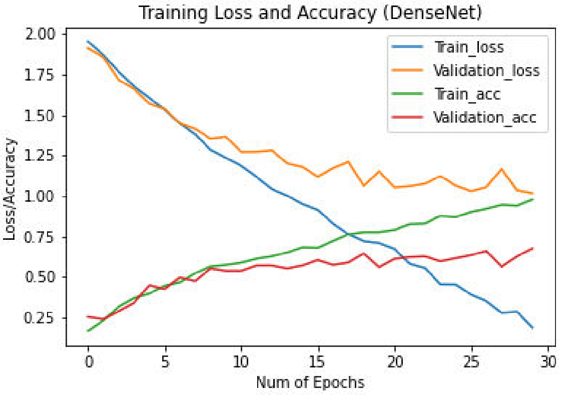
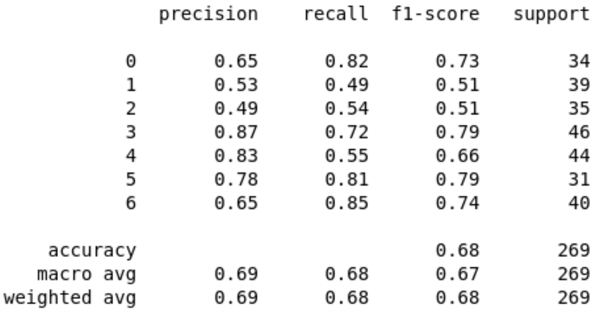


Figure-4. Model behavior based on training and validation data.



```
              precision    recall  f1-score   support

           0       0.65      0.82      0.73        34
           1       0.53      0.49      0.51        39
           2       0.49      0.54      0.51        35
           3       0.87      0.72      0.79        46
           4       0.83      0.55      0.66        44
           5       0.78      0.81      0.79        31
           6       0.65      0.85      0.74        40

    accuracy                           0.68       269
   macro avg       0.69      0.68      0.67       269
weighted avg       0.69      0.68      0.68       269
```

Figure-5. Model metrics.

Observing the behavior of the model, we can see that there is no over-fitting of the network. The error of the validation data is reduced in less proportion to the error of the training data from the fifth epoch onwards. The same behavior is observed in the accuracy of the validation data, which is reduced concerning the accuracy of the training data from the tenth epoch onwards.

Considering the training data, the accuracy of the model reached 69%, a value that is confirmed by the recall, f1-score, and support metrics. The categories corresponding to angry, fearful, and disgust reached the highest precision values (87%, 83%, and 78% consecutively). This may be because they correspond to negative emotions with a marked emotional response in people, which facilitates their identification of the images. Similarly, the lowest values of accuracy were for the sad and happy categories (49% and 53% consecutively), emotions that seem to print fewer distinctive features in the images.

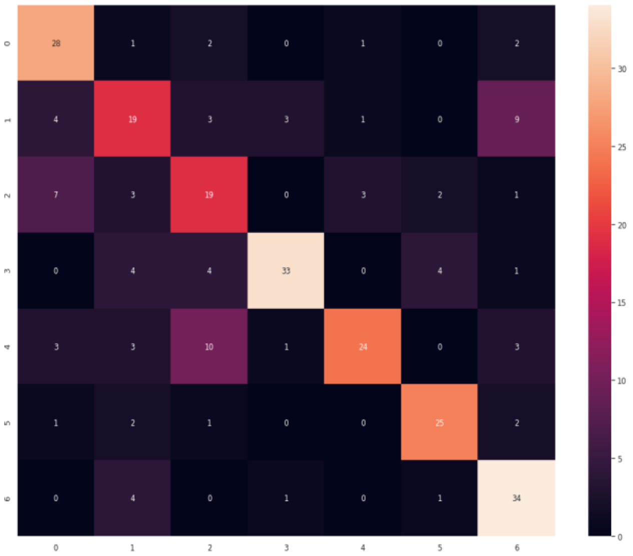


Figure-6. Model confusion matrix.

For the validation data, the highest number of hits was for the surprised (34 positive hits out of 40) and angry categories (33 positive hits out of 40). The worst categories were again sad and happy, each with 19 hits out of 40. The confusion matrix confirms the model behavior for unknown data, and visually a good model behavior is observed.

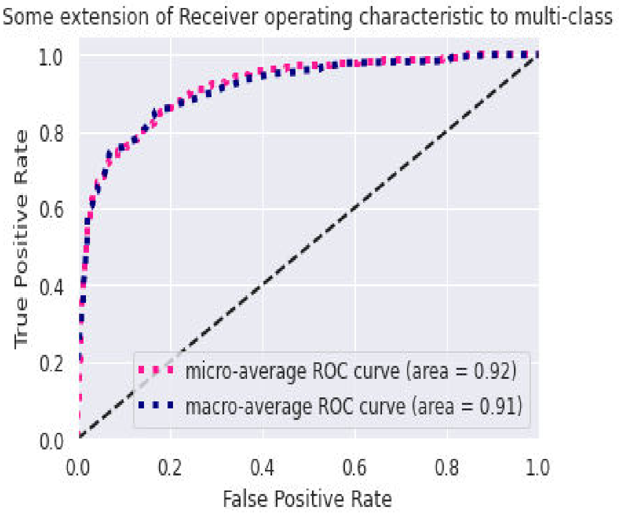


Figure-7. ROC curve (average behavior).

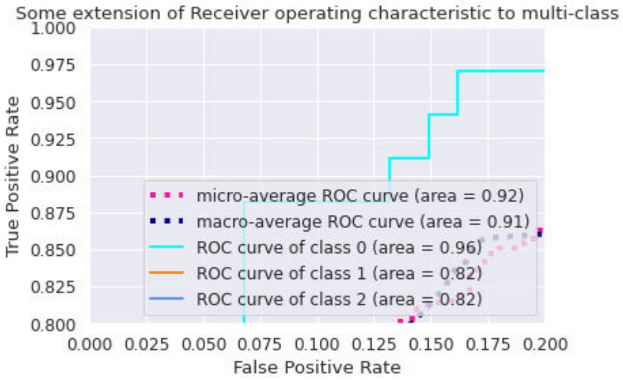


Figure-8. ROC curve (behavior by class).

The ROC curves show more detail of the behavior observed in the confusion matrix. The average behavior shows a true positive rate close to 70%, with some individual categories reaching 85%. In general terms, as a support to our emotion detection algorithm based on the face, the behavior of this audio detection algorithm is very good for our application, and overall increases the performance of our robot to 96.5% (our previous algorithm had a success rate of 92%).

5. CONCLUSIONS

In this paper, we show the design of PyHRE, our Psychoacoustic Model for Human-Robot Emotional Integration. This model was designed to support our emotion identification algorithm that works from the characterization of the person's face in front of the robot. Our previous algorithm had problems identifying the emotions of the person when it could not capture his face in a frontal way, so we proposed an alternative algorithm that uses as an input parameter the voice of the person, and from it, it could identify the particular characteristics of each one of the emotions of interest. The emotions that conform to the identification categories are calm, happy, sad, angry, fearful, disgusted, and surprised. We used a public database that contains audios of professional actors who excite these seven emotions. The database is processed to generate an image for each audio file. The images correspond to MEL spectrograms that consider human perception in the conversion range. A total of 192 images were obtained for each category (1344 images in total). The model was built based on a DenseNet convolutional neural network. The training was performed using categorical cross-entropy as a loss function and stochastic gradient descent as an optimization function. To evaluate performance, accuracy, recall, f1-score, and support metrics were calculated, as well as the confusion matrix and the ROC curve. The model obtained individually achieved an overall success rate of 69% and in conjunction with our previous system, 96.5%. The results shown correspond to laboratory tests developed on our ARMOS TurtleBot robot.

ACKNOWLEDGEMENT

This work was supported by the Universidad Distrital Francisco José de Caldas, in part through CIDC, and partly by the Facultad Tecnológica. The views expressed in this paper are not necessarily endorsed by Universidad Distrital. The authors thank the research group ARMOS for the evaluation carried out on prototypes of ideas and strategies.

REFERENCES

[1] S. Kiesler and M. Goodrich. 2018. The science of human-robot interaction. ACM Transactions on Human-Robot Interaction, 7(1): 1-3. ISSN 2573-9522. doi:10.1145/3209701.

[2] M. Decker, M. Fischer and I. Ott. 2017. Service robotics and human labor: A first technology assessment of substitution and cooperation. Robotics and Autonomous Systems, 87(1): 348-354. ISSN 0921-8890. doi:https://doi.org/10.1016/j.robot.2016.09.017

[3] M. Mataric. 2019. Human-machine and human-robot interaction for long-term user engagement and behavior change. In: The 25th Annual International Conference on Mobile Computing and Networking. ACM. doi:10.1145/3300061.3300141.

[4] V. Tung and R. Law. 2017. The potential for tourism and hospitality experience research in human-robot interactions. International Journal of Contemporary Hospitality Management, 29(10): 2498–2513. ISSN 0959-6119. doi:10.1108/ijchm-09-2016-0520.

[5] V. Kaptelinin, A. Kiselev, A. Loutfi and T. Hellstrom. 2017. Robots in contexts. In: Proceedings of the European Conference on Cognitive Ergonomics 2017 - ECCE 2017. ACM Press. doi:10.1145/3121283.3121424.

[6] G. Du, M. Chen, C. Liu, B. Zhang and P. Zhang. 2018. Online robot teaching with natural human–robot interaction. IEEE Transactions on Industrial Electronics, 65(12): 9571-9581. ISSN 0278-0046. doi:10.1109/tie.2018.2823667.

[7] M. Jung. 2017. Affective grounding in human-robot interaction. In: 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2017). 1-6.

[8] J. Nasir, U. Norman, W. Johal, J. Olsen, S. Shahmoradi and P. Dillenbourg. 2019. Robot analytics: What do human-robot interaction traces tell us about learning? In: 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN 2019). 1-6. doi:10.1109/RO-MAN46459.2019.8956465.

[9] A. Thomaz, G. Hoffman and M. Cakmak. 2016. Computational human-robot interaction. Foundations and Trends in Robotics, 4(2-3): 104-223. ISSN 1935-8253. doi:10.1561/2300000049.

[10] P. Tsarouchi, S. Makris and G. Chryssolouris. 2016. Human-robot interaction review and challenges on task planning and programming. International Journal of Computer Integrated Manufacturing, 29(8): 916-931. ISSN 1362-3052. doi:10.1080/0951192x.2015.1130251.

[11] B. Alenljung, J. Lindblom, R. Andreasson and T. Ziemke. 2019. User experience in social human-robot interaction. In: Rapid Automation, IGI Global. 1468-1490. doi:10.4018/978-1-5225-8060-7.ch069.

[12] S. Lemaignan, M. Warnier, E. Sisbot, A. Clodic and R. Alami. 2017. Artificial cognition for social human-robot interaction: An implementation. Artificial Intelligence, 247(1): 45-69. ISSN 0004-3702. doi:10.1016/j.artint.2016.07.002.

[13] T. Sheridan. 2016. Human–robot interaction. Human Factors: The Journal of the Human Factors and Ergonomics Society, 58(4): 525-532. ISSN 0018-7208. doi:10.1177/0018720816644364.

[14] F. Martinez, C. Hernandez, and A. Rendon. 2020. Identifier of human emotions based on convolutional neural network for assistant robot. TELKOMNIKA (Telecommunication Computing Electronics and Control), 18(3): 1499-1504. ISSN 1693-6930. doi:10.12928/telkomnika.v18i3.14777.

[15] H. Admoni and B. Scassellati. 2017. Social eye gaze in human-robot interaction: A review. Journal of Human-Robot Interaction, 6(1): 1-25. ISSN 2090-9888. doi:10.5898/jhri.6.1.admoni.

[16] O. Bertel, C. Moreno and E. Toro. 2009. Aplicación de la transformada wavelet para el reconocimiento de formas en visión artificial. Tekhnê, 6(1): 3-8. ISSN 1692-8407.

[17] M. Dent. 2017. Animal psychoacoustics. Acoustics Today, 13(3): 19-26. ISSN 1557-0215.

[18] K. Tsunekawa, F. Leiva and J. Ruiz. 2018. Visual navigation for biped humanoid robots using deep reinforcement learning. IEEE Robotics and Automation Letters, 3(4): 3247-3254. doi:10.1109/lra.2018.2851148.

[19] R. Rodrigues, M. Basiri, A. Aguiar and P. Miraldo. 2018. Low-level active visual navigation: Increasing robustness of vision-based localization using potential fields. IEEE Robotics and Automation Letters, 3(3): 2079-2086. doi:10.1109/lra.2018.2809628.

[20] B. Calli, W. Caarls, M. Wisse and P. Jonker. 2018. Active vision via extremum seeking for robots in unstructured environments: Applications in object recognition and manipulation. IEEE Transactions on Automation Science and Engineering, 15(4): 1810-1822. ISSN 1545-5955. doi:10.1109/TASE.2018.2807787.

[21] F. Martínez, E. Jacinto and F. Martínez. 2020. Obstacle detection for autonomous systems using stereoscopic images and bacterial behaviour. International Journal of Electrical and Computer Engineering, 10(2): 2164–2172. ISSN 2088-8708. doi:http://doi.org/10.11591/ijece.v10i2.pp2164-2172.

[22] K. Lee, J. Gibson and E. Theodorou. 2020. Aggressive perception-aware navigation using deep optical flow dynamics and PixelMPC. IEEE Robotics and Automation Letters, 5(2): 1207-1214. doi:10.1109/lra.2020.2965911.

[23] A. Sek and B. Moore. 2020. Psychoacoustics: Software package for psychoacoustics. Acoustical Science and Technology, 41(1): 67-74. ISSN 1347-5177. doi:https://doi.org/10.1250/ast.41.67.

[24] P. Balazs, N. Holighaus, T. Necciari and D. Stoeva. 2017. Frame theory for signal processing in psychoacoustics. Applied and Numerical Harmonic Analysis, 5(1): 225-268. ISSN 2296-5009. doi:https://doi.org/10.1007/978-3-319-54711-4_10.

[25] J. Kennedy, S. Lemaignan, C. Montassier, P. Lavalade, B. Irfan, F. Papadopoulos, E. Senft and T. Belpaeme. 2017. Child speech recognition in human-robot interaction. In: Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction. ACM, 82-90. doi:10.1145/2909824.3020229.

[26] S. Livingstone and F. Russo. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391.