Bridging the Semantic Gap with Affective Acoustic Scene Analysis: an Information Retrieval-based Approach

Clara Luis-Mingueza*, Esther Rituerto-González*, Carmen Peláez-Moreno

Department of Signal Theory and Communications, University Carlos III of Madrid, Spain
clmingueza@tsc.uc3m.es, erituert@ing.uc3m.es, carmen@tsc.uc3m.es

Abstract

Human emotions induce physiological and physical changes in the body and can ultimately influence our actions. Their study belongs to the field of Affective Computing, to improve human-computer interaction tasks. Defining an ‘affective acoustic scene’ as an acoustic environment that can induce specific emotions, in this work we aim to characterize acoustic scenes that elicit affective states regarding the acoustic events occurring and the available acoustic information. This is achieved by generating emotion embeddings to define the ‘affective acoustic fingerprint’ of such affective acoustic scenes. We use YAMNet, an acoustic events’ classifier trained in AudioSet to classify acoustic events in the WEMAC Audiovisual stimuli dataset. Each video in this dataset is labelled by crowd-sourcing with the categorical emotion it elicits. Thus we determine the relevance of the detected acoustic events that induce each emotion by performing an affective acoustic mapping, creating interpretable acoustic fingerprints of such emotions, by means of the well-known information-retrieval-based TF-IDF algorithm. This paper intends to shed light on the path to the definition of emotional acoustic embeddings.

Index Terms: acoustic scene, affective computing, acoustic events, emotion embeddings

1. Introduction

Acoustic Scene Analysis and Interpretation aims to explain the acoustic information in the environment often captured by a multi-microphone acquisition system [1]. In this paper, we want to contribute to the field by investigating the relationship between acoustic scenes and the emotions they can elicit, in a new task that could be termed as affective acoustic scene analysis.

Any real-life environment where we find ourselves in can elicit emotions in us, e.g., a traffic jam, a paradisiacal beach, a thunderstorm, birdsong,... Something similar happens when we visualize and listen to a situation occurring in an environment. Eliciting emotions that we experience in our lives by means of audiovisual stimuli and, in particular, those drawn from popular culture (e.g., films, music videos, etc.) is well established, since according to Ellis “the reality effect of moving images and sounds invokes many of the emotions that we experience during direct encounters in our lived spaces” [2]. This has been used to determine the acoustics of emotion in audio [3] in a pioneering work to find an holistic computational model of affect in sound. On it, the high degree of cross-domain consistency found that encoding the two main dimensions of affect–arousal and valence–was attributed to the co-evolution of speech and music from multimodal affect bursts, including the integration of nature sounds for expressive effects. However, these findings where established on the basis of enacted and spontaneous emotional speech, music and general sound events [4] in isolation. Targeting an holistic model that is able to explain affect in sound, we aim to characterize the emotions elicited by being immersed in a specific acoustic scene, taking into account the acoustic information in the environment as a whole. In particular, we adopt a setup based on information retrieval classical methods to produce a representation of the affective acoustic scene based on the well-known tf-idf (term-frequency–inverse document frequency) algorithm [5]–an approach based on information theory– [6], where we build the vector space of acoustic events occurring in a scene balancing the acoustic event frequency and the inverse scene frequency.

To infer the emotions embeddings space, we use the UC3M4Safety audio-visual stimuli dataset [7] designed to collect the multimodal dataset WEMAC recently released [8] and specifically designed to portray the emotion of fear. By using the cosine similarity function, we find that the tf-idf representation embeddings show the acoustic similarity of emotions as expressed in the dataset. Note that this emotional categorization is different (and could be complementary) to the classical Acoustic Scene Classification and Detection where scenes are typically related to the physical places to be characterized, e.g. airport, metro station or urban park.

The remainder of this paper is organised as follows. Section 2 outlines the state of the art regarding acoustic scenes and emotions. In Section 3, we introduce the dataset and techniques used. Sections 4 and 5 describe the application of the tf-idf method to determine the affective acoustic fingerprint or emotion embeddings and the results, and Section 6 contains the closing conclusions and further work.

2. Related Work

One of the barriers for research in Affective Computing is the scarcity of labeled datasets and the moderate size of the few available, due to the difficulty to label emotions. There is no universal agreement on how to categorize or measure emotions, and self-assessment annotations by a specific subject can differ from those annotated by observer external evaluators.

Specifically for emotional speech, there are a number of data resources (for example, SUSAS [9], UT-Scope [10], VOCE [11], BioSpeech [12], HEU Emotion [13]) that even contain other types of biosignals. However, there are not annotated databases—to the knowledge of the authors—that include acoustic events and speech recorded consequentially. In [14], we enriched BioSpeech with acoustic events to fill that gap.

As films are one of the most effective ways to elicit
emotions [15], our team UC3M4Safety1 recorded a database for fear recognition in the context of gender-based violence in the EMPATIA-CM2 project, and the first release, WEMAC, is freely available [8]. It includes the set of audiovisual stimuli [7, 16] which were displayed to users and we use in this paper. They were selected by a rigorous validation procedure of crowd-sourced annotations that rely on experts in diverse fields.

The common underlying representation of emotion triggers of sounds, music and speech is discussed in [3], but in spite of the abundant literature, pointing to the relevance of the acoustic environment and human emotions in the cognitive sciences (e.g. [17]), there are very few studies that investigate the relationship between acoustic events and the elicitation of emotions [4] and scarcely any investigate the relationship between fear and sounds [18].

Fortunately, the DCASE community3 has been releasing several datasets for the detection and classification of acoustic scenes and events since 2013. This has fostered a wealth of research contributions in this field. Moreover, a large-scale dataset of manually annotated audio events, AudioSet [19], triggered the investigation in deep learning models, several opened to the research community such as YAMNet [20]. This offers a robust alternative for the representation of the acoustic environment that can be transferred to other domains and tasks such as affective acoustic scene analysis, the one we target.

In the particular domain of Affective Sound Datasets, a large-scale dataset called ARAUŞ [21] allows classification with high-parameter models. However the perceptual labels available do not refer to a specific emotion but rather to the pleasantness, calmness or vibrancy of the sounds, in agreement with ISO/TS 12913-2[22]. Being able to train high-parameter models was desirable but not required for our purposes.

The Emo-Soundscapes dataset [23] consists on 1,213 synthetic mixes of 6s duration. The audio events are extracted from Freesound4 and each mix includes affective annotations based on the Self-Assessment Manikin (SAM) [24] (valence and arousal). However, for the purpose of this work it was preferable to consider a dataset that also provides labels related to the dominance of the emotion.

To the authors’ knowledge, no prior work focuses on the intrinsic emotional information of a soundscape and proposes a method to find direct and unsupervised relations between the audio events of an acoustic scene and its elicited emotion.

1https://www.uc3m.es/instituto-estudios-genero/UC3M4Safety
2https://www.uc3m.es/instituto-estudios-genero/EMPATIA
3https://dcase.community/
4https://freesound.org/

3. Methodology

3.1. Dataset

In this research, 43 from a total of 79 videos of the UC3M4Safety audio-visual stimuli dataset [7] [16] collection are used to create a standard representation of acoustic information and events that induce certain emotions . Each stimuli lasts between 30 – 120 seconds, and the collection consists of movie clips, ambience scenarios, and video compilations. In this subset from the first release, each video is assigned an emotion label by crowd-sourcing, corresponding to the emotion that it elicits in the viewers. Of such videos, 19 are categorized as fear and the 24 remaining are labeled with categories of other 9 discrete emotions.

3.2. Acoustic Events Classifier

The data we use is the audio component only from the audiovisual stimuli collection. It contains different types of sounds –speech, music, sound effects– that, along with the visual information, induce in the viewers the labeled emotions. To identify the acoustic events occurring in the data we employ a pre-trained sound event classification model: YAMNet [20].

3.3. Term Frequency - Inverse Document Frequency

TF-IDF (term frequency - inverse document frequency) [5] is a statistical method widely applied in Information Retrieval that evaluates how relevant a word is to a document in a collection of documents. This importance is given by a score, result of multiplying two metrics: the number of times such word appears in a document (TF), and the inverse document frequency of the word across a set of documents (IDF).

It works by increasing the score proportionally to the number of times a word appears in a document, but decreasing it when the number of documents that contain such word are high. The higher the TF-IDF score of a word, the more relevant it is in that particular collection of documents. These TF-IDF scores could be fed to machine learning algorithms as word vectors as they are a representation of the data.

In this paper we make use of the TF-IDF algorithm, taking the acoustic event labels predicted by YAMNet as words, and the audio stimuli eliciting emotions as documents, where our set of audio stimuli is equivalent to the collection of documents. We obtain a vector of tf-idf scores per clip –with one value per acoustic event label– which represents the effective acoustic fingerprint of potential emotional triggers of each video.

3.4. Cosine Distance

With the purpose of computing the similarity between each pair of TF-IDF vectors of each video of the UC3M4Safety dataset collection, we use a similarity metric based on the cosine distance (detailed in Eq. 1). 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 our videos have different lengths, we rely on this distance to measure the similarity between the affective acoustic embeddings represented by the tf-idf vectors.

4. Experimental Setup

In order to perform the affective acoustic scene analysis we first extract sound events from the audio waveforms that allow us to characterize the acoustic scene with YAMNet. Our goal here is to obtain a corpus of weighted label scores that represent the occurring sound events per time window, so that we can later establish a metric that measures how close these representations are within the gathered video stimulus of the UC3M4Safety audio-visual stimuli dataset. In this paper we are concerned with the construction and evaluation of the vector space of acoustic events and the vectors that represent the directions of the different emotions. Thus, the following pipeline has been applied and is publicly available on GitHub.

Each of the 43 videos from the collection elicits one emotion, validated by more than 50 users each\(^7\). Both the acoustic and the visual modality are the ones inducing these emotions. However it is the acoustic scene and context what we would like to further analyze, since apart from speech, these audios also contain information about the acoustic scene that causes such emotions. Thus, we first extract the audio only with the command-line tool flipnpe.

At the preprocessing stage, the audio signal is normalized, and converted to 16 kHz mono. Then a log-mel spectrogram of 64 bins is computed to extract a time-frequency representation of the audio signal as an image. Next we use YAMNet to detect and classify the acoustic events present in the audio signals of all the video stimuli. The time window is set to 0.96s and the patch hop is set to 1s. The network outputs one multi-event label matrix of 521 multi-class scores for each second.

As YAMNet is a general sound event classifier, it may produce very detailed class labels which may not provide useful information to our acoustic characterization given the audiovisual stimuli used, but would only make the task and descriptions more complex. So considering the Audioset Ontology, the children classes of Music and Animal labels are filtered out, except the classes of Music Mood and Wild Animals, where all subclasses are kept. From the total of 521 classes that YAMNet classifies, the filtered ones result in 351. Figs. 1 and 2 represent the word cloud of labeled audiovisual stimuli in the dataset, resulting in a scores matrix of dimensions (43, 351).

In Figure 4, as a heatmap, we represent the result of computing Eq. 1 for each audiovisual stimuli with its labeled emotion with respect to the rest of audiovisual stimuli, with a total of 37, after removal of outliers. Lighter colours on the heatmap represent higher similarity, and darker colours show lower similarity, between affective acoustic embeddings. As each video aims to trigger a single emotion, in this manner we can understand how

![Figure 3: YAMNet output of a sample of the UC3M4Safety audio-visual stimuli dataset with the temporal representation (top), spectrogram (middle, bands spanning 125 to 7500 Hz) and top events found (bottom).](https://example.com/figure3)

5. Results

In Figure 4, as a heatmap, we represent the result of computing Eq. 1 for each audiovisual stimuli with its labeled emotion with respect to the rest of audiovisual stimuli, with a total of 37, after removal of outliers. Lighter colours on the heatmap represent higher similarity, and darker colours show lower similarity, between affective acoustic embeddings. As each video aims to trigger a single emotion, in this manner we can understand how

\[ \text{similarity}(x, y) = 1 - \cos(\theta) = 1 - \frac{x \cdot y}{\|x\| \|y\|} \]
The outliers detection and removal was performed after comparing each affective acoustic embedding with the rest of the embeddings of the same emotion category. For instance, V32 was identified as an outlier considering that its embedding had a big dissimilarity with respect of the rest of embeddings labeled with disgust. Further analysis reveals that the acoustic context does not match the visual information, since V32—which is a video compilation—contains mostly classical music, similar to videos labeled in the calm category, and therefore its embedding is similar to these later emotion embeddings.

We can observe that similar emotions present alike colored clusters in Figure 4, meaning that videos labeled with the same emotion, have a similar acoustic characterization. On Figure 4, 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 (Please, Arousal, Dominance) space on the Valence and Arousal axes [25] of these emotions.

Afterwards, we performed the mean of the tf-idf matrix for every audiovisual stimuli labeled with the same emotion category. In that manner we can understand how each acoustic label impacts in the classification of each emotion. In Figure 5 we present the resulting heatmap, from the acoustic point of view, of emotion embeddings. We can observe how the results are promising, as similar emotions present a greater similarity between them (e.g. calm and tenderness), than emotions that humans categorize 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.

In Figure 6 we plotted the affective acoustic embeddings using the t-sne algorithm. We can observe that the distances and clustering between them are somehow similar to the grouping happening in Fig. 4.

The relationship between fear and anger is peculiar, as contrary than what we would expect present a great similarity. This could be explained taking in to account the gender bias [16], that states that in certain situations, people can feel different emotions to the same stimuli depending on their gender. This deserves further investigation.

6. Conclusions

In this paper we try to answer the question of whether it is possible to characterize an acoustic scene with respect to the emotions it elicits. We draw from the premise that characterizing the affective acoustic scene involves taking into account the acoustic context. And regarding the results presented, we seem to have achieved a favourable emotional characterization of the acoustic scene in audiovisual material, being a first start to affective acoustic scene analysis in real-world environments.

Other indicators besides the acoustic context—such as information from other modalities (e.g., bio-signals from the subject in which the emotion is induced)—could provide additional insight to differentiate and accurately characterize affective acoustic scenes of emotional triggers.

Two factors may be influencing the robustness of this analysis, first, the agreement among the annotators that labeled each video and their gender, and second, the amount of videos per each emotion category. In [16] the determination of the emotional label per audiovisual stimuli when there was a strong disagreement between the annotators, was selected to be the one annotated by the majority of women. Thus, the agreement is influenced by the gender of the annotators. Thus, as future work, a more insightful analysis with a more in-depth study could be carried out using the original set of videos—up to 79—or other databases of acoustic scenes with emotional annotations. The annotators agreement per gender as a variable can also be taken into account to study its relevance. Furthermore, using the tf-idf vectors as features, machine learning models could be fed with such data and predict emotional labels in supervised learning.

We conclude that using tf-idf with acoustic events’ labels is a promising method with interpretable results for characterizing an acoustic scene with respect to the emotional information. Robust embeddings that acoustically characterize emotions can be used to measure the emotional load of—or the emotion to be elicited by—the acoustic information in other databases.
7. 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), Grant PDC2021-121071-I00 funded by MCINAEI10.13039/501100111033 and by the European Union “NextGenerationEU/PRTR” and Grant PID2021-125780NB-100 funded by AEI. 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.

8. References