Implementing Emotion Detection from Speech for Psychological Assessment of Elderly People: A Comparative Study of Python-based Approaches and Existing Solutions

I. Warnants[1], N. Tsiogkas[1], J. Roca Gonzalez[2], F.J. Ortiz Zaragoza[2], I. Méndez[3], JA. Vera Repullo[2], J.P. Serna[4]

[1] Catholic University of Leuven (KU Leuven), Leuven, Belgium, ismael.warnants@student.uhasselt.be, nikolaos.tsiogkas@kuleuven.be

[2] Departamento de Tecnología Electrónica, Universidad Politécnica de Cartagena, Cartagena, España, {jroca.gonzalez, francisco.ortiz, jose.vera}@upct.es

[3] Department of Developmental and Educational Psychology, Universidad de Murcia, Espinardo, Murcia, 30100, España, inmamendez@um.es

Development Systems & Services, S.L. Algezares, 30157 Murcia, España. juanpedroserna@gmail.com

Abstract

In the last ten years, the number of people over 65 has increased 30% in Spain. This trend is anticipated to grow and require more healthcare personnel. To prevent this, people should live longer independently instead of in care homes. The ADDIM system will assist them in living independently. The research presented in this paper is part of the mood detection of the user in the ADDIM (Asistencia Domiciliaria Digital Integral para Mayores) system. This is a Digital platform for monitoring older people's health, safety, companionship, and emotional support at home based on robotics, artificial intelligence, and ambient assisted living.

To detect user emotions, the right speech corpus, feature extraction methods, preprocessing methods, and machine learning models have to be selected. Based on the detected emotion, the robot will interact with the user to perform predefined actions. The final mood of the user will be estimated using this output in conjunction with visual feedback and the sensors in the user's home with the ADDIM system.

Three speech corpora are selected with retraining to achieve personalized detection based on the user's previous recordings. In addition, this will ensure that the detection is improved over time, which has yet to be implemented in other research. Finally, the implementation uses dimensional emotion detection instead of discrete emotion detection. This augments the number of detectable emotions.

1. Introduction

Over the last decade, Spain witnessed a 30% increase in its population aged 65 and above. This is attributed to a rise in life expectancy from 79.1 years in 2000 to 83.2 years in 2019, making it one of the countries in the European Union and worldwide with the highest life expectancy among older adults [1]. While this is a positive outcome, it has implications for the country's healthcare system and the elderly population's mental health. The growth in elderly people living alone may lead to mental disorders impacting their quality of life. In this regard, several experts have spoken out in favor of a change in the model of dependency care in Spain, pointing out among the priorities the availability of comprehensive and integrated care in the home to relieve the pressure on healthcare personnel [2]. As an example, during COVID-19 there was a shortage of care response capacity in nursing homes, causing a dramatic situation. Therefore, person-centered care and promoting adapted assisted living should be implemented to support elderly people living alone in Spain.

In this research, we propose the ADDIM system, a digital platform for monitoring the health, safety, companionship, and emotional support of older people at home. In this system, a Speech Emotion Recognition (SER) model will detect the user's emotions and work together with the rest of the ADDIM system's components to personalize the elderly user's care. The implementation goal is to be a proof-of-concept to check the feasibility of triggering actions in a robot based on the result provided by the SER model. Future research will explore the possible ways to improve the SER model and its real-world limitations. The aim is to detect emotions in audio fragments instead of continuous speech.

2. Framework

The ADDIM system extends mental assessment and mood monitoring with pro-active coaching capabilities, provided by a companion social robot, which suggests activities to older people to improve their mood and mental well-being.

The proposed companion robot and Ambient Assisted Living (AAL) system comprises several components, summarized below. It should be remembered that different deployments of the system are possible, from the simplest (only AAL) to the most complex model, including the social robot. Using low-cost commercial components (such as low-cost commercial sensors), open-source software, and communication standards (ROS, Node-RED, MQTT, Home Assistant) also favor the incorporation of new hardware and software for future developments.

- Assistive mobile robot: Designed to navigate autonomously around the house to attend to the users' needs or suggest an activity according to their mood.
- Home automation sensor ecosystem: Includes a set of low-cost sensors, both commercial and self-designed, to monitor the user's lifestyle.

ISBN: 978-84-17853-76-1

- Speech Emotion Recognition (SER): A Python implementation of a machine learning algorithm to detect the emotion of the user from an audio file sent by the robot to the computer.
- Application for psychological data acquisition: Carries out a psychological study and thus relates the information obtained from the Empatica E4 to the user's state of mind, such that basic questions about activity/well-being can be answered. This is another input for the machine-learning algorithm.
- Central home-assistant unit: Collects home automation data on the person's routines at home and activates the necessary devices. This unit, an embedded PC to house the artificial intelligence algorithms for the coordination and functioning of the various well-being devices. ROS, Home Assistant, and IoT were used to make the information transparent among the system's different elements. The SER model will run in this part of the system.

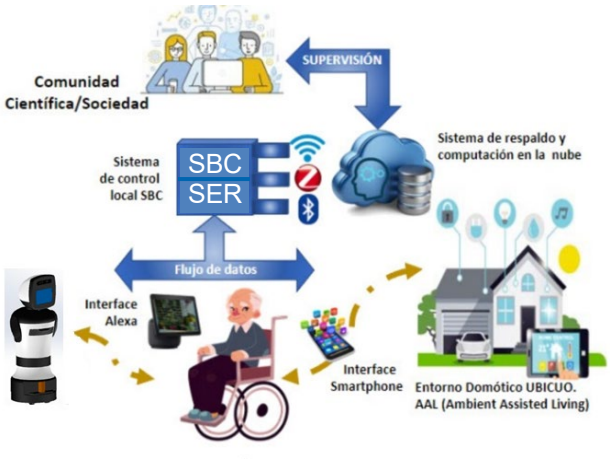

Figure 1. *A scheme of the main elements.*

3. State-of-the-art

Following the WHO's [3] recommendations on ethics and governance in the field of health, artificial intelligence, as used in this work, holds great promise for public health practice, as it has great benefits, such as protecting human autonomy, promoting the well-being and safety of individuals and the public, ensuring transparency, clarity and intelligibility, and guaranteeing inclusiveness, equity and sustainability.

The field of mood recognition has gained in importance in the world of artificial intelligence. Successful studies have been conducted with wearables [4], [5], such as IoT sensors for behavioral inference [6]. However, putting the system to practical use is still the challenge we are facing with the proposed system.

Different sensor systems have been proposed to collect physiological and activity data and analyze them to automatically assess people's moods. Smart monitoring systems use smartphones and self-reported depression reports to assess people's moods, as in [7]. However, the latter system does not provide counseling. In the H2020 Help4mood project, data on daily activities are converted into graphical, textual, and conceptual summaries that can be communicated to clinicians providing external assessment.

Automatic mood monitoring and assessment are important components of support therapies and daily life coaching. Ecological momentary assessment (EMA) can ease patients' mental burdens since the traditional assessment tests are too long to be used repeatedly. In addition, the latest smartphones and wearable devices make the EMA approach much more feasible as a solution for monitoring mental illness and offer economically friendly solutions.

The above-described method of gathering information on mental well-being has been used in several works. In [8], a system is proposed to monitor potential depressive patterns in elderly people living alone. Emotional states are assessed from diverse sources, such as surveys, smartwatches, and EMA questionnaires. EMA methods for obtaining emotional information can expand this research area to practical applications. For example, in [9], a system that gathers information on a person's activities can detect long-term stress patterns. Stress detection during daily real-world tasks through advances in intelligent systems is described in [10]. Techniques such as machine learning were also applied [11] to monitor elderly people's moods via intelligent sensors on wristbands.

An issue to be addressed is the usability of these modern assistive technologies by the elderly. In [12], the authors present a user-centered design for a web-based multi-modal user interface tailored for elderly users of near-future multi-robot services.

3.1. Empathic vs. non-empathic Robot

In [13] they look at the effects of the robot named Ryan with or without emotional assessment in a conversation with the user. They noticed little difference in the empathic or non-empathic version of the robot in conversation with the user. However, when the word count was measured, and the exit survey of all participants was finished, they noticed that the emotional version of Ryan was more engaging and likable by the participants. And that the empathic version of the robot encourages users to have longer conversations compared to the non-empathic version. This was suggested to have the potential to decrease depression in the users.

3.2. Speech Emotion Recognition (SER)

In the state of the art [14], there are two methods to detect emotions: discrete and dimensional. The dimensional was chosen as it has two parameters to plot any emotion on the Speech Affective Space Model (SASM) [15]. Figure 2 shows the Speech Affective Space Model used in the state of the art. The SASM is not limited to the number of labels on which the model is trained and can, because of this predict more emotions compared to the available emotions in the speech corpus. The Euclidean distance is used to measure the distance of the predicted emotion to the closest discrete emotion. For each dimension (Valence and Arousal) the concordance correlation coefficient (CCC) is calculated to measure the accuracy and precision of the predicted results compared to the expected results. In the

ISBN: 978-84-17853-76-1

state of the art the CCCV is 0.28 ± 0.05 and the CCCA is 0.58 ± 0.05 [14].

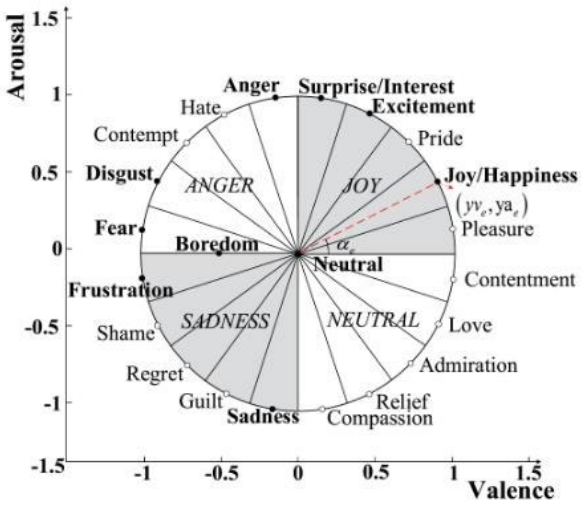

Figure 2. *Representation of emotions on the Speech Affective Space Model (SASM) by using Arousal-Valence values [14]*

4. Methodology

To create a SER model, the first step is to select a speech corpus, but as there are no speech corpora with dimensional values, the discrete values are mapped onto the SASM and the model output will be compared to these converted Valence-Arousal values. The selected speech corpora was EmoDB to compare our results to the state of the art. Our results were similar for other speech corpora. It has to be noted that most speech corpora are artificial and the expression of emotions can vary across a lot of factors (ethnicity, age, etc.) [16].

The second step is to extract the features from the audio files from the speech corpora. For this, a wide range of methods and features can be used. The best results were achieved using a library by Renovamen (Github) [17]. It consists of 320 features that are normalized. The ones with the highest correlation to the output are chosen with principal component analysis (PCA).

With these features, the model can be trained and optimized. The selected models are deep learning models as they improve with the more data they get. This is beneficial as the goal is to retrain later on with the final emotion from the user we get from all the inputs of the ADDIM system that will detect the user's emotion and can retrain the Emotion Recognition (ER) models like this one. A CNN, LSTM and CNN+LSTM model are used to compare to the state of the art. The CNN+LSTM model combines CNN layers as the input and LSTM layers as the output. These models are optimized by changing the amount and types of layers and other parameters.

The models will be evaluated on their accuracy, MSE, RMSE, R^2, CCCV, CCCA, and the number of inputs. The amount of inputs is also correlated with the amount of calculations that have to be performed and, thus also the time to train and predict.

5. Results & discussion

The models were trained and used to predict emotions on a CPU and were able to predict in a matter of seconds.

Table 1 compares our models with the State Of The Art [14]. The CNN model and CNN+LSTM model show great performance and outperform the state of the art. The LSTM results are not promising, but still in line with the state of the art. The training and validation loss of the models are shown in Figure 3-5.

The results with the CNN+LSTM model are achieved with less inputs compared to the other models. This means that there are less features required and thus also less computing time.

Model	CNN	LSTM	CNN+ LSTM	SOTA
Accuracy (%)	78.78	36.36	78.78	/
MSE	0.1110	0.2501	0.1269	/
RMSE	0.3333	0.5001	0.3562	/
R^2	0.6553	0.2429	0.6413	/
CCCV	0.8531	0.5573	0.7615	0.28±0.05
CCCA	0.8331	0.5253	0.8901	0.58±0.05
Inputs	46	41	40	/

Table 1. *Model training results for CNN, LSTM and CNN+LSTM with EmoDB, compared to the the State-Of-The-Art (SOTA)*

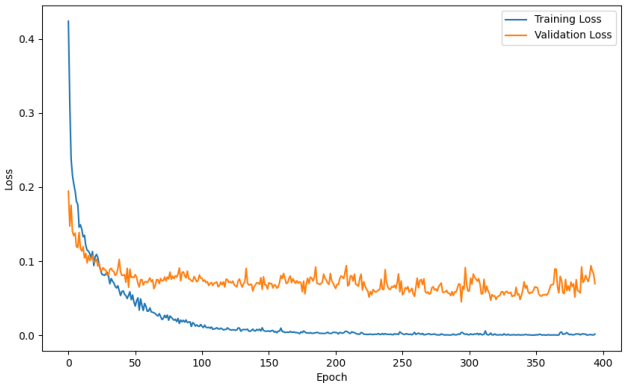

Figure 3. *Training and validation loss during training of the CNN model on the EmoDB speech corpus*

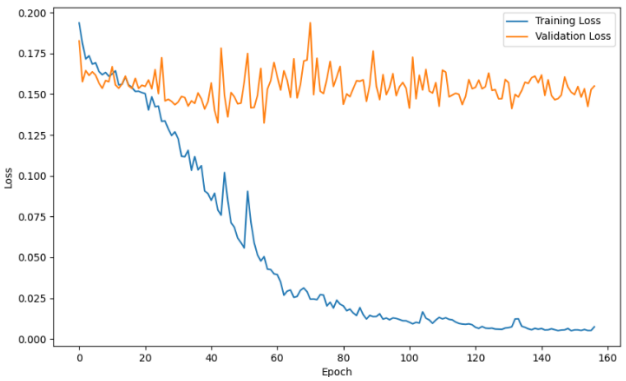

Figure 4. *Training and validation loss during training of the LSTM model on the EmoDB speech corpus*

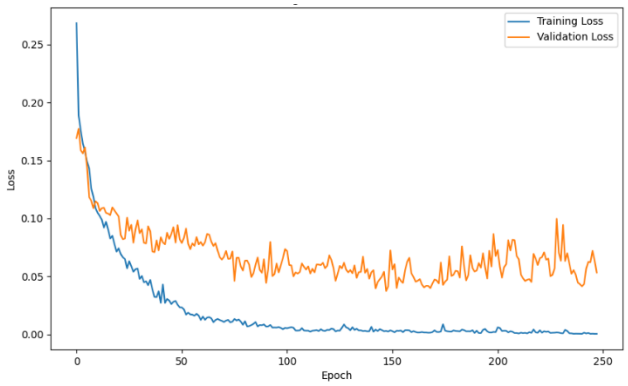

Figure 5. *Training and validation loss during training of the CNN+LSTM model on the EmoDB speech corpus*

The features that were selected with principal component analysis (PCA) as inputs for the models were consistent during testing while comparing with other speech corpora. This seems to show that there are features that indicate the amount of Valence and Emotion regardless of the language, recording method, and other variables.

The models only get features that can be extracted from the audio files, so there is no speech context used for these results. This means that conversations cannot be recorded by the model, which is again better for privacy.

The results of the SER will be added with the other models into the system to get an even more accurate result for the user's emotion. This emotion will be used for retraining the model to adapt to the user and become better over time. This is possible because we use deep learning models which thrive with the increased amount of data.

There are however some limitations. The EmoDB speech corpus is limited and artificial in nature, which is the same for other speech corpora. It is also difficult to find speech corpora in Spanish, especially for elderly people.

6. Conclusion

The CNN and CNN+LSTM models have proven to outperform the state of the art. This while maintaining privacy by running locally on a CPU and predicting fast.

These results should be retrained on a Spanish speech corpus, if possible, but the features that are used to train and predict seem to be consistent across speech corpora and languages.

Future work must prove the advantages of retraining in a real-world scenario. Results for audio filtering could also be explored for real-world scenarios. The model could also be adapted to work in real-time to respond even faster compared to the speed we have now.

These improvements will benefit the promotion and monitoring of the emotional state and therefore the quality of life of older people.

Acknowledgements

The HIMTAE project, Robwell subproject (reference RTI2018-095599-A-C22) has been funded by: Programa Estatal de Investigación, Desarrollo e Innovación Orientada a los Retos de la Sociedad, en el marco del Plan Estatal de Investigación Científica y Técnica y de Innovación 2013-2016.

The pilot project ADDIM-Poncemar (Asistencia Domiciliaria Digital Integral para Mayores en Poncemar) funded by the "Fundación Integra Digital" of the Region of Murcia is currently being developed in the Day Care Centers of the Poncemar Foundation, in the Health Campus of Lorca, University of Murcia. We thank them for their collaboration and involvement in the testing of the solutions presented in this communication.

Finally, we would like to thank the previous CASEIB organising committees for their kindness in allowing us to use their style guides as a reference for this document.

References

[1] "Spain data | World Health Organization." Accessed: Aug. 18, 2023. [Online]. Available: https://data.who.int/countries/724

[2] V. Davey, *Situación en España de la evaluación de sistemas de atención a personas mayores en situación de dependencia. Informe en Red nº 28.* 2021. doi: 10.13140/RG.2.2.28439.09128.

[3] "Ethics and governance of artificial intelligence for health." Accessed: Sep. 26, 2023. [Online]. Available: https://www.who.int/publications-detail-redirect/9789240029200

[4] O. M. Mozos *et al.*, "Stress Detection Using Wearable Physiological and Sociometric Sensors," *Int. J. Neural Syst.*, vol. 27, no. 02, p. 1650041, Mar. 2017, doi: 10.1142/S0129065716500416.

[5] M. Koutli, N. Theologou, A. Tryferidis, and D. Tzovaras, "Abnormal Behavior Detection for Elderly People Living Alone Leveraging IoT Sensors," in *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*, Oct. 2019, pp. 922–926. doi: 10.1109/BIBE.2019.00173.

[6] Calatrava, F.M.; Ortiz, F.J.; Vera, J.A.; Roca, J.; Jiménez, M.; Martínez, O., "Heterogeneous system for monitoring daily activity at home and the well-being of the elderly.," in *In Proceedings of the XLII Conference on Automatic: Minutes Book*, Castelló, Spain, 2021, pp. 632–639.

[7] A. Doryab, "Detection of behavior change in people with depression," Jul. 2014.

[8] H. Kim, S. Lee, S. Lee, S. Hong, H. Kang, and N. Kim, "Depression Prediction by Using Ecological Momentary Assessment, Actiwatch Data, and Machine Learning: Observational Study on Older Adults Living Alone," *JMIR MHealth UHealth*, vol. 7, no. 10, p. e14149, Oct. 2019, doi: 10.2196/14149.

[9] R. Kocielnik, N. Sidorova, F. M. Maggi, M. Ouwerkerk, and J. H. D. M. Westerink, "Smart technologies for long-term stress monitoring at work," in *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*, Jun. 2013, pp. 53–58. doi: 10.1109/CBMS.2013.6627764.

[10] M. V. Gómez-Gómez, M. V. Bueno-Delgado, C. Albaladejo-Pérez, and V. Koch, "Augmented Reality, Virtual Reality and Mixed Reality as Driver Tools for Promoting Cognitive Activity and Avoid Isolation in Ageing Population," in *Smart Objects and Technologies for Social Good*, I. M. Pires, S. Spinsante, E. Zdravevski, and P. Lameski, Eds., in Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Cham: Springer International Publishing, 2021, pp. 197–212. doi: 10.1007/978-3-030-91421-9_15.

[11] "EnrichMe - Collaborative Projects | PAL Robotics." Accessed: Sep. 26, 2023. [Online]. Available: https://pal-robotics.com/collaborative-projects/enrichme/

[12] A. Di Nuovo *et al.*, "The multi-modal interface of Robot-Era multi-robot services tailored for the elderly," *Intell. Serv. Robot.*, vol. 11, no. 1, pp. 109–126, Jan. 2018, doi: 10.1007/s11370-017-0237-6.

[13] H. Abdollahi, M. Mahoor, R. Zandie, J. Sewierski, and S. Qualls, "Artificial Emotional Intelligence in Socially Assistive Robots for Older Adults: A Pilot Study," *IEEE Trans. Affect. Comput.*, pp. 1–1, 2022, doi: 10.1109/TAFFC.2022.3143803.

[14] S. Jing, X. Mao, and L. Chen, "Automatic speech discrete labels to dimensional emotional values conversion method," *IET Biom.*, vol. 8, no. 2, pp. 168–176, 2019, doi: 10.1049/iet-bmt.2018.5016.

[15] N. Kamaruddin and A. Wahab, "Human behavior state profile mapping based on recalibrated speech affective space model," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug. 2012, pp. 2021–2024. doi: 10.1109/EMBC.2012.6346354.

[16] A. Hanjalic, "Extracting moods from pictures and sounds: towards truly personalized TV," *IEEE Signal Process. Mag.*, vol. 23, no. 2, pp. 90–100, Mar. 2006, doi: 10.1109/MSP.2006.1621452.

[17] X. Zou, "Speech Emotion Recognition." Aug. 17, 2023. Accessed: Aug. 18, 2023. [Online]. Available: https://github.com/Renovamen/Speech-Emotion-Recognition