EEG-based Music Cross-modal Retrieval

Ana Rafaela Gravelho Saraiva
rafaela.saraiva@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

May 2019

Abstract

Music Information Retrieval (MIR) is an essential subject nowadays, aiming to aid users in solving the problem of having a vast number of choices and finding music by a similarity criterion. Nonetheless, the human perception of similarity is highly dependent on the user and MIR systems lack user focus. In order to assess the perception and subjective experience of the user while listening to music, we explore the potential of using EEG signals, in addition to audio signals, to capture this information. For this, we leverage multi-view deep models to learn a shared embedding space between audio and Electroencephalography (EEG) signals, and the resulting embeddings are evaluated on cross-modal retrieval tasks, i.e., given a piece of audio we aim to retrieve the corresponding EEG and vice-versa. Two audio-EEG datasets were collected for training the proposed models: the INESC-ID dataset with 22 participants, a mean of 30 minutes of music per participant repeated twice recorded using the 16-channel OpenBCI, and the CENC dataset containing recordings of 19 participants and recorded with the BrainVision QuickAmp with an electrode-cap of 64 electrodes. Through the extensive cross-modal retrieval tasks performed, we concluded that the learned embeddings encode the intrinsic associations between audio and EEG and should reflect the auditory concepts that are perceived and processed by the brain, suggesting the existence of a subject-specific neural signature able to distinguish music pieces. In addition, we demonstrated the viability of using the commercial device OpenBCI for this application.

Keywords: EEG, Music, Cross-modal retrieval, Multi-view deep models

1. Introduction

Technological advances in several fields of science have facilitated the increasing growth in music production and distribution. In 1991, music started to be digitalised, promoting its transmission over digital networks and encouraging people to try new ways of consuming and sharing music. Later on, streaming audio technology appeared, making music more popular than ever [8]. Thus, the resulting size of music databases and its constant expansion make it remarkably difficult for listeners to recall a particular song, and causes an additional effort and time for users to browse for suitable music, making the task of music recommendation more challenging. For that reason, MIR becomes an essential subject, aiming to aid users in solving the problem of having a vast number of choices and helping them to find music by a particular similarity criterion.

When recommending music, the human perception of similarity is highly dependent on the user and is influenced by several factors such as lyrics, beat, the opinion of the users friends and the mental state of the user [26]. Nonetheless, previous MIR systems lack focus on the user [26, 34]. In this work, we investigate the potential of using EEG signals, in addition to audio to assess the perception, cognition and the subjective experience of the user while listening to music. EEG is a recording technique that measures the electrical brain activity obtained from the scalp, and presents a high temporal resolution, on the order of milliseconds, being very precise in measuring brain responses to particular stimuli, such as audio. This signal could be seen as a mediator between the music and the subjective experience of the user [29]: it is able to reflect a user's state of mind [23], may capture intrinsic responses to music, and reflect emotional responses.

Retrieving music information based on brain signals is a very recent field, and the area is mostly dominated by studies that explored emotion recognition based on EEG recorded while listening to music [1, 6, 10, 13–15, 32, 33]. These studies proved the viability of using EEG signals to codify the emotional content induced by music and allowed to establish common practices regarding the experimental setup of EEG recordings.

The feasibility of using EEG signals recorded during auditory presentations to identify perceived rhythms was studied by Stober et al. [27, 28], being one of the first studies that explored deep learning towards that goal. The study [27] aimed to identify whether a participant listened to East African or

Western rhythm and to predict each rhythm they listened to, out of 24 rhythms. Two approaches were considered: pre-train a multilayer perceptron as a stacked denoising autoencoder, and a Convolutional Neural Network (CNN) with two layers. Both cases yielded encouraging results for classifying chunks of 1-2 seconds from a single EEG channel into African and Western rhythms and above chance results were achieved for classifying individual rhythms (8.3%-21.4%). Later on, Stober et al. [28] continued the work of classifying the earlier mentioned rhythms through CNNs, where different subject-specific configurations of the CNN were studied. The obtained accuracies varied between subjects (18.1%-36.1%), which was justified by the strong differences in rhythm perception and different quality in the EEG recordings. These results showed that the perceived rhythms can be identified through CNNs using as input the frequency spectrum of EEG recordings with only one channel, showing the potential of using EEG to distinguish different brain rhythms, despite the limitations identified, e.g. each EEG channel was studied individually and that each stimulus was only recorded once.

Raposo et al. proposed a framework for modelling music audio semantics that focused on the perception of the listener, through EEG and audio data. It consisted of an end-to-end two-view Neural Network (NN) architecture using a CNN for each view and Deep Canonical Correlation Analysis (DCCA) as a loss function for correlating audio and EEG. The maximally correlated learned embeddings obtained for audio and EEG were used on as an audio feature extractor on a transfer learning task for performing audio-lyrics cross-modal retrieval, whose results (with 3 hours of audio-EEG pairs) achieved a comparable performance with other embeddings which were trained with much more data (2083 hours).

In order to address the potential of using EEG signals to retrieve music information, this work explores a deep learning-based approach for learning the association between audio and EEG signals, evaluated through EEG- and audio-based music cross-modal retrieval tasks. For this, the framework proposed by Raposo et al. [21] is used. To train the proposed models, a music audio-EEG pairs database was built and two EEG recording devices were used: 16-channel OpenBCI [19] and BrainVision QuickAmp [5] with a 64 electrodes-cap. Hence, another objective consisted of studying the viability of using a non-professional device for this application. Moreover, several variations of the multi-view models are evaluated on audio-EEG cross-modal retrieval tasks.

2. Cross-modal Retrieval Methods

Cross-modal retrieval encompasses retrieval tasks where the retrieved items are of a different type than the query, for instance, retrieving the caption of a given image [37] or the audio of a certain video [30]. A popular approach for cross-modal retrieval consists in defining transformations that transform the data from different modalities into a joint embedding space. Afterwards, a query is projected into the embedding space, and the corresponding counterpart from the other modality is retrieved using nearest neighbour search [9]. For learning these transformations, several subspace learning methods have been proposed such as Canonical Correlation Analysis (CCA), partial least squares, among others. Nonetheless, the mentioned methods are linear, thus, not able to learn nonlinear representations neither capture high-level associations between multi-view data. To surpass this drawback, other approaches have been proposed such as DCCA [2] that combines deep NNs with the CCA loss function, and other multi-view deep learning methods which make use of models optimised with specific loss functions.

Fig.1 presents the high-level architecture of the cross-modal retrieval network used in this work, which has been previously considered in related studies [9, 11, 21, 22, 37, 38]. It is constituted by a two-view NN, where the data of each modality (audio and EEG) is inputted and optimised through one of the loss functions presented in Section 2.1. When using Pairwise Ranking Loss (PRL), the aim is to decrease the distance between the matching training instances and increase the distance of mismatching instances, and when using the DCCA loss function, the aim is to maximise the correlation between the two views. After obtaining the learned joint embeddings, if the loss function was the DCCA, they have to be passed through a linear CCA because the output embeddings are not the ones being optimised, but their CCA projection, and data is projected into the embedding space. At test time, queries were projected to the embedding space and retrieved using cosine distance. Then, the similarity between the components was calculated, and the results were ranked in decreasing order.

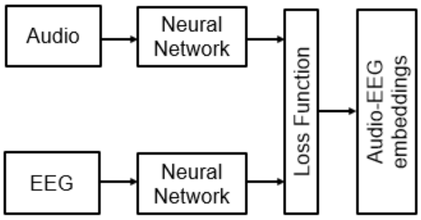

Figure 1: High-level architecture of the cross-modal retrieval network.

The type of NNs used for each view depends on the input data (waveform, 2D data), including: Dense Neural Networks (DNNs), One-Dimensional Convolutional Neural Network (1D-CNN) and Two-

Dimensional Convolutional Neural Network (2D-CNN). The 1D-CNN architecture used is constituted by sequences of convolutional blocks composed by a layer of batch normalisation, a convolutional layer (conv1D) with filter size x, kernel size y (length of the window of the convolution), stride size z (amount by which the filter is moved at each step), and a max-pooling layer with window size w that replaces the output of the layer by the maximum value of the nearby outputs. When using a 2D-CNN, the convolutional block followed the same architecture, except that the convolutional and the max-pooling layers were performed in two dimensions.

2.1. Objective Functions

2.1.1 Deep Canonical Correlation Analysis

DCCA aims to learn simultaneously two deep non-linear mappings of two views that are maximally correlated, using two deep NNs and CCA as objective function [2]. CCA is a technique for learning representation of two data views where linear transformations of each view are such that the correlations between the transformed variables are maximal:

$$(w_x^*, w_y^*) = \underset{(w_x, w_y)}{\operatorname{argmax}} \quad \operatorname{corr}(w_x^T x, w_y^T y) = \underset{(w_x, w_y)}{\operatorname{argmax}} \frac{(w_x^T C_{xy} w_y)}{\sqrt{w_x^T C_{xx} w_x \cdot w_y^T C_{yy} w_y}} \tag{1}$$

where w_x and w_y are the estimated linear projections, C_{xx} and C_{yy} are the covariances of x and y, respectively, and C_{xy} is the cross-covariance [2].

The CCA objective can be written as: maximise $\operatorname{tr}(W_x^T C_{xy} W_y)$, subject to $W_x^T C_{xx} W_x = W_y^T C_{yy} W_y = I$ since in a multivariate representation, the subsequent features need to be uncorrelated with the original ones and correlation is invariant to scaling. One of the solutions for this objective is to define $T \triangleq C_{xx}^{-1/2} C_{xy} C_{yy}^{-1/2}$ and perform singular value decomposition on T, resulting in $T = UDV'$. W_x and W_y are the matrices that project x and y into a common subspace and can be computed as: $(W_x^*, W_y^*) = (C_{xx}^{-1/2} U_k, C_{yy}^{-1/2} V_k)$. The covariances and cross-covariances need to be positive definite and to guarantee that the covariances are larger than zero, a regularisation term is added, which also reduces overfitting [2].

Concerning DCCA, the goal is to jointly learn non-linear mappings (φ_x, φ_y) and canonical weights (w_x, w_y) for both views such that the correlation after the non-linear mapping is maximised:

$$(\mathrm{w}_x^*, w_y^*, \varphi_x^*, \varphi_y^*) = \underset{(w_x, w_y, \varphi_x, \varphi_y)}{\operatorname{argmax}} \quad \operatorname{corr}(w_x^T \varphi_x(x), w_y^T \varphi_y(y)) \tag{2}$$

Considering that $X \in \mathbb{R}^{o \times m}$ and $Y \in \mathbb{R}^{o \times m}$ are matrices whose columns are the top-level representations obtained from the deep models, o is the number of units of the final layer and m is training set size, the total correlation of the top k components if $k = o$ corresponds to the matrix trace norm of T: $\operatorname{corr}(X, Y) = \operatorname{tr}(T^T T)^{1/2}$, the value being maximised. Then, the gradients are computed and propagated along the two branches of the NN and all the parameters are learned jointly via backpropagation.

2.1.2 Pairwise Ranking Loss

The PRL loss function is defined as:

$$\text{loss} = \sum_x \sum_k \max\{0, \alpha - s(x, y) + s(x, y_k)\} \tag{3}$$

where x and y are the embedded spaces of the first and second modality, y_k is the mismatching embedding sample of the second modality, and α is a margin value [9]. This margin value encourages the loss of an embedding space of matching samples to be lower than the loss of mismatching samples. For the scoring function, defined as $s(x, y)$, we considered the cosine of the angle between the vectors x and y and the Euclidean distance. Regarding cosine, $s(x, y) = \cos(x, y)$, forcing the cosine distance between matching samples to be lower than the cosine distance of mismatching samples, and α can take values between 0 and 1. $s(x, y)$ is substituted by: $\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$ when considering the Euclidean distance, and alpha take values between 0 and $2 * \sqrt{d}$, where d is the dimension of the output embeddings.

2.2. Evaluation Metrics

In order to evaluate the effectiveness of the models in cross-modal retrieval tasks, the following metrics were used: Mean Reciprocal Rank (MRR), Mean Average Precision (MAP) and Rank Accuracy (RA). For each of the metrics, instance-level and class-level retrieval performances were computed. For instance-level, a segment of EEG is retrieved when a segment of audio is given, and vice-versa, whilst for class-level, any segment of EEG belonging to the same class (song) can be retrieved and vice-versa.

3. EEG Dataset Collection

To build the audio-EEG database composed by EEG signals recorded while listening to music, two datasets were constructed: CENC dataset recorded at Centro de Electroencefalografia e Neurofisiologia Clínica (CENC) and INESC-ID dataset recorded at Laboratório de Sistemas de Língua Falada (L^{2}F).

3.1. Participants

23 participants (11 females), with ages comprised between 19 and 50 years (mean age: 24 years). 21 participants were right-handed, and all participants reported normal hearing. All participants took part in the recording sessions of both datasets, with some exceptions due to technical problems.

3.2. Stimuli

The musical stimuli were composed by a playlist of songs chosen by each participant with the restrictions that they must appreciate the chosen songs, and that the playlist should last about 30 minutes, with a minimum of 6 songs and a maximum of 12.

3.3. INESC Dataset

The EEG acquisitions took place in a laboratory room in L^{2}F, under dim lighting condition and prepared with a computer with the required software, and an armchair. EEG signal was recorded using the OpenBCI 32bit Board with the OpenBCI Daisy Module and 16 gold cup electrodes with dry silver-silver chloride electrodes (to penetrate the hair and avoid the usage of conductive gel) were used. These electrodes were placed on a homemade elastic cap according to the extended international 10-20 system (see Fig. 2), and a linked-earlobe electrode served as reference [31]. The EEG signals were visualised and recorded using the open-source software of OpenBCI [3] and digitised with a 125 Hz sampling rate. In order to synchronise the recorded EEG with the stimuli, we programmed a custom GUI widget that identified the stimulus in the output EEG file.

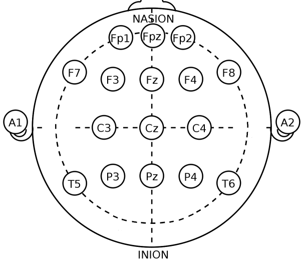

Figure 2: Location of the 16 electrodes positioned according to the 10-20 international system [36].

The experiment was divided into 3 sessions that were performed in different days or moments of the day to avoid tiredness and possible discomfort introduced by the bumps of the dry electrodes. The set of stimuli was constituted by the participant playlist duplicated, and the songs were randomly distributed in 3 subsets, with no repeated songs per subset. A 15-second silence was presented at the beginning in order to establish a baseline and the stimuli were randomly presented to mitigate possible subject-expectancy effects, with a 10-second interval between them to reestablish the baseline. Each listening session lasted about 20 minutes, and the EEG was recorded continuously during the session.

The procedures of these experiments were approved by the Instituto Superior Técnico (IST) ethics committee.

Session 1. The experiment began with an explanation of the objectives of the experiment and the course of the sessions, followed by the participant reading and signing the Informed Consent. Afterwards, a self-report questionnaire was used to collect personal information and musical preferences. Then, the participant sat in a comfortable armchair with the headphones placed, and the volume level of the headphones was adjusted to maximise the comfort, followed by placing the electrode-cap, and the quality of each electrode was visualised in the OpenBCI software. The electrodes that were placed on the earlobe were two gold cup electrodes fixed with tape, and using conductive electrode paste. The participants were instructed to remain as still as possible to reduce muscular artefacts and to keep their eyes closed to avoid ocular artefacts. The preparation steps lasted on average 30 minutes.

Sessions 2 and 3. The course of these sessions occurred similarly to the first session and only a small question asking how participants were feeling in the day of the experiment, according to the arousal-valence model [25] was asked.

3.4. CENC Dataset

The experimental procedure took place in a room located at CENC, under dim lighting conditions, equipped with two computers and a stretcher. The amplifier used was a 72-channel BrainVision QuickAmp and an electrode-cap with 64 Ag/AgCl electrodes. The amplifier had specific connectors (DB37) for the electrode-cap and because some of the connections of the cap's cable did not match the amplifier connections as each pin of the amplifier was assigned to a specific electrode, only the 57 electrodes of Fig. 3 were available. The amplifier did not require a reference channel since it used a built-in common average referencing of the recorded channels [5] and the participant ground was hard-wired in the amplifier (channel AFZ). Data was recorded and visualised online using the BrainVision Recorder with 50 Hz notch filter and bandpass filter 1-70 Hz applied just for visualisation, and signal was acquired at a 2000 Hz sampling rate. To synchronise the audio stimuli with the EEG signal, we used OpenSesame [16] and the plugin [24], and an experiment that played the audio stimuli and sent a trigger to the parallel port was programmed. The trigger was received by the BrainVision Recorder and represented as markers in the EEG data.

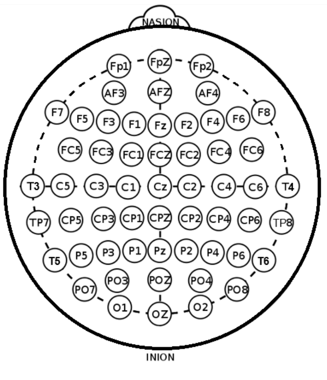

Figure 3: Location of the 57 electrodes positioned according to the 10-10 international system [35].

The experiment was divided in 2 sessions that were performed in the same day with a break period between them. It began with an explanation of the setup and the session and in each session, the participant listened to all the stimuli of their playlist, randomly presented with 10 seconds between them and 15 seconds in the beginning.

The preparation stage started by cleaning the head of the participants with cotton soaked with alcohol and an exfoliant to remove dead skin and decrease its impedance. Afterwards, the electrode-cap was placed, and each electrode was filled with a conductive gel to decrease the impedance between the skin and the electrodes. The value of the impedances was decreased to lower than 20 kΩ and electrodes whose impedances were larger than 100 kΩ were turned off. Then, the headphones were placed and the volume level was adjusted. EEG was recorded continuously during the sessions.

4. Signal Processing

4.1. EEG Processing

The preprocessing pipeline proposed for the EEG signals is presented in Fig. 4. The first step comprises the correction of the power line noise by applying a notch filter at 50 Hz. In addition, a band-pass filter (2^{nd} order Butterworth) between 1 and 70 Hz was applied for the CENC dataset to remove direct current slow drifts and other typical slow artefacts such as electrogalvanic signals [20], and high-frequency noise components. For the INESC-ID dataset, the high cut-off frequency was set to 62 Hz because we were limited to the 125 Hz sampling rate. Afterwards, flat channels were removed and two criteria were used: the standard deviation of the signal for a particular channel being equal to zero or the amplitudes being very close to 0 (less than 10^{-6}) in 80% of the samples. Then, bad channels were removed based on the following criterion: for each channel, the Euclidean distance between the PSD of the channel and the mean of the PSD of all the channels was calculated and the resulting distance was subtracted by the mean of the Euclidean distances of all the channels and divided by their standard deviation. If this value was larger than a specific threshold (set to 3 for both datasets), the channel was removed. The next step comprises the segmentation of the files by stimulus. The EEG files from the CENC dataset were downsampled to a sampling frequency of 250 Hz. Afterwards, for removing the remaining artefacts such as muscular and electrode popping, Artefact Subspace Reconstruction (ASR) method was used, an automatic component-based artefact removal method that is capable of removing transient or large amplitude artefacts. It identifies clean portions of data to determine thresholds for rejecting components, and data is reconstructed from the remaining components [18]. This method requires the definition of a cutoff parameter k that corresponds to the rejection threshold and this threshold was set to $k = 15$ for both datasets as any value within the interval $k \in [15, 30]$ appears to present a good compromise between the percentage of data modified and the variance reduced. In the last step of the preprocessing stage, signals were normalised between -1 and 1. After the preprocessing steps, EEG features were calculated, including the Power Spectral Density (PSD), bandpower, spectrogram and the topographical map.

4.2. Audio Processing

The preprocessing pipeline proposed for the audio signal is presented in Fig. 5. After loading the audio signals, they were downsampled to 22050 Hz, transformed to mono-channel, and normalised between -1 and 1. After the preprocessing steps, audio features were calculated including the Mel spectrogram and Mel Frequency Cepstral Coefficients (MFCCs).

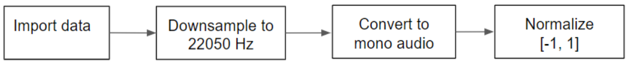

Figure 5: Audio preprocessing pipeline.

5. Cross-Modal Retrieval Tasks

The task in focus during these experiments is the retrieval task that given a query element aims to retrieve the correct counterpart in the other modality, i.e., given an EEG segment, it should return the corresponding piece of audio and vice-versa. In Fig. 6, the complete pipeline necessary to accomplish the cross-modal retrieval tasks is presented. For all the experiments performed in the following sections, except in Section 5.6, we ran each experiment 3 times with 5-fold cross-validation and averaged the results across folds and across runs, where data was randomly shuffled across folds.

5.1. Baseline

In order to establish a baseline, we designed an experiment similar to the one proposed by Raposo et al. [21] that takes raw audio and EEG as input of a two-view NN constituted by a 1D-CNN for each branch, and the DCCA loss function. The CENC dataset was used, and the audio and EEG signals were segmented into non-overlapping segments of 2 seconds. Regarding the configuration of the network, an automated approach using Bayesian optimisation [4] was used for defining the following hyperparameters: batch size, dimension of the embeddings, learning rate and the hyperparameters of the 1D-CNN (the filter, kernel and window sizes). The batch size was limited to a maximum value imposed by memory restrictions: 8, 16, 32 and for the DCCA loss function, it needed to be at least as large as the dimensions of the CCA components to have stability in the covariance estimates. The dimen-

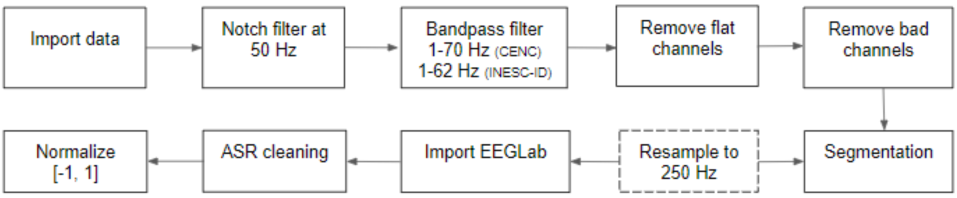

Figure 4: Preprocessing pipeline for the EEG signals. The dashed area is a step of the CENC dataset.

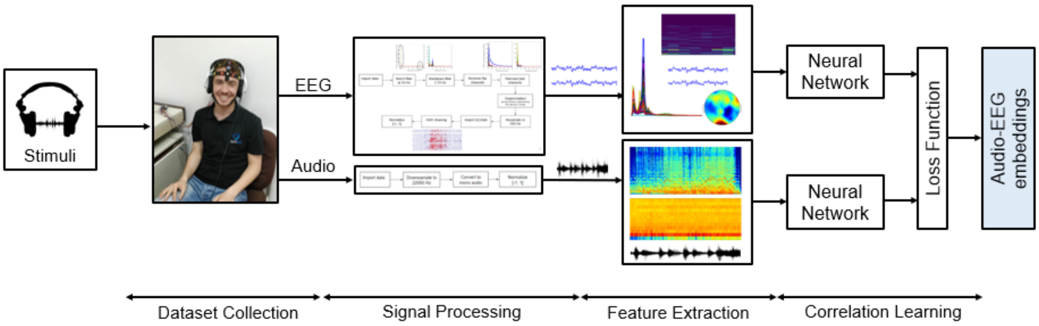

Figure 6: Pipeline for obtaining the audio-EEG embeddings. It starts with the collection of the dataset, followed by passing the audio and EEG data through the preprocessing pipelines. Afterwards, features for are extracted and input into a multi-view NN, which aims to learn the audio-EEG representations.

sions of the embeddings were considered to be the same for each network and equal to the number of canonical components. We did not tune the number of epochs because an early stopping callback, which interrupts training when the validation loss did not improve for 15 epochs was used. For choosing the hyperparameters, we ran the optimiser 50 times for participant 0 (DCCA column of Table 2). For all the models, weights were randomly initialized, the activation function used was Rectified Linear Unit (ReLU), and Adam optimiser was considered. The 1D-CNN of each branch is composed by a sequence of convolutional blocks, except for the last layer that the size of the filter is equal to the dimension of the embeddings.

The reported results are presented in the DCCA row of Table 1. For the instance-level case, as we only had 2 relevant items out of 200, we are interested in knowing the position of the first relevant item, which is answered by the MRR and RA metrics. A value within 10.53/10.77% was obtained for the MRR, depending on the query. Despite at first glance this value appeared to be discouraging, the achieved performances were almost 3 times higher than the random performance for the MRR metric (3.92%), suggesting that some relationship between EEG and audio signals is being captured. Nonetheless, these experiments also suggested that this relationship was hard to establish, which may be due to the fact that EEG signals reflect a lot of brain processes further than those related with the stimulus.

The class-level results were superior than those of the instance-level, but the number of correct answers was also superior. In fact, the results were closer to the random performance than in the instance-level case, showing the increased difficulty of the task, which assumed that different instances of EEG recorded for the same music would share some information characteristic of the music piece that is reflected on the EEG signal, such as the genre of the music, the rhythm, and so on.

Table 1: Average of the cross-modal retrieval results(%) across participants obtained for the different loss functions, evaluated on 200 instances.

Loss		Instance			Class	
function	**Query**	**MRR**	**MAP**	**RA**	**MRR**	**MAP**
DCCA	Audio	10.53	10.11	84.50	35.54	17.94
	EEG	10.77	10.25	85.22	35.51	17.99
Euclidean	Audio	14.29	13.77	90.73	37.15	18.27
	EEG	14.50	13.94	91.21	37.56	18.33
Cosine	Audio	14.22	13.62	90.28	37.53	18.27
	EEG	14.51	13.89	90.68	37.52	18.30

In the following sections, further experiments were performed in order to enrich the information contained in the learned audio-EEG embeddings and, consequently, improve the performance of the tasks.

5.2. Impact of the Loss Function

In order to evaluate the impact of the loss function in the retrieval tasks, we performed experiments using the DCCA loss function and the PRL with cosine and Euclidean distances.

The results showed that the PRL outperformed the DCCA loss function, independently of the scoring function. The best performances obtained when using the PRL over the DCCA could be justified by the fact that the PRL function is focused on the retrieval tasks; thus, the retrieval objective is taken into account during training, in opposite to the DCCA loss function that seeks to maximise the correlation in the embeddings space, not specific to the performed task. Hence, we chose the PRL as the loss function for the experiments performed from now on.

Table 2: Configuration of the models.

	Loss Function		
	DCCA	**Cosine**	**Euclidean**
Dimension of embeddings	8	16	16
Batch size	32	32	32
Learning rate	0.00001	0.0001	0.0001
Type of network	1D-CNN	1D-CNN	1D-CNN
Kernel size audio	[2, 3, 5, 5, 6, 7, 7]	[2, 2, 3, 3, 5, 5, 7, 7]	[4, 5, 5, 7, 7, 9]
Kernel size EEG	[5, 5, 20]	[4, 5, 5, 5]	[2, 2, 5, 5, 5]
Filter size audio	[128, 128, 256, 256, 512, 1024, 1024]	[128, 128, 256, 256, 512, 512, 1024, 1024]	[128, 128, 256, 256, 512, 1024]
Filter size EEG	[128, 256, 512]	[128, 128, 256, 512]	[128, 256, 512, 512, 1024]
Alpha	-	0.74320	2.79296

5.3. Performance of CENC and INESC-ID Datasets

To compare the performance of CENC and INESC-ID datasets on the cross-modal retrieval tasks, we varied the input signal as follows:

1. CENC dataset resampled at 250 Hz and considering all the available electrodes;
2. CENC dataset at 125 Hz and considering only the electrodes in common of both datasets;
3. INESC-ID dataset at 125 Hz with all the available electrodes.

These experiments were performed for all the participants using the PRL with Euclidean distance loss function. The hyperparameters are presented in Table 2, optimised for participant 0. When reducing the sampling rate, the first window size of the EEG branch was reduced by a factor of 2, because of the different dimensions of the input.

Table 3: Average of the cross-modal retrieval results(%) across participants, evaluated on 200 instances.

EEG Input	**Query**	**Instance**			**Class**	
		MRR	**MAP**	**RA**	**MRR**	**MAP**
1	Audio	13.63	13.14	90.73	37.15	18.27
	EEG	14.40	13.73	91.21	37.56	18.33
2	Audio	13.59	13.04	90.27	37.46	18.38
	EEG	13.92	13.34	90.76	37.47	18.44
3	Audio	13.41	12.81	90.94	36.97	18.20
	EEG	13.88	13.25	91.38	37.07	18.21

The differences in the obtained performances were very small between the experiments, allowing to validate the use of the OpenBCI device with 16 electrodes in this task.

5.4. Study of Audio and EEG Input Representations

These experiments explored handcrafted audio and EEG features as input in order to understand which are more discriminative. We studied the audio input representation, while maintaining the EEG branch fixed, followed by studying EEG representations, fixing the audio branch. In the end, different audio and EEG representations data were combined.

Regarding the audio input representations, MFCCs and Mel spectrogram were considered and a 2D-CNN was used for the audio branch since the convolution could be performed in two axes. The studied EEG features included the spectrogram, PSD, bandpower, topographical map (calculated based on the bandpower values for only one band containing the full spectrum) and some variations of these features.

The combination of the best audio and EEG input representations among those explored in this work corresponded to the bandpower features normalised by the bandpower of the baseline condition (silence) for the EEG input and MFCCs for audio. These results confirmed the effectiveness of using MFCCs to characterise the audio content and suggested that the normalisation of the bandpower feature with the baseline bandpower may have mitigated the effect of the individual intrinsic physiological processes that are unrelated with the stimulus, helping the model to better associate the brain signals with the corresponding audio.

5.5. EEG Embedder

With the aim to mitigate the noise that is affecting the EEG signals of different systems and sessions, the architecture presented in Fig. 7 was proposed. This network has an additional preprocessing step for the EEG branch, that we called EEG embedder, which jointly optimises two views of EEG data such that the output embeddings contain the shared information between the signals, and the artefacts specific of a recording session or system are mitigated. In these experiments, the impact of the EEG embedder when using the EEG signal of different systems or sessions as input was investigated.

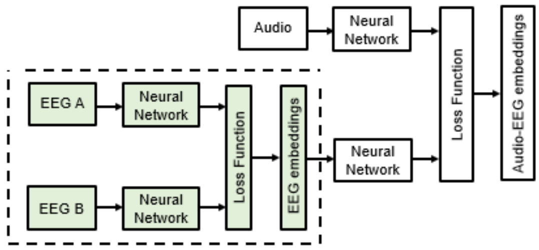

Figure 7: Architecture of the cross-modal retrieval network with the EEG embedder (dashed area).

Sessions The EEG signal of each session of the CENC dataset was fed in a multi-view NN with two 1D-CNN branches, trained with the Euclidean loss function. The obtained EEG embeddings were input into a DNN and trained together with the audio branch composed by a 1D-CNN, followed by a DNN. The ReLU activation function was also used with the dense layers, except for the last layer of both networks that corresponded to a dense layer with the number of units equal to the dimension of the embeddings and linear activation function. The hyperparameters were selected based on the results of the optimiser and are presented in Table 4.

Table 4: Configuration of the models for the architecture with the EEG embedder.

Dimension of embeddings	16
Batch size	128
Learning rate	0.00001
Type of network	1D-CNN
Filter size	[128, 256, 512, 1024]
Kernel/Window size	[2, 5, 5, 5]
Alpha	3.45328
Dropout	0.2
Dimension of embeddings	32
Batch size	32
Learning rate	0.0001
Type of network EEG	DNN
Number of units EEG	[64, 128]
Type of network audio	1D-CNN+ dense
Filter size audio	[128, 128, 256, 256, 512, 1024]
Kernel/Window size audio	[4, 5, 5, 7, 7, 9]
Number of units audio	[64, 128]
Alpha	2.79296
Dropout	0.5

Table 5: Average of the cross-modal retrieval results (%) across participants with and without the proposed EEG embedder, evaluated on 200 instances.

EEG	**Instance**			**Class**	
emb.	**MRR**	**MAP**	**RA**	**MRR**	**MAP**
without	13.62	13.10	90.32	37.32	18.33
	13.99	13.42	90.97	37.49	18.37
with	29.15	27.86	88.87	46.32	18.30
	31.76	30.35	89.04	48.00	18.41

When using the EEG embedder, the yielded results for the MRR and MAP metric increased more than two-fold for the instance-level evaluations, and a less remarkable improvement was noticed for the class-level. Nonetheless, the results for the RA metric decreased. As an attempt to understand this behaviour, we analysed the distribution of the positions of the first relevant item for the participant 0, where we concluded that when using the EEG embedder, almost 300 out of 377 instances were classified at positions under 25 and the remaining were distributed more or less uniformly by the following positions (cf. Fig. 8). However, the number of instances classified above 100 is higher than in the experiment without the EEG embedder, which decreased the RA metric as it gives more weight to higher positions than the MRR metric.

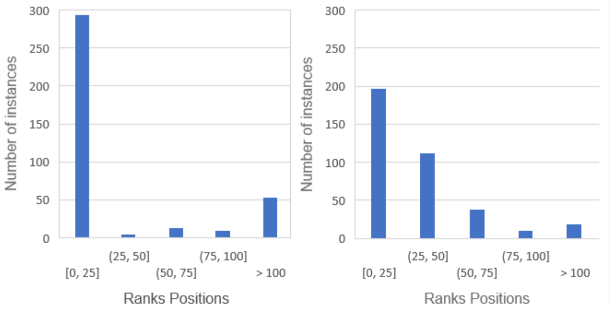

Figure 8: Distribution of the number of instances by the rank positions with (left) and without (right) the EEG embedder.

Systems In order to evaluate the impact of using the EEG embedder with different systems, one branch of the network received the signal of the INESC-ID dataset and the other branch the signal from the CENC dataset and the produced embeddings were then input in the EEG branch and trained as before together with the audio branch to obtain the audio-EEG embeddings.

Table 6: Cross-modal retrieval results (%) obtained with and without the proposed EEG embedder for the INESC-ID and CENC dataset as input for participant 0, evaluated on 377 instances.

EEG	**Instance**			**Class**	
emb	**MRR**	**MAP**	**RA**	**MRR**	**MAP**
without	8.42	7.44	91.96	33.37	14.96
	8.70	7.71	92.15	33.00	14.94
with	11.00	9.57	84.93	33.43	15.05
	11.11	9.62	85.13	33.62	14.99

Again, the EEG embedder was able to improve the obtained results when considering the EEG signal of different datasets, showing an improvement of about 30% for MRR and MAP metrics for the instance-level evaluations. Nonetheless, the enhancement of the results was less pronounced than for sessions, which may be justified by the fact that systems presented more variable conditions, for instance: equipment, number and location of the electrodes, material, and the impedances between the electrodes and the scalp.

Overall, these results allowed to conclude that the proposed method better associated the information from the brain waves with the audio signal since more instances were classified at the top positions of the ranking list, showing its effectiveness in reducing the variability between sessions and systems. In a practical setting, this approach has the disadvantage of requiring pairs of EEG signal recorded with the same music stimuli to produce the EEG embeddings and it is more time-consuming. These experiments were repeated for the INESC-ID dataset, where the results also proved the efficacy of this method.

5.6. Global Model

In this section, we aim to build a general model that considers the EEG and audio data of all the participants together. The number of channels for the EEG signal of the CENC dataset was set to 48 as it was the maximum number of channels in common to all the participants. Each branch of the network was constituted by a 1D-CNN and the PRL with Euclidean distance loss function. The configuration of the model is detailed in Table 2. We ran these experiments once with 10-fold cross-validation. The models were trained with 31500 instances and tested with 2000 and 200 instances, for comparing the results with the obtained for the subject-specific model.

Table 7: Cross-modal retrieval results (%) obtained for the global model for the CENC dataset.

	Shuffle				**Songs**			
Nb.	**Instance**		**Class**		**Instance**		**Class**	
test	**MRR**	**RA**	**MRR**	**MAP**	**MRR**	**RA**	**MRR**	**MAP**
2000	19.39	98.20	20.85	2.70	1.51	74.67	14.48	4.75
	20.70	98.92	21.96	2.79	1.71	74.67	14.29	4.75
200	50.67	96.65	51.34	29.68	7.11	69.76	16.99	7.98
	53.18	97.16	53.80	31.12	8.15	70.01	18.59	8.09

When randomly distributing the data by folds, we noticed an improvement of the results over the obtained for the average of the subject-specific models (cf. first row of Table 5). Our initial interpretation was that the model could have adapted to match the audio and EEG modalities through the artefacts introduced by each participant, i.e., the model would be differentiating the data from different participants and not the different stimuli. Nonetheless, when the test set contained data from only one participant, the results even surpassed the previous ones, allowing to conclude that the model was able to differentiate the different pieces of songs belonging to the same participant. Aiming to further understand these results, we imposed that the songs considered for training were not in the test set and a drastic reduction of the performance in relation to the experiments with shuffled distribution was observed, which suggests that the models trained with shuffled distribution were benefitting from the fact that instances belonging to the same song, session and participant were seen during training. Then, we imposed that data from a participant was never seen during training, allowing to evaluate the ability of the model to generalise to other participants. As it was already expected, the obtained results were very poor, indicating that the model was not able to learn representations that share commonality across individuals because the EEG signal is strongly dependent on the participant as well are the physiological responses to music. In fact, Melnik et al. [17] showed that participants account for 32% of the variance in EEG studies.

6. Conclusions

This thesis demonstrated the existence of a subject-specific neural signature that may reflect the perceptual and cognitive brain mechanisms underlying the process of listening to a particular music, and that, together with the audio information are able to generate meaningful representations, findings that were supported by the audio-EEG cross-modal retrieval tasks carried out. Nonetheless, the results also indicated that this neural signature presents a strong variability as different neurophysiological and psychological processes occur simultaneously and the EEG response to music is driven not only to sources related to the auditory stimulus but also by unrelated sources, being the information of interest only a small fraction of the overall neural activity [7, 12].

The proposed EEG embedder approach greatly improved the cross-modal retrieval results both for the session and system experiments, allowing to conclude that the variable conditions presented when recording the same stimulus in different sessions or systems, caused by technical aspects or conditions of the participant, were negatively affecting the performance of the models, also supporting the existence of neural signatures able to distinguish music stimuli as stronger associations were established. The experiments with the global model indicated that, although the EEG response to a stimulus is subject-specific, the information gathered from the other participants helped in better discriminate between stimuli, suggesting the need of a larger dataset per participant.

Finally, we concluded that the commercial device OpenBCI is suitable for this task, which increases the feasibility of using EEG signals on future real-world music applications as it is a wireless device, with lower preparation time and fewer electrodes.

Future work includes improving the model through several possible approaches: pre-training the layers of each side using denoising autoencoders [2]; exploring other loss functions such as the soft intra-modal loss [11] and the CCA layer [9]; and adding a time shift before segmenting the audio and EEG as an attempt to synchronise the stimulus with its auditory and conceptual processing. Moreover, the learned audio-EEG embeddings can be used in different scenarios such as in the improvement of music recommendation systems, especially for therapeutically purposes such as stress-relieve and sleep systems, as they strongly rely on psychological characteristics of the user.

Acknowledgements

This document was written and made publically available as an institutional academic requirement and as a part of the evaluation of the MSc thesis in Biomedical Engineering of the author at IST. The work described herein was performed at INESC-ID's Spoken Language Systems Laboratory, IST, Universidade de Lisboa, and CENC, between February 2018 and May 2019. The work was carried out under the supervision of Prof. David Matos and Prof. Teresa Paiva to whom I would like to express my appreciation for their guidance throughout this work.

References

[1] E. O. Altenmüller. How many music centers are in the brain? *Annals of the New York Academy of Sciences*, 930(1):273–280, 2001.

[2] G. Andrew et al. Deep canonical correlation analysis. In *International Conference on Machine Learning*, pages 1247–1255, 2013.

[3] C. Audette et al. The OpenBCI GUI. `https://github.com/OpenBCI/OpenBCI\_GUI`, 2018. Accessed on 2018-04-16.

[4] J. S. Bergstra et al. Algorithms for hyper-parameter optimization. In *Advances in neural information processing systems*, pages 2546–2554, 2011.

[5] Brain Products. *User Manual for the 40 and 72-channel physiological measurement system QuickAmp-40 and QuickAmp-72 Version.* Brain Products, version 3.0 edition, May 2014.

[6] I. Daly et al. Neural correlates of emotional responses to music: an EEG study. *Neuroscience letters*, 573:52–57, 2014.

[7] A. De Cheveigné et al. Decoding the auditory brain with canonical component analysis. *NeuroImage*, 172: 206–216, 2018.

[8] I. Dobie et al. *The impact of new technologies and the Internet on the music industry, 1997-2001.* PhD thesis, University of Salford, UK, 2001.

[9] M. Dorfer et al. End-to-end cross-modality retrieval with CCA projections and pairwise ranking loss. *International Journal of Multimedia Information Retrieval*, 7 (2):117–128, 2018.

[10] S. K. Hadjidimitriou and L. J. Hadjileontiadis. Toward an EEG-based recognition of music liking using time-frequency analysis. *IEEE Transactions on Biomedical Engineering*, 59(12):3498–3510, 2012.

[11] S. Hong et al. CBVMR: Content-Based Video-Music Retrieval Using Soft Intra-Modal Structure Constraint. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 353–361. ACM, 2018.

[12] L. Jäncke et al. Time course of EEG oscillations during repeated listening of a well-known aria. *Frontiers in human neuroscience*, 9:401, 2015.

[13] Y.-P. Lin et al. Multilayer perceptron for EEG signal classification during listening to emotional music. In *TENCON 2007-2007 IEEE Region 10 Conference*, pages 1–3. IEEE, 2007.

[14] Y.-P. Lin et al. Support vector machine for EEG signal classification during listening to emotional music. In *Multimedia Signal Processing, 2008 IEEE 10th Workshop on*, pages 127–130. IEEE, 2008.

[15] Y.-P. Lin et al. EEG-based emotion recognition in music listening. *IEEE Transactions on Biomedical Engineering*, 57(7):1798–1806, 2010.

[16] S. Mathôt et al. Opensesame: An open-source, graphical experiment builder for the social sciences. *Behavior research methods*, 44(2):314–324, 2012.

[17] A. Melnik et al. Systems, subjects, sessions: to what extent do these factors influence EEG data? *Frontiers in human neuroscience*, 11:150, 2017.

[18] T. R. Mullen et al. Real-time neuroimaging and cognitive monitoring using wearable dry EEG. *IEEE Transactions on Biomedical Engineering*, 62(11):2553–2567, 2015.

[19] OpenBCI. Openbci hardware. `http://docs.openbci.com/Hardware/01-OpenBCI_Hardware`, 2018.

[20] S. S. Patil and M. K. Pawar. EOG artifact correction from EEG signals for biomedical analysis. *Quality Advancement of EEG by Wavelet Denoising for Biomedical Analysis*, 57:9, 2012.

[21] F. Raposo et al. Towards Deep Modeling of Music Semantics using EEG Regularizers. *arXiv preprint arXiv:1712.05197*, 2017.

[22] F. Raposo et al. Learning Embodied Semantics via Music and Dance Semiotic Correlations. *CoRR*, abs/1903.10534, 2019.

[23] R. Richer et al. Real-time Mental State Recognition using a Wearable EEG. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5495–5498. IEEE, 2018.

[24] B. Rosbag. OpenSesame Plug-in: Parallel Port Trigger. `https://github.com/dev-jam/opensesame_plugin_-_parallel_port_trigger`, 2016. Accessed on 2018-06-08.

[25] J. A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[26] M. Schedl et al. The neglected user in music information retrieval research. *Journal of Intelligent Information Systems*, 41(3):523–539, 2013.

[27] S. Stober et al. Classifying EEG Recordings of Rhythm Perception. In *ISMIR*, pages 649–654, 2014.

[28] S. Stober et al. Using Convolutional Neural Networks to Recognize Rhythm Stimuli from Electroencephalography Recordings. In *Advances in neural information processing systems*, pages 1449–1457, 2014.

[29] I. Sturm. Analyzing the perception of natural music with EEG and ECoG. 2016.

[30] D. Surís et al. Cross-modal Embeddings for Video and Audio Retrieval. *arXiv:1801.02200*, 2018.

[31] M. Teplan et al. Fundamentals of EEG measurement. *Measurement science review*, 2(2):1–11, 2002.

[32] N. Thammasan et al. EEG-Based Emotion Recognition during Music Listening. In *The 28th annual conference of the Japanese society for artificial intelligence*, volume 1215, 2014.

[33] N. Thammasan et al. Continuous music-emotion recognition based on electroencephalogram. *IEICE TRANSACTIONS on Information and Systems*, 99(4):1234–1241, 2016.

[34] Z. Wei. *Utilizing EEG signal in music information retrieval.* PhD thesis, 2010.

[35] Wikipedia. 21 electrodes of International 10-20 system for EEG. `https://commons.wikimedia.org/wiki/File:International_10-20_system_for_EEG-MCN.svg`, 2017. Accessed on 2018-09-23.

[36] Wikipedia. 1020 system (EEG). `https://en.wikipedia.org/wiki/10-20_system_(EEG)`, 2018. Accessed on 2018-09-23.

[37] F. Yan and K. Mikolajczyk. Deep correlation for matching images and text. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3441–3450, 2015.

[38] Y. Yu et al. Deep Cross-modal Correlation Learning for Audio and Lyrics in Music Retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 15(1), 2019. doi: 10.1145/3281746.