LISTEN TO YOUR MIND'S (HE)ART: A SYSTEM FOR AFFECTIVE MUSIC GENERATION VIA BRAIN-COMPUTER INTERFACE

Marco TIRABOSCHI (marco.tiraboschi@unimi.it) (0000-0001-5761-4837) [1],
Federico AVANZINI (federico.avanzini@unimi.it) (0000-0002-1257-5878) [1], and
Giuseppe BOCCIGNONE (giuseppe.boccignone@unimi.it) (0000-0002-5572-0924) [2]

[1] *Laboratorio di Informatica Musicale (LIM)*, *Department of Computer Science*, **Università degli Studi di Milano**, Italy
[2] *PHuSe Lab*, *Department of Computer Science*, **Università degli Studi di Milano**, Italy

ABSTRACT

We present an approach to the problem of real-time generation of music, driven by the affective state of the user, estimated from their electroencephalogram (EEG). This work is aimed at exploring strategies for real-time music generation applications using sensor data. Applications can range from responsive music for x-reality to art installations, and music generation as feedback in pedagogical contexts. We developed a Brain-Computer Interface in the open-source platform OpenViBE. It manages communication with the EEG device and computes the relevant features. A benchmark dataset is used to evaluate the performance of supervised learning methods on the binary classification task of valence and arousal. We also assessed the performance using a reduced number of electrodes and frequency-bands, in order to address the problems of lower budgets and noisy environments. Then, we address the requirements for a real-time music generation model and propose a modification to Magenta's MusicVAE, introducing a parameter for controlling inter-batch memory. In the end, we discuss possible strategies to map desired music features to a model's native input features. We present a Probabilistic Graphical Model to model the mapping from valence/arousal to MusicVAE's latent variables. We also address dataset dimensionality problems proposing three probabilistic solutions.

1. INTRODUCTION

Interest and curiosity towards exploiting the EEG to control sound and music with one's own brain have always been strong in the sound and music community. It was not until 1973 that EEG gained currency as a means for setting a direct communication between brains and computers [1], and it was almost 20 years later that the first BCI successfully allowed the users to control the cursor on a computer screen. But the first reported use of EEG in music is "Music for Solo Performer" by Alvin Lucier (1965) [2]. He had met researcher Edmond Dewan who asked Lucier if he would be interested in using his equipment to detect alpha waves for a piece of music. Alpha waves have frequencies around 8-13 Hz, and would not be audible as audio: so he thought of them as rhythms, thus suitable to create a piece for percussion, by amplifying the alpha bandwidth to drive loudspeakers placed on top of drums membranes [3]. Subsequently, other pioneers of EEG musical applications, such as Richard Teitelbaum, David Rosemboom and Roger Lafosse, exploited brainwaves and other biological signals (e.g. the ECG) to control sound synthesisers, as in the experimental piece "Spacecraft" by Richard Teitelbaum, presented at *Musica Elettronica Viva* in 1967 [4]. In recent years, interest has grown around so-called *Brain-Computer Music Interfaces* [5] and general BCIs because of the increasing affordability of reliable EEG equipment.

The chief concern of this work is to present a system for generating music that reflects the users' affective state, which is estimated from their EEG. We also propose solutions to some practical problems of real BCI systems. Our system is composed of four main modules. Each module is discussed in its own section and sections follow the design order, rather than the data-flow order.

- Brain-Computer Interface (Sec. 3): the EEG hardware and software for acquisition and preprocessing
- EEG Affect Recognition (Sec. 2): the model of valence and arousal correlates of an EEG signal
- Musical Affect Model (Sec. 5): the generative model of musical features conditioned on affective states
- Music Generator (Sec. 4): the model for generating music conditionally to a set of musical features

2. EEG AFFECT RECOGNITION

The EEG Affect Recognition module is the subsystem that is responsible for associating an affective label to the electrical activity of the brain. Affect denotes the mental counterpart of bodily sensation and affective features, such as valence and arousal, capture what a given instance of experience feels like [6]. Valence refers to the feeling of pleasure or displeasure; arousal refers to a feeling of activation or sleepiness. It is worth remarking that in the literature concerning the computational modelling of emotions, the term "affect" is often used interchangeably with that of "emotion" but they should not be confused; emotions are constructed from affect, emotional events being specific instances of affect that are linked to the immediate situation and involve intentions to act [6]. Indeed, the system presented here deals with affect. However, in what follows, markedly when discussing related work, we will occasionally adopt such convention for the sake of simplicity.

2.1 Related Work

There are several works that address EEG emotion recognition. The DEAP dataset [7] is a popular dataset used as a benchmark for this task. It is a dataset collecting EEG and physiological signals recorded from 32 subjects over 40 trials per subject. The authors also presented some approaches to the emotion recognition task. They used a Gaussian Naive Bayes classifier trained on band-power features. They report a Leave-One-Out Cross-Validation (LOOCV) F_1 score of 0.56 for valence and 0.58 for arousal. Jatupaiboon et al. [8] assessed the problem of real-time valence estimation. They used SVM with power-spectral features for the binary classification of valence. Their work is not directly comparable to the DEAP paper because they use their own dataset, but they show some very important points. First, the average performance of subject-independent models is significantly lower than subject-dependent models (0.65 against 0.75 accuracy). Second, by using only the pair of channels T7 and T8, the performance achieved is comparable to that attained by using all the 14 channels (0.73 accuracy). Menezes et al. exploited the DEAP dataset as a benchmark for emotion recognition in virtual environments [9]. They evaluated band-power features, temporal statistics and the Higher Order Crossing (HOC) features [10] via SVM and Random Forest. They found that band-powers and their statistics performed similarly while the HOC were less predictive.

2.2 Approach

2.2.1 Features

We chose band-powers as features for several reasons: they are the most common features used in EEG emotion recognition; they have a good predictive power; they can be computed efficiently and online; they have been shown to have neurobiological significance [11, 12] in describing the brain activity. For consistency, we adopted the same band definitions as in the DEAP paper: *theta* (4 to 8 Hz), *slow alpha* (8 to 10 Hz), *alpha* (8 to 12 Hz), *beta* (12 to 30 Hz), *gamma* (over 30 Hz). We compute the logarithm of the RMS of the signal for each band and channel. Then, we also compute the difference of each band-power for 14 pairs of symmetric electrodes.

2.2.2 Evaluation Protocol

We divided the 32 subjects randomly into a training set and a test set (60:40). For each subject, every model is evaluated via LOOCV. We will refer to the LOOCV scores on the training subjects as the *validation* scores and to the LOOCV scores on the test subjects as the *test* scores. Comparison between validation scores is performed with related samples tests. The paired-sample T-test [13] for normally distributed samples and the Wilcoxon signed-rank test [14] otherwise. Comparison between validation scores and test scores is performed with an independent samples test, (independent-samples T-test if normal or Wilcoxon rank-sum otherwise). Normality is checked via the Shapiro-Wilk test [15]. We computed validation scores for common machine learning approaches from the BCI literature: Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis [16] (QDA), and Support Vector Machine [17] (SVM). We used QDA with diagonal covariance: this can be called Gaussian Naive Bayes (as in the DEAP dataset paper). As to SVM, we adopted linear, polynomial and RBF kernels. Two "dummy" classifiers were used as reference: one that always predicts the majority class (*Majority*) and another that predicts at random with a probability determined by the label ratio (*Ratio*).

	Valence		Arousal	
Model	mean	std	mean	std
Majority	0.315	0.138	0.310	0.163
SVM (rbf)	0.395	0.133	0.340	0.156
Ratio	0.472	0.062	0.485	0.074
Naive Bayes	0.620	0.095	0.535	0.100
LDA	0.609	0.079	0.560	0.106
SVM (linear)	0.590	0.093	**0.579**	0.078
SVM (poly 7)	**0.626**	0.114	0.557	0.115
Test	0.622	0.092	0.550	0.089

Table 1. Validation F_1 scores for different models and test scores for the best-scoring model (SVM with polynomial kernel of seventh degree) for binary classification of valence and arousal in ascending order of average score. Suboptimal polynomial SVMs are omitted.

2.2.3 Feature Sets

Performance on the full feature set is summarized in Tab. 1. Overall, the results on arousal match the ones in the DEAP dataset paper, but not for valence: we obtained 0.62 accuracy against their 0.56 (this could be due to the slightly different feature set). The best-scoring model (on average and on valence) is the SVM with a 7th degree polynomial kernel. Linear SVM performs better on arousal. Rbf-kernel SVM performs worse than the dummy *Ratio* predictor. The performance of SVM (7th degree polynomial kernel) is significantly different from the dummy predictors and rbf-kernel SVM ($p \leq 0.01$), but not from other models ($p > 0.1$). Test scores are not significantly different from validation.

In a real-time BCI setting, it would be impractical to use an EEG headset with 32 channels, as it would require a very long setup time. The minimal set of channels to be able to use information about band power asymmetry is 2. The brain activity is known to correlate with valence if measured on T7-T8 [8]. Also, activity at CP5-CP6 [7] correlates with arousal. Because of our hardware, we are interested in the specific case of a feature set built using 6 channels. Thus, we evaluated two different setups, adding two pairs of electrodes to either one of the two pairs T7-T8 and CP5-CP6. We wanted one pair at the front of the brain and one at the back, to diversify the information: the pairs FC5-FC6 (frontal-central) and PO3-

6c	T7-T8				CP5-CP6				T7-T8 $(\alpha+\beta)$				CP5-CP6 $(\alpha+\beta)$			
	Valence		Arousal		Valence		Arousal		Valence		Arousal		Valence		Arousal	
Model	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
LDA	0.512	0.066	0.538	0.113	0.563	0.085	0.522	0.084	0.532	0.074	0.539	0.091	0.550	0.070	0.562	0.104
NB	0.597	0.128	0.525	0.098	**0.622**	0.093	0.531	0.080	0.589	0.118	0.522	0.104	0.603	0.093	0.549	0.086
SVM-L	0.611	0.117	0.553	0.115	0.615	0.101	**0.557**	0.079	0.592	0.123	0.518	0.096	**0.616**	0.083	0.530	0.093
SVM-P	0.606	0.118	0.537	0.111	0.609	0.097	0.550	0.114	0.559	0.106	**0.575**	0.114	0.556	0.092	0.528	0.095

Table 2. Validation F_1 scores for different models on the 6-channels feature sets: using the T7-T8 as central channels or CP5-CP6, and using all frequencies or just *alpha* and *beta* bands. Mean and standard deviation of the F_1 scores are reported for binary classification of valence and arousal.

2c	T7-T8				CP5-CP6				T7-T8 $(\alpha+\beta)$				CP5-CP6 $(\alpha+\beta)$			
	Valence		Arousal		Valence		Arousal		Valence		Arousal		Valence		Arousal	
Model	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
LDA	0.601	0.125	0.503	0.128	**0.612**	0.060	**0.555**	0.092	0.587	0.113	0.487	0.092	0.617	0.087	0.514	0.099
NB	0.584	0.131	0.509	0.105	0.610	0.071	0.529	0.094	0.561	0.137	0.487	0.078	**0.593**	0.092	0.544	0.087
SVM-L	0.619	0.136	0.494	0.150	0.640	0.072	0.535	0.104	0.578	0.171	0.435	0.073	0.571	0.122	0.485	0.123
SVM-P	0.650	0.070	0.515	0.086	0.602	0.101	0.538	0.090	0.577	0.119	0.529	0.119	0.561	0.114	0.512	0.105

Table 3. Validation F_1 scores for different models on the 2-channels feature sets: using the T7-T8 channels or CP5-CP6, and using all frequencies or just *alpha* and *beta* bands. Mean and standard deviation of the F_1 scores are reported for binary classification of valence and arousal.

PO4 (parieto-occipital) show correlations with affective labels [7]. We evaluated performance on two 6-channels feature sets: FC5, FC6, PO3, PO4 with T7-T8 or with CP5-CP6. In real-world operation, low-frequency components can be subject to noise from muscle movement. Also, high-frequency components can be affected by power-line interference (50 Hz or 60 Hz). Hence, we also assessed the performance of the models trained only on the central frequency bands (*slow alpha*, *alpha* and *beta*). Validation scores on 6-channel feature sets are summarized in Tab. 2 for all four configurations: two channel choices, both with all frequency bands or only with central frequency bands $(\alpha+\beta)$. We also assessed the performance on 2-channels feature sets, using T7-T8 or CP5-CP6, exploiting all frequency bands or just the central bands (Tab. 3).

2.3 Results

Using 6 channels, F_1-scores for valence classification are very similar to the ones obtained with the full feature set. Scores on arousal slightly decreased. Using all frequency bands, the best model on average is the Linear SVM (0.62 on valence and 0.56 on arousal). Using only the central bands, the best model on average is Naive Bayes (0.60 on valence and 0.55 on arousal). In almost every case, performance is better using CP5-CP6. Employing 2 channels with all frequency bands, F_1-scores are still very similar to the previous ones. However, the two best validation scores on valence (Polynomial SVM with T7-T8 and Linear SVM with CP5-CP6) are significantly different from the test scores: the test score for Linear SVM is 0.54 ($p < 0.05$). Thus, we consider LDA as the best model (0.61 on valence and 0.56 on arousal), since it is consistent across the two partitions. We observed that performance on arousal using T7-T8 is not significantly different from the dummy predictor (Ratio). We surmise that T7-T8 is not a sufficient configuration for arousal classification, in contrast to its use for valence classification (as in previous literature [8]). Using 2 channels and only the central bands, none of the models is significantly better than the dummy predictors for arousal classification (all $p < 0.05$). Valence classification is still possible, but the validation score of LDA with CP5-CP6 (0.62) is significantly different from the test score: 0.52 ($p < 0.01$). So, we consider Naive Bayes as the best model, with an F_1-score of 0.59.

Based on these observations, we determined three different setups to address different requirements: FC5-FC6-CP5-CP6-PO3-PO4 $(\alpha+\beta)$ for robust features, CP5-CP6 for minimal hardware and CP5-CP6 $(\alpha+\beta)$ for robust features and minimal hardware, but for valence only.

3. BRAIN-COMPUTER INTERFACE

The Brain-Computer Interface module is the system that allows for sensing the brain activity of the user and extracting the relevant features.

3.1 Related Work

Brain-Computer Interfaces for music are becoming more and more popular. BCIs are often categorized as “passive” (using arbitrary brain activity without the purpose of voluntary control), “reactive” (using brain activity arising in reaction to external stimulation), and “active” (using brain activity that is consciously controlled by the user, independently from external events) [18]. A popular technique for reactive BCIs is based on *steady state visually evoked potentials* (SSVEP). It consists in presenting images on screen that flicker at different rates, and detecting electrical potentials on the visual cortex to determine which object the user is watching. SSVEP-BCIs have been used for

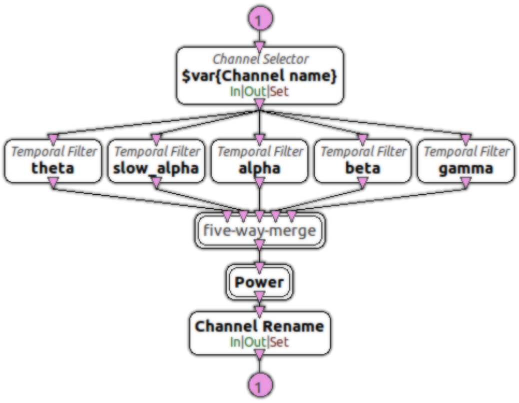

Figure 1. The *Band-powers DEAP* metabox. This metabox selects one channel (named *Channel name*) out of the multi-channel signal and computes its band-powers.

music writing, music navigation [19], and sonic expression [20]. A common approach to active BCIs is *motor imagery*. It consists in detecting patterns correlated to the imagination of motor activity [21] (μ *rhythms*). MI-BCIs are becoming popular for video game control [22]. Passive BCIs are often exploited in affective computing, e.g. for monitoring attention, stress and affective states [23].

3.2 Approach

We used the g.tec g.Sahara active dry EEG electrode system headset. We developed a feature extraction software for the BCI in OpenViBE, a cross-platform open-source environment [24]. It is headset-independent, because the OpenViBE *acquisition server* handles incoming signals. Thus, it can be used with any EEG headset, provided that the drivers are available. The OpenViBE designer is a visual programming environment. An executable is called a *scenario* and its components *boxes*. A scenario can be used within another scenario as a *metabox*.

3.2.1 Feature Extraction

For each channel, we compute the band-powers. We defined the band-power of a frequency band as the mean-square of the band-passed signal (see Sec. 2.2.1). We defined the metabox *Power* to compute the power of a signal over overlapping temporal windows. Then, we developed the metabox *Band-powers DEAP*, that selects one channel and splits it into 5 different bands using a time-domain filterbank. It uses the *Power* metabox to compute the band-powers. The band-power signals are rearranged into a single multi-channel signal and renamed. The metabox *Band-Power Features* instantiates a *Band-powers DEAP* metabox for each channel to extract its band-powers and rearranges the band-power signals into a single multi-channel signal.

3.2.2 Data Transmission

After computing the band-power features, we need to be send them to the EEG affect recognition process. We use the Open Sound Control (OSC) [25] network protocol for this. Due to the great interest of the sound and music community in BCIs, OSC has also become a common protocol for BCIs and some companies that develop BCIs provide OSC utilities, such as Emotiv's *Mind Your OSCs*. OpenViBE also has a rudimentary OSC client *box*. Each feature is sent on a different OSC method. The address pattern is `/eeg/<channel>/<band>`.

4. ONLINE MUSIC GENERATION

The Online Music Generation module is the system that generates music in real-time. We want to be able to control it with parameters that can change over time. We propose a transfer-learning approach for generating music via affective parameters.

4.1 Related Work

The recent developments in deep learning enabled new approaches that are now the state-of-the art of music generation. Deep learning for music mainly exploits advancements in natural language modelling. Especially, the introduction of *attention-based* Recurrent Neural Networks [26] (RNN) allowed for longer time-scale coherence than before. The Magenta project by Google Brain has developed several deep learning models for computational creativity. Melody RNN [27] employs an attention-based LSTM (*Long Short-Term Memory*, a type of RNN) for generating melodies. To improve long-term structure, they developed a hierarchical RNN decoder for an autoencoder called MusicVAE [28]. The latest architecture in language modelling is the transformer [29], a deep neural network that does not use recurrence, but relies entirely on attention, such as OpenAI's GPT-2 [30] and GPT-3 [31]. Magenta's Music Transformer [32] is a transformer that employs relative self-attention for music generation. OpenAI's MuseNet [33] is based on their GPT-2 and is a large-scale transformer for symbolic generation that supports up to 10 different instruments. On the other hand, their Jukebox [34] generates raw musical audio of fully arranged compositions with singing voice. It can be conditioned on either artist, genre or lyrics. The main drawback of deep learning is the amount of data required for training: as an example, Music Transformer has been trained with 10 000 hours of piano music retrieved from YouTube and converted from audio to MIDI by another neural network, Onsets and Frames [35].

4.2 Modified MusicVAE

We chose MusicVAE for real-time music generation. One reason is that pre-trained checkpoints are available for download. Also, operations in its latent space have semantic effects on the output (e.g. interpolation in the latent space results in the semantic interpolation of the MIDI content). Finally, it supports multi-instrument pieces. A pre-trained checkpoint is available for use with *trios* (drums, bass and melody). We propose a small modification that does not require re-training for real-time parameter modification. MusicVAE is a variational auto-encoder (VAE)

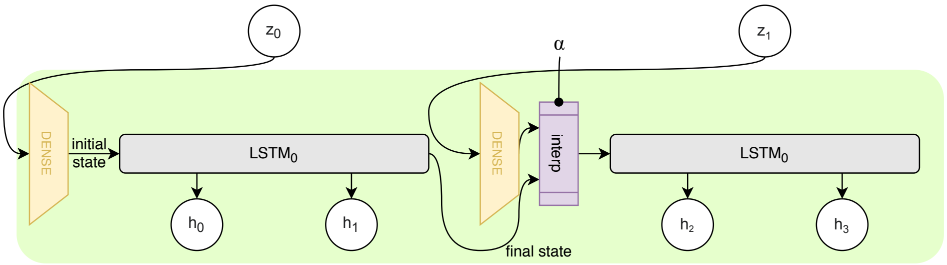

Figure 2. Top layer of the restructured decoder RNN for online music generation. The final state of the LSTM is interpolated with the new state to obtain a new initial state that is *in-between* the two. This results in a transition between the two parts.

composed of a deep Bidirectional LSTM (BLSTM) probabilistic encoder and a hierarchical LSTM probabilistic decoder. The encoder encodes an entire input MIDI into one single latent vector. It can only be executed batch because of its bidirectionality (the output at any time step depends both from past and future inputs). The decoder decodes a latent vector into any length of musical content. It can be executed in real-time because it is monodirectional. Also, memory cost does not increase with time because the output depends only on the current state of the network. The drawback of MusicVAE for our intended application is that the decoder doesn't allow any input for transitioning to a different part without starting over. Thus, we restructured the network for sequentially generating music with smooth transitions. The first layer of the network is visualized in Fig. 2. When sampling conditionally on a new encoding z_{t+1}, we compute the corresponding new initial state for the RNN using the same fully-connected layer as the original network. We then compute the new state for the LSTM as a convex combination of that output and the final state from the previous execution of the LSTM. We introduce a *memory* coefficient $\alpha \in [0, 1]$. Defining f_t the final state after decoding z_t and $d(z_t)$ the result of applying the dense layer to z_t, the new initial state will be

$$s_{t+1} := \alpha \cdot f_t + (1 - \alpha) \cdot d(z_{t+1}). \qquad (1)$$

Setting $\alpha = 0$ results in independent samples (the network forgets the previous state), whilst setting $\alpha = 1$ the network ignores new inputs. Setting the memory to an in-between value allows for adjusting the trade-off between coherence ($\alpha \to 1$) and change ($\alpha \to 0$). We implemented this modification in Python by extending Magenta's own class for MusicVAE pre-trained models and overriding the definition of the decoding operation.

4.3 Multiprocess System

Decoding a MIDI section from MusicVAE is not a fast operation in a musical context. It can take several seconds on a laptop using a CPU. We developed a dual-process system to overcome this problem. The two processes involved are a MIDI sequencer client and a MusicVAE server. Since they are local processes, we handle their communication with pipes. We implemented the MIDI sequencer using the Python bindings for FluidSynth [36]. The sequencer is never blocked waiting for a request of a MIDI sequence. Instead, every time the client callback is called, if there is a pending request, the MIDI is read from the pipe and scheduled for the synthesizer. Then, a new MIDI sequence is requested from the client.

5. MUSICAL AFFECT MODEL

The module named Musical Affect Model is the subsystem that maps musical features to affective labels. Here, musical features are encoding vectors in the MusicVAE latent space (see Sec. 4) and an affective state is a pair of binary labels for valence and arousal (see Sec. 2).

5.1 Related Work

Affect correlates of music features have always been a subject of great interest for musicologists, although they are not as often computationally exploited for music generation. Williams *et al.* propose a taxonomy for what they refer to as Affective Algorithmic Composition (AAC) systems [37]. AAC systems can be compositional if they generate music (e.g. Robertson *et al.* [38]) or performative if they execute a musical piece in a way that reflects the target emotion (such as RaPScoM [39]). Briefly, they are generative if they write new music or transformative if they modify a given input, such as a music production system [40]. They can be real-time or in batch: real-time systems are adaptive if they can adjust their output during execution, as our modified MusicVAE (see Sec. 4.2). Williams *et al.* later presented an AAC system that targeted affective states by means of lookup table of musical features compiled from literature review [41]. They specify a discrete mapping from the affective-state space to the set of musical features, then a neural network outputs MIDI. Kirke and Miranda developed an AAC system for communicating the affective state detected from an EEG signal [42]. Valence is computed as the difference of logarithmic alpha-band energy between left and right frontal regions of the brain. Arousal is computed as the negative of the sum of such logarithmic alpha-band energies. Binary valence and arousal are used to transform a pre-written musical score by changing key, pitch and tempo. Galvanic skin response (GSR) is another biological signal that is known to correlate with affective states and Daly *et al.* developed a system where GSR serves as input for affective music generation [43].

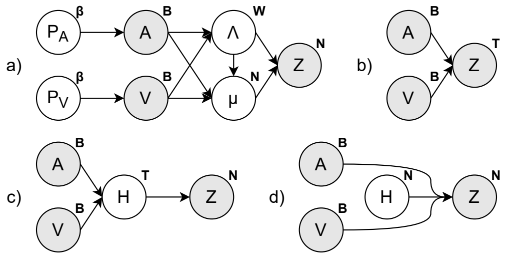

Figure 3. Probabilistic Graphical Models discussed in text. A PGM is a directed graph where nodes represent random variables and arcs stand for conditional probabilities. Hidden variables are in white and observed variables are in grey. Annotations on top of each variable denote the type of its conditional distribution (Beta, Bernoulli, Wishart, Normal or T). The PGM (a) is simplified in a graph only including observed variables (b). PPCA introduces a new hidden variable H: PPCA can be performed on the entire dataset (c) or for each class (d). See text for details.

5.2 Approach

We developed a directed Probabilistic Graphical Model (PGM) to map affective labels to MusicVAE encodings.

5.2.1 Dataset

The *MIREX-like mood* dataset is a dataset for multimodal music emotion recognition [44]. It collects 903 audio samples, 193 of which with lyrics and MIDI. Affective tags are adjectives, grouped in 5 clusters. First, we preprocessed the dataset to get the MusicVAE encodings from the MIDI files: 151 of the 193 MIDI files were compatible with the MusicVAE *trio* model. We used a dataset containing the "*Norms of valence, arousal, and dominance for 13 915 English lemmas*" [45] for converting adjectives to valence and arousal values. We observed that the 5 pre-defined clusters did not map to clusters in the valence-arousal 2-D space. We partitioned the samples into four classes by binarizing valence and arousal, median values being the thresholds.

5.2.2 Model

We used a directed PGM to model the interdependency of the different variables. As a consequence of MusicVAE's ELBO loss function, the prior distribution of the latent codes Z is a standard *multivariate Normal distribution* (MVN) [28]. Therefore, we model the conditional distribution of Z given an affective state (a, v) as a MVN, as well. The mean and precision parameters given each affective state $(\mu_{a,v}, \Lambda_{a,v})$ are unknown. The joint posterior distribution of the unknown mean and precision parameters of a MVN is a Normal-Wishart distribution. However, they are never observed and we are not interested in inferring them. So, we can directly model the distribution of latent codes given an affective state as a multivariate Student's T distribution: this is the distribution of the samples of a MVN whose mean and the precision are Normal-Wishart distributed. Valence and arousal only assume binary values, so, we model them as Bernoulli variables. The mean parameters (p_a and p_v) are unknown. The posterior distribution of the mean of a Bernoulli variable is a Beta distribution. As previously, their inference is not of interest. The distribution of a Bernoulli variable whose mean is Beta distributed is still a Bernoulli distribution. The full PGM and its simplified version are visualized in Fig. 3a and Fig. 3b.

5.2.3 Dimensionality Reduction

The dimensionality of the latent space is much larger than the available data points. The MusicVAE trio model has 512 latent variables and only 151 points are in the dataset: when partitioned into 4 classes, it amounts to an average of 38 points per class. This is not sufficient for inferring the parameters of the multivariate Student's T distribution because the sample covariance matrix is singular. We present three approaches for overcoming this problem.

We can make a Naive Bayes assumption, imposing all features to be independent from each other given the class. Often Naive Bayes is applied in contexts where the independence assumption is not supported [46]. In our case, it could be partially motivated by the fact that the prior distribution of the encodings is a standard MVN, for which the assumption holds. We applied Naive Bayes to our graphical model by setting to zero all non-diagonal values of the sample covariance matrix.

We can use probabilistic PCA (PPCA) [47] to map our samples to a lower-dimensional space. There are two possible ways of applying PPCA to our graph. We can use a *class-independent* PPCA to map all the latent codes to a lower-dimensional space, where we can estimate the parameters of the class-wise T distributions. The resulting PGM is visualized in Fig. 3c. Alternatively, we can use PPCA as the model of the distribution of the latent codes of each class. In this setting, each affective state corresponds to different values for the PPCA parameters. The corresponding BBN is shown in Fig. 3d. [1]

6. CONCLUSIONS

We have presented a pipeline for generating affectively-driven music using the EEG. Main results so far achieved can be recapped as follows. We have shown that a reduced number of EEG channels can still be effective for the binary classification of valence and arousal, resulting in cheaper and more practical BCIs. We have also developed an online feature extraction algorithm using the OpenViBE platform. This software is cross-platform and headset independent. We used the OSC protocol for the communication with the affect classification module. We discussed the requirements for an online music generation algorithm and made a modification to a pre-trained neural autoencoder (MusicVAE). This modification allows control over the latent codes when generating music, so that the output MIDI is the result of a trajectory in the latent space, instead of a single static code. Finally, we proposed a probabilistic model for mapping affective labels to music features, in the

[1] Examples generated with a *class-dependent* PPCA (6 principal components) are publicly available at `https://chromaticisobar.github.io/ListenToYourMindsHeArt`

form of MusicVAE latent codes. Thee different probabilistic solutions are presented for dimensionality problems.

In the future, we plan to make a thorough investigation in the relationship between the number of EEG channels and the affect classification performance. Also, we plan to collect a dataset of affectively labelled MIDI to evaluate the dimensionality reduction models we exploited when mapping affective labels to music features.

Acknowledgments

Part of this work has been possible thanks to the collaboration with the *Interdisciplinary Centre for Computer Music Research* of University of Plymouth. This collaboration was funded by "Thesis Abroad" scholarship of University of Milan. We would especially like to thank Prof. Eduardo Miranda, Dr. Edward Braund and Mr. Satvik Venkatesh.

7. REFERENCES

[1] J. J. Vidal, "Toward direct brain-computer communication," *Annual Review of Biophysics and Bioengineering*, 1973.

[2] V. Straebel and W. Thoben, "Alvin Lucier's music for solo performer: experimental music beyond sonification," *Organised sound*, 2014.

[3] V. Rusche and H. Harder, "No ideas but in things: The composer Alvin Lucier," Movie, 2012.

[4] B. Arslan, A. Brouse, J. Castet, J.-J. Filatriau, R. Lehembre, Q. Noirhomme, and C. Simon, "From biological signals to music," in *2nd International conference on enactive interfaces*, 2005.

[5] E. R. Miranda and J. Castet, *Guide to brain-computer music interfacing*. Springer, 2014.

[6] L. F. Barrett, "The theory of constructed emotion: an active inference account of interoception and categorization," *Social cognitive and affective neuroscience*, vol. 12, no. 1, pp. 1–23, 2017.

[7] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 18–31, 2011.

[8] N. Jatupaiboon, S. Pan-ngum, and P. Israsena, "Real-time eeg-based happiness detection system," *The Scientific World Journal*, vol. 2013, 2013.

[9] M. L. R. Menezes, A. Samara, L. Galway, A. Sant'Anna, A. Verikas, F. Alonso-Fernandez, H. Wang, and R. Bond, "Towards emotion recognition for virtual environments: an evaluation of eeg features on benchmark dataset," *Personal and Ubiquitous Computing*, vol. 21, no. 6, pp. 1003–1013, 2017.

[10] P. C. Petrantonakis and L. J. Hadjileontiadis, "Emotion recognition from eeg using higher order crossings," *IEEE Transactions on information Technology in Biomedicine*, vol. 14, no. 2, pp. 186–197, 2009.

[11] G. Deuschl and A. Eisen, *Recommendations for the practice of clinical neurophysiology: guidelines of the International Federation of Clinical Neurophysiology*. Elsevier Health Sciences, 1999, no. 52.

[12] R. Yuvaraj, M. Murugappan, N. M. Ibrahim, M. I. Omar, K. Sundaraj, K. Mohamad, R. Palaniappan, E. Mesquita, and M. Satiyan, "On the analysis of eeg power, frequency and asymmetry in parkinson's disease during emotion processing," *Behavioral and brain functions*, vol. 10, no. 1, p. 12, 2014.

[13] Student, "The probable error of a mean," *Biometrika*, pp. 1–25, 1908.

[14] F. Wilcoxon, "Individual comparisons by ranking methods," in *Breakthroughs in statistics*. Springer, 1992, pp. 196–202.

[15] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3/4, pp. 591–611, 1965.

[16] D. Garrett, D. A. Peterson, C. W. Anderson, and M. H. Thaut, "Comparison of linear, nonlinear, and feature selection methods for eeg signal classification," *IEEE Transactions on neural systems and rehabilitation engineering*, vol. 11, no. 2, pp. 141–144, 2003.

[17] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, pp. 1–27, 2011.

[18] T. O. Zander, C. Kothe, S. Jatzev, and M. Gaertner, "Enhancing human-computer interaction with input from active and passive brain-computer interfaces," in *Brain-Computer Interfaces - Applying Our Minds to Human-Computer Interaction*, D. S. Tan and A. Nijholt, Eds. Springer, 2010, pp. 181–199.

[19] S. Venkatesh, "Investigation into stand-alone brain-computer interfaces for musical applications," Master's thesis, University of Plymouth, 2019.

[20] S. Venkatesh, E. Braund, and E. R. Miranda, "Designing brain-computer interfaces for sonic expression," in *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*. Birmingham City University, 2020, pp. 525–530.

[21] A. S. Aghaei, M. S. Mahanta, and K. N. Plataniotis, "Separable common spatio-spectral patterns for motor imagery bci systems," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 1, pp. 15–29, 2015.

[22] D. Coyle, J. Garcia, A. R. Satti, and T. M. McGinnity, "Eeg-based continuous control of a game using a 3 channel motor imagery bci: Bci game," in *2011 IEEE Symposium on Computational Intelligence, Cognitive Algorithms, Mind, and Brain (CCMB)*. IEEE, 2011, pp. 1–7.

[23] A.-M. Brouwer, J. van Erp, D. Heylen, O. Jensen, and M. Poel, "Effortless passive bcis for healthy users," in *Universal Access in Human-Computer Interaction. Design Methods, Tools, and Interaction Techniques for eInclusion*, C. Stephanidis and M. Antona, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 615–622.

[24] Y. Renard, F. Lotte, G. Gibert, M. Congedo, E. Maby, V. Delannoy, O. Bertrand, and A. Lécuyer, "Openvibe: An open-source software platform to design, test, and use brain–computer interfaces in real and virtual environments," *Presence: teleoperators and virtual environments*, vol. 19, no. 1, pp. 35–53, 2010.

[25] A. Freed, "Open sound control: A new protocol for communicating with sound synthesizers," in *International Computer Music Conference (ICMC)*, 1997.

[26] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[27] E. Waite *et al.*, "Generating long-term structure in songs and stories," *Web blog post. Magenta*, vol. 15, 2016.

[28] A. Roberts, J. Engel, C. Raffel, C. Hawthorne, and D. Eck, "A hierarchical latent vector model for learning long-term structure in music," *arXiv preprint arXiv:1803.05428*, 2018.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[30] A. Radford, J. Wu, D. Amodei, D. Amodei, J. Clark, M. Brundage, and I. Sutskever, "Better language models and their implications," *OpenAI Blog https://openai. com/blog/better-language-models*, 2019.

[31] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[32] C.-Z. A. Huang, A. Vaswani, J. Uszkoreit, N. Shazeer, I. Simon, C. Hawthorne, A. M. Dai, M. D. Hoffman, M. Dinculescu, and D. Eck, "Music transformer," *arXiv preprint arXiv:1809.04281*, 2018.

[33] C. M. Payne, "Musenet," OpenAI Blog, 2019.

[34] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, "Jukebox: A generative model for music," *arXiv preprint arXiv:2005.00341*, 2020.

[35] C. Hawthorne, E. Elsen, J. Song, A. Roberts, I. Simon, C. Raffel, J. Engel, S. Oore, and D. Eck, "Onsets and frames: Dual-objective piano transcription," *arXiv preprint arXiv:1710.11153*, 2017.

[36] J. Newmarch, "Fluidsynth," in *Linux Sound Programming*. Springer, 2017, pp. 351–353.

[37] D. Williams, A. Kirke, E. R. Miranda, E. B. Roesch, and S. J. Nasuto, "Towards affective algorithmic composition," in *The 3rd International Conference on Music & Emotion, Jyväskylä, Finland, June 11-15, 2013*. University of Jyväskylä, Department of Music, 2013.

[38] J. Robertson, A. de Quincey, T. Stapleford, and G. Wiggins, "Real-time music generation for a virtual environment," in *Proceedings of ECAI-98 Workshop on AI/Alife and Entertainment*. Citeseer, 1998.

[39] J. Doppler, J. Rubisch, M. Jaksche, and H. Raffaseder, "Rapscom: towards composition strategies in a rapid score music prototyping framework," in *Proceedings of the 6th Audio Mostly Conference: A Conference on Interaction with Sound*, 2011, pp. 8–14.

[40] A. P. Oliveira and A. Cardoso, "Automatic manipulation of music to express desired emotions," in *Proceedings of the 6th Sound and Music Computing Conference. Porto: University of Porto*. Citeseer, 2009, pp. 265–270.

[41] D. Williams, A. Kirke, E. Miranda, I. Daly, F. Hwang, J. Weaver, and S. Nasuto, "Affective calibration of musical feature sets in an emotionally intelligent music composition system," *ACM Transactions on Applied Perception (TAP)*, vol. 14, no. 3, pp. 1–13, 2017.

[42] A. Kirke and E. R. Miranda, "Combining eeg frontal asymmetry studies with affective algorithmic composition and expressive performance models," in *ICMC*, 2011.

[43] I. Daly, A. Malik, J. Weaver, F. Hwang, S. J. Nasuto, D. Williams, A. Kirke, and E. Miranda, "Towards human-computer music interaction: Evaluation of an affectively-driven music generator via galvanic skin response measures," in *2015 7th Computer Science and Electronic Engineering Conference (CEEC)*. IEEE, 2015, pp. 87–92.

[44] R. Panda, R. Malheiro, B. Rocha, A. Oliveira, and R. P. Paiva, "Multi-modal music emotion recognition: A new dataset, methodology and comparative analysis," in *International Symposium on Computer Music Multidisciplinary Research*, 2013.

[45] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 english lemmas," *Behavior research methods*, vol. 45, no. 4, pp. 1191–1207, 2013.

[46] H. Zhang, "The optimality of naive bayes," *AA*, vol. 1, no. 2, p. 3, 2004.

[47] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.