2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Speech Based Affective Analysis of Patients Embedded in Telemedicine Platforms*

Athanasios Kallipolitis, Michael Galliakis, Andreas Menychtas and Ilias Maglogiannis, Senior Member, IEEE

Abstract—**Speech is a basic means of human expression, not only due to the combination of words that exits our mouth, but also because of the different way we express these words. Apart from the main objective of speech, which is the communication of information, emotions flow in human speech as various vocal characteristics (prosodic, spectral, tonal). By processing these characteristics, Speech Emotion Recognition aims to analyze and assess the human emotional status to complement medical data captured during telemedicine sessions. Driven by the latest developments in Computer Vision concerning Deep Learning techniques, EfficientNets are exploited to extract features and classify imagery representations of human speech into emotions as a web service along with an interpretation scheme. The developed web service will be consumed during video conferences between medical staff and patients for the near real-time assessment of emotional status of patients during video teleconsultations.**

I. Introduction

For decades, the emotional transactions of the human side when interacting with a machine have been neglected, thus leading to a poor communication 'protocol'. In the same way that humans can communicate better, when understanding the emotional status of the other side, information systems including Emotional Artificial Intelligence (EAI) capabilities can provide more efficient results during their interaction with humans [19]. Integrating EAI capabilities in electronic healthcare and particularly remote homecare systems has a very prosperous potential due to the influence of the emotional status of a patient in the treatment process. In this context, the paper proposes a near real-time speech emotion recognition system that is intended to be consumed as a web service during telemedicine sessions. The measurement and assessment of a patient's emotional status during teleconferences is of high value to the medical experts especially for patients that suffer from chronic diseases. Their emotional status is recorded and analyzed along with other personal health and clinical data for the extraction of important patterns, while timely intervention of specialists can improve their quality of life due to the prevention of emotional breakdowns. Our contribution lies in the incorporation of the emotion analysis service in a healthcare platform (Fig. 1) for the enhancement of the medical record and in the fact that the reported accuracy is of the highest reported in literature while handling datasets from different sources. This increases the ability of the predictive model for generalization. Furthermore, the implementation provides an interpretation scheme of the predictive model, linking the outcome of the prediction to the specific visual patterns found in the image representation of the sound clip. This connection provides useful insight for the decision-making factors of the predictive model.

II. Related Work

EAI is widely utilized in commerce and industry for a variety of applications ranging from evaluating employees' engagement to work [9], assessing movies from the analysis of audience's emotions [10], quantifying high engagement to commercial advertisements [11] to improving gaming experience [12]. Throughout this variety, healthcare is one of the most interesting fields for implementing affective computing systems. The prospect of decoding human emotion is highly appreciated in healthcare services due to the sense of lack of control over body and psyche that patients experience. 'Arguably, there is no other service setting in which emotions are more relevant than in health care' [7]. According to the expression means utilized to assess the emotional status, Affective Computing is divided in three main categories. Facial Emotion Recognition (FER) refers to the analysis of human facial expressions, Speech Emotion Recognition (SER) refers to the analysis of human speech sound or/and words and Pose Emotion Recognition (PER)

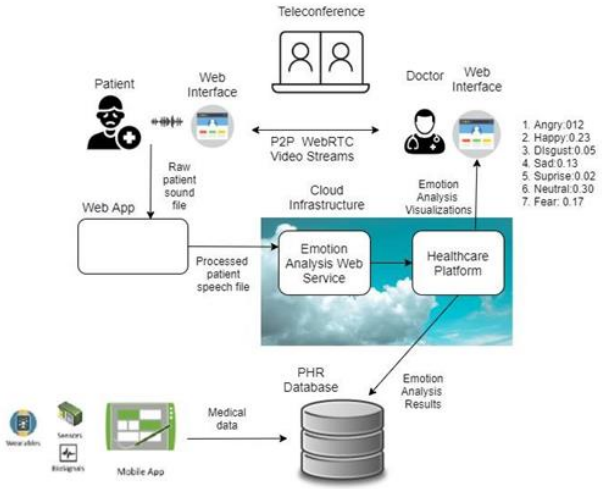

Figure 1. Overview of the homecare platform after the integration of the emotion analysis service

*This research has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship, and Innovation, under the call RESEARCH – CREATE – INNOVATE (project code: MediLudus - Personalized home care based on game and gamified elements T1EDK-03049). A. Kallipolitis, M. Galliakis and I. Maglogiannis are with University of Piraeus, Dept. of Digital Systems, Greece (e-mail: nasskall@unipi.gr, michaelgalliakis@unipi.gr, imaglo@unipi.gr).

refers to the analysis of human pose. In many cases multimodal approaches [16, 17] have been utilized for the quantification of human emotions from different sources and their fusion towards accurate recognition. Speech Emotion Recognition (SER) systems have been discussed for many years throughout the literature as standalone implementations [6] for the quantification of human emotional status. However, rare are the cases that the provided functionality is incorporated in a healthcare information system for the assessment of the psychological condition of patients. Many approaches have been proposed for the representation of speech and among them representing speech as a mel- spectrogram has been utilized widely [14, 15] with a main purpose to exploit the well-established capabilities of deep convolutional networks when applied on images. In such implementations, single datasets or combination of different datasets are utilized. While enhancing generalization properties, the case of combinatory utilization of different datasets poses a significant challenge due to the numerous characteristics' variations from dataset to dataset that make the discovery of efficient global patterns an arduous task.

III. METHODOLOGY

The proposed system consists of a speech emotion recognition scheme and an interpretation scheme that are both incorporated in a web service. The web service is based on the RESTful architectural standard that provides a stateless communication between a server (back-end) and a client (front-end). In the front-end section, human speech detection and speech file preprocessing takes place, whereas in the back- end, Speech to Image (mel-spectrogram) Transformation and Classification – Interpretation is being implemented.

A. Front End

The audio signal is captured via a microphone during the telemedicine session and is sent to the patient's interface to provide baseline notifications. Consequently, the signal enters the speech detection module, where certain events are emitted to the speech recorder to highlight the voiced-unvoiced windows in time. This occurs by transforming the audio to Fast Fourier Transform and selecting all samples that are above a predefined threshold [18]. The parts of the audio file, labeled as voice are transformed into .wav files, encapsulated in the body of an http request and sent to the back end for analysis.

B. Back End

Once the speech file is received by the back end, the short-time Fourier transform is applied upon to extract the spectrogram of the file that will, in turn, be transformed to a mel- spectrogram by converting the frequency scale to the mel scale. Mel scale is the result of a non-linear transformation that serves the cause of placing different sounds in the distance that corresponds to the human perception unlike Hz scale. The length of windowed signal is dictated by the utilized Python library (Librosa), as the appropriate for speech processing and is chosen to be a power of two (512) for speed optimization of the Fast Fourier Transform. By setting the number of FFT samples to 512 for a sampling rate of 22050Hz, we get 23 milliseconds which is the

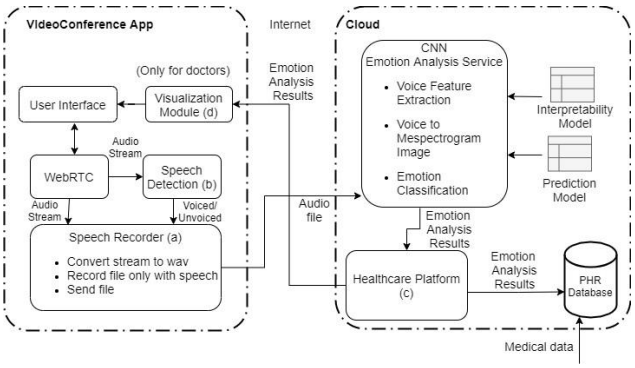

Figure 2. Emotional Analysis Operational Scenario

selected physical duration of the window. Before entering the next step, the vector is converted to a 128 by 216 by 3 matrices by repeating two times one dimension to take the appropriate form to be processed by the pre-trained deep learning model that assumes that the image has 3 channels. For the classification of speech files, the B0 EfficientNet is selected due to its capability of achieving high accuracy with considerably few parameters (5.3M) [13]. The B0 EfficientNet model is pre-trained on the ImageNet and utilizes the adjusted weights to extract features from the mel-spectrogram images. The B0 EfficientNet utilizes various inverted residual-squeeze and excite blocks to extract from the input image the final flattened feature vector that consists of 1280 values. Consequently, the fully connected layers lead to seven neurons, each one corresponding to the probability of the input image depicting one of the seven emotions. The interpretation scheme is added on the existing convolutional neural network (CNN) architecture and measures the gradients with respect to feature maps activations for each predicted class without interfering with the functionality of the predictive model according to the Grad-CAM technique proposed in [6]. Grad-Cam utilizes the spatial information which is maintained throughout all convolutional layers to highlight class discriminative regions that play an important role to decision making.

C. Integration in the telemedicine platform

The functionality of the proposed EAI web service can be easily exploited by a client without further procedures, thus, resulting in the effortless integration of the speech emotion recognition system to different platforms. In this context, theproposed system is seamlessly integrated in an existing telemedicine platform [20] to enhance the medical record of a patient with emotion-related information. There are four distinct integration points: a) on the client application of the patient to capture the audio during the telemedicine session, b) on the client application to distinguish voice from unvoiced segments, c) on the backend of the platform for storing of theresults, and d) on the client application of the doctor for presenting in near real-time the emotion analysis (Fig. 2). The telemedicine session is implemented using the WebRTC [21] technology, which is available nowadays in all popular web and mobile platforms. Therefore, for the integration point (a) specific probes are created to extract the audio from the WebRTC stream and for point (b) the speech detection

module intervenes to separate voice from noise, while for the point (d) adaptations in video UI took place to handle the visualization of the emotion analysis results. The integration point (c) is in data level and addresses the requirement of storing the results to the health record of the patient that is maintained in the platform. Concerning the emotion analysis process, the emotion analysis service, the healthcare platform, and the web application communicate by exchanging 11 basic messages in the sequence depicted in Fig. 3.

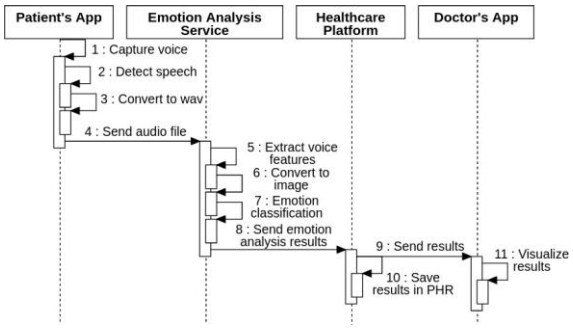

Figure 3. Overview of the emotion analysis process.

IV. Experimental Results

Throughout the total workflow of training and evaluating the proposed predictive model the following three (3) different open-source datasets have been utilized: The Toronto emotional speech set [1], the Ryerson Audio-Visual Database of Emotional Speech and Song RAVDESS [2] and the SAVEE database [3, 4]. All the above-mentioned samples are added (4720 in total) to create a seven (7) categories dataset consisting of seven emotions (anger, happiness, fear, surprise, sadness, disgust and neutral). The duration of each sample varies from 1 to 2.5 seconds. Therefore, the samples that last less than 2.5 sec are padded with zeros before and after. The whole dataset is utilized for the training and evaluation of the predictive model in the following proportions: 66,6% of the dataset as training set, 33,3% of the dataset as validation set. The participation of independent datasets in the formation of the final dataset mitigates the risk of overfitting, which is also verified by a bootstrapping test. The procedure is repeated 100 times with replacing the already used population. The learningrate is reduced according to a custom step learning rate scheduler, adjusted after thorough observation of measurements of the loss function during the training of the model to improve its performance.

By repeatedly sampling from the population of sounds in proportion 66,6-33,3%, the accuracy metric is calculated as a distribution. The mean value of the accuracy metrics is 0.88% and varies between 0.86% and 0,92%, as shown in Fig. 4. The result is amongst the highest reported in the field of Speech Emotion Recognition when compared with the state of the art discussed in [5]. The confusion matrix in Fig. 5, on the other hand, highlights the misclassification cases that are detected more often between sad and neutral samples and angry and disgust samples. These results agree with the human perception about what sound can be easier mistaken for a different emotion. The confusion matrix is randomly chosen from the 100 iterations of the experiment. The performance of the predictive model is evaluated by choosing different durations for the sound files. Consequently, the whole file is initially divided in clips of 1 sec, then 2 secs and finally 2.5 secs. As a result, the predictive model demands at last 2 secs to better identify patterns of emotion in clips. Concerning the interpretation scheme, the thorough examination of the generated heatmaps, shown in Fig. 6, results in the three conclusions. First, emotions are detected in a narrow specific time window of less than the total time of 2.5 sec apart from the case of neutral sounds. Secondly, the areas in the generated images that are responsible for decision making are concentrated in frequencies lower than 1500Hz. Moreover, the decision for neutral sounds is based on the monitoring of the whole clip for the verification of absence of the patterns that are indicative of the six emotions.

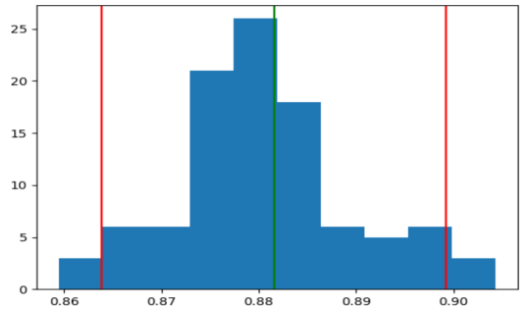

Figure 4. Results of accuracy by bootstrapping the whole dataset (70-30%) 100 times with replacement.

V. Conclusion And Future Work

In this paper we proposed a methodology for the assessment of human emotion from speech based on the visual representations of sound clips and the exploitation of state-of- the art deep CNNs. This work is a follow-up of the developed FER system that aims to analyze real-time facial expressions of patients during teleconferences [8]. As a future step we intend to design a multimodal approach for the assessment of human emotion. The results are promising, but there is a need of combining multiple sources of human emotion to create a more robust representation of human emotional status. Throughout the whole work, explainability remains an important aspect since it provides useful knowledge for the decision-making factors of the predictive model.

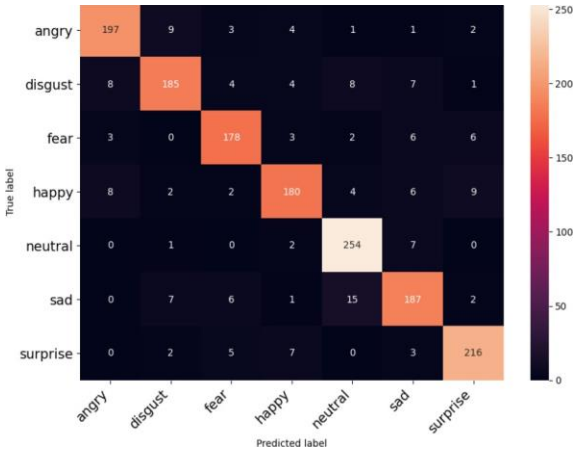

Figure 5. Confusion matrix of results by performing classification once withEfficientNet B0.

The generated features can be further utilized for the classification of videos in emotional or psychological conditions as time series.

Compliance with ethical standards: This study uses public datasets as presented in [1-4]

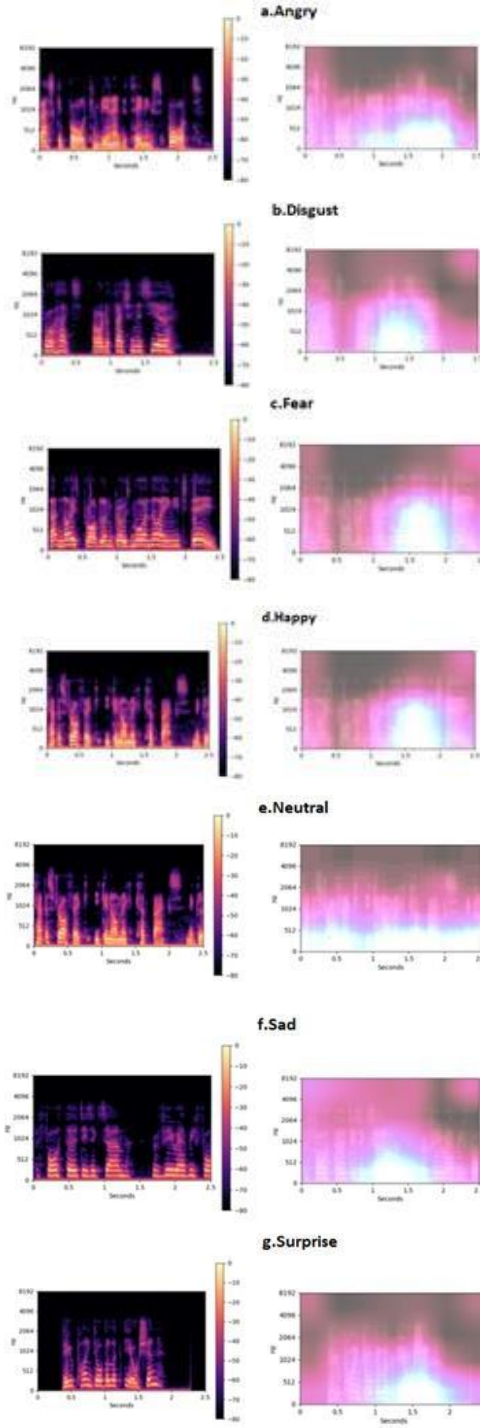

Figure 6. Mel-spectrogram image (on the left) and with the generated heatmap (on the right) for a) angry, b) disgust, c) fear, d) happy, e) neutral, f) sad and g) surprise emotion. The most important areas for the classification result are highlighted with white color, in contrary to the less important that are shown with darker colors. The vertical axis represents frequency (Hz) in log scale, whereas the horizontal axis represents time in seconds.

REFERENCES

[1] Pichora-Fuller, M. Kathleen; Dupuis, Kate, 2020, "Toronto emotional speech set (TESS)", https://doi.org/10.5683/SP2/E8H2MF, Scholars Portal Dataverse, V1W.-K. Chen, *Linear Networks and Systems* (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.

[2] Livingstone SR, Russo FA (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.

[3] S. Haq and P.J.B. Jackson, "Multimodal Emotion Recognition", In W. Wang (ed), Machine Audition: Principles, Algorithms and Systems, IGI Global Press, ISBN 978-1615209194, chapter 17, 2010, pp. 398-423.

[4] S. Haq, P.J.B. Jackson, and J.D. Edge. Audio-Visual Feature Selection and Reduction for Emotion Classification. In Proc. Int'l Conf. on Auditory-Visual Speech Processing, 2008, pp. 185-190.

[5] Mehmet Berkehan Akçay, Kaya Oğuz. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Communication,Volume 116,2020,pp. 56-76.

[6] Selvaraju, R.R., Cogswell, M., Das, A. et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Int J Comput Vis 128, 336–359 (2020).

[7] Janet R. McColl-Kennedy, Tracey S. Danaher, Andrew S. Gallan, Chiara Orsingher, Line Lervik-Olsen, Rohit Verma. How do you feel today? Managing patient emotions during health care experiences to enhance well-being, Journal of Business Research,Volume 79, 2017, pp. 247-259.

[8] Kallipolitis, A., Galliakis, M., Menychtas, A., & Maglogiannis, I. (2019). Emotion Analysis in Hospital Bedside Infotainment Platforms Using Speeded up Robust Features. AIAI.

[9] Radhakrishnan, Dr.Subhashini & Niveditha, P.R.. (2015). Analyzing and Detecting Employee's Emotion for Amelioration of Organizations. Procedia Computer Science. 48. 530-536.

[10] Oliveira, E., Martins, P. & Chambel, T. Accessing movies based on emotional impact. Multimedia Systems 19, 559–576 (2013).

[11] Otamendi FJ and Sutil Martín DL (2020) The Emotional Effectiveness of Advertisement. Front. Psychol. 11:2088.

[12] Yannakakis G.N., Karpouzis K., Paiva A., Hudlicka E. (2011) Emotion in Games. In: D'Mello S., Graesser A., Schuller B., Martin JC. (eds) Affective Computing and Intelligent Interaction. ACII 2011. Lecture Notes in Computer Science, vol 6975. Springer, Berlin, Heidelberg.

[13] Tan, M., & Le, Q.V. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. ArXiv, abs/1905.11946.

[14] Badshah, A. M., Ahmad, J., Rahim, N., Baik, S. W. Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network, 2017 International Conference on Platform Technology and Service (PlatCon), Busan, Korea (South), 2017, pp. 1-5.

[15] Zhao, J., Mao, X., Chen, L. Speech emotion recognition using deep 1D & 2D CNN LSTM networks, Biomedical Signal Processing and Control, Volume 47, 2019, pp. 312-323.

[16] Lee, C. et al. "Convolutional Attention Networks for Multimodal Emotion Recognition from Speech and Text Data." ArXiv abs/1805.06606 (2018): n. pag.

[17] Mittal, T., Guhan, P., Bhattacharya, U.,Chandra, R., Bera, A., Manocha. D."EmotiCon: Context-Aware Multimodal Emotion Recognition Using Frege's Principle," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 14222-14231, doi: 10.1109/CVPR42600.2020.01424.

[18] "otalk/hark," GitHub. [Online]. Available: https://github.com/otalk/hark. [Accessed: 25-Feb-21].

[19] Egger M., Ley M., Hanke, S. Emotion Recognition from Physiological Signal Analysis: A Review, Electronic Notes in Theoretical Computer Science, Volume 343, 2019, pp.35-55.

[20] A. Menychtas, M. Galliakis, P. Tsanakas and I. Maglogiannis, "Real-time integration of emotion analysis into homecare platforms," 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 2019, pp. 3468-3471.

[21] "Webrtc" [Online]. Available: https://webrtc.org/ [Accessed:20-Mar-21]