AFFECTIVE ACOUSTIC SCENE ANALYSIS TO DETECT THREATENING SCENARIOS

PACS: 43.60.NP, 43.60.+D, 43.60.−C, 43.60.BF, 43.66.LJ, 43.75.ZZ, 43.90.+V

Rituerto-González, Esther; Luis-Mingueza, Clara; Peláez-Moreno, Carmen
erituert@ing.uc3m.es; carmen@tsc.uc3m.es
University Carlos III of Madrid, Av. de la Universidad 30, Leganés, Madrid, Spain

Keywords: acoustic events, affective computing, emotions, affective acoustic scene, sound analysis

ABSTRACT. (Arial, línea 25, tamaño 10, alineado izquierda, interlineado sencillo, max 200 words).

Affective Acoustic Scene Analysis is an emerging research topic which can be framed as a subfield of Affective Computing, aiming at analysing an acoustic environment when it elicits emotions or affective states in a person from a cognitive standpoint. As part of a project to combat gender-based violence, where we analyse physiological variables, speech and the acoustic scene looking for life-threatening scenarios, we present a methodology for the analysis of acoustic scenes and emotions. It is derived from a use case of an unsupervised system where the emotional information linked to an acoustic scene emerges from the analysis of its acoustic events and sounds, in order to contextualise and complement the emotional information retrieved by other sensors or the detection of risk situations.

RESUMEN.

El Análisis de Escenas Acústicas Afectivas (*Affective Acoustic Scene Analysis*, *AASA*) es un tema de investigación emergente que puede enmarcarse como un subcampo de la Computación Afectiva (*Affective Computing*) encargado de analizar un entorno acústico cuando éste provoca emociones o estados afectivos en una persona desde un punto de vista cognitivo. En el marco de un proyecto de lucha contra la violencia de género, en el que se analizan variables fisiológicas, el habla y la escena acústica para detectar situaciones que pongan en peligro la vida de las personas, presentamos una metodología para el análisis de escenas acústicas y emociones. Está derivado de un caso de uso previo, que utiliza un sistema no supervisado en el que la información emocional vinculada a una escena acústica emerge del análisis de los eventos acústicos y sonidos, con el fin de contextualizar y complementar la información emocional capturada por otros sensores, para la detección de situaciones de riesgo.

1. INTRODUCTION

The environments or scenarios in which we find ourselves and the situations we experience can generate emotions in us, as for example, attending a football match can generate excitement, sitting in a hospital waiting room can generate nervousness, hiking in a natural park can produce joy and calm, and hearing steps when walking down a lonely street can generate anxiety. Humans capture stimuli through their senses, and emotions and human senses are very tightly intertwined. It is through the sight that we can appreciate a landscape or a nicely-decorated room, through smell that we can get in the mood for eating, and the acoustics of a setting can make the difference between feeling comfortable in it or not.

The acoustics of a scene can influence our mood. In the field of Affective Computing [1], the branch of study of artificial intelligence that refers to the interpretation and understanding of human emotions, research fields are emerging regarding the affective auditory stimulus analysis and affective sound processing which study the emotions in sounds.

As a consequence, the question arises as to how to analyse real-world acoustic environments and their effects and influence in our affect and emotions. In previous works we already analysed the effects of synthetically added acoustic events to stress detection in speech [2], but in this paper we focus on realistic events composing an acoustic scene. We define an acoustic scene as a scenario that consists of the combination of different audio signals such as speech, sounds, acoustic events and ambient noise, and we aim to study and characterise a scene with respect to its acoustics and the emotions it provokes in the people who are immersed in it. By achieving this, we could be able to extract the information related to the emotions that resides in the acoustic signals from an acoustic environment, what we could name as its *affective acoustic fingerprint.*

In this work we make a brief literature overview and aim to describe a methodology to analyse how real-world acoustic environments can affect and influence human emotions. From the background of a project to combat gender-based violence, we propose a novel methodology to the processing of acoustic scenes to characterise them affectively, that is, regarding the affective state they induce. We also collect, under the Related Work section, similar studies under the analysis of emotions and acoustic scenes. Besides, we present a case of use of this methodology for the UC3M4Safety Audiovisual stimuli database [3], a database collected for emotion elicitation in women as part of a project to protect them against gender-based violence [4], [5]. We use this dataset with the objective of characterising audio clips acoustically to the emotion they elicit, with the ultimate goal of using them for the detection of gender-based violence life-threatening situations.

The analysis of *Affective Acoustic Scenes* ought to be bi-directional, used both for the prediction of emotions and affects induced by an acoustic environment, as well as for the design of such acoustic scenarios with the aim of setting a mood in the subjects immersed in it. Together with artificial intelligence, this field can be used to generate wellbeing and calm in scenarios for the healthcare system, to create immersive experiences together with emotions of excitement for entertainment, as well as for appropriately eliciting emotions which is especially important for affective computing research.

2. RELATED WORK

Although some work on the relationship between acoustic scenes and emotions exists in the literature, it has not been collectively identified or specifically defined. There is not a single title or acronym, as for example with the widely known field of Speech Emotions Recognition (SER) where a solid corpus of work is being developed. Thus, we found related work on acoustic scenes and emotions under different names: Assessments of Acoustic Environments by Emotions, Emotions in Soundscapes [6], Emotional [Acoustic] Scene Understanding, Induced Emotions in Sonification [7], Evoked Emotion Recognition by General Sound Events, Sound Design Theory [8] or Acoustic Design of Virtual Environments [9] among others. However, despite the disunite and limited number of works, there is still some promising research in the field. Our purpose with this paper is to provide an overarching view of this subfield, collecting it under the term of *Affective Acoustic Scenes Analysis* (AASA*).*

The motivation of such works in the literature is to provide machines with the ability of understanding what a person is experiencing from her acoustic frame of reference. This includes her acoustic contextual information, meaning the situation and auditory surroundings of the person.

This work [7] points out the two types of emotions associated with sounds: (1) “perceived” emotions, in which listeners recognize the emotions expressed by the sound, and (2) “induced” emotions, in which listeners feel emotions induced by the sound. Listeners may widely agree on the perceived emotion for a given sound, however they often do not agree about the induced emotion, so it is difficult to model them. The study develops machine and deep learning models that predict the perceived and induced emotions associated with sounds, and achieves moderate results concluding that predicting emotions is a difficult task but it is easier for perceived than for induced emotion.

Other studies have a more theoretical approach, as for example [6] includes a thorough overview of the theoretical background that applies emotion theory to soundscapes in the context of emotions that are elicited by them. The study concludes by stating that a deeper understanding of emotions elicited by soundscapes and their measurability would be a significant step forward for research.

2.1 Datasets

The data used for the analysis of emotions in acoustic scenes can be either audio of realistic environments, audio from movie or video clips, or virtual ambiences mimicking real environments, among others.

Some datasets have been captured with the aim of understanding what a person is experiencing from her situations and surroundings but based on the Computer Vision field. One example is the EMOTions In Context (EMOTIC) Dataset [10] which, in order to estimate emotion in the wild from visual information, it contains images with people in real environments, annotated with the apparent emotions both categorically (26 emotion categories) and in continuous dimensions (Valence, Arousal and Dominance). The work in [11] presents a learning-based algorithm for context-aware perceived human emotion recognition using 3 interpretations of emotions, including the semantic context of such images, achieving better results than previous state-of-the-art models for the categorical emotions.

In the field of emotions and acoustic scenes, the authors in [12] targeted sound emotion recognition of realistic acoustic environment conditions. They captured the Emotional Sound Database by selecting 390 sound clips from different areas of the daily human environment and modelled them using the dimensional approach in the emotional arousal and valence space, which were annotated by four persons instructed with the perceived emotion. Their findings indicate that sound perception is thus wrapped up with emotional response and affect, and they find spectral features to be most important as a group after individual prosodic features for Arousal and Valence prediction in the different sound categories.

Another example of a database that includes sounds and emotional labels is the IADS-E database [13]. It contains 935 digitally recorded natural sounds common in daily life, such as babies crying, typing, footsteps, background music, and sound effects. The sounds in IADS-E are divided into ten semantic categories (containing ‘animals’, ‘people’ and ‘scenarios’ among others). Each sound is labelled by annotators according to the induced emotion in the affective dimensions of valence, arousal, or dominance/control, using the Self-Assessment Manikin and on three basic emotion-rating scales (happiness, sadness, and fear).

Additionally, the Emo-Soundscapes database [14] contains 1,213 6-second audio clips labelled using a crowdsourcing listening experiment, and ranks the audio clips according to the perceived valence/arousal of the listener. This dataset allows studying soundscape emotion recognition and how the mixing of various soundscape recordings influences the perceived emotion.

The database we used in our use case for this methodology is the UC3M Audiovisual Stimuli database [3]. It consists of movie clips, videos of indoors and outdoors scenarios, and video compilations, annotated with the perceived emotional labels according to more than 1,300 raters by crowd-sourcing. Of such videos, 19 are categorised as *fear* and the 24 remaining are labelled with categories of other 9 discrete emotions. Separating audio from video can be used to create standard representations of acoustic soundscapes and emotions.

3. METHODOLOGY

In this section we detail our proposed methodology for *Affective Acoustic Scenes Analysis* (*AASA)* step by step. We put forward that this is a more comprehensive alternative to the classical machine learning setting that extracts features from audio signals and then plugs them directly into a machine learning model for inference, that also facilitates interpretability and accountability. Figure 1 illustrates such methodology in a block diagram.

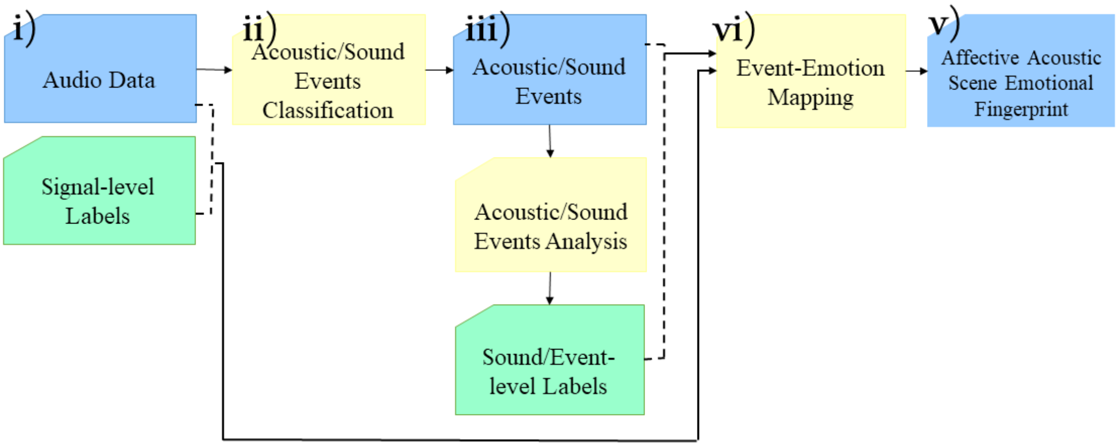

i)
Audio Data
Signal-level Labels
ii)
Acoustic/Sound Events Classification
iii)
Acoustic/Sound Events
Acoustic/Sound Events Analysis
Sound/Event-level Labels
vi)
Event-Emotion Mapping
v)
Affective Acoustic Scene Emotional Fingerprint

Figure 1 – Block Diagram of Affective Acoustic Scene analysis methodology

Starting on the blocks from the left, i) the first step is to use audio data, specially useful if being from realistic conditions or a synthetic mixture that imitates such (such as a Virtual Reality Environment, movie clip, realistic video game,...). Ideally, such data would be labelled according to affective states or emotions perceived by the users that actively listen to it. Affect labels such as arousal, valence or dominance, pleasure, categorical emotions and liking, could be used. The aim of these labels is to reflect the emotion or affective state perceived by a person that is immersed in such acoustic environment.

As the next step, which could be optional, ii) an acoustic events detection or classification module can be applied, that identifies the acoustic events or sounds from the audio signal. Such module might be a pre-trained machine learning model with databases that include emotional labels from sounds, iii) and so a relationship or alignment between the detected sounds and their emotional component annotated could be found.

Once the acoustic events or sounds have a corresponding emotional label, iv) the mapping between the two needs to be analysed, either in a supervised or unsupervised manner, with an algorithm that can evaluate the relation between the acoustic events or sounds and the emotional labels. This step can be performed with any pair or data-label, for instance, the separate acoustic events together with the whole original signal-level emotional label. Finally, v) an acoustic emotional fingerprint or embedding is extracted from the analysis, which condenses the emotional information from the analysed audio.

3.1. Challenges

An important challenge that arises is related to the intensity of the emotional event, that is, its emotional saliency. This is a biologically adaptive cue that can influence how an event is remembered and possibly how it is integrated in memory.

Moreover, different sounds or acoustic environments can lead to different emotions elicited in the listeners based on their previous experience and the memory associations that the sounds evoke. Thus, there may be a majority emotional reaction, but we should not forget the individual differences in each person, specifically in the case of women who have suffered or are suffering from gender-based violence.

As we have already mentioned, the associative and relational memory component of an acoustic event can also play a role in the emotional reaction of a person. The sounds of keys opening a door may be a sound of joy meaning welcoming a loved person, but for a victim of gender-based violence it may mean that her abuser has arrived. The emotional effect can be completely different even though the acoustic event may be the same. Therefore, the need for a method that can be adapted and personalised is of paramount importance in this field with such a high level of subjectivity.

3.2. Acoustic Events Detection and Classification

As part of the block diagram represented in Fig. 1, optionally we can aim to classify the acoustic events occurring in the audio data available, for such, we could employ pre-trained sound event classification models, able to detect acoustic events or sounds. We refer to this step as optional because a direct analysis of the complete acoustic signal and its emotional label could be also performed, but we believe this step to be key to identify the acoustic events composing an audio signal so that our later interpretation is more transparent and direct, more explainable.

The field of Sound Event Detection (SED) is [15] a research field of AI in which different approaches have been developed and used for the detection of sound events, oftentimes imitating the human auditory system, and including different feature sets and detection algorithms. An example of such algorithms is YAMNet [16], a Convolutional Neural Network (CNN) pre-trained on 521 classes of AudioSet [17]. This data is a large-scale collection of human-labelled sound clips drawn from YouTube videos. The network is ready to perform inference over audio files to classify occurring sound-events. YAMNet outputs are multilabel, and it is able to classify between 521 weak annotations of sound events.

This sound detection can help us to emotionally characterise an audio signal, relating the acoustic events that appear with the emotion the audios tend to elicit, finding this relationship in other sound databases that have emotional sounds and labels.

3.3 Algorithms for Sound-Emotion Mapping

This is the core part of the methodology, an algorithm that analyses the relationship between acoustic events or sounds and elicited emotions.

Somehow, we have to extract from the audio signals the most salient or relevant moments, which allow us to condense the information, within the audio signal, that can trigger an emotion in a listener. One way would be to extract audio features from the sound signal, as if the task were a speech emotion recognition task, for subsequent emotional classification using ML prediction models.

Another type of process that can be used and which is the one we use in our use case, is the TF-IDF (Term Frequency - Inverse Document Frequency) algorithm. It is a statistical method widely applied in Information Retrieval that evaluates how relevant a word is to a document in a collection of documents. Taking the acoustic events as words, the audio signal as a document, and the whole dataset of audio signals is equivalent to the collection of documents, we obtain a vector of tf-idf scores per audio signal which represents the affective acoustic fingerprint of each potential emotional trigger of each audio.

3.4 Acoustic Emotional Fingerprints or Embeddings

Once we have extracted the acoustic emotional embeddings or fingerprints, they could be used as input for machine learning models. These can be supervised – re-using the emotional labels as ground truth labels, as for ML regression or classification models – or unsupervised, using some kind of clustering or similarity metric to be applied. We consider it is also key that the results could be visualised, with the help of explainability models (XAI), to verify and interpret the accountability of the results gathered.

4. CASE OF USE

In our use case [18] of AAS analysis, 43 audios detached from videos of the UC3M Audiovisual Stimuli database [3] collection are used to create a standard representation of acoustic information that induces certain emotions. Each video – or audio, in this case – has been annotated with a discrete emotion label from 10 categories by more than 50 raters via crowd-sourcing, corresponding to the emotion that it elicits.

The audio component contains different types of sounds – speech, music, sound effects – and the first step is to detach the audio from the video component. Then, to identify the acoustic events occurring in the audio data we employ the pre-trained sound event classification model YAMNet. Afterwards we make use of the TF-IDF algorithm, taking the acoustic event labels predicted by YAMNet as words, and the audio signals eliciting emotions as documents, where our dataset of audio is equivalent to the collection of documents. We obtain a vector of tf-idf scores per audio clip – with one value per acoustic event label – being the representation of the affective acoustic fingerprint, which is directly related to the emotional label.

Finally, with the purpose of computing the similarity between each pair of affective acoustic fingerprints of each audio on the dataset collection for its visualisation and understanding, we use a similarity metric based on the cosine distance. Cosine similarity is widely used in information retrieval as a simple and effective way of providing a useful measurement of how similar two documents are likely to be, independently of the length of such documents. Thus, as the audios used have different lengths, we rely on this distance to measure the similarity between the tf-idf vectors.

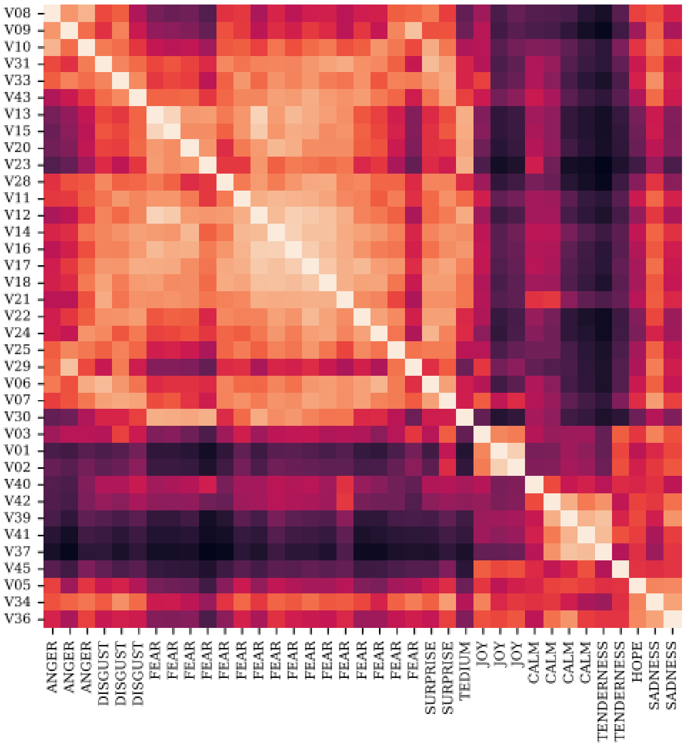
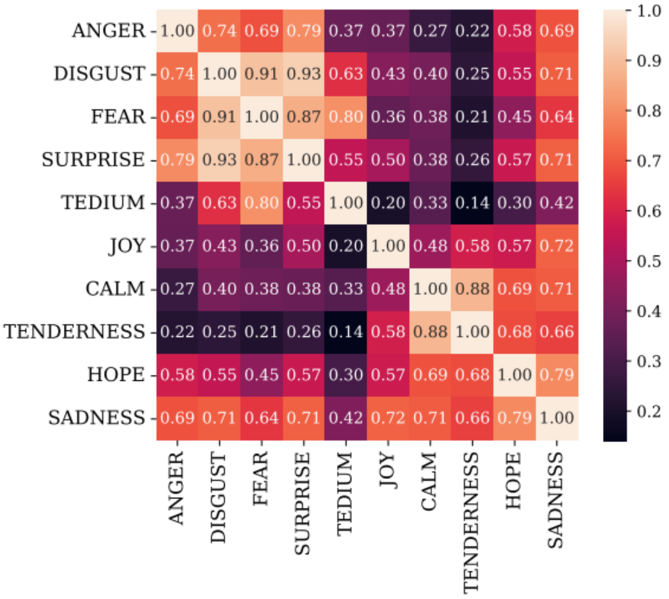


Figure 2 – Heatmap of cosine distance similarity between affective acoustic fingerprints, sorted by emotions, after removing outliers. Taken from [18]



.

Figure 3 – Heatmap of cosine distance similarity between affective acoustic fingerprints. Taken from [18]

In Figure 2 we represent as a heatmap the result of computing the cosine similarity for each audio with its labelled emotion with respect to the rest of audio signals, with a total of 37, after removal of outliers. Lighter colours on the heatmap represent higher similarity, and darker colours show lower similarity, between the affective acoustic fingerprints. As each video aims to trigger a single emotion, in this manner we can understand how each audio is related to its emotion and also to the representation of their corresponding emotions from the rest of audios.

We observe that similar emotions present similar coloured clusters in Figure 2, meaning that audios labelled with the same emotion, have a similar acoustic characterization. On Fig. 2, four clusters can be roughly observed: a big cluster including *anger, disgust, fear, tedium* and *surprise*, another cluster for *joy*, and another cluster for *calm* and *tenderness*, and the last one including *hope* and *sadness*. These four groupings are to some extent consistent with the similarity in the PAD space on the Valence and Arousal axes [19] of these emotions.

Afterwards, we performed the mean of the tf-idf matrix for every audio labelled with the same emotion category. In that manner we can understand how each acoustic label impacts the categorization of each emotion. In Figure 3 we present the resulting heatmap. We observe how the results are promising, as similar emotions present a greater similarity between them (e. g. *calm* and *tenderness*), than emotions that humans categorise as more different (e. g. *tedium* and *joy*). In particular, the *fear* category lays close to the *disgust* and *surprise* labels, which hinders the discrimination between them if we only take into account the acoustic context.

The relationship between *fear* and *anger* is peculiar, as contrary to what we would expect, they present a great similarity. This could be explained taking into account the gender bias [3], that states that in certain situations, people can feel different emotions to the same stimuli depending on their gender. This deserves further investigation.

5. CONCLUSIONS

In this paper we have proposed a step-by-step methodology for the analysis of acoustic scenarios, with events or sounds, and the perceived emotions within the scope of emotions in sound research. We exemplified its use by a case of use, representing the emotional acoustic fingerprints of sound clips by using the cosine similarity, the tf-idf method, the acoustic events classifier YAMNet and the novel multimodal UC3M Audiovisual Stimuli Dataset with favourable results. We recapitulate the field work although under different headings, which for future work a more exhaustive compilation could be made. And we opened a path for the exploration of the emotions that certain acoustical environments can elicit in us. The presented case of use could be extended to all the crowdsourcing annotated clips of the initial stage of UC3M4Safety, being that a total of 80 clips.

ACKNOWLEDGEMENTS

This work has been partially supported by the Dept. of Research and Innovation of Madrid Regional Authority, in the EMPATIA-CM research project (reference Y2018/TCS-5046), SAPIENTAE4BINDI Project by Grant PDC2021-121071-I00 funded by MCIN/AEI/10.13039/501100011033 and by the European Union “NextGenerationEU/PRTR". The authors also thank the Spanish Ministry of Science, Innovation and Universities (MCIU) for the FPU grant FPU19/00448 and the rest of the members of UC3M4Safety for their support.

REFERENCES

[1] R. W. Picard, *Affective Computing*. MIT Press, 2000.

[2] E. Rituerto-González, C. Luis-Mingueza, y C. Pelález-Moreno, «Using Audio Events to Extend a Multi-modal Public Speaking Database with Reinterpreted Emotional Annotations», en *IberSPEECH 2021*, mar. 2021, pp. 61-65. doi: 10.21437/IberSPEECH.2021-13.

[3] M. Blanco-Ruiz, C. Sainz-de-Baranda, L. Gutiérrez-Martín, E. Romero-Perales, y C. López-Ongil, «Emotion Elicitation Under Audiovisual Stimuli Reception: Should Artificial Intelligence Consider the Gender Perspective?», *Int. J. Environ. Res. Public. Health*, vol. 17, n.º 22, Art. n.º 22, ene. 2020, doi: 10.3390/ijerph17228534.

[4] J. A. Miranda *et al.*, «WEMAC: Women and Emotion Multi-modal Affective Computing dataset». arXiv, 8 de junio de 2022. doi: 10.48550/arXiv.2203.00456.

[5] J. Miranda *et al.*, «Bindi: Affective Internet of Things to Combat Gender-based Violence», *IEEE Internet Things J.*, pp. 1-1, ene. 2022, doi: 10.1109/JIOT.2022.3177256.

[6] A. Fiebig, P. Jordan, y C. C. Moshona, «Assessments of Acoustic Environments by Emotions – The Application of Emotion Theory in Soundscape», *Front. Psychol.*, vol. 11, 2020, Accedido: 21 de septiembre de 2022. [En línea]. Disponible en: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.573041

[7] F. Abri, L. F. Gutiérrez, P. Datta, D. R. W. Sears, A. Siami Namin, y K. S. Jones, «A Comparative Analysis of Modeling and Predicting Perceived and Induced Emotions in Sonification», *Electronics*, vol. 10, n.º 20, Art. n.º 20, ene. 2021, doi: 10.3390/electronics10202519.

[8] T. Goerne, «The Emotional Impact of Sound: A Short Theory of Film Sound Design», ene. 2019. doi: 10.29007/jk8h.

[9] D. Västfjäll, «The Subjective Sense of Presence, Emotion Recognition, and Experienced Emotions in Auditory Virtual Environments», *Cyberpsychology Behav. Impact Internet Multimed. Virtual Real. Behav. Soc.*, vol. 6, pp. 181-8, may 2003, doi: 10.1089/109493103321640374.

[10] R. Kosti, J. M. Alvarez, A. Recasens, y A. Lapedriza, «Context Based Emotion Recognition Using EMOTIC Dataset», *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, n.º 11, pp. 2755-2766, nov. 2020, doi: 10.1109/TPAMI.2019.2916866.

[11] T. Mittal, P. Guhan, U. Bhattacharya, R. Chandra, A. Bera, y D. Manocha, «EmotiCon: Context-Aware Multimodal Emotion Recognition Using Frege’s Principle», en *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, jun. 2020, pp. 14222-14231. doi: 10.1109/CVPR42600.2020.01424.

[12] B. Schuller, S. Hantke, F. Weninger, W. Han, Z. Zhang, y S. Narayanan, «Automatic recognition of emotion evoked by general sound events», en *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, mar. 2012, pp. 341-344. doi: 10.1109/ICASSP.2012.6287886.

[13] W. Yang *et al.*, «Affective auditory stimulus database: An expanded version of the International Affective Digitized Sounds (IADS-E)», *Behav. Res. Methods*, vol. 50, n.º 4, pp. 1415-1429, ago. 2018, doi: 10.3758/s13428-018-1027-6.

[14] J. Fan, M. Thorogood, y P. Pasquier, «Emo-soundscapes: A dataset for soundscape emotion

recognition», en *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, oct. 2017, pp. 196-201. doi: 10.1109/ACII.2017.8273600.

[15] T. K. Chan y C. S. Chin, «A Comprehensive Review of Polyphonic Sound Event Detection», *IEEE Access*, vol. 8, pp. 103339-103373, 2020, doi: 10.1109/ACCESS.2020.2999388.

[16] M. Plakal and D. Ellis, «YAMNet», *GitHub*, enero de 2020. https://github.com/tensorflow/models/tree/master/research/audioset/yamnet (accedido 21 de septiembre de 2022).

[17] J. F. Gemmeke *et al.*, «Audio Set: An ontology and human-labeled dataset for audio events», New Orleans, LA, 2017.

[18] Luis-Mingueza, Clara, Rituerto-Gonzaléz, Esther, y Peláez-Moreno, Carmen, «Bridging the Semantic Gap with Affective Acoustic Scene Analysis: an Information Retrieval-based Approach», presentado en IBERSPEECH [accepted], Granda, Spain, nov. 2022.

[19] A. Mehrabian, *Basic dimensions for a general psychological theory: implications for personality, social, environmental, and developmental studies / Albert Mehrabian.* Cambridge, Mass: Oelgeschlager, Gunn & Hain, 1980.