A Statistical Inference Framework for Understanding Music-Related Brain Activity

Stavros Ntalampiras and Ilyas Potamitis

***Abstract*—Following the success in Music Information Retrieval (MIR), research is now steering towards understanding the relationship existing between brain activity and the music stimuli causing it. To this end, a new MIR topic has emerged, namely Music Imagery Information Retrieval, with its main scope being to bridge the gap existing between music stimuli and its respective brain activity. In this paper, the encephalographic modality was chosen to capture brain activity as it is more widely available since of-the-shelf devices recording such responses are already affordable unlike more expensive brain imaging techniques. After defining three tasks assessing different aspects of the specific problem (stimuli identification, group and meter classification), we present a common method to address them, which explores the temporal evolution of the acquired signals. In more detail, we rely on the parameters of linear time-invariant models extracted out of electroencephalographic responses to heterogeneous music stimuli. Subsequently, the probability density function of such parameters is estimated by hidden Markov models taking into account their succession in time. We report encouraging classification rates in the above-mentioned tasks suggesting the existence of an underlying relationship between music stimuli and their electroencephalographic responses.**

***Keywords*—Music information retrieval; music imagery information retrieval; electroencephalography; music signal processing**

I. Introduction

Even though the field of Music Information Retrieval (MIR) is relatively new, it has attracted the interest of a plethora of researchers occupied in heterogeneous disciplines ranging from musicology to computer science and signal processing. MIR research addresses various applications based on processing musical information, such as music genre recognition [1], [2], music emotion prediction [3], [4], automatic transcription and instrument analysis [5], etc. Interestingly, music emotion recognition is gaining high popularity as shown by a recent article in BBC news advertising the potential discovery of the so-called 'saddest' song ever [6].

Since the literature already includes mature solutions to the above-mentioned tasks, MIR is expanding towards the exploitation of signals representing brain activity while listening or imagining music pieces. This sub-area of MIR is called Music Imagery Information Retrieval (MIIR) and has only recently emerged [7]. It aims to support existing MIR solutions in applications such as query by singing, humming, tapping, or beat-boxing, to name but a few, with the ultimate goal being complete reconstruction of the music stimuli based on its respective brain activity, similarly to what has been recently proposed in the visual stimuli case [8], [9].

This work investigates classification of musical content via the respective EEG responses. During the last decades, such a line of research has gained popularity giving birth to a series of interesting approaches used to study brain responses to music [10]–[12]. However, there is still a great need for collaboration between MIR and neuroscience researchers towards constructing a systematic framework able to bridge the gap existing between musical signals and the way these are encoded and finally understood by the human brain.

In this context, Brain-Computer Interfaces (BCIs) could benefit from MIIR as they could incorporate capabilities that go well beyond music composing and playlist creation (and sometimes automatics adaptation based on the user preferences) offered by existing BCIs [13]–[16]. MIIR is based on the assumption that different music stimuli activate the brain in a different way, while the recording equipment is capable of capturing such differences. Then, in principle, one could use the obtained recordings to derive a data-driven transfer function transforming the brain activity to music stimuli and vice-versa. However, this process has to face several obstacles, such as the limitations of the recording equipment, interference of unrelated processes carried out by the brain, etc. Overall, MIIR research paves the way for observing, capturing, and potentially understanding and modeling the way the human brain responds, analyzes, and encodes music stimuli.

A fundamental step of MIIR research is reported in [17], where an open-access dataset in support of MIIR research was developed. In brief, the collected data are Electroencephalography (EEG) recordings taken during music perception (more details are given in Section V-A). The follow-up work [7] describes a systematic approach including a classification scheme for 3 vital MIIR tasks. It is based on a linear support vector machine classifier fed on the output of the pre-trained encoder pipeline. This article builds on these findings and investigates the inclusion of the temporal dimension in the modeling algorithm. Motivated by the possibilities offered by neuroimaging methods for MIR purposes [10], we aim at a framework able to make inferences regarding the structure and content of a musical signal via processing the respective EEG responses. More in detail, we exploit the parameters of linear time-invariant (LTI) models capturing the evolution of recorded EEG responses. Building on the multivariate Gaussian distribution of such parameters [18], we propose to model the LTI parameters based on hidden Markov models (HMMs) with each state being characterized by Gaussian mixtures. We demonstrate how such a framework can be used to model

S. Ntalampiras is with Università degli studi di Milano, Department of Computer Science, via Celoria 18, 20135, Milan, Italy, stavros.ntalampiras@unimi.it, https://sites.google.com/site/stavrosntalampiras/home. I. Potamitis is with the Technological Educational Institute of Crete, Department of Music Technology & Acoustics, potamitis@staff.teicrete.gr

music-related brain activity and address three classification tasks as defined in [7] able to assess thoroughly the relationship existing between heterogeneous music stimuli and their EEG responses. Lastly, we report encouraging classification rates highlighting the relevance of the temporal dimension.

The rest of this work is organized as follows: Section II formulates the problems designed to evaluate various aspects of the relationship between the music stimuli and the corresponding EEG responses. Section III describes the proposed framework including the feature extraction and pattern recognition modules. Moving on, section V provides a thorough analysis of the dataset, the parameterization of the proposed method, the obtained results and how they compare to the state of the art. Finally, in section VI we draw our conclusions and outline our future research goals.

II. Problem Formulation

Let us denote the music signal as m^t and the associated encephalographic response as $e^t_{i,c}$ where t is the time instant, i the subject exhibiting the specific response, and c the encephalographic channel. In this context, there are three problems of interest, each one designed to reveal different aspects of the underlying relationship between m^t and e^t in a user-independent setting [7]. These are the following:

- *stimuli-specific classification*,
- *group classification*, i.e. songs recorded with lyrics, songs recorded without lyrics, and instrumental pieces, and
- *meter classification*, i.e. 3/4 vs. 4/4 meter.

Towards assessing the relationship between musical patterns and their encephalographic responses, this work approaches the three above-mentioned classification problems. The overall aim is to provide a *common* solution with the lowest possible amount of misclassifications.

III. The proposed solution for music-related brain activity analysis

This section analyses the proposed framework able to process the EEG responses and address the classification tasks described in section II. The block diagram of the propose solution is shown in Fig. 1. Initially, the EEG responses associated with each music signal are captured. Subsequently, the mean of the acquired EEG responses is modeled through an autoregressive (AR) process. The selection of the mean statistical moment was motivated by the literature, e.g. [19], [20]; however other types of channel integration could be explored. The parameters of the AR model are modeled by hidden Markov models (HMM) addressing the characteristics of each classification problem. Finally, class prediction is achieved based on the maximum log-likelihood criterion. The following subsections describe the components of the proposed framework.

A. Autoregressive feature extraction

This section explains the method used for modeling the electroenchephalographic responses to the available music signals coming from every different class. The method is inspired by [21] and its output forms the feature vector feeding the HMM.

Let us denote by $X_i : \mathbb{N} \rightarrow \mathbb{R}$ the stream of data acquired by the c-th channel of the i-th user, i.e. $e^t_{i,c}$. In the following we assume that the temporal evolution of such datastream can be characterized by a process $\mathcal{P}$ which is time-invariant or that every class of interest can be approximated by a sequence of models even if it is time-variant (e.g. through a Markov process operating in the parameter space).

Therefore, to construct a temporal evolution model, we consider the general *discrete-time linear Single Input Single Output (SISO)* structure:

$$A(z)X_i(t) = \frac{B(z)}{F(z)}X_i(t) + \frac{C(z)}{D(z)}d(t),$$

where $d(t)$ is an independent and identically distributed random variable accounting for the noise, z is the time-shift operator while $A(z), B(z), C(z), D(z)$ and $F(z)$ represent z-transfer functions, whose parameter vectors are $\theta_A, \theta_B, \theta_C, \theta_D$ and θ_F respectively. Consequently, an element f_θ in the approximating model family $M(\theta)$ is fully described with a $\theta \in \mathbb{R}^p$ which comprises the above parameter vectors. Following the logic of [22], we create an ensemble of dynamic models with various orders and select the one which best fits the datastreams (i.e. lowest reconstruction error) while low-order models are preferred. The model search algorithm minimizes a robustified quadratic prediction error criterion.

The utilization of linear models ensures that the regularity assumptions imposed by [18], [23] are satisfied. Thus, our framework is placed on a solid mathematical background despite the introduced model bias $||f_\theta - \mathcal{P}||$ suggesting that the underlying distribution of the parameters is a multivariate Gaussian (the bias here is seen as a time-invariant ”difference” between the predicted and the true process). However various models are needed to describe a specific source of data, the number and the connections of which is not known a priori. A hidden Markov model is appropriate for dealing with this type of bias since it can break the problem into a specific number of states which are connected in a probabilistic way (see Fig. 2).

It should be stressed out that in the proposed brain activity understanding framework, LTI models are employed only as a tool explaining the temporal evolution of EEG signals. At this stage we emphasize on avoiding high complexity even if this process does not offer the highest possible prediction accuracy. The parameter vectors θ are modeled via HMMs in correspondence to a specific class of music signals. The probability density functions of their EEG responses are associated with the various classes and thus an analysis in the parameter space can be used for music understanding via the related classification problems (see Section II).

Feature values extracted out of the encephalographic responses of user no. 10 are depicted in Fig. 3. We can see that the model search process minimizing the prediction error reveals different optimal points w.r.t the responses to the available music signals. Higher feature dimensionality demonstrates greater model order, thus increased complexity in modeling existing temporal dependencies.

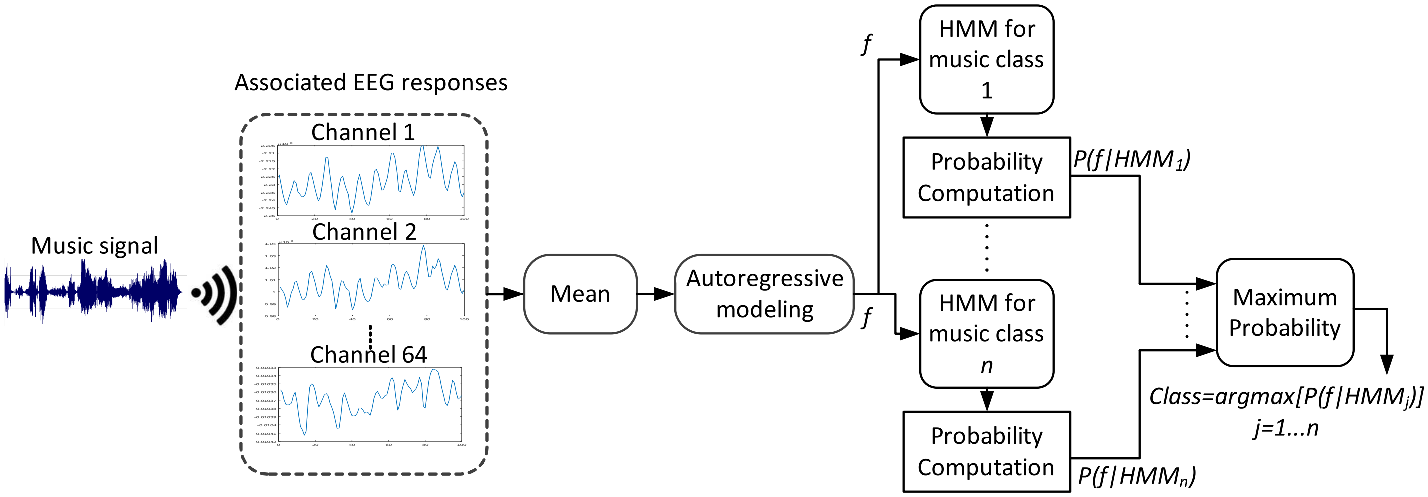

Fig. 1. The pipeline of the proposed method. The encephalographic responses of music signals are modeled by means of HMMs able to address the classication problems explained in Section II.

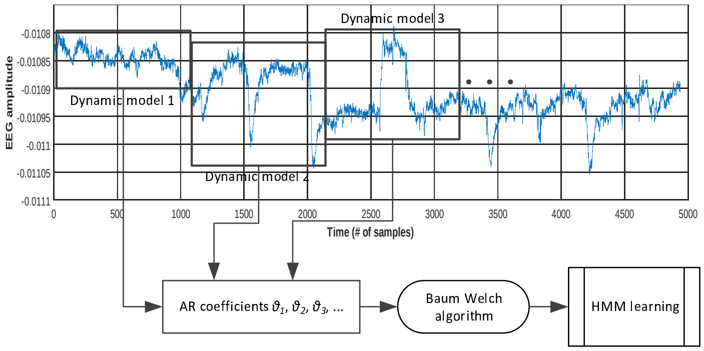

Fig. 2. The process leading to HMM learning. The EEG responses are windowized and modeled by autoregressive processes whose parameters θ are input to the Baum-Welch algorithm outputting the HMM.

B. Hidden Markov Models

Hidden Markov models constitute an extension of the discrete Markov processes while the main focus is placed on real-world problems. HMMs have been proposed in [24], where the observation is a probabilistic function of the state. The resulting model includes two stochastic processes, one of which is not observable (hidden) and can only be observed through another set of stochastic processes which produce the sequence of observations. In an HMM, the states are referred to as hidden because the system we wish to model may have underlying causes that cannot be observed.

An HMM is characterized by the following components:

- the number of states N,
- the probability density function associated with each state modelled as a mixture of Gaussians (GMM), $P(x|\theta) = \sum_{k=1}^{K} p_k p(x|\theta_{(k)})$, where $p_k s$ are the mixture weights, x is a continuous-valued data vector (e.g. measurements or features), $\theta_{(k)}$ represents the $k-th$ component of the vector, $\theta = [\sum, \mu]$, $p(x|\theta_{(k)}) = \frac{1}{(2\pi)^{d/2}|\sum_k|} e^{-\frac{1}{2}(x-\mu_k)^t \sum_k^{-1}(x-\mu_k)}$
- the state transition probability matrix $A = \{a_{ij}\}$ where entry a_{ij} represents the probability of moving from state j at time t to state i at time t+1. For example, the transition probability of moving from state 1 to state 2 is represented by a_{12}. For the case where the system may transit to any state at a given time instant, we have $a_{ij} > 0, \forall i, j$. In case some transitions are not allowed, the respective a_{ij}s should be set to zero.
- the initial state distribution $\pi = \{\bar{\pi}_i\}$, where $\bar{\pi}_i$ corresponds to the probability that the HMM starts in state i, i.e. $\pi_i = P[S_1], 1 \leq i \leq N$.

C. HMM Training

Model parameters, that is, *the transition probabilities, emission probabilities and the initial state probability* need to be adjusted so as to maximize the probability of the observed sequence and adequately represent the training set. The Baum-Welch algorithm [25] is a method that uses an iterative approach and provides a solution to this problem. It starts with preassigned probabilities and tries to adjust them based on the observed sequences in the training dataset.

The HMM parameters can be initialized to predetermined values or to a constant before applying the Baum-Welch algorithm. As the path taken is not known, the algorithm counts the number of times each component is used when the observed set of elements in the training sequence is given to the present HMM. Each iteration of the algorithm includes two steps, the Expectation step (E Step) and the Maximization step (M Step). The Maximization step uses the counts of the number of times an element is seen at a state and the number of times a transition occurs between two states which were obtained from the Expectation step to update

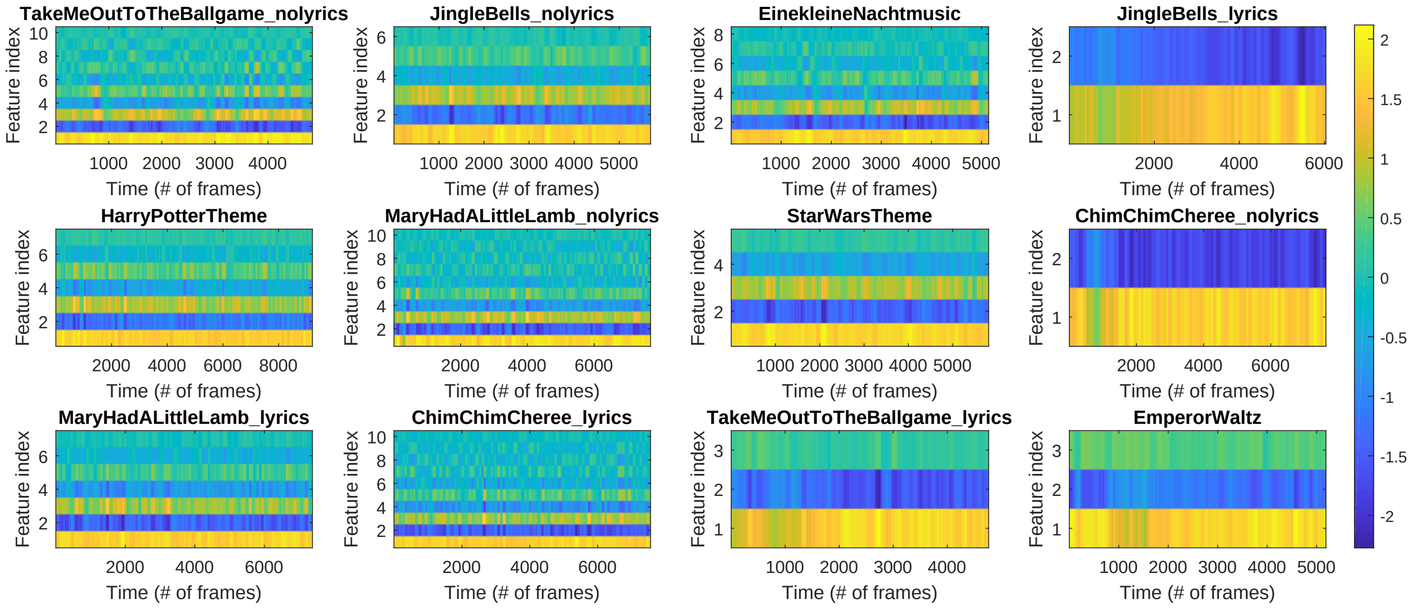

Fig. 3. Values of the proposed features describing the enchephalographic signals. During feature extraction the LTI model search process minimizes the prediction error which may lead to different model orders depending on the complexity of the available EEG responses. Higher dimensionality demonstrates need for increased model order.

the transition and emission probabilities in order to maximize the performance. The algorithm stops when the convergence criterion is satisfied (the log-likelihood between subsequent iterations is under a threshold) or when the maximum number of permitted iterations is reached.

D. Log-likelihood Computation of Unknown Data

The Viterbi algorithm is used to find the most probable path taken across the states in the HMM. The algorithm checks all possible paths leading to a state and gives the most probable based on dynamic programming. It keeps track of the best state during a transition using pointers. The most probable path is found by moving through the pointers backwards starting from the end state to the start state. In case there are more than one paths exhibiting the highest probability, a random selection is made. The Viterbi Algorithm is analytically explained in [26].

IV. The Music-related Brain Activity Pattern Classification Algorithm

Towards addressing the three problems mentioned in Section II, the training phase of the proposed methodology creates one HMM per class while the testing one examines the probability that the novel data sequence was produced by the created HMMs. Finally, the system assigns the class associated with the HMM producing the highest log-likelihood to the unknown data. Based on the specific logic we essentially try to quantify the statistical similarity between the unknown data and the one available during training. The higher the similarity with an HMM, the more probable that this data sequence belongs to the class represented by the specific HMM.

The music-related brain activity pattern classification algorithm is summarized in Algorithm 1. We assume a training set corresponding to $O_{i,T_0,1\leq i\leq N}$ associated with each music class. We compute the d model coefficients over a predefined window of the encephalographic responses of size M. They are used to train the HMM which is to characterize the specific class (line 1, Alg. 1). In order to identify the HMMs with the best classification capabilities, we build a variety of HMMs with different parameters (number of states and Gaussian components) and we select the HMM based on the highest recognition rate criterion. The set of the constructed HMM represents the set of classes in a 1-1 sense.

When unknown data is processed, it is first windowized (line 2, Alg. 1) and the model coefficients with respect to each window are computed (line 4, Alg. 1) and inserted into the trained HMM. The log-likelihood vector is then calculated for window W_j (line 5, Alg. 1) and its maximum element is discovered (line 6, Alg. 1) revealing the HMM which best ”explains” W_j. Finally the class represented by this HMM is assigned to W_j. The classification process is also demonstrated in the latter part of Fig. 1.

V. Experimental Design and Analysis of the Results

This section describes the dataset used to assess the proposed methodology, the experimental protocol, and the obtained results.

A. The dataset

A systematic attempt towards a corpus satisfying the specifications required by the current study is the OpenMIIR dataset

```
1. Build one HMM per music class,
   $H_{f_1-f_N} = \{S_{f_1-f_N}, P_{f_1-f_N}, A_{f_1-f_N}, \pi_{f_1-f_N}\}$ from
   the vectors of parameters $\theta_1 ... \theta_d$ each of which
   associated with a linear dynamic model applied to the
   training data $O_{i,T_0,i=1,...,d}$ windowized using length
   $M$ overlapping by $M-1$;
2. Windowize the incoming novel data as above, which
   results in windows $W = W_1 ... W_x$;
repeat
    3. j=1;
    4. Compute the parameter vectors of the $j-th$
       dynamic model $\theta_j$ with respect to $W_j$;
    5. Compute the vector of log-likelihoods
       $L_{W_{1...j}} = P(\theta_1 \dots \theta_j | H_{f_1 \dots f_N})$;
    6. Compute $argmax(L_{W_{1...j}})$ and assign the class
       with the highest log-likelihood to window $W_j$;
    7. $j = j + 1$;
until $(1)$;
```

Algorithm 1: The general-purpose user-independent music-related brain activity pattern classification algorithm.

[17]. It includes Electroencephalography (EEG) recordings taken during *music perception* and *imagination*[1]. Following the findings of the work reported in [7], where the poor relationship between the music and EEG signals recorded in the imagination setting, this work employs only the music perception part. The OpenMIIR dataset was employed as provided by [17] without any modification.

The OpenMIIR dataset includes response data of 10 subjects who listened to 12 music fragments with duration ranging from 7s to 16s coming from popular musical pieces. EEG modality was deemed effective as well as useful for scientists working in MIR since there exist of-the-shelf electronics able to record such responses. On top of that, acquiring functional Magnetic Resonance Imaging data is still characterized by high cost. Interestingly, EEG responses provide high temporal resolution, while they 1) reflect the way music is perceived by the subjects, and 2) enable finding and revealing potential temporal correlations existing between music signals and their EEG responses including rhythmic characteristics.

The music stimuli come from various genres while care was taken so that important music aspects (i.e. meter, tempo, presence/absence of lyrics) were covered allowing representation of heterogeneous music retrieval and classification problems.

The characteristics of the OpenMIIR dataset contents are tabulated in Table I. There exist three main groups, i.e. *Songs recorded with lyrics*, *Songs recorded without lyrics*, and *Instrumental pieces*, each one composed of 4 tracks. The average length and tempo are 10.4s and 176BPM respectively. Interestingly, the meter values are perfectly balanced across the entire dataset, which is useful while setting up a classifier addressing the specific task.

In more detail, the dataset includes:

- Music stimuli with lyrics, i.e. a singing voice,
- Music stimuli including only melody, i.e. without the singing voice. These are different recordings of the songs used in the previous point without a singing voice, and
- Music stimuli of instrumental pieces, i.e. no lyrics are present nor the possibility of singing along.

It is important to note that all recordings are of similar length (see Table I) and include complete musical phrases starting from the beginning of the piece, while they are normalized in volume. Moreover, recordings coming from the same music stimuli with and without lyrics are matched in the tempo dimension.

The OpenMIIR dataset includes EEG recordings from 10 participants (3 male, 7 female), i.e. $i \in \{1, 10\}$ (see Section II). Their ages range from 19 to 36 years without any history of hearing impairment and/or brain injury. The recordings were taken by means of a BioSemi Active-Two system with 64 EEG channels sampled at 512Hz, i.e. $c \in \{1, 64\}$ (see Section II). Eye movements are captured via Horizontal and vertical Electrooculography (EOG).

The raw EEG and EOG recordings were preprocessed as per [17] using the MNE-python toolbox [27] eliminating potential inaccurate measurements. Further anomalies were removed by manual visual inspection [7]. Subsequently, the recordings were passed through a bandpass filter keeping the frequencies between 0.5 and 30Hz and canceling slowly occurring drifts exhibited by the EEG response. On top of that, independent component analysis [28] achieved the removal of artifacts caused by eye blinks. Having the artifacts removed, all 64 EEG channels were reconstructed to the same dimensionality. The last preprocessing stage includes signal normalization to zero mean and [-1, 1] range.

For more information on the dataset, the interested reader is referred to [17].

B. Parameterization of the proposed solution

The HMMs have been configured in a fully connected topology (ergodic HMM), which means that the algorithm permits every possible transition across states. This approach was followed since EEG data do not always follow a consistent pattern which may indicate a specific state ordering, e.g. a left-right topology (where cyclic transitions in the automaton are not allowed). Lastly, the distribution of each state is modeled by a GMM with a diagonal covariance matrix.

We employed the Torch machine learning framework [29] during both learning and validation phases. The maximum number of *k*-means iterations for cluster initialization was set to 50 while the Baum-Welch algorithm used to estimate the transition matrix was bounded to 100 iterations with a threshold of 0.001 between subsequent iterations. The number of explored states ranges from 3 to 7 while the number of Gaussian components used to build each GMM comes from the {2, 4, 8, 16, 32, 64, 128, 256 and 512} set. Parameter optimization is performed on a validation set which is part of the training one. Standard normalization techniques, i.e. mean removal and variance scaling were applied. Finally, the window length M was 100, a value which provided satisfactory reconstruction error during the preliminary experimental phase.

[1] The dataset is publicly available at https://openmiir.github.io

TABLE I. CHARACTERISTICS OF THE CONTENTS OF THE OPENMIIR DATASET. THE AVERAGE LENGTH AND TEMPO ARE 10.4S AND 176BPM RESPECTIVELY. FOR MORE INFORMATION ON THE DATASET AND THE COMPLETE TABLE, THE INTERESTED READER IS REFERRED TO [17].

Group	Piece (id)	Meter	Length	Temp (BPM)
Songs recorded with lyrics	Chim Chim Cheree (1)	3/4	13.3s	212
	Take Me Out to the Ballgame (2)	3/4	7.7s	189
	Jingle Bells, lyrics (3)	4/4	9.7s	200
	Mary Had a Little Lamb (4)	4/4	11.6s	160
Songs recorded without lyrics	Chim Chim Cheree (11)	3/4	13.5s	212
	Take Me Out to the Ballgame (12)	3/4	7.7s	189
	Jingle Bells (13)	4/4	9s	200
	Mary Had a Little Lamb (14)	4/4	12.2s	160
Instrumental pieces	Emperor Waltz (21)	3/4	8.3s	178
	Hedwigs Theme, Harry Potter (22)	3/4	16s	166
	Imperial March, Star Wars Theme (23)	4/4	9.2s	104
	Eine Kleine Nachtmusik (24)	4/4	6.9s	140

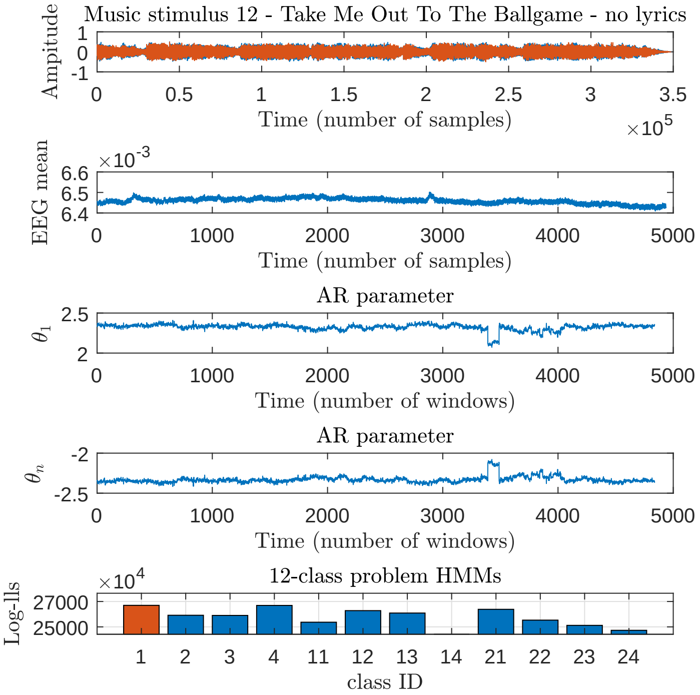

Fig. 4. An illustrative example of the operation of the proposed method on the 12-class stimuli identification problem. Starting from the audio waveform, the respective EEG responses are obtained. The AR model providing the lowest reconstruction error is identified and its parameters are modeled by means of an HMM. Finally, the class of the novel EEG responses is discovered based on the maximum log-likelihood criterion (in this example, the winning class with ID 1 is depicted in red).

Moving to autoregressive modeling, their orders are selected by exhaustive search based on the lowest mean squared error criterion. Early experimentations showed that search space [1,10] is sufficient to identify clear minima in the MSE.

C. Contrasted approach

To the best of our knowledge, there is only one work addressing the present problem as MIIR is a relatively new scientific field. In [7], the author describes a classification scheme based on a linear support vector machine classifier fed on the output of the pre-trained encoder pipeline.

The present work employs a 10-fold cross-validation scheme across subjects, i.e. training on 9 and testing on the 10th subject. Care was taken so that each one of the 600 trials is included in the testing set, while obtaining subject-independent classification results.

D. Results

This subsection includes a characteristic example of the operation of the proposed methodology, the results achieved in the classification tasks explained in section II, as well as their analysis.

Before presenting and analyzing the recognition rates achieved by the proposed scheme, Fig. 4 provides an illustrative example of its operation regarding the 12-class classification problem. On the top row, we can see both channels of the acoustic signal belonging to the music stimuli no. 12 entitled *Take Me Out To The Ballgame* without lyrics. In the following, we observe the mean of the respective EEG responses. The subsequent subplots demonstrate the evolution of AR model parameters $\{\theta_1, \ldots, \theta_n\}$. Finally, during the classification stage each class-specific HMM outputs a log-likelihood each one demonstrating the probability that the specific series of parameters was produced by each HMM. Classification is achieved via the maximum log-likelihood criterion [30], [31]. The specific example was categorized to class ID 1 as the corresponding HMM emitted the highest log-likelihood.

a) Task 1: Music Stimulus Identification: Task 1 is concerned with identification of each stimulus used in the current study, i.e. it comprises a 12-class problem. Despite the limited dataset, such a task is able to provide an indication of the efficacy of the proposed classification methodology as well as insights on the connection existing between each acoustic stimuli and the respective EEG response.

The results are tabulated in two tables, i.e. Table II includes the average classification rates for all test folds, i.e. for data coming from each subject and Table III demonstrates the confusion matrix averaged across all test folds/subjects.

In Table II, we can see that the best rate (50.8%) is achieved for subject 2 and the worst (38.3%) for subject 4, while the overall average rate is 42.7%. This table essentially shows how

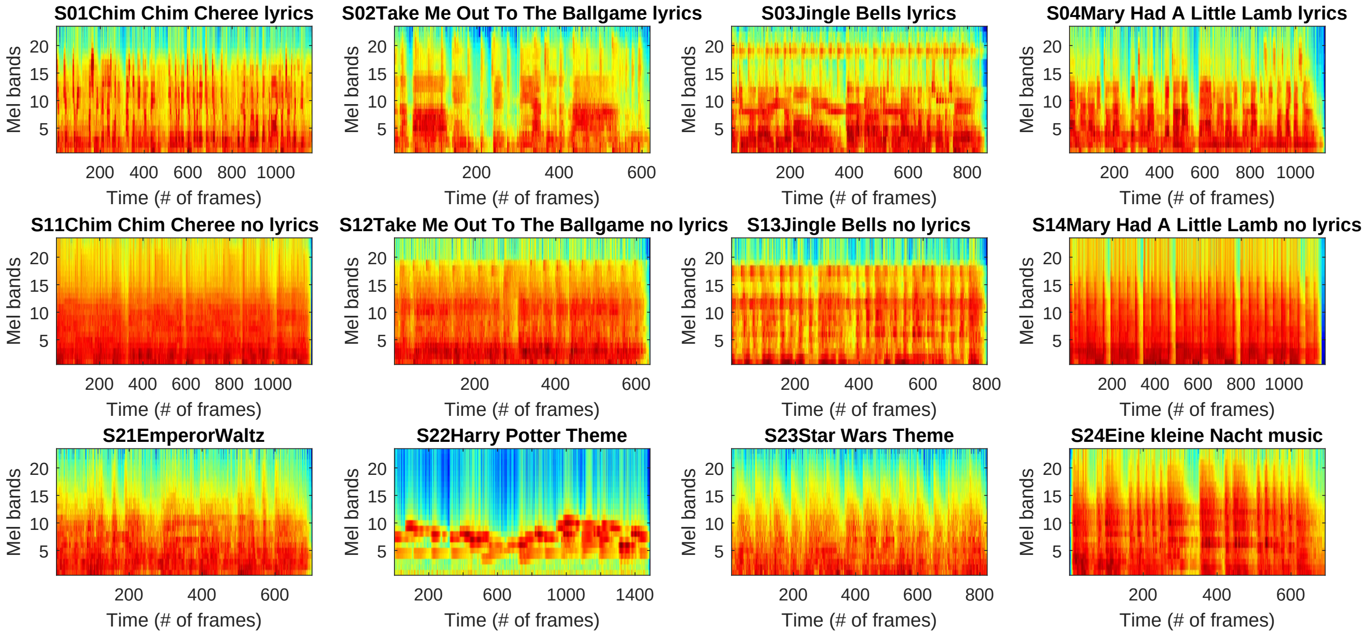

Fig. 5. The Mel-scaled spectrograms of the music stimuli employed in the current study. The first line includes songs with lyrics, the second songs without lyrics, and the third one instrumental pieces.

TABLE II. THE CLASSIFICATION RATE (IN %) FOR EACH FOLD, WHILE DATA COMING FROM EACH USER COMPRISE PART OF THE TESTING SET.

Problem	**Subject 1**	**Subject 2**	**Subject 3**	**Subject 4**	**Subject 5**	**Subject 6**	**Subject 7**	**Subject 8**	**Subject 9**	**Subject 10**	**Average** $\pm\sigma$**)**
Task 1	45.8	50.8	40.3	38.3	40.2	40.8	39.4	41.9	44.7	45.2	42.7±3.8
Task 2	49	48.7	49.6	49	49	50.2	49.3	50.8	50.2	50.2	49.6±0.7
Task 3	62.4	71.5	66.7	69.2	69.5	68.8	72.8	68.7	70	67.4	68.7±2.8

TABLE III. THE CONFUSION MATRIX OF TASK 1. THE HIGHEST CLASSIFICATION RATES PER CLASS ARE EMBOLDENED. THE AVERAGE CLASSIFICATION RATE IS 42.7%. HMMS COMPOSED OF 5 STATES AND 64 GAUSSIAN FUNCTIONS PROVIDED THE HIGHEST RECOGNITION RATES.

Responded **Presented**	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Chim Chim Cheree (lyrics) 1	**34.1**	3.9	6.2	1.8	2	6	4	6.1	5.9	18.3	2	9.7
Take Me Out to the Ball game (lyrics) 2	2	**44**	2	1.5	2.5	6.1	3.9	2	10.3	5.7	6.2	13.8
Jingle Bells (lyrics) 3	2	10	**40.2**	1.8	2	9.5	2	6	6.5	7.7	-	12.3
Mary Had a Little Lamb (lyrics) 4	-	2	2	**43.7**	6.4	6.1	1.5	2	6	12.4	-	17.6
Chim Chim Cheree 5	2	12	4	2	**24**	10	2.5	7.5	4.1	14	1.9	16
Take Me Out to the Ball game 6	4	2	4	-	2	**52.5**	5.5	4.9	4	9.1	-	12
Jingle Bells 7	-	-	6	-	-	17.2	**49.5**	4	3.7	4	2	13.6
Mary Had a Little Lamb 8	-	8	2	4	2	8	2	**46**	2.1	11.9	2.9	11.1
Emperor Waltz 9	2	5.9	2.1	-	-	8.6	10	5.4	**50.8**	7.4	5.6	2
Hedwigs Theme (Harry Potter) 10	2	9.4	2.6	4	-	10.1	2	9.9	6	**42**	2.5	9.5
Imperial March (Star Wars Theme) 11	-	1.1	-	2	4.4	21.6	6	8.1	3.9	2	**36**	14
Eine Kleine Nachtmusik 12	4.1	5.9	4	4.1	3.9	4	7.7	8.3	2.5	3.5	2	**49.2**

well data concerning the responses of different subjects match those of another and that is the source of variance among the subjects. Subject-depended rates are kept in relatively high values considering the task difficulty (limited dataset, 12-class problem and inference on a different modality), while they are significantly above chance. This fact is encouraging and shows that there are common patterns in the way that music is perceived by different subjects. The specific experiment highlighted the importance of the temporal dimension as the achieved classification rate is higher than the state of the art [7] (27.6%) where a solution which does explicitly look for temporal patterns is employed.

Moving to the confusion matrix (Table III), we see that the best rate (52.5%) is achieved for Take Me Out to the Ball game (song recorded without lyrics) and the worst one (24%) for Chim Chim Cheree (song recorded without lyrics). Furthermore, the main diagonal includes the highest rates for all classes. Unlike the results reported in [7], stimuli 1-4 are not misclassified with their corresponding tempo-matched versions without lyrics (stimuli 11-14). This may be due to the consideration of temporal patterns which are very evident in these signals, especially when speech is included in the

TABLE IV. CLASSIFICATION RESULTS OF THE APPROACH EMPLOYING SOLELY THE MUSIC SIGNALS.

Task - ID	**Classification rate (%)**
Music stimuli identification - 1	99.8±0.1
Group classification - 2	99.7±0.07
Meter classification - 3	99.8±0.04

stimuli [32]. Moreover, it suggests that the EEG signals exhibit differences based on the music stimuli, and more importantly, these are captured by the proposed classification scheme indicating insignificant chances of horse classifier behavior [33], where classification might use signal characteristics unrelated to music information. More in detail, the classification rates demonstrate that there is indeed a consistent pattern characterizing the underlying relationship between the audio stimuli and the associated response.

In Table III, the average classification rates according to group, i.e. music stimuli 1-4, 5-8 and 9-12, are 40.5%, 43%, and 44.5% respectively. Interestingly, instrumental pieces are generally recognized with higher accuracy than the rest showing that their temporal patterns are captured by HMMs trained on the parameters of linear time-invariant models are quite distinctive.

An interesting complementary experiment was conducted in [7] where 8 subjects required on average 1-3s to recognize the individual music stimuli. Following the same line of thought, in this work we construct automatic classification systems using the music signals for addressing all three tasks described in Section II. This series of experiments assesses the level up to which the music signals are distinguishable by an approach following the same line of thought w.r.t the one elaborating on the EEG responses. In brief, the classifier is based on the Mel-frequency cepstral coefficients along with their dynamics (velocity and acceleration), while each probability density function is estimated by HMMs [30]. The first half of each song was used for training, while testing was conducted on the remaining part of the song. Table IV shows the respective classification rates. As we can see, in all tasks there are clearly recognizable patterns as excellent recognition rates were achieved. Ultimately, the aim of MIIR would be to achieve similar rates while based on EEG responses to musical stimuli. Nonetheless, towards understanding the feasibility of such a logic, it is crucial to augment the current dataset. Music-only classification is successful as it is applied directly on the signal of interest. However, when trying to perform the same classification task via data coming from a different modality, larger data quantity is needed to capture and reveal the relationships existing between the two different modalities. The specific task could benefit from the transfer learning technology [34], a direction we aim to follow in our future work.

b) Task 2: Group classification: The second classification task concerns the following 3 classes: *a)* music stimuli including lyrics, *b)* music stimuli excluding lyrics, and *c)* instrumental pieces.

TABLE V. CONFUSION MATRIX ACHIEVED BY THE PROPOSED APPROACH AS REGARDS TO THE GROUP CLASSIFICATION TASK. THE HIGHEST RATES ARE EMBOLDENED. THE AVERAGE CLASSIFICATION RATE IS 49.6%. HMMS COMPOSED OF 7 STATES AND 16 GAUSSIAN FUNCTIONS PROVIDED THE HIGHEST RECOGNITION RATES.

Responded **Presented**	*1*	*2*	*3*
Songs recorded with lyrics 1	**52.3**	20.7	27
Songs recorded without lyrics 2	35	**42.7**	22.3
Instrumental pieces 3	22.8	23.2	**54**

Tables II and V include the classification rates achieved w.r.t each subject and the confusion matrix respectively. As we can see the lowest rate is achieved in correspondence to data coming from subject 2 (48.7%) and the highest one to data coming from subject 8 (50.8%). The average rate is 49.6% which is very close to the one achieved in [7], i.e. 48.9%, and not the expected one after the rate achieved in the previous 12-class problem. This may indicate that the parameters of LTI models are relevant, but more data is needed to identify and capture consistent temporal patterns in such a classification task. Moreover, the problem is of increased complexity as HMMs with more states (7 for Task 2 vs. 5 for Task 1) are needed to model the respective probability density functions. Nonetheless, the rate is again significantly above chance. The highest one is achieved for the instrumental pieces (54%) and lowest for the music stimuli without lyrics (43%).

Interestingly, the rates included in the confusion matrix are in line with the ones presented in [7] suggesting that the classifiers conducted similar errors. Even so, the two approaches could be thought as complementary (a generative and a discriminative classification scheme), thus in the future it would be interesting to set-up a synergistic framework benefiting from both solutions. Last but not least, the music-only classifier performs similarly to the one described in the first task.

c) Task 3: Meter Classification: The specific task is concerned with classifying music stimuli in 3/4 meter v.s those in 4/4 meter. The corresponding results are shown in Tables II and VI. The mean rate is 68.7% while stimuli in 4/4 is identified slightly better (0.9% difference). Similar results are reported in [7] with the corresponding rate being 69.4%. However, having to deal with a 2-class problem one would expect a higher rate. This may suggest that the proposed method alone is not adequate to identify distinctive patterns, thus the above-mentioned synergistic framework could be a an interesting path to follow. At the same time, augmenting the current dataset could give useful insights towards modeling and recognizing such classes effectively.

In this case, HMMs composed of 7 states and 16 Gaussian functions provided the highest recognition rates. Lastly, the classifier based on the acoustic modality reports excellent recognition rates (see Table IV).

As a final remark, it is worth noting that the inferred transition matrices w.r.t all tasks exhibit higher values across their diagonal suggesting that such a connecting topology is able to explain the features' evolution. As an example, Fig. 6 depicts the transition matrices associated with the group classification task, where the diagonal behavior is evident.

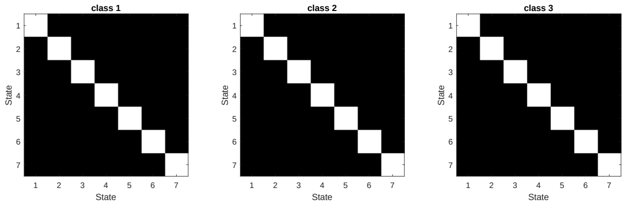

Fig. 6. The HMM transition matrices w.r.t the group classification task. We see that high probabilities are associated with transitions corresponding to the diagonal of each matrix. Such probabilities favor same-state transitions.

VI. Conclusions

This paper comprises one of the first attempts systematically exploring the existence of connections between music stimuli and their EEG responses, while considering the temporal evolution of the latter. We showed that parameters of LTI models can capture the dynamics of such signals, and when HMMs learn their evolution in time, a successful inference framework can be structured. Importantly, such a classifier achieved state of the art results in three classification tasks designed to investigate heterogeneous aspects of the underlying relationship between music stimuli and their EEG responses. Interestingly, the proposed solution is a comprehensible classification scheme, since its operation does not follow the black-box logic, while one is able to 'open' the classifier, and by inspecting the misclassifications, understand the reasons leading to the specific errors. Overall, we observed that a higher number of HMM states is needed for modeling tasks associated with *group* and *meter* classification. This suggests a higher complexity probably caused by organizing data based on underlying cognitive characteristics which are harder to capture and model. In the future, we intent to apply and evaluate the proposed solution on the imagination part of the dataset and compare the achieved results.

Progress in the new and exciting scientific field of MIIR is heavily relying onto bringing together researchers from MIR and music cognition areas. A close interdisciplinary synergy is required to grasp and identify the relationship existing between brain activity and the music stimuli causing it. This paper is a primary step towards this direction and we hope that it will encourage scientists working in the MIR domain to apply their methods and explore the possibilities offered by this emerging research field.

Acknowledgement

The authors would like to thank the creators of the OpenMIIR dataset for making it publicly available.

TABLE VI. Confusion matrix achieved by the proposed approach as regards to the meter classification task. The highest rates are emboldened. The average classification rate is 68.7%. HMMs composed of 6 states and 2 Gaussian functions provided the highest recognition rates.

Responded **Presented**	*3/4*	*4/4*
3/4	**68.2**	31.8
4/4	30.9	**69.1**

References

[1] M. Soleymani, M. N. Caro, E. M. Schmidt, C.-Y. Sha, and Y.-H. Yang, "1000 songs for emotional analysis of music," in *Proceedings of the 2Nd ACM International Workshop on Crowdsourcing for Multimedia*, ser. CrowdMM '13. New York, NY, USA: ACM, 2013, pp. 1–6. [Online]. Available: http://doi.acm.org/10.1145/2506364.2506365

[2] K. Markov and T. Matsui, "Music genre and emotion recognition using gaussian processes," *IEEE Access*, vol. 2, pp. 688–697, 2014.

[3] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, "On the acoustics of emotion in audio: What speech, music and sound have in common," *Frontiers in Psychology*, vol. 4, no. 292, 2013. [Online]. Available: http://www.frontiersin.org/emotion_science/10.3389/fpsyg.2013.00292/abstract

[4] Y.-H. Yang and H. H. Chen, "Machine recognition of music emotion: A review," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 3, pp. 40:1–40:30, May 2012. [Online]. Available: http://doi.acm.org/10.1145/2168752.2168754

[5] A. Klapuri, "Multipitch analysis of polyphonic music and speech signals using an auditory model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 255–266, Feb 2008.

[6] M. Quick, *Can data reveal the saddest number one song ever?*, 2018 (accessed Sept. 2, 2018), http://www.bbc.com/culture/story/20180821-can-data-reveal-the-saddest-song-ever.

[7] S. Stober, "Toward studying music cognition with information retrieval techniques: Lessons learned from the OpenMIIR initiative," *Frontiers in Psychology*, vol. 8, aug 2017. [Online]. Available: https://doi.org/10.3389/fpsyg.2017.01255

[8] A. S. Cowen, M. M. Chun, and B. A. Kuhl, "Neural portraits of perception: Reconstructing face images from evoked brain activity," *NeuroImage*, vol. 94, pp. 12–22, jul 2014. [Online]. Available: https://doi.org/10.1016/j.neuroimage.2014.03.018

[9] T. Horikawa, M. Tamaki, Y. Miyawaki, and Y. Kamitani, "Neural decoding of visual imagery during sleep," *Science*, vol. 340, no. 6132, pp. 639–642, apr 2013. [Online]. Available: https://doi.org/10.1126/science.1234330

[10] B. Kaneshiro and J. Dmochowski, "Neuroimaging methods for music information retrieval: Current findings and future prospects," in *ISMIR*, 2015.

[11] N. Sankaran, W. F. Thompson, S. Carlile, and T. A. Carlson, "Decoding the dynamic representation of musical pitch from human brain activity," *Scientific Reports*, vol. 8, no. 1, jan 2018. [Online]. Available: https://doi.org/10.1038/s41598-018-19222-3

[12] C. Foster, D. Dharmaretnam, H. Xu, A. Fyshe, and G. Tzanetakis, "Decoding music in the human brain using EEG data," in *2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)*, Aug 2018, pp. 1–6.

[13] A. Pinegger, S. C. Wriessnegger, and G. R. Mller-Putz, "Sheet music by mind: Towards a brain-computer interface for composing," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Aug 2015, pp. 1053–1056.

[14] B. Arora, T. Choudhury, P. Kumar, and Mukherjee, "An intelligent way to play music by brain activity using brain computer interface," in *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, Oct 2016, pp. 223–228.

[15] Z. Li and G. Xuhong, "EEG control of music player," in *2012 Fifth International Conference on Intelligent Networks and Intelligent Systems*, Nov 2012, pp. 189–192.

[16] B. Hamadicharef, M. Xu, and S. Aditya, "Brain-computer interface based musical composition," in *2010 International Conference on Cyberworlds*, Oct 2010, pp. 282–286.

[17] S. Stober, A. Sternin, A. M. Owen, and J. A. Grahn, "Towards music imagery information retrieval: Introducing the openmiir dataset of eeg recordings from music perception and imagination," in *ISMIR*, 2015.

[18] L. Ljung, “Convergence analysis of parametric identification methods,” *Automatic Control, IEEE Transactions on*, vol. 23, no. 5, pp. 770 – 783, oct 1978.

[19] E. P. Lana, B. V. Adorno, and C. J. Tierra-Criollo, “Detection of movement intention using EEG in a human-robot interaction environment,” *Research on Biomedical Engineering*, vol. 31, no. 4, pp. 285–294, nov 2015. [Online]. Available: https://doi.org/10.1590/2446-4740.0777

[20] N. P. L. . S. Nunez, M. D., *Electroencephalography (EEG): neurophysics, experimental methods, and signal processing.* In Ombao, H., Linquist, M.,Thompson, W. and Aston, J. (Eds.)Handbook of Neuroimaging Data Analysis (pp. 175-197), Chapman and Hall/CRC.Advance online publication, 2016.

[21] C. Alippi, S. Ntalampiras, and M. Roveri, “An HMM-based change detection method for intelligent embedded sensors,” 2012.

[22] P. P. Bonissone, F. Xue, and R. Subbu, “Fast meta-models for local fusion of multiple predictive models,” *Appl. Soft Comput.*, vol. 11, no. 2, pp. 1529–1539, Mar. 2011.

[23] L. Ljung and P. E. Caines, “Asymptotic normality of prediction error estimators for approximate system models,” in *Decision and Control including the 17th Symposium on Adaptive Processes, 1978 IEEE Conference on*, vol. 17, jan. 1978, pp. 927 –932.

[24] L. R. Rabiner, “A tutorial on hidden markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, pp. 257–286, 1989.

[25] L. R. Rabiner and B. H. Juang, “An introduction to hidden markov models,” *IEEE ASSP Magazine*, pp. 4–15, January 1986.

[26] E. S. R. K. A. Durbin, R. and G. J. Mitchison, “Biological sequence analysis: Probabilistic models of proteins and nucleic acids,” *Cambridge University Press, London*, 1998.

[27] A. Gramfort, M. Luessi, E. Larson, D. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, and M. Hmlinen, “MEG and EEG data analysis with MNE-python,” *Frontiers in Neuroscience*, vol. 7, p. 267, 2013. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnins.2013.00267

[28] T.-W. Lee, M. A. Girolami, and T. J. Sejnowski, “Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources,” *Neural Computation*, vol. 11, pp. 417–441, 1999.

[29] Torch machine learning library. [Online]. Available: http://torch.ch/

[30] S. Ntalampiras, “A novel holistic modeling approach for generalized sound recognition,” *IEEE Signal Processing Letters*, vol. 20, no. 2, pp. 185–188, Feb 2013.

[31] ——, “A classification scheme based on directed acyclic graphs for acoustic farm monitoring,” in *Proceedings of the 23rd Conference of Open Innovations Association FRUCT*, ser. FRUCT’23. Helsinki, Finland, Finland: FRUCT Oy, 2018, pp. 37:276–37:282. [Online]. Available: http://dl.acm.org/citation.cfm?id=3299905.3299942

[32] P. M. A. Nambi, Y. Mahajan, N. Francis, and J. S. Bhat, “Temporal fine structure mediated recognition of speech in the presence of multitalker babble,” *The Journal of the Acoustical Society of America*, vol. 140, no. 4, pp. EL296–EL301, oct 2016. [Online]. Available: https://doi.org/10.1121/1.4964416

[33] B. L. Sturm, “A simple method to determine if a music information retrieval system is a horse,” *IEEE Transactions on Multimedia*, vol. 16, no. 6, pp. 1636–1644, Oct 2014.

[34] S. Ntalampiras, “A transfer learning framework for predicting the emotional content of generalized sound events,” *The Journal of the Acoustical Society of America*, vol. 141, no. 3, pp. 1694–1701, 2017. [Online]. Available: https://doi.org/10.1121/1.4977749

Stavros Ntalampiras is an Assistant Professor at the Department of Computer Science, University of Milan, Italy. He received the engineering and Ph.D. degrees from the Department of Electrical and Computer Engineering, University of Patras, Greece, in 2006 and 2010, respectively. He has carried out research and/or didactic activities at Politecnico di Milano, the Joint Research Center of the European Commission, the National Research Council of Italy, and Bocconi University. Currently, he is an Associate Editor of IEEE Access and PLOS One journals, as well as member of the IEEE Computational Intelligent Society Task Force on Computational Audio Processing. His research interests include content-based signal processing, audio pattern recognition, machine learning, and cyber-physical systems.

Ilyas Potamitis received the B.Eng. and Dr.Eng. degrees in electrical engineering from the Department of Electrical and Computer Engineering. University of Patras, Patras, Greece, in 1995 and 2002, respectively. He is a Full Professor at the Department of Music Technology and Acoustics of the Technological Educational Institute of Crete, Greece. His research interests are audio signal processing, biacoustics, speech enhancement, speech feature extraction for robust speech recognition, and one-channel signal separation.