Combating Gender-based Violence Through the Perception of Acoustic Scenes: What is Most Informative in Audio for Soundscape Classification?

Claudia Montero-Ramírez
clmonter@pa.uc3m.es
Department of Signal Theory and Communications, Universidad Carlos III de Madrid Leganés, Spain

Esther Rituerto-González
Esther.Rituertogonzalez@med.uni-muenchen.de
Department of Psychiatry and Psychotherapy, Section for Precision Psychiatry, Ludwig-Maximilian University & Max Planck Institute of Psychiatry Munich, Bavaria, Germany

Carmen Peláez-Moreno
cpelaez@ing.uc3m.es
Department of Signal Theory and Communications, Universidad Carlos III de Madrid Leganés, Spain

ACM Reference Format:

1 PROBLEM STATEMENT

The automatic detection of anomalous or risky situations in real-world scenarios could represent a significant advancement in the prevention of gender-based violence (GBV) in our society. In 2022, our team UC3M4Safety1 developed the first multimodal dataset designed to address GBV in real-world scenarios: WE-LIVE. Using internet of things (IoT) wearable devices, WE-LIVE captures various multimodal signals, including audio, that are used to characterize the environments in a usual life setting [8].

One of the elements that constitute such an environment is context. Context is the interrelationship between the person, activity, and place in space and time. The context relates to the perception of sounds through what we call a soundscape. According to the International Organization for Standardization (ISO), a soundscape is the acoustic environment perceived, experienced, and understood by individuals within a given context [3].

By automatically classifying soundscapes through audio, we not only gather information about the context but also capture people’s perceptions based on their previous experiences. In the case of the detection of GBV situations, we aim to characterise acoustic scenes that elicit affective states, particularly fear, as it appears in violent or risky situations, to provide context to other artificial intelligence systems. To achieve this, the first step is to identify the most informative audio features for soundscape classification.

2 METHODS

Firstly, we conduct a review of the state-of-the-art in Acoustic Scene Classification (ASC). This review focuses particularly on the DCASE challenge [1] and the recently published article by Ding et al. (2024) [2], as well as previous research conducted by our research group [7] [5]. We selected three feature groups:

1) Librosa. It includes 38 hand-crafted features that capture the general characteristics of sound. These consist of cepstral features (MFCCs), time-domain features (zero crossing rate and root mean square), and frequency-domain features (spectral centroid, roll-off, flatness, and pitch). In total, there are 19 features from which statistical parameters (mean and standard deviation) are extracted, resulting in 38 features.

2) Timbral. It is a timbral characterisation tool from AudioCommons2 for semantically annotating non-musical content. It extracts 8 hand-crafted features that include psychoacoustic parameters, features related to sound quality and environmental descriptors.

3) Event Detection and Acoustic Unit Descriptors. These are vector representations derived through feature learning, based on the acoustic events or sounds present in an audio. Two models for acoustic event detection (YAMNet [6] and PANNs CNN14 [4]) are compared, followed by the use of information retrieval algorithms to obtain associated relevant numerical representations. These representations capture semantic information, along with others based on the occurrence of acoustic events (TF-IDF). Additionally, an end-to-end training is performed adding an additional layer on the end of SED pre-trained models.

Using static vector representations, a vanilla MLP (Multi-layer perceptron) with a single hidden layer is trained for each group of features, to evaluate how informative they are for the classification of soundscapes. The objective is to classify 8 soundscapes for 14

---

1https://www.uc3m.es/institute-gender-studies/UC3M4Safety
2https://github.com/AudioCommons/timbral_models
different users using audio recordings from their real-life situations, utilizing a highly unbalanced dataset. A nested cross-validation with 5 outer folds and 4 inner folds is conducted using a stratified k-fold data partition strategy.

3 RESULTS
Regarding static vector representations, we find that Librosa is the most informative feature group for classifying soundscapes in our dataset using a vanilla MLP model. Timbral features yield similar results, a significant finding as this feature group had not been previously used explicitly for ASC. Acoustic Unit Descriptors do not achieve good results using this architecture, presumably because dealing with the audio from this in-the-wild dataset is challenging for pre-trained Sound Event Detection (SED) models.

However, fitting the pre-trained models for our database, this assumption changes. PANNs CNN14 has the best performance for our database in ASC task. It is capable of obtaining information from the spectrogram, capturing temporal and frequency relationships.

4 SIGNIFICANCE
This work contributes to the fight against GBV by groundbreaking research on the automatic detection of acoustic scenes by using smart IoT devices. This is the first study to conduct soundscape classification on a real-world database specifically aimed at combating GBV, to capture the bidirectional relationship between the acoustic environment and the emotions it evokes in women, particularly following prior experiences.

REFERENCES