Applied Intelligence (2023) 53:27096–27109
https://doi.org/10.1007/s10489-023-04967-w

Automatic music emotion classification model for movie soundtrack subtitling based on neuroscientific premises

Maria Jose Lucia-Mulas[1] · Pablo Revuelta-Sanz[2] · Belen Ruiz-Mezcua[1] · Israel Gonzalez-Carrasco[1]

Accepted: 11 August 2023 / Published online: 1 September 2023

Abstract
The ability of music to induce emotions has been arousing a lot of interest in recent years, especially due to the boom in music streaming platforms and the use of automatic music recommenders. Music Emotion Recognition approaches are based on combining multiple audio features extracted from digital audio samples and different machine learning techniques. In these approaches, neuroscience results on musical emotion perception are not considered. The main goal of this research is to facilitate the automatic subtitling of music. The authors approached the problem of automatic musical emotion detection in movie soundtracks considering these characteristics and using scientific musical databases, which have become a reference in neuroscience research. In the experiments, the Constant-Q-Transform spectrograms, the ones that best represent the relationships between musical tones from the point of view of human perception, are combined with Convolutional Neural Networks. Results show an efficient emotion classification model for 2-second musical audio fragments representative of intense basic feelings of happiness, sadness, and fear. Those emotions are the most interesting to be identified in the case of movie music captioning. The quality metrics have demonstrated that the results of the different models differ significantly and show no homogeneity. Finally, these results pave the way for an accessible and automatic captioning of music, which could automatically identify the emotional intent of the different segments of the movie soundtrack.

Keywords Music emotion recognition · Automatic subtitling · Convolutional neural network

1 Introduction

Captioning is the reference assistive tool for hearing impairment. Captions are based on speech subtitling but include additional information such as sound effects, speaker identification, and other essential non-speech features. Captions are the “audio” for the deaf and hard of hearing. Special regulations have been issued to guarantee its application[1] and quality, considering factors such as synchronism, presentation speed, or accuracy, among others [1]. Pre-recorded captioning is the standard mode of captioning movies. Pre-recorded captions are produced after the movie has been created and are carefully checked for accuracy using specific software frameworks that ease tasks such as video file editing, audio frame localization, caption editing or preview. Based on deep learning, speech recognition technologies significantly reduce speech captioning time by automatically proposing the corresponding transcript for voice frames [2].

This study aims to evaluate other deep learning technologies that could be added to these frameworks to ease the task of music captioning. The capacity of music to generate emotions is widely used in movie soundtracks [3] as a support to the narrative [4, 5].

For example, the meaning of a wordless scene in which a character is seen from behind looking out a window is

✉ Israel Gonzalez-Carrasco
igcarras@inf.uc3m.es

Maria Jose Lucia-Mulas
maluciam@inf.uc3m.es

Pablo Revuelta-Sanz
revueltapablo@uniovi.es

Belen Ruiz-Mezcua
bruiz@inf.uc3m.es

[1] Computer Science Department, Universidad Carlos III de Madrid, Av. Universidad, 20, 28915 Leganés, Madrid, Spain

[2] University of Oviedo, C Luis Ortiz Berrocal s/n, 33203 Gijón, Spain

[1] In Spain, the General Law of Audiovisual Communication requires captioning for at least 90% of all public television broadcasts.

changed by a few seconds of happy, sad, or frightening music. Accessible captioning must include music information whenever it is important to help understand the plot, with a text summarizing the type of music, sensation transmitted, or identification of the piece, e.g., "(Horror Music)". The professional responsible for captioning the film decides when the music should be captioned and the feeling the author intended to convey. Deep learning technologies which could detect significant musical fragments, and propose the corresponding musical emotion, could contribute to the automation of this task.

Emotion investigation is a field of neuroscience research that has begun in relatively recent decades, and much remains to be known. Since the end of the last century, research has been developed based on two basic paradigms: the categorical model and the dimensional model of emotion [6].

The categorical model of emotion presupposes the existence of a limited number of basic, innate, and universal emotions. The Ekman model is the best known and considers seven basic emotions: fear, sadness, anger, happiness, surprise, disgust, and contempt [7]. Subsequently, studies have reduced this set to four "basic" emotions: happiness, sadness, fear, and anger [8]. On the other hand, the dimensional model of emotion states that emotions may be represented in a continuous space, generally of 2 or 3 dimensions. The hybrid model, "Circumplex model of affect", proposes that all affective states arise from cognitive interpretations of central neural sensations that are the product of two independent dimensions: one related to valence (positive/negative stimuli) and one related to arousal (activation) [9]. The discrete emotions would be subjective psychological "labels" that can be identified with points of that continuum space Valence-Arousal. In both models, it has been widely assumed that emotions are the subjective representations of primary neural circuits, basic for survival, that have evolved from the earliest complex animals [10]. Emotion would be a primitive adaptive mechanism that is triggered by critical stimuli for survival, prompting action. Thus, some authors consider that music could activate biologically important emotional circuits for processing sounds [11–13]. This primitive origin would explain the immediacy and universality of musical emotion concerning the basic emotions of happiness, sadness, and fear, which are the most identifiable in musical extracts when expressed with intensity [6, 14, 15]. The recognition of these basic emotions in music is consistent among listeners from the same culture [16], and among listeners from different cultures [17, 18].

In addition, this recognition is immediate, occurring in less than two seconds, with a simple chord or a few notes, when the music expresses the basic emotions of happiness, sadness, or fear [11, 14, 15]. In [15], the authors found an average time of 483 ms, 1446 ms, 1737 ms and 1261 ms for correctly recognising the happy, sad, scary, and peaceful excerpts, respectively. In [14], the authors, using a set of very short musical clips of on average 1.6 seconds, showed that the experimental subjects correctly categorised, and with great precision, the emotions associated with these clips and that even 250 milliseconds from the start of the music were enough in some cases to distinguish sad music from happy music. With a weaker musical expression of these emotions or concerning other emotions, the consensus among listeners decreases significantly [6].

One of the problems that neuroscientists encountered in these studies were deciding on the musical stimuli. [6, 15] have created standard scientific musical databases rigorously validated for musical emotion research, which have become a reference, and that are precisely based on movie soundtracks, as this is music composed to transmit powerful emotional stimuli. Musical fragments are labelled on the perceived basic emotions of joy, sadness, fear, and a fourth emotion, peacefulness/tenderness, which is not considered a primary emotion but is easily identified as a perceived musical emotional state. In these studies, evaluators are instructed to evaluate perceived emotions (the emotion music intends to represent) rather than induced emotion (the felt emotion). However, the border between the two is very diffuse and empirical studies show great similarity in both emotions [6].

The relationship between musical parameters and emotion is also gaining much interest. Many studies, in general, focused on the basic emotions of joy, sadness, and fear, show that mode, tempo, register, dynamics, articulation, and timbre are the most critical parameters that affect musical emotion and that these parameters operate additively (see Table 18). The relative importance of these parameters varies for each emotion. For example, the mode is extremely important for happy and sad emotions or articulation for fear [14, 19, 20].

The ability of music to induce emotions has also risen in the field of computer science and affective computing to a field of research dedicated to identifying the characteristics of music that generate different emotional states. This field, called Music Emotion Recognition (MER), has been arousing a lot of interest in recent years, mainly due to the boom in music streaming platforms and the use of automatic music recommenders [21–23]. MER is based on the analysis of low or medium-level characteristics of music. These characteristics are obtained from digital audio samples using the techniques of another nearby field of research, the so-called Music Information Retrieval (MIR). According to the review performed in [21], the first article in this field was published in 2003 [24]. In this work, the authors proposed a system for classifying songs into four emotional categories: happiness, sadness, anger, and fear, based on two musical characteristics, tempo (fast or slow) and articulation (staccato or legato). Since then, many studies on music emotion classification algorithms have been produced.

The typical scheme of development of an MER model is based on three steps: selection and labelling of digital musical samples (ground truth), selection and extraction of characteristics (features) from the audio digital samples, and application of supervised machine learning to map emotion with features. Each phase entails significant limitations [21, 25].

First, there is an absence of public, consensual and adequately validated datasets. In general, MER datasets are labelled in a variable and poorly controlled environment, without previous training of the evaluators and control of the evaluation process, evaluators or even labels changing throughout the evaluation process. For example, commonly used data sets such as Million Song Dataset [26], MTurk [27], or MagnaTagATune [28] are the result of a free annotation open to any user.

Second, regarding features, there is no agreement on the significant audio features to capture the musical emotion nor certainty in the validity of the algorithms used to extract them. Tools like Librosa,[2] Essentia,[3] or MirToolBox[4] allow the extraction of a large amount of audio information from which it is difficult to choose the significant parameters. Thus, different feature sets, grouping many of these characteristics, have been used to establish a predictive model [21, 25]. Still, it is unclear if the audio features used are sufficiently relevant to the problem [25].

1.1 Related work

Classification algorithms such as Gaussian Mixture Models (GMM), K-Nearest Neighbour (KNN), Support Vector Machines (SVM), and Support Vector Regression (SVR) are generally used [29–31] with SVM being the classifier that would obtain the best results[25, 32]. The review in [21] indicates that the highest accuracy achieved in emotion classification was 69.5% considering five emotional categories. In [25], the authors compared different results using SVM to classify musical fragments in the four quadrants of the "Circumplex model of affect", obtaining accuracies of up to 76.4%.

Another difficulty is choosing the length of the musical segments to be evaluated. In the case of a song a few minutes long, the emotional content can fluctuate temporarily. Usually, the song is divided into small segments to get more accurate results, detecting the emotion for each segment. For example, the typical segmentation length for popular music is usually 25-30 seconds [21]. For classical music, results obtained were optimal with lengths of 8-16 seconds (4-8-16-32 were tested) [33].

Recently, models based on feature selection have been replaced by Convolutional Neural Networks (CNN) models with promising results. The success of neural networks in image recognition has aroused interest in applying these networks using as input images the spectrograms obtained from audio samples, such as Short-time Fourier transform(STFT) or Mel spectrograms. [34] used a CNN network as a novel approach for music genre classification using MFCC (Mel Frequency Cepstral Coefficients) spectra, showing that CNNs had great potential to extract features from audio samples. In the review, [21], the best accuracy obtained with CNN was 69,5%. In [35], the authors have benchmarked the latest CNN architectures proposed in classifying musical genres. The results showed that the most straightforward architectures applied to short musical fragments (about 3 seconds) obtained the best results.

However, the field of automatic MER is wide open, and the mentioned accuracy rates are yet far from being fair enough to be used in automatic emotional labelling and captioning. In general, MER approaches are based on combining multiple audio parameters and machine learning techniques without considering the problem's main characteristics based on human musical perception and emotions. Computational models seem anchored in the labyrinth of MIR and machine learning algorithms, distancing themselves from the neuroscientific foundations of musical emotion perception.

1.2 Objectives and hypothesis

In this study, the authors aimed to approach the problem of emotion detection in movie soundtracks considering neuroscience results in emotion perception as a basis for approaching the problem.

Hence, as the first approach to a music captioning tool, it was decided to develop an automatic classification model to extract emotions from film music, starting from the following conclusions mentioned before and supported by neuroscientific studies:

- Use a basic classification of the emotions of happiness, sadness, and fear (expressed to an intense degree), which are best recognized in music, with consensus among subjects. In addition, the ones of most significant interest when it comes to movie soundtracks.
- Consider musical segments of 2 seconds, enough time to generate immediate musical emotion.
- Use CNN models, as there is no agreement on the significant audio characteristics for capturing musical emotion. CNN models allow working without prior selection of characteristics.
- Use the scientific musical datasets, Film Music Excerpts [6] and Musical Excerpts [15], as they are the only ones

[2] https://librosa.org/doc/latest/index.html

[3] https://essentia.upf.edu/

[4] https://www.jyu.fi/hytk/fi/laitokset/mutku/en/research/

based on the film musical genre, labelled with scientific rigour in terms of emotion from the field of neuroscience.

In summary, in this work, the authors propose a novel approach based on the latest neuroscientific evidence, and the results show an improvement in the state-of-the-art results.

The remainder of this paper proceeds as follows. Section 2 outlines the dataset used in the study. Section 3 discusses the classification methods of the research, and the two experimentation sets are defined. Section 4 presents the results obtained and a discussion about them. Finally, the paper ends with a summary of research findings, limitations and concluding remarks.

2 Materials and methods

2.1 Dataset description

It was decided to use only scientifically contrasted musical samples corresponding to intensely expressed emotions, which is the aim of music captioning. Thus, standard scientific musical databases mentioned above [6, 15] have been included in the research. Both datasets' musical excerpts are based on movie music and were labelled in a controlled experimental environment with different cross-tests to obtain validated results.

The Musical Excerpts dataset comprises 40 excerpts composed specifically in the film music genre [15]. The fragments are qualified based on four emotions (10 for each type of emotion): happiness, sadness, threat, and peacefulness. Each emotion's recognition rates are 99%, 84%, 72% and 94%. The average duration is 12.5 seconds. The copyright owner of these excerpts is Bernard Bouchard, and their use is permitted [15]. All 40 excerpts were included in the dataset used in this research.

The Film Music Excerpts dataset comprises a first set of 360 musical excerpts from 60 film soundtracks [6]. The excerpts offer examples of the emotions of happiness, sadness, fear, anger, and peacefulness, expressed in high and moderate intensity. These excerpts have been scored by experimental participants who evaluated the expressed emotion and the intensity of this emotion in a range from 1 to 7. In total, 94 fragments were selected. The selection criteria were to include only excerpts scored $>=6$ in the intensity of the expressed emotion. With less score, the consensus among participants on the expressed emotion decreased. The average duration of these fragments is 16 seconds. In total, 30 excerpts for happiness, 21 for fear, 24 for sadness, and 19 for peacefulness were selected.

The 10 peacefulness excerpts from [15] and the 19 tenderness excerpts from [6] were combined in a unique group under the label of peacefulness. As results in [6] and [15] show, peacefulness and tenderness are very close emotions and overlap in the continuum space Valence - Arousal, with little Arousal and positive Valence.

The selected musical fragments generated 976 samples of two seconds duration. Two seconds is enough time to generate immediate musical emotion, according to neuroscience results (Table 1).

The format of the samples was MP3, with an original sampling rate of 44,1KHz. Samples were down-sampled to 16KHz and divided into two-second samples (see Fig. 1). It was found that the results were not affected by reducing the sampling rate, while processing time was improved (see Table 4).

To be able to apply CNN models frequency spectrograms were used. For each 2-second sample, and using the Python Librosa library, three types of spectrograms were generated: STFT (frequency spectrograms), Mel (frequency spectrograms converted to the Mel scale), and Chromagram or Constant-Q-Transform (frequencies are represented on a logarithmic scale, corresponding to the different notes and octave bands C1, C2, C3, C4, etc.). Overlapping windows of 512 samples (length corresponding to about 31 milliseconds) were considered, with an overlapping length of 50% (see Figs. 2 and 3). It is noteworthy that the STFT spectrogram could be considered analogous to the Fourier analysis performed in the ear at the level of the basilar membrane within the cochlea, the Mel spectrogram to the non-linear human perception of frequencies, and the Constant-Q-Transform (CQT) spectrogram to the relative perception of the relationship between frequencies [36, 37].

This dataset was used in all experiments. Table 2 details the size of the input data corresponding to each type of spectrogram, considering two seconds of audio at 16KHz, and sliding windows of 512 samples and 50% overlapping.

Table 1 Emotion distribution in musical fragment

Emotion	Fragments	% from [15]	% from [6]	Total length (seconds)
Happiness	40	25%	75%	589.68
Fear	31	32%	68%	471.34
Sadness	34	29%	71%	558.99
Peacefulness	29	34%	66%	454.42

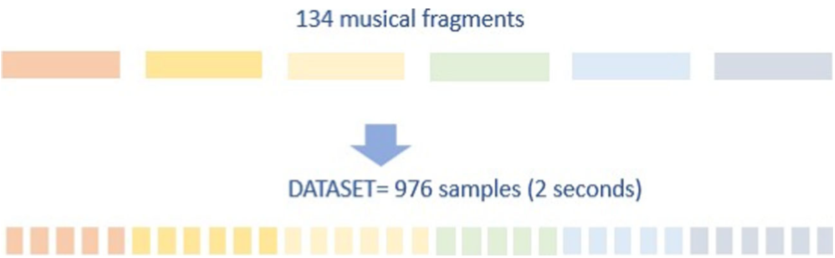

Fig. 1 Dataset description (musical fragments and samples)

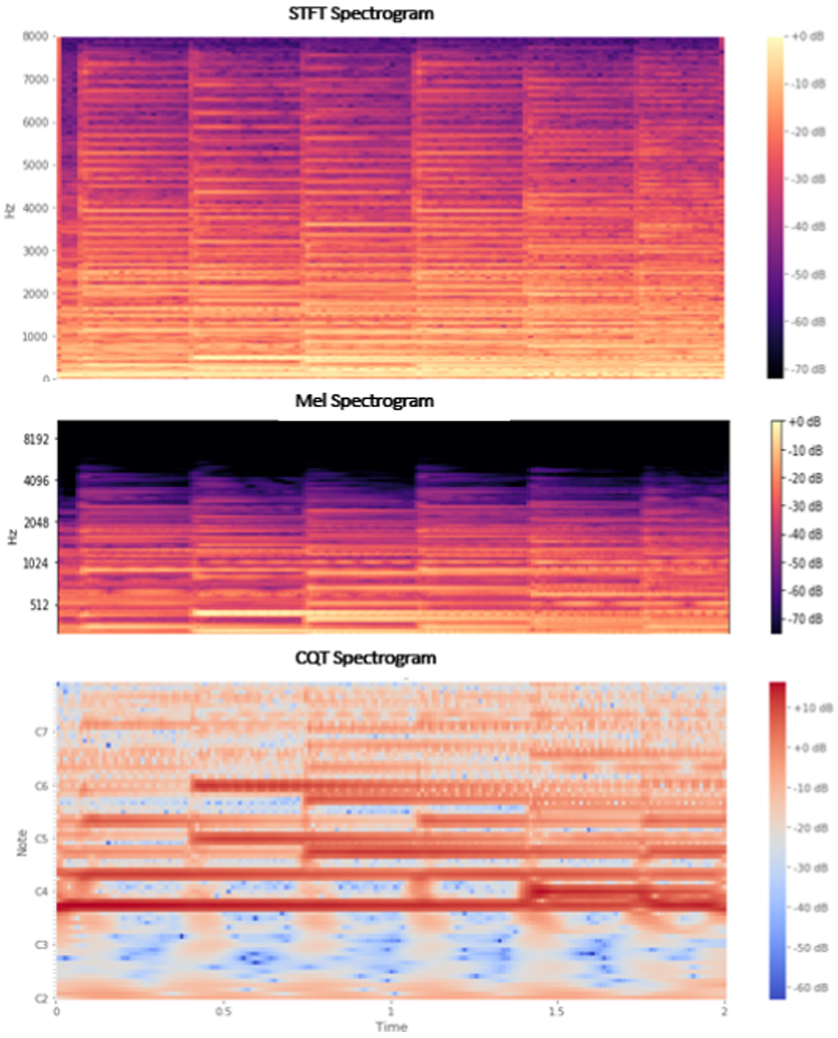

Fig. 2 STFT, Mel and CQT spectrograms corresponding to the six eighth-notes of the first measure of the excerpt in Fig. 3

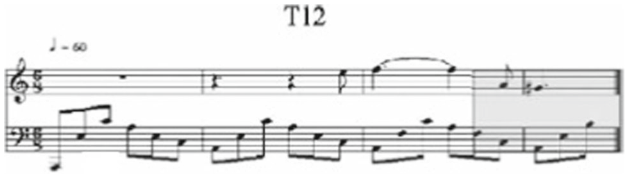

Fig. 3 Sad excerpt score from [15], recorded with digital synthesiser set to piano timbre

Table 2 Input data size for each type of spectrogram

Spectrogram	Input Size
STFT	257x126
Mel	128x126
CQT	82x126

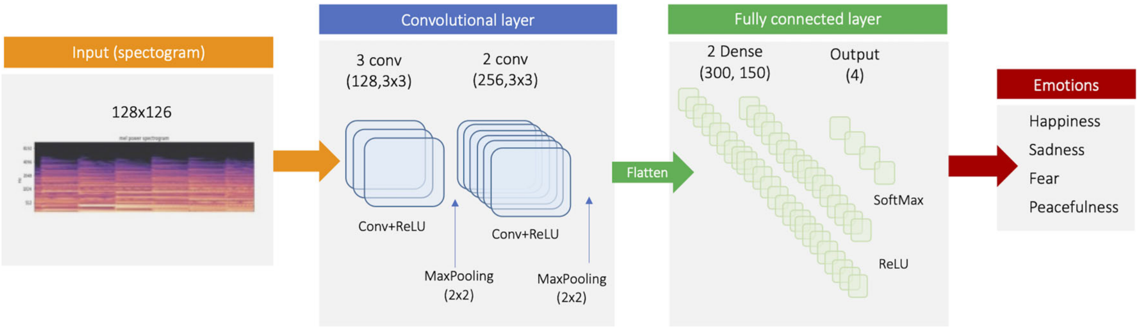

Fig. 4 CNN model for experimentation 1

3 Classification methods

3.1 Experimentation 1

In this first experimentation, the development of a basic CNN model was sought that would achieve recognition rates in line with state of the art to determine the most suitable type of spectrogram for the classification of the basic emotions of happiness, sadness, and fear.

The CNN model used was a standard convolutional network with few layers, with a typical structure Filters (3x3) + BatchNormalization + Activation ReLu + MaxPooling (2,2), inspired in [22], and adapted to obtain a classification based on four labels corresponding to the emotions of happiness, sadness, fear, and peacefulness. Several tests were conducted to refine the model with the three sets of STFT, Mel, and CQT spectrograms, adding and adjusting layers and hyperparameters.

Figure 4 summarizes the resulting model. Five convolutions with 3x3 filters and ReLU activation are considered in the convolutional layer. BatchNormalization is performed after each convolution, and Max Pooling is 2x2 every two convolutions. The result resized to a one-dimensional vector, is processed in the fully connected layer by two dense networks of 300 and 150 neurons, with ReLU activation and a final output layer with Softmax activation. Dropout is performed on these dense layers to avoid overfitting (Fig. 5).

Table 3 summarizes the parameters with which the best results were obtained and the parameters selected for the next experiments.

Once this base model was established, the different types of spectrograms were re-evaluated. For the model evaluation on each set of spectrograms, k-fold cross-validation was used (with k=10). k-fold cross-validation implies the partition of the data in k separate sets. This separation can be generally realized randomly in k disjunctions of a similar size. The training process repeats itself as many times as the sets which have been generated until all of the sets have been used to train and validate the network [38].

3.2 Experimentation 2

The second experimentation is based on [35], which shows the results of a benchmarking carried out with the most representative CNN models of state of the art for music genre classification, and it is especially of interest as it shows that the simplest architectures achieved better results in music genre classification with short musical segments. From this work, the authors have selected a standard CNN architecture and adapted it to classify the basic emotions of happiness, sadness, fear, and the musical emotion of peacefulness. This model is entitled CNN-4 model, it was adjusted with the hyperparameters defined in experimentation one (Table 3), and CQT spectrograms were used as input data. The model was also adapted to classify only the three basic emotions of happiness, sadness, and fear. In this case, the model is defined as the CNN-3 model.

Two variants of CNN-4, including Resnet and Inception modules, were also considered: (1) CNN + ResNet and (2)

Fig. 5 Training and validation datasets

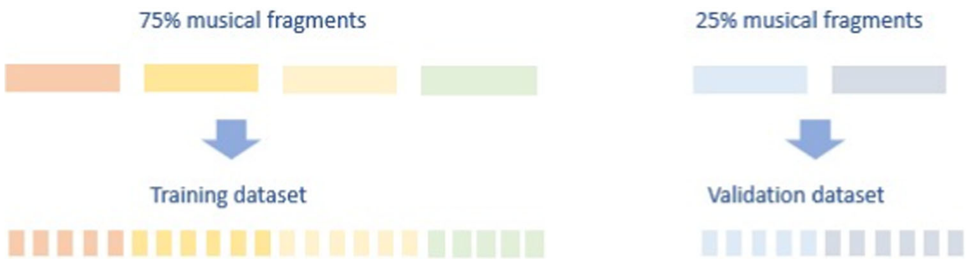

Table 3 Tuned hyperparameters (tested and selected)

Parameter	Tested Values	Selected Values
Sampling rate	16KHz, 32KHz and 44.1 KHz	16KHz
Learnig Rate	1e-2, 1e-3 and 1e-4	1,00e-03
Dropout	Layers Conv / Dense	Only in Dense Layers (0.25)
Optimizer	Stochastic gradient descent (SGD) Adaptive Moment Estimation (Adam)	SGD
Epochs	50,100 and 200	50

CNN+Inception, In the CNN + ResNet model, each convolution layer was replaced by a ResNet block (with the same number of filters and filter size). In CNN+Inception, the first convolutional layer was composed of parallel convolutions with different filters size.

Figure 6 summarizes the CNN-4 model. The model is very similar to the model developed in the previous experimentation but with more depth. In the convolutional layer, eight convolutions with 3x3 filters and ReLU activation are considered. BatchNormalization and 2x2 Max Pooling are performed after each convolution. The result, resized to a one-dimensional vector, is processed in the fully connected layer of 512 neurons, with ReLU activation, and a final output layer with Softmax activation. Dropout is performed in the Dense layer to reduce overfitting. The training data set for this model was the 976 CQT spectrograms based on the 2-second audio fragments with 16KHz sampling and size 82x126. The CNN-3 model was the same but with the output reduced to the three emotions of happiness, sadness, and fear.

To evaluate both models, k-fold cross-validation was first used (k=10). The evaluation was carried out with 50 epochs and repeated with 100 epochs with better results. Subsequently, the musical fragments were divided into a training set (75%, 732 samples) and a validation set (25%, 244 samples). Samples in the validation set belonged to musical fragments not included in the training set (see Fig. 5). This operation was repeated four times to have different combinations of musical fragments in the training and validation sets. This operation is called the training/validation phase, whose objective was to evaluate the generalizability of the models.

The entire process was applied to both CNN-4 and CNN-3 models.

Finally, k-fold cross-validation (k=10) was also applied to the variant models CNN + ResNet and CNN + Inception

4 Results

Regarding experimentation 1, the results depicted in Tables 4 and 5 were in line with state-of-the-art. In addition, it was observed that the best results were obtained with the CQT spectrogram. Slightly lower were the results with spectrograms Mel, and STFT. However, the training time was much longer with Mel and especially STFT spectrograms.

The training time increases with size. STFT inputs (257x126) require approximately double the time of Mel inputs (128X126), which approximately require double the time of CQT inputs (82x126). CQT spectrograms reduce dimensionality, hence less computational resources, while keeping the main characteristics of the musical sample Figs. 2 and 3.

Table 6 shows the average results obtained for the f1-score with CQT, Mel and STFT spectrograms in classifying the different emotions. It was again observed that the results

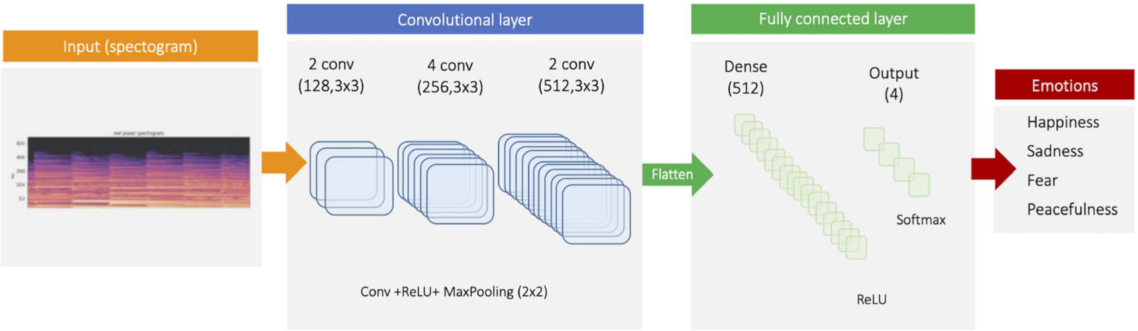

Fig. 6 CNN model for experimentation 2

Table 4 Experimentation 1: Mean Accuracy (MA) results and Processing times per spectrogram (PT)

	16KHz		32KHz		44,1KHz	
Spectrogram	MA	PT (mn)	MA	PT (mn)	MA	PT (mn)
CQT	0.789 (std=0.028)	38.00	0.786 (std=0.031)	80.00	0.788 (std=0.031)	99.00
Mel	0.763 (std=0.032)	60.00	0.713 (std=0.022)	130.00	0.621 (std=0.029)	165.00
STFT	0.766 (std=0.025)	144.00	0.701 (std=0.048)	240.00	0.703 (std=0.016)	360.00

with the type of spectrogram CQT were better than with Mel or STFT spectrograms.

Based on these results, CQT spectrograms were selected for Experimentation 2.

Results for the CNN-4 model are summarized in Tables 7 and 8. Cross-validation results (0.825 mean accuracy) improved from experimentation 1 (0.789 mean accuracy). But the results obtained in the training / validation phase were much lower (0.58 mean accuracy). These results show that the CNN-4 model does not generalize well (Tables 9, 10, and 11).

Results for the CNN-3 model are summarized in Tables 12 and 13. Cross-validation results (0.920 mean accuracy) improved compared to CNN-4 model (0.825 mean accuracy). This improvement is also observed in the generalization capacity, reaching a mean accuracy value of 0.79 (compared to 0.59 with CNN-4) in the training/validation phase.

CNN-3 model shows similar scores in precision and recall, that is, both in the ability not to classify negative samples as positive and in the ability to recognize all positive examples.

Tables 9 and 10 show the results obtained with CNN+ResNet and CNN + Inception models with kfold cross-validation (k=10, 100 epochs).

Table 14 compares the k-fold cross-validation results obtained with CNN+ResNet and CNN + Inception models. CNN+ResNeT and CNN+Inception models showed worst results than the CNN-4 model.

Table 5 Experimentation 1: Results for CNN Model with CQT

Cross-Validation (k=10, 50 epochs)	
k	Accuracy
1	0.78
2	0.78
3	0.74
4	0.72
5	0.84
6	0.81
7	0.83
8	0.78
9	0.86
10	0.74
Mean	0.79

Finally, Table 11 shows the k-fold cross-validation results obtained with CNN+ResNet and Mel spectrogram, which again show worst results than the CNN-4 model with CQT spectrograms. This test was performed in order to check if a deeper neural network could behave better with a spectrogram with higher dimensions.

4.1 Statistical analysis

In any empirical scientific work, when repeating an experiment in conditions which are indistinguishable to the researcher, it is very common for the results to show some variability; this is known as experimental error. Therefore, in any scientific experimental study, it is crucial to compare and evaluate the characteristics of the different sets of samples and the results obtained. In this research, and following the steps defined in [39], the validity of the results has been validated from a statistical standpoint, thereby reducing the appearance of experimental errors or the appearance of possible randomness.

Figure 7 includes the scatter-plot, box-plot, analysis of means and residual plot associated with the results. The scatter plot describes the behaviour of the set of samples obtained for each classifier through a point cloud. The box plot allows, through a simple visual inspection, to have an approximate idea of the central tendency (through the median), the dispersion (through the interquartile range), the symmetry of the distribution (through the symmetry of the plot) and the possible outliers of each classifier. The central line within the box describes the location of the sample median and the mean is represented by a cross. The graph also includes a notch for the median, the width of which roughly indicates the 95% confidence interval. In the analysis of the means plot, all the models are compared together with the overall

Table 6 Experimentation 1: Average results per emotion (F1 score)

Emotion	CQT	MEL	STFT
Happiness	0.84	0.85	0.81
Fear	0.86	0.83	0.84
Sadness	0.74	0.75	0.74
Peacefulness	0.70	0.54	0.58

Table 7 Experimentation 2: Results for CNN-4 Model

Cross-Validation (k=10, 100 epochs)	
k	Accuracy
1	0.81
2	0.81
3	0.82
4	0.85
5	0.79
6	0.79
7	0.83
8	0.90
9	0.86
10	0.80
Mean	0.82

Table 8 Experimentation 2: Results for CNN-4 Model

Accuracy	Precision	Recall	f1-score
0.60	0.62	0.61	0.58
0.66	0.67	0.68	0.65
0.54	0.59	0.55	0.52
0.62	0.63	0.66	0.62
0.58	**0.63**	**0.62**	**0.59**

Training/Validation (4 rounds, 100 epochs)

Table 9 Experimentation 2: Results for CNN-4 Model + ResNet Model

Cross-Validation (k=10, 100 epochs)	
k	Accuracy
1	0.75
2	0.75
3	0.79
4	0.52
5	0.78
6	0.72
7	0.68
8	0.79
9	0.73
10	0.71
Mean	0.72

Table 10 Experimentation 2: Results for CNN-4 Model + Inception Model

Cross-Validation (k=10, 100 epochs)	
k	Accuracy
1	0.78
2	0.61
3	0.67
4	0.78
5	0.50
6	0.67
7	0.59
8	0.70
9	0.65
10	0.68
Mean	0.66

Table 11 Experimentation 2: Results for CNN-4 Model + ResNet + Mel Spectrogram

Cross-Validation (k=10, 100 epochs)	
k	Accuracy
1	0.69
2	0.76
3	0.70
4	0.79
5	0.80
6	0.75
7	0.74
8	0.75
9	0.64
10	0.71
Mean	0.73

Table 12 Experimentation 2: Results for CNN-3 Model

Cross-Validation (k=10, 100 epochs)	
k	Accuracy
1	0.91
2	0.99
3	0.90
4	0.88
5	0.93
6	0.88
7	0.94
8	0.93
9	0.97
10	0.87
Mean	0.92

Table 13 Experimentation 2: Results for CNN-3 Model

Accuracy	precision	recall	f1-score
0.79	0.79	0.79	0.79
0.79	0.79	0.81	0.79
0.81	0.82	0.79	0.8
0.76	0.77	0.77	0.77
0.79	**0.79**	**0.79**	**0.79**

Training/Validation (4 rounds, 100 epochs)

mean and the 95% decision limits. The samples outside the decision limits, CNN-3 and CNN-4 model + Inception model are significantly different from the overall mean. Finally, the residual plot shows the residuals obtained for each of the alternatives. The residuals are equal to the values of the percentage correct minus the mean value for the group from which they come and show that the variability within each alternative is approximately the same.

In the box plot, the different boxes show an asymmetry in the distribution of the sample. In this case, the widths of the median notches, for a 95% confidence interval, are not similar. This suggests that there is a statistically significant difference between the medians at this confidence level.

Hence, in order to be able to apply a comparison of different models, it is necessary to ensure that there are no significant differences between the variances of the populations. Therefore, a variance check has been performed. The three statistics displayed in Table 15 test the null hypothesis that the standard deviations of the results within each of the five levels of models are the same. Since the smaller of the p-values is less than 0.05, there is a statistically significant difference between the standard deviations at a 95.0% confidence level. This violates one of the important assumptions underlying the analysis of variance and will invalidate most standard statistical tests (e.g. ANOVA or Analysis Of Variance method).

Once it has been determined that there is a statistically significant difference between the variances, the Kruskal-Wallis test is the most appropriate method for comparing populations whose distributions are not normal [40]. It is the non-parametric method, derived from the F-test, for testing the equality between the medians of a group of populations. The reason for using the median is that it is robust, i.e. not very sensitive to atypical data, while the mean is very sensitive. If the distribution is normal, the mean and median coincide, but if there is a discrepancy between the two, the median is preferable. Therefore, in the absence of normality, contrasts are relevant not on the mean, but on the median.

Table 14 Average Accuracy results of cross-validation in the different models evaluated

Model	100 epochs
CNN-4	0.82
CNN + ResNet	0.72
CNN + INCEPTION	0.66
CNN + ResNet + MEL	0.73
CNN-3	0.92

Table 15 Variance check

Contrast	Value	p-value
Cochran’s C test	0.3789	0.1387
Bartlett’s test	1.2790	0.0314
Levene’s test	0.9499	0.4441

The Kruskal-Wallis method shown in Table 16, tests the null hypothesis of equality of the medians within each of the 5 models. The data from all columns are first combined and ordered from smallest to largest. The median is determined by ranking the observations and finding the observation at the number [N + 1] / 2 in the ranked order. Then the median rank is calculated for the data in each column. The mean rank is the average of the ranks for all observations within each sample. Since the p-value is less than 0.05, there is a statistically significant difference between the medians at a confidence level of 95.0%. In Fig. 7, the Box and Whisker plot shows which medians are significantly different from each other (each box has a median notch).

Finally, the quality metrics have demonstrated that the results obtained with the CNN-3, as well as being higher on average, are significantly different and show no homogeneity with other models. This allows the researcher to incline towards this approach with no doubt about its fitness for this problem.

5 Discussion

The different experiments carried out show, on the one hand, that the CQT spectrograms, the ones that best represent the relationships between musical tones from the point of view

Table 16 Kruskal-Wallis Test

Model e	Mean Range
CNN3	45.2
CNN4	35.0
Inception	10.0
Mel	18.35
ResNet	18.95
Statistic = 38.2738	P-value = $9.8388 * 10^{-8}$

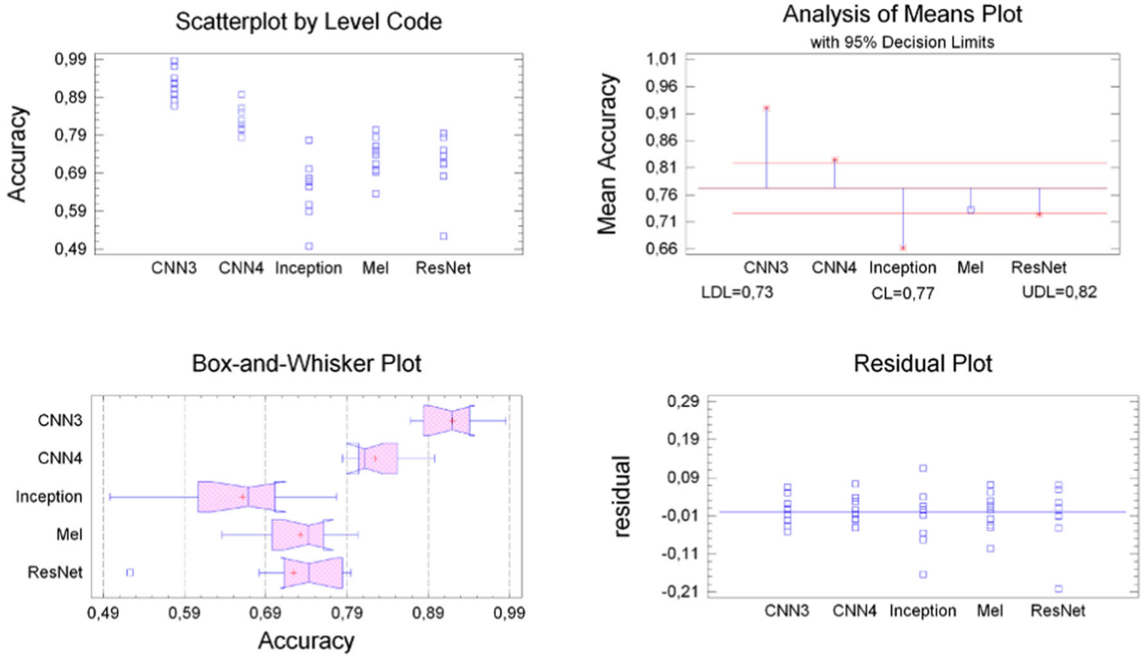

Fig. 7 Scatter and box plots (left) and Residuals and analysis of means plot (right)

of human perception, are the ones that offer the best results when used as input data to the CNN model. In addition, the processing time they require is much less than the rest of the spectrograms (see Table 4).

Moreover, the different experiments carried out show that the CNN models with a simpler architecture, relatively deep (8 convolutional layers), with convolutions with a simple structure, offer better results than other more complex models that include, for example, ResNet or Inception blocks (deeper networks require larger data samples for more effective training, and our dataset had a limited size, which could explain the worst result in the case of Resnet or Inception networks).

The architecture of the CNN model follows this structure:

$$Filter(3x3) + BN + AR + MaxPooling(2, 2) \tag{1}$$

where is BN is Batch Normalization and AR is Activation ReLu.

Table 17 Confusion matrix generated in the training/validation phase of CNN-4

	happiness	sadness	fear	peacefulness
happiness	50	1	3	5
fear	1	41	9	9
sadness	5	3	24	34
peacefulness	5	4	17	38

Thus, the CNN model is the one that obtains the best results, and the classification improves when only the three basic emotions of happiness, sadness, and fear, were considered, discarding the emotion of peacefulness. As mentioned, peacefulness is not considered a basic emotion but is frequently perceived as a musical emotion and is characterized by musical parameters partly shared with sadness (see Table 18). Therefore, as observed by [15], in human perception, peacefulness tends to be more often confused, particularly with sadness, than the basic emotions of happiness or fear. This same tendency is observed in model CNN-4, as we can see in the example of the confusion matrix in Table 17, where sadness and peacefulness are more often confused than happiness or fear and explains the better performance of the CNN-3 model, which only includes the basic emotions of happiness, sadness and fear (Table 18).

Indeed, the CNN-3 model classifies the three basic emotions of happiness, sadness, and fear that are the most interesting from the point of view of music characterizing

Table 18 Emotion and musical parameters

	happiness	sadness	fear	peacefulness
mode	major	minor	minor	major
tempo	fast	slow	fast	slow
register	high	low	high-low	medium
dynamics	loud	quiet	loud	quiet
articulation	staccato	legato	staccato	legato
timbre	brilliant	mellow	brilliant	mellow

Table 19 Comparison of results with CNN-3 model versus [15]

	Vieillard et al. [15]	Accuracy
Happiness	0.99	0.78
Sadness	0.84	0.84
Fear	0.72	0.75

in movie soundtracks, as these emotions are the most used in music to inform the development of the dramatic action.

The review of the current research in automatic emotion recognition [23], published in 2022, shows that the best accuracy obtained with CNN neural networks is 0.695. Table 14 shows that CNN-4 model gets 0.82 accuracy, and CNN-3 model achieves 0.92 accuracy. Finally, in Table 19, the authors compare the mean accuracy values per emotion obtained with CNN-3 model and the results obtained in [15] with experimental participants during the elaboration of the Musical Excerpts dataset used in this research. These results show that CNN-3 results are close to those obtained by human perception.

6 Conclusions

The capacity of music to generate intense emotions is widely used in movie soundtracks, especially regarding happiness, sadness, and fear, to support the dramatic plot. The objective of this study was to evaluate deep learning technologies that could be added to movie pre-recording captioning frameworks to ease the task of music captioning for accessibility purposes. Contrary to MER approaches based on combining multiple audio parameters and machine learning techniques with a somewhat random approach, the authors aimed to approach the problem of emotion detection considering the latest neuroscientific evidence (only happiness, sadness, and fear expressed with intensity are consistently and universally recognized by listeners, and immediately, in less than 2 seconds), and using only scientifically labelled film music datasets.

Taking into account the results, it can be considered that CQT spectrograms combined with a simple CNN architecture result in an efficient emotion classification model for 2-second musical audio fragments representative of intense basic emotions of happiness, sadness, and fear. These emotions are precisely the most interesting ones to identify in the case of movie music captioning, approaching the results of neuroscientific experiments with subjects.

In addition, compared to other models, it has the great advantage of not requiring a previous selection of the characteristics of the audio samples, which makes its application easier.

It also shows that movie music can be automatically classified based on basic emotions. This paves the way for accessible and automatic captioning of music, which could automatically identify the emotional intent of the different segments of the movie soundtrack. Starting with these results, next steps will consider the automatic analysis of the whole film soundtrack, to detect musical fragments of intense emotion that possibly should be captioned, mark the location of these segments, and propose the corresponding emotion.

Within future work, it is proposed to apply the designed models to a complete or partial soundtrack of films to detect maximum emotional intensity.

Finally, the authors want to point out the importance of combining neuroscience, musical theory, and computational models in this type of study.

Acknowledgements This work was supported by the Research Program of the Ministry of Science and Innovation - Government of Spain (ACCESS2MEET project-PID2020-116527RB-I00). This work has been supported by the Madrid Government (Comunidad de Madrid-Spain) under the Multiannual Agreement with UC3M in the line of Excellence of University Professors (EPUC3M17) and in the context of the V PRICIT (Regional Programme of Research and Technological Innovation). Funding for APC: Universidad Carlos III de Madrid (Read & Publish Agreement CRUE-CSIC 2023)

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Data Availability The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of Interests The authors declare that they do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

References

1. AENOR. (2012) Norma UNE 153010 Subtitulado para personas sordas y personas con discapacidad auditiva [Norm UNE 153010. Subtitling for deaf and hearing-impaired persons]

2. Che X, Luo S, Yang H, Meinel C (2017) Automatic Lecture Subtitle Generation and How It Helps. In Proceedings - IEEE 17th International Conference on Advanced Learning Technologies, ICALT 2017, pages 34–38
3. Donnelly KJ (2005) The Spectre of Sound. British Film Institute, London
4. Thompson WF, Russo FA, Sinclair D (1994) Effects of underscoring on the perception of closure in filmed events. Psychomusicol J Res Music Cogn 13(1–2):9–27
5. Pehrs C, Deserno L, Bakels J H, Schlochtermeier L H, Kappelhoff H, Jacobs A M, Fritz T H, Koelsch S, Kuchinke L (2014) How music alters a kiss: Superior temporal gyrus controls fusiform-amygdalar effective connectivity. Soc Cognitive Affect Neurosci 9(11):1770–1778 11
6. Eerola T, Vuoskoski JK (2011) A comparison of the discrete and dimensional models of emotion in music. Psychol Music 39(1):18–49
7. Ekman P (1992) An Argument for Basic Emotions. Cogn Emot 6(3-4):169–200, 5
8. Jack R E, Garrod O GB, Schyns P G (2014) Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. Curr Biol 24(2):187–192
9. Posner J, Russell JA, Peterson BS (2005) The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. Dev Psychopathol 17(3):715–734
10. Lang P J, Bradley M M (2010) Emotion and the motivational brain 7
11. Peretz I, Gagnon L, Bouchard B (1998) Music and emotion: Perceptual determinants, immediacy, and isolation after brain damage. Cogn 68(2):111–141, 8
12. Peretz I (2012) Towards a Neurobiology of Musical Emotions. Handbook of Music and Emotion: Theory. Research, Applications, pp 99–126
13. Koelsch S (2014) Brain correlates of music-evoked emotions. Nat Rev Neurosci 15(3):170–180
14. Paquette S, Peretz I, Belin P (2013) The "Musical Emotional Bursts": A validated set of musical affect bursts to investigate auditory affective processing. Front Psychol 4(AUG):509
15. Vieillard S, Peretz I, Gosselin N, Khalfa S, Gagnon L, Bouchard B (2008) Happy, sad, scary and peaceful musical excerpts for research on emotions. Cogn Emot 22(4):720–752
16. Vieillard S, Gilet A-L Age-related differences in affective responses to and memory for emotions conveyed by music: a cross-sectional study. Front Psychol 4:711
17. Balkwill L L, Thompson W F (1999) A cross-cultural investigation of the perception of emotion in music: Psychophysical and cultural cues. Music Percept 17(1):43–64, 10
18. Balkwill LL, Thompson WF, Matsunaga R (2004) Recognition of emotion in Japanese, Western, and Hindustani music by Japanese listeners. Japanese Psychol Res 46(4):337–349
19. Gabrielsson A, Lindström E (2012) The Role of Structure in the Musical Expression of Emotions. Handbook of Music and Emotion: Theory. Research, Applications, pp 367–400
20. Eerola T, Friberg A, Bresin R (2013) Emotional expression in music: Contribution, linearity, and additivity of primary musical cues. Front Psychol 4(JUL):487, 7
21. Yang X, Dong Y, Li J (2018) Review of data features-based music emotion recognition methods. Multimed Syst 24(4):365–389
22. Zhang W, Lei W, Xu X, Xing X (2016) Improved music genre classification with convolutional neural networks. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, volume 08-12-Sept, pages 3304–3308
23. Han D, Kong Y, Han J, Wang G (2022) A survey of music emotion recognition. Front Comput Sci 16(6):1–11
24. Feng Y, Zhuang Y, Pan Y (2003) Popular Music Retrieval by Detecting Mood. In SIGIR Forum (ACM Special Interest Group on Information Retrieval), number SPEC. ISS., pages 375–376, ACM nEW yORK
25. Panda R, Malheiro R M, Paiva R P (2020) Audio Features for Music Emotion Recognition: a Survey. IEEE Trans Affect Comput pages 1–1
26. Bertin-Mahieux T, Ellis D PW, Whitman B, Lamere P (2011) The million song dataset. In Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011, pages 591–596
27. Speck J A, Schmidt E M, Morton B G, Kim Y E (2011) A comparative study of collaborative vs. Traditional musical mood annotation. In Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011, pages 549–554
28. Law E, West K, Mandel M, Bay M, Downie J S (2009) Evaluation of algorithms using games: The case of music tagging. In Proceedings of the 10th International Society for Music Information Retrieval Conference, ISMIR 2009, pages 387–392
29. Wu Z (2022) Research on Automatic Classification Method of Ethnic Music Emotion Based on Machine Learning. J Math 2022
30. Seo Y S, Huh J H (2019) Automatic emotion-based music classification for supporting intelligent IoT applications. Electron (Switzerland) 8(2)
31. Medina YO, Beltrán JR, Baldassarri S (2022) Emotional classification of music using neural networks with the MediaEval dataset. Person Ubiquitous Comput 26(4):1237–1249
32. Han B J, Rho S, Dannenberg R B, Hwang E (2009) SMERS: Music emotion recognition using support vector regression. In Proceedings of the 10th International Society for Music Information Retrieval Conference, ISMIR 2009, pages 651–656
33. Xiao Z, Dellandrea E, Dou W, Chen L (2008) What is the best segment duration for music mood analysis ? In 2008 International Workshop on Content-Based Multimedia Indexing, CBMI 2008, Conference Proceedings. IEEE, pages 17–24, 6
34. Li T LH, Chan A B, Chun A HW (2010) Automatic musical pattern feature extraction using convolutional neural network. In Proceedings of the International MultiConference of Engineers and Computer Scientists 2010, IMECS 2010, pages 546–550, Hong Kong
35. Won M, Ferraro A, Bogdanov D, Serra X (2020) Evaluation of CNN-based automatic music tagging models. Proceedings of the Sound and Music Computing Conferences, 2020-June:331–337
36. Schellenberg E G, Trehub S E (1994) Frequency ratios and the perception of tone patterns. Psychon Bull Rev 1(2):191–201, 6
37. Gold T, Pumphrey R J, Gray (1948) Hearing. I. The cochlea as a frequency analyzer. Proc Royal Soc B: Biol Sci 135(881):462–491, 12
38. Gonzalez-Carrasco I, Garcia-Crespo A, Ruiz-Mezcua B, Lopez-Cuadrado J L (2011) Dealing with limited data in ballistic impact scenarios: An empirical comparison of different neural network approaches. Appl Intell 35(1):89–109, 12
39. Gonzalez-Carrasco I, Garcia-Crespo A, Ruiz-Mezcua B, Lopez-Cuadrado JLJL, Colomo-Palacios R (2014) Towards a framework for multiple artificial neural network topologies validation by means of statistics. Exp Syst 31(1):20–36
40. Kruskal WH, Wallis WA (1952) Use of Ranks in One-Criterion Variance Analysis. J Am Stat Ass 47(260):583–621

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

María J. Lucía was born in Ponferrada, León, Spain, in 1962. She received the M.S. degree in Psychology in 1985 and the M.S. degree in Physics in 1989, both from the Universidad Complutense de Madrid. From 1989 to 2016, she worked at Telefónica España as consultant and manager for ICT projects. In 2017, she joined the Universidad Carlos III de Madrid, where she is currently an associate professor and collaborates with the Spanish Center for Captioning and Audio Description (CESyA) as a researcher. She received the Ph.D. degree in Computer Sciences at the Carlos III University of Madrid.

Pablo Revuelta was born in Madrid, Spain, in 1980. He received the B.S degree in telecommunications engineering in 2006, the M.S.degree in advanced electronics in 2008, and the Ph.D. degree in electronic engineering in 2013, all from Universidad Carlos III de Madrid. Since 2006, he has been a researcher with the Displays and Photonic Applications Group of the Carlos III University, the Spanish Center for Captioning and Audio Description (CESyA), and the start-up Visión Táctil S.L. He also worked with the Digital Signal Processing Lab 5 (EPFL, Switzerland) in 2010 and the Georgia Tech Sonification Lab (United States) in 2012. He teaches at the Oviedo University since 2018 and collaborates with CESyA in Madrid, Spain.

Belén Ruiz was born in Caracas, Venezuela in 1959. She received the M.S. degree in physics from the Complutense University of Madrid, in 1983, and the Ph.D. degree in computer science in 1998 from Universidad Carlos III de Madrid. From 1984 to 1996, she worked in ICT companies, such as ALCATEL and INDRA. In 1996 she joined Universidad Carlos III de Madrid, where she has been teaching at the Computer Sciences Department as professor since 2018, and dean of the Computer Sciences Degree. She was director of the Spanish Center for Captioning and Audio Description (CESyA). She has led several national and international research projects on voice recognition, human-machine interface, software engineering, systems analysis, mobile communications, and audiovisual accessibility.

Israel Gonzáles-Carrasco holds his PhD degree in Computer Science by Universidad Carlos III de Madrid since 2010. He is an associate professor and the assistant manager of the Computer Science Department of this University. He is co-author of several papers in international journals (indexed in ISI-JCR) and international conferences. His main lines of research are Soft Computing, Software Engineering and Accessibility. He has been involved in different national and international projects. Moreover, he is a member of the editorial reviewer board of international journals (indexed in ISI-JCR) and member of the organizing committee at international conferences. Currently, he is the assistant manager of the Spanish Center for Captioning and Audio Description (CESyA)