Use of CNN to assess emotions evoked by auditory stimuli in videos

Douglas H. S. Abreu, Lucas H. Ueda, Marta D. Fernandez, Vítor Y. Shinohara, Bruno S. Masiero, Paula D. P. Costa

Abstract— **Consumption of immersive audio content, e.g., via binaural reproduction through headphones, has increased over time. However, objective models that allow classifying audio systems through human perception are still poorly researched. Therefore, this study aims to evaluate the change in emotional state caused by immersive musical content. The evaluation was conducted subjectively, using a *Mean Opinion Score - MOS* model and objectively, by classifying the subjects' facial reaction with a Convolutional Neural Network, based on the VGG-16. The results obtained are favorable to the expected objectives, and may serve as a trigger for further studies.**

Keywords— **Emotions, Affective Computing, Immersive Audio, Convolutional Neural Networks, MOS.**

I. Introduction

Among the various existing definitions, music can be understood as the art and science of combining sounds in a pleasant way. It is believed that it arose when human beings began to organize themselves into tribes on the African continent around 50,000 years ago, from then on, it became part of human history. [13].

However, despite being present in everyday human life since the most remote antiquity, it was only in 1877 that the hypothesis of commercialization of sounds arose through an invention of Thomas Edison, the phonograph. Despite having a low quality of sound reproduction, it promoted a noticeable impact on the form of music consumption, giving rise to the phonographic industry. [10].

The music industry is always looking for improvements capable of adding value to the final artistic product, not only assimilating the emergence of modern technologies, but also encouraging their constant development. Therefore, the latest technological innovations experienced by the modern world and provided by the advent of electronics contributed to changing the systems and techniques of sound reproduction radically. [15].

Currently, there are several techniques available for sound reproduction, which were developed with the aim of creating more realistic audio reproduction systems, i.e., systems capable of reproducing the spatial characteristics of sound and, thus, meeting the needs of various modern applications, among which stand out those that encompass virtual reality technology [5], [11], [20].

The spatial sound reproduction itself can be performed using headphones or loudspeakers. If more than two loudspeakers are used, they can be driven in such a way as to recreate spatial impressions in a particular region of space. This area, usually in the center of the array, is called Sweet Spot.

Regarding spatial sound reproduction via headphones, it seeks to recreate spatial impression by presenting the listener with monaural and interaural cues, both Interaural Level Difference (ILD) and Interaural Time Difference (ITD). Binaural signals can be either captured with the aid of a recording mannequin or synthesized with the aid of a dataset of Head-related Transfer Functions (HRTFs) [12]. Another way to position virtual sources is via stereophonic techniques, presenting a correlated signal with different amplitude gains or temporal delays for each ear, resulting in a lateralization of the phantom source [16].

Audio quality presents a multidimensional concept and can be treated in different domains: physical, sensory, and/or affective, according to the application of interest and the context to be analyzed [8]. In this sense, it is hypothesized that sound stimuli reproduced by different modes of sound reproduction via headphones can generate a change in the individual's emotional state. The conceptual basis of this experiment is the field known as Affective Computing applied with Machine Learning techniques.

The emerging field of affective computing (AC) has been working on machines that automatically measure affective states for two decades. [14] defines AC as computing that involves or arises from human emotion, and says that it is an interdisciplinary field that integrates affective and computational sciences. AC contributes to affective science because of its advanced emphasis on detection, computational modeling, scalable objective measures, and real-world applications [6]. Therefore, the AC approach is conceptually grounded in affective science, in the studies of psychophysiology and nonverbal behavior. Its computational foundation lies in digital signal processing [19] and machine learning [4].

Thus, it is expected that this work will serve as a trigger for future research on the development of objective methods to evaluate sound systems that provide sound immersion. It is timely that these methods are based on human perception.

Finally, to validate this statement, an experiment was developed involving volunteers in a suitable environment for this purpose. All participants signed an Informed Consent Term (ICF), and the research was approved by the Ethics Committee (CEP-3743451). In this report, we describe the details of the development of the work, the algorithm used for such analysis, the results, and future work possibilities.

Douglas H. S. Abreu, DECOM/FEEC, Unicamp, Campinas-SP, e-mail: douglashsabreu@gmail.com

II. Fundamentals

It is essential for the understanding and future reproduction of this work to present a brief review of the database and the machine learning algorithm used for this work.

A. Database for modeling

When working with machine learning models, one of the most crucial aspects that influence the quality of the trained algorithm is the proper choice of the database to be used, better known as a dataset.

The Cohn-Kanade Extended (CK+) dataset [9] is commonly and widely used when it comes to dealing with facial expressions in images and videos [21]. The CK+ database contains 593 video sequences from a total of 123 different subjects, ranging from 18 to 50 years of age, with a variety of genders and heritage. Each video shows a facial shift from the neutral expression to a targeted peak expression — one of Ekman's six emotions. Figure 1 illustrates the progression of emotion along with the progression of video frames.

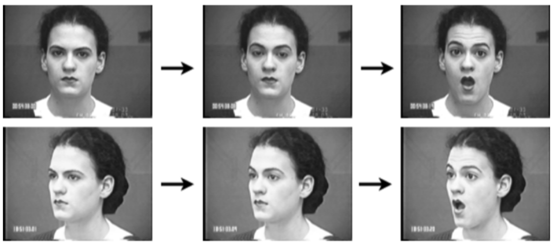


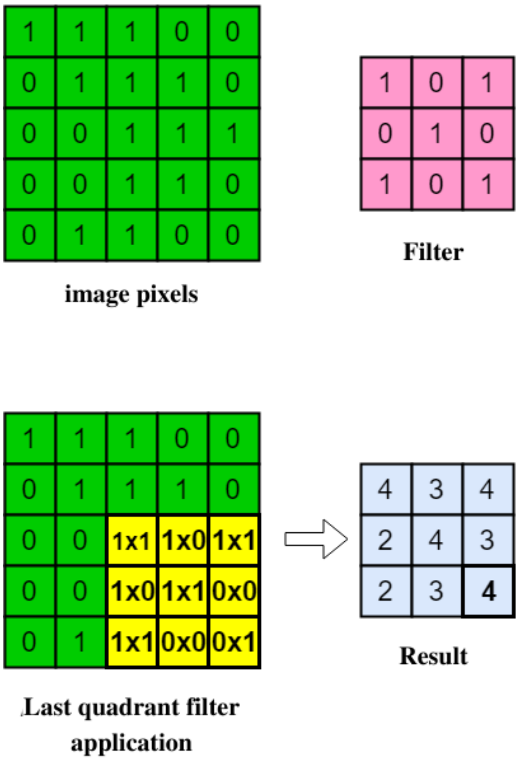
Fig. 1
Video progression of an actor sketching the emotion of surprise.

B. Modeling using Convolutional Neural Networks (CNN)

Convolutional Neural Networks (CNN) is a machine learning algorithm usually used for classifying videos and images due to its operation. This algorithm is like conventional neural networks, where there are several layers with one or more neurons. However, only CNNs have layers that allow the application of filters that give importance and enhance different patterns present in images, such as shapes, edges, and texture [1].

The presence of two types of layers in a CNN can be mentioned. The convolutional layer and the pooling layer. The convolutional layer is responsible for applying filters that enhance specific patterns. On the other hand, the pooling layer makes the neural network invariant to distinct positions of the object present in the image, that is, the position whose object is arranged in the image does not affect the ranking or result of the algorithm.

1) Convolutional Layer: The convolutional layer aims to elaborate and apply a particular filter or kernel to the image introduced in the layer. The filter is an $n \times m$ array with different weights, which can be optimized by the convolutional neural network to emphasize distinct parts of the image, enhancing shapes, edges, and textures in the image. Subsequently, the output of the convolutional layer is used as input for other layers of the neural network architecture. Figure 2 illustrates the convolution process in an image, represented in pixels.



1	1	1	0	0
0	1	1	1	0
0	0	1	1	1
0	0	1	1	0
0	1	1	0	0

image pixels

1	0	1
0	1	0
1	0	1

Filter

1	1	1	0	0
0	1	1	1	0
0	0	1x1	1x0	1x1
0	0	1x0	1x1	0x0
0	1	1x1	0x0	0x1

Last quadrant filter application

4	3	4
2	4	3
2	3	**4**

Result

Fig. 2
Convolution process in pixels of any image.

As can be seen in Figure 2, a 3×3 filter is defined, represented by a matrix. Then, the scalar product of the filter is performed on a 3×3 set on the image, represented in pixels that was provided for the layer. Afterward, the filter is shifted by one pixel and the product is computed again. The process is repeated until all pixels are covered by the filter. [1].

2) Pooling Layer: The second type of layer found in a CNN is the pooling layer, whose function is to avoid that the rotation or displacement of an image influences the functioning of the algorithm. Among the types of Pooling layers, we can mention Max Pooling, Average Pooling, and Sum Pooling.

Figure 3 illustrates the operation of the max pooling and average pooling layers. These layers perform the summarization of a set of pixels provided as input. The max pooling type obtains the maximum value of a given quadrant, representing it in just one number. On the other hand, the average pooling type represents the average of the quadrant in a single element.

C. Mean Opinion Score - MOS

The method known as the "benchmark quality indicator" is a popular indicator of perceived quality used in several types of media. It can be used with a variety of approaches, both subjective and objective. The opinion score is defined by the

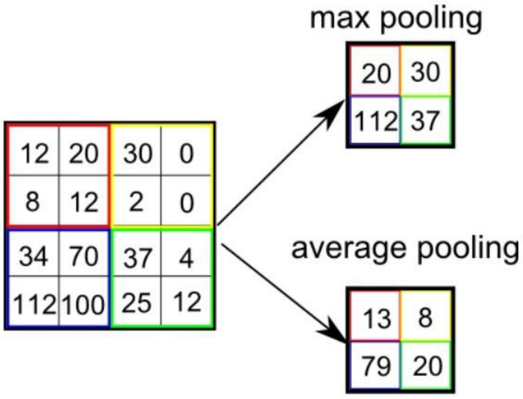


Fig. 3
TWO EXAMPLES OF POOLING APPLICATION: MAX POOLING AND AVERAGE POOLING

International Telecommunications Union (ITU)[1], as a "value on a predefined scale that a subject assigns to his opinion about the performance of a system". Therefore, when averaging this assessment, we have the well-known Mean Opinion Score (MOS). The MOS is considered a metric to quantify the perceived media quality, with its scale defined in five points: excellent, good, fair, poor, and bad [18].

III. METHODOLOGY

To carry out the experiment, several steps and procedures were necessary. Initially, volunteers who attend FEEC/UNICAMP were recruited. The set-up environment aimed to be as neutral as possible so that the participant did not have a deviation of attention during the tests.

Four distinct binaural audios were selected, where two of them were songs and the other two were soundscapes captured from famous places in big cities[2]. For each binaural audio, a monophonic version of it was also generated. In possession of the set of eight audios, an algorithm was created to randomly select, for each test subject, the order of execution of the audios. In parallel with the process of selection and generation of audios, a questionnaire was delivered for the participant to respond, for each presented audio, the following question: *"How do you feel when you listen to each of this audio?"*

When being directed to sit on the chair, the participant received a headset and was invited to sit in front of the high-resolution camera responsible for capturing the image of their face, as shown in Figure 4. After the reproduction of each audio, a pause was given to allow annotation in the form. Finally, the volunteer was required to inform the criteria used in the subjective assessment described above.

The videos with the volunteer's facial expressions were cut and labeled to identify which audio it belonged to, thus allowing an analysis to be made using computational tools. However, when subjectively analyzing the videos, it was noticed that the participant's facial expression when receiving auditory stimuli showed only minor variation, with subtle or non-existent movements. Because of this, the objective of the design and modeling of the machine learning algorithm was focused on the identification of valence, being positive, neutral, or negative, instead of emotion.

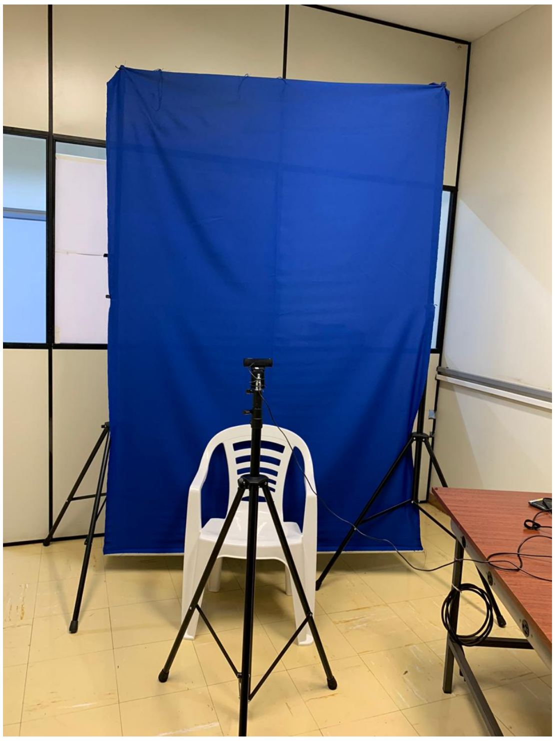


Fig. 4
TESTING SCENARIO.

A. Emotional Modeling

Regarding the emotional modeling related to this research, it was initially assumed the use of 6 emotions proposed by Ekman, being anger, joy, sadness, surprise, fear, and disgust, since they are universally recognizable and easily interpreted through facial expressions. [2].

However, it was analyzed through preliminary results that audio presentation yielded insignificant variation in facial expression, with subtle and often imperceptible movements. In addition, there is the difficulty of identifying the emotion of disgust in audios. Based on the problems discussed, valence recognition was adopted instead of Ekman's six emotions, since it can be identified and analyzed more easily and fits in the context of facial expressions together with audio.

B. Database

To carry out the project, the CK+ database was used. However, it was necessary to pre-process it, since the database provides videos categorized into six Ekman emotions, but the objective of the work was to identify valence. Therefore, an adaptation of the database was carried out: the first four frames of each video were labeled as neutral, the emotion "Joy" was labeled as positive, Surprise was disregarded, and

[1] `https://www.itu.int/`
[2] `http://urban-soundscapes.org/`

all other emotions were labeled as negative, thus modeling the categorical database (six emotions) on data classified by valence.

C. Modeling the Convolutional Neural Network

For the modeling of the machine learning algorithm, a well-known convolutional network (CNN) architecture was used in the field of image classification, called VGG-16 [17]. The primary motivation for using the VGG-16 architecture was its high hit rate in the ImageNet database [3], reaching 90% accuracy in labeling more than 3 million images.

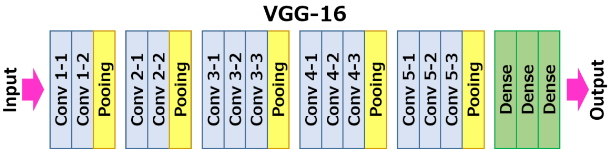


Fig. 5

VGG-16 Convolutional Neural Network Architecture

Figure 5 illustrates the sequence of convolution layers, pooling layers, and finally, dense layers. It is important to emphasize that the parameters of the dense layers were adjusted for the problem addressed in this report, which consists of identifying valence through facial images.

Finally, the videos were submitted to the computational model using a library called Dlib [7], which identifies the face and thus makes it possible to disregard the rest of the image. The extracted face was sent as input to the CNN, which returns three values as output, classified as positive, neutral, or negative, for each frame.

IV. Results

After the completion of the experiments, it is possible to perform a brief analysis of the results obtained. Initially, the MOS was calculated for each of the audios, thus allowing the comparison of subjective and objective responses (result from CNN output). Figure 6 shows how each of the audios was subjectively analyzed by the participants and makes it clear that the audios containing music present a better evaluation in relation to the audios with soundscapes. However, as objectified by this work, a comparison between binaural and mono reproduction is necessary, and it is noticeable that higher values in the valence representation were obtained for all binaural audios compared to their respective monaural versions.

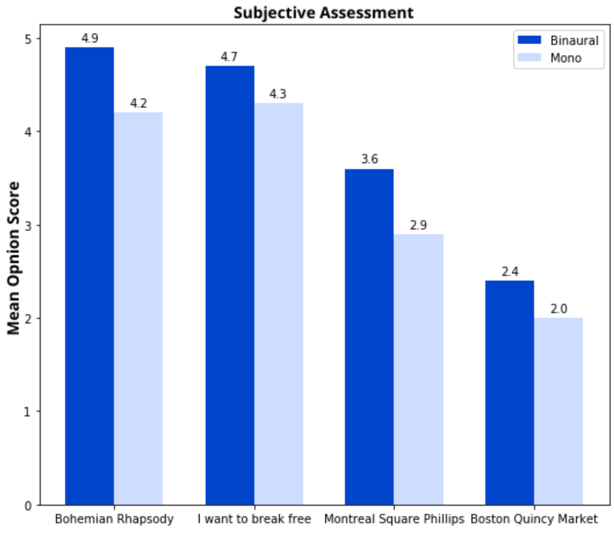


Fig. 6

Subjective Assessment.

By supplying the videos as input to the algorithm modeled in the development of this work, it was possible to obtain three output values, labeled as positive, neutral, and negative. These outputs are complementary, forming a total equal to 1, in other words, a probability distribution is performed for each analyzed frame. It was defined to use the results of the output classified as positive to perform the analysis. In the audios with the presence of music, it is possible to identify a difference, with the positive evaluation showing higher values for binaural reproduction in relation to mono reproduction. In the audios with soundscapes, the difference was less than 1%, and therefore it can be considered that there was no difference between them. Figure 7 details the analysis described.

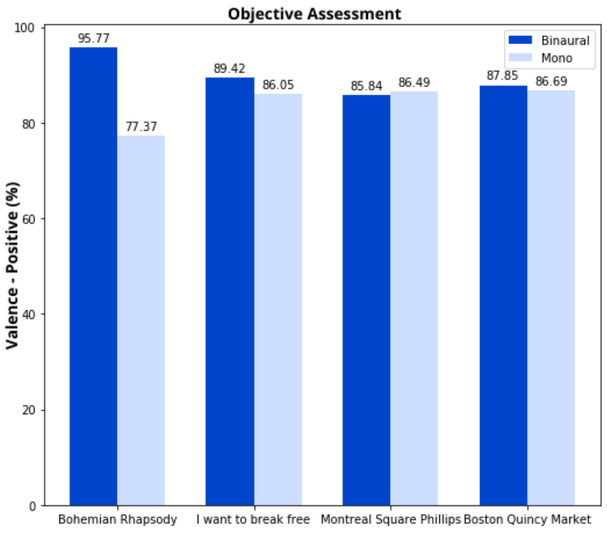


Fig. 7

Objective Assessment.

For both subjective and objective analysis it can be said that binaural reproduction presents a better evaluation than mono reproduction, which can serve as a trigger for more detailed analysis.

V. Conclusions

This research aimed to evaluate the existence of a change in the individual's emotional state when presented with immersive sounds. For this, a supervised machine learning algorithm was modeled to analyze facial images, and finally, to identify the change in the emotional state of a volunteer when presented with immersive sound or not.

Through the analysis of the videos obtained by the experiment, it can be observed that the variation of the facial expression of an individual listening to binaural and mono audios is minimal. Therefore, it was necessary to model the machine learning algorithm for recognizing micro-expressions. However, there were great difficulties finding public datasets to perform the model training and, later, the tests.

Even not contemplating the concept of facial microexpressions, the CK+ database aims at the variation of expressions, which consists of easily detectable emotions. And, even with this limitation, the results obtained were satisfactory.Thus, it was possible to identify changes in the emotional state through the tendency of binaural reproduction to have a more positive classification than mono reproduction.

Finally, for future work, we consider two main points of attention. One is to develop a different neural network architecture, which can model the frame sequences carrying temporal information. The other would be to use a database focused on micro-expression in conjunction with the convolutional neural network architecture referenced in this report.

The codes used in this work are available at `https://github.com/douglashsabreu/CodeSBrT2022Paper`, however, it is important to note that the videos containing images of the participants were not shared for privacy reasons.

Acknowledgment

This study was partially financed by the Coordination of Superior Level Staff Improvement (CAPES) - Finance Code 001, and by the National Council for Scientific and Technological Development (CNPq) - grant #141782/2019-6.

References

[1] Christopher M Bishop. *Pattern recognition and machine learning*. Springer Science+ Business Media, 2006.

[2] Randolph R Cornelius. *The science of emotion: Research and tradition in the psychology of emotions.* Prentice-Hall, Inc, 1996.

[3] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009.

[4] Pedro M Domingos. A few useful things to know about machine learning. *Commun. acm*, 55(10):78–87, 2012.

[5] Regis Rossi Alves Faria. *Auralização em ambientes audiovisuais imersivos.* PhD thesis, Universidade de São Paulo, 2005.

[6] Arvid Kappas. Smile when you read this, whether you like it or not: Conceptual challenges to affect detection. *IEEE Transactions on Affective Computing*, 1(1):38–41, 2010.

[7] Davis E King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(Jul):1755–1758, 2009.

[8] Gaëtan Lorho. *Perceived quality evaluation: an application to sound reproduction over headphones.* PhD thesis, Aalto University, Espoo, Finland, 2010.

[9] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 94–101. IEEE, 2010.

[10] Bruno Masiero. Qual o futuro do mp3? áudio espacial e codificação orientada a objetos. *14º Congresso de Engenharia de Áudio*, pages 96–105, 2017.

[11] Bruno Sanches Masiero and Michael Vorländer. Spatial Audio Reproduction Methods For Virtual Reality. In *Anais del 42o Congreso Español de Acústica*, pages 1–8, Cáceres, Spain, 2011.

[12] Henrik Møller. Fundamentals of binaural technology. *Applied acoustics*, 36(3-4):171–218, 1992.

[13] J. Paulino. Música. *INTERESPE - Interdisciplíniaridade e Espiritualidade na Educaão*, 1(3):99–100, 2013.

[14] Rosalind W Picard. *Affective computing*. MIT press, 2000.

[15] Ville Pulkki and Matti Karjalainen. *Communication acoustics: an introduction to speech, audio and psychoacoustics*. John Wiley & Sons, 2015.

[16] Francis Rumsey and Tim McCormick. *Sound and recording: An introduction.* Focal Press, 2009.

[17] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[18] Robert C Streijl, Stefan Winkler, and David S Hands. Mean opinion score (mos) revisited: methods and applications, limitations and alternatives. *Multimedia Systems*, 22(2):213–227, 2016.

[19] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing: Survey of an emerging domain. *Image and vision computing*, 27(12):1743–1759, 2009.

[20] Michael Vorländer. From acoustic simulation to virtual auditory displays. In *Proceedings of the 22nd International Congress on Acoustics (ICA), Buenos Aires, Argentina*, 2016.

[21] Yacine Yaddaden, Mehdi Adda, Abdenour Bouzouane, Sebastien Gaboury, and Bruno Bouchard. Facial expression recognition from video using geometric features. In *8th International Conference of Pattern Recognition Systems (ICPRS 2017)*, pages 1–6. IET, 2017.