Does the Beat go on?

Identifying Rhythms from Brain Waves Recorded after Their Auditory Presentation

Sebastian Stober, Daniel J. Cameron and Jessica A. Grahn
Brain and Mind Institute, Department of Psychology
Western University
London, Ontario, Canada, N6A 5B7
{sstober,dcamer25,jgrahn}@uwo.ca

ABSTRACT

Music imagery information retrieval (MIIR) systems may one day be able to recognize a song just as we think of it. As one step towards such technology, we investigate whether rhythms can be identified from an electroencephalography (EEG) recording taken directly after their auditory presentation. The EEG data has been collected during a rhythm perception study in Kigali, Rwanda and comprises 12 East African and 12 Western rhythmic stimuli presented to 13 participants. Each stimulus was presented as a loop for 32 seconds followed by a break of four seconds before the next one started. Using convolutional neural networks (CNNs), we are able to recognize individual rhythms with a mean accuracy of 22.9% over all subjects by just looking at the EEG recorded during the silence between the stimuli.

Categories and Subject Descriptors

I.2.6 [**Computing Methodologies**]: Pattern Recognition—*Models – Neural nets*; J.4 [**Computer Applications**]: Social and behavioral sciences—*Psychology*

General Terms

Algorithms, Experimentation

Keywords

Electroencephalography, Music Information Retrieval, Deep Learning, Convolutional Neural Networks

1. INTRODUCTION

Will it one day be possible to just think of a song and the music player will start its playback? As we have already motivated in [22], music imagery information retrieval (MIIR) – i.e., retrieving music by imagination – has the potential to overcome the expressivity bottleneck of current music information retrieval (MIR) systems, which require their users to somehow imitate the desired song through singing, humming, or beat-boxing [24] or to describe it using tags, meta-data, or lyrics fragments. Furthermore, music imagery appears to be a very promising means for driving brain-computer interfaces (BCIs) as proposed by Schaefer et al. [17]. They argue that *"music is especially suitable to use here as (externally or internally generated) stimulus material, since it unfolds over time, and electroencephalography (EEG) is especially precise in measuring the timing of a response."* This allows us to exploit temporal characteristics of the signal such as rhythmic information.

Therefore, as a first step towards MIIR technology, we analyze EEG recordings of rhythm perception collected during a rhythm perception study in Kigali, Rwanda. In this study, we tested 13 adults, mean age 21, who performed three behavioral tasks using rhythmic tone sequences derived from either East African or Western music. For the EEG testing, 24 rhythms – half East African and half Western with identical tempo and based on a 2-bar 12/8 scheme – were each repeated for 32 seconds with a 4s break in between. During presentation, the participants' brain waves were recorded via 14 EEG channels. In a previous experiment with this data [21], we found that by using deep learning techniques – specifically stacked denoising autoencoders (SDAs) and convolutional neural networks (CNNs) – the rhythm type could be classified with an overall average accuracy of 72.2% by looking at 2s-segments from a single channel. For some subjects, the accuracy was even higher than 90%. Moreover, using the same classifier configuration, we were able to recognize the correct rhythm out of the 24 candidates in a preliminary test with an accuracy of over 20% for one specific subject.

In this paper, we elaborate and improve our original classification approach. However, this time we are not attempting to classify the EEG data recorded *during* the rhythm perception trials but *between* them. In the following, we review related work in Section 2. Section 3 covers the data acquisition and pre-processing. In Section 4, we describe our experiments. The obtained results are presented in Section 5. Finally, we conclude and point out further steps in Section 6.

2. RELATED WORK

How the brain responses to auditory rhythms has already been investigated in several studies using EEG and magnoencephalography (MEG): Oscillatory neural activity in the gamma (20-60 Hz) frequency band is sensitive to accented tones in a rhythmic sequence and anticipates isochronous tones [20]. Oscillations in the beta (20-30 Hz) band increase in anticipation of strong tones in a non-isochronous sequence [10, 5, 6]. Another approach has measured the magnitude of steady state evoked potentials (SSEPs) (reflecting neural oscillations entrained to the stimulus) while listening to

AM '14, October 01 – 03 2014, Aalborg, Denmark

ACM 978-1-4503-3032-9/14/10 ...$15.00.
http://dx.doi.org/10.1145/2636879.2636904.

rhythmic sequences [15, 16]. Here, enhancement of SSEPs was found for frequencies related to the metrical structure of the rhythm (e.g., the frequency of the beat). In contrast to these studies investigating the oscillatory activity in the brain, other studies have used EEG to investigate event-related potentials (ERPs) in responses to tones occurring in rhythmic sequences. This approach has been used to show distinct sensitivity to perturbations of the rhythmic pattern vs. the metrical structure in rhythmic sequences [7], and to suggest that similar responses persist even when attention is diverted away from the rhythmic stimulus [12]. Further, Will and Berg [26] observed a significant increase in brain wave synchronization after periodic auditory stimulation with drum sounds and clicks with repetition rates of 1–8Hz.

Furthermore, there is also strong evidence that perception and imagination of music share common processes in the brain. As Hubbard concludes in his recent review of the literature on auditory imagery, *"auditory imagery preserves many structural and temporal properties of auditory stimuli"* and *"involves many of the same brain areas as auditory perception."* This is also underlined by Schaefer [17, p. 142] whose *"most important conclusion is that there is a substantial amount of overlap between the two tasks* [music perception and imagery]*, and that 'internally' creating a perceptual experience uses functionalities of 'normal' perception."* Thus, brain signals recorded while listening to a music piece could serve as reference data for a retrieval system in order to detect salient elements in the signal that could be expected during imagination as well.

For imagined rhythms, a method for detecting temporal patterns has already been sketched in [3]. Encouraging preliminary results for classifying purely imagined music fragments were reported in [18] where four out of eight participants produced imagery that was classifiable (in a binary comparison) with an accuracy between 70% and 90% after 11 trials. Vlek et al. [25] further showed that imagined auditory accents can be recognized from EEG. They asked ten subjects to listen to and later imagine three simple metric patterns of two, three and four beats on top of a steady metronome click. Using logistic regression to classify accented versus unaccented beats, they obtained an average single-trial accuracy of 70% for perception and 61% for imagery. These results are very encouraging to further investigate the possibilities for retrieving information about the perceived rhythm from EEG recordings.

In the field of deep learning, there has been a recent increase of works involving music data. However, to our knowledge, no prior work has been published yet on using deep learning to analyze EEG recordings related to music perception and cognition. However, there are some first attempts to process EEG recordings with deep learning techniques. Wulsin et al. [27] used deep belief nets (DBNs) to detect anomalies related to epilepsy in EEG recordings of 11 subjects by classifying individual "channel-seconds", i.e., one-second chunks from a single EEG channel without further information from other channels or about prior values. Their classifier was first pre-trained layer by layer as an autoencoder on unlabelled data, followed by a supervised fine-tuning with backpropagation on a much smaller labeled data set. They found that working on raw, unprocessed data (sampled at 256Hz) led to a classification accuracy comparable to hand-crafted features. Langkvist et al. [13] similarly employed DBNs combined with a hidden Markov model (HMM) to classify different sleep stages. Their data for 25 subjects comprises EEG as well as recordings of eye movements and skeletal muscle activity. Again, the data was segmented into one-second chunks. Here, a DBN on raw data showed a classification accuracy close to one using 28 selected features.

Table 1: Rhythmic sequences in groups of three that pairings were based on. All 'x's denote onsets. Larger, bold 'X's denote the beginning of a 12 unit cycle (downbeat).

Western Rhythms																								
1	**X**	x	x	x		x	x		x	x			**X**	x	x	x		x	x		x	x		
2	**X**			x		x	x			x		x	**X**			x		x	x			x		x
3	**X**		x	x		x	x		x	x	x	x	**X**		x	x		x	x		x	x	x	x
4	**X**		x	x		x	x			x		x	**X**		x	x		x	x			x		x
5	**X**	x	x	x			x	x		x			**X**	x	x	x			x	x		x		
6	**X**		x	x		x	x		x	x	x	x	**X**		x	x		x	x		x	x	x	x
East African Rhythms																								
1	**X**		x	x	x	x	x		x	x	x	x	**X**		x	x	x	x	x		x	x	x	x
2	**X**		x		x		x			x		x	**X**				x		x		x			
3	**X**		x				x		x				**X**		x		x		x					
4	**X**		x	x	x		x	x	x		x	x	**X**		x	x	x		x	x	x		x	x
5	**X**		x	x		x	x		x	x		x	**X**		x	x		x	x		x	x		x
6	**X**		x	x		x	x		x		x		**X**		x	x		x	x		x		x	

3. DATA SET & PRE-PROCESSING

3.1 Stimuli

The African rhythm stimuli were derived from recordings of traditional East African music [1]. The author (DC) composed the Western rhythmic stimuli. Rhythms were presented as sequences of sine tones that were 100ms in duration with intensity ramped up/down over the first/final 50ms and a pitch of either 375 or 500 Hz. All rhythms had a temporal structure of 12 equal units, in which each unit could contain a sound or not. For each rhythmic stimulus, two individual rhythmic sequences were overlaid whereby one sequence was played at the high pitch and the other at the low pitch. There were two groups of three individual rhythmic sequences for each cultural type of rhythm as shown in Table 1. With three combinations within each group and two possible pitch assignments, this resulted in six rhythmic stimuli for each group, 12 per rhythm type and 24 in total. Finally, rhythmic stimuli could be played back at one of two tempi: having a minimum inter-onset interval of 180 or 240ms.

We used sequences of sine tones (as opposed to real music or musical instrument sounds) in order to avoid any cultural familiarity of aspects of the stimuli besides the strictly temporal structure of the rhythms. The frequencies of the sine tones were selected to be in the normal range of music and with a pitch interval (a perfect fourth, 375 and 500 Hz) common to both Western and East African music [14]. Tempi were selected to be in the preferred tempo range (around 100-120 beats per minute) [4].

3.2 Study Description

Sixteen East African participants were recruited in Kigali, Rwanda (3 female, mean age: 23 years, mean musical training:

3.4 years, mean dance training: 2.5 years). Thirteen of these participated in the EEG portion of the study as well as the behavioral portion. All participants were over the age of 18, had normal hearing, and had spent the majority of their lives in East Africa. They all gave informed consent prior to participating and were compensated for their participation, as per approval by the ethics boards at the Centre Hospitalier Universitaire de Kigali and the University of Western Ontario.

The participants first completed three behavioral tasks: In the *rhythm discrimination task*, they heard a rhythm twice and then a test rhythm that was either the exact same as or a slightly altered version of the first. They had to decide if they were the same or different. The trial order was randomized (all 24 East African and Western rhythms). The stimuli in this task (and reproduction) consisted of two individual rhythms presented simultaneously. One rhythm was a sequence of sine tones, the other a sequence of snare drum sample sounds. The rhythm consisting of sine tones was to be discriminated between original and test rhythms.

In the *rhythm reproduction task*, participants heard a rhythm twice and then were prompted to tap the rhythm on a laptop keyboard. If they tapped the correct number of intervals and the duration of each interval was within 20% of the presented interval, then the attempt was considered accurate. If an attempt was inaccurate, they repeated the procedure until they reproduced the rhythm accurately up to a maximum of five attempts. After the five attempts or accurate reproduction, they continued with the next trial. The trial order was randomized (all 24 East African and Western rhythms). Stimuli were as in the rhythm discrimination task, and the rhythm consisting of tones was to be reproduced.

In the *beat tapping task*, participants heard trials of paired rhythms (both sine tone sequences) repeated for 32 seconds. They were instructed to tap along with their sense of the beat. (These were the exact same stimuli participants listened to during the EEG recording.) There were 24 trials in randomized order.

After completion of the behavioral tasks, electrodes were placed on the participant's scalp. They were instructed to sit with eyes closed and without moving for the duration of the recording, and to maintain their attention on the auditory stimuli. All rhythms were repeated for 32 seconds, presented in counterbalanced blocks (all East African rhythms then all Western rhythms, or vice versa), and with randomized order within blocks. 12 rhythms of each type were presented – all at the same tempo, and each rhythm was preceded by 4 seconds of silence. EEG was recorded via a portable Grass EEG system using 14 channels at a sampling rate of 400Hz and impedances were kept below 10kΩ.

3.3 Data Pre-Processing

EEG recordings are usually very noisy. They contain artifacts caused by muscle activity such as eye blinking as well as possible drifts in the impedance of the individual electrodes over the course of a recording. Furthermore, the recording equipment is very sensitive and easily picks up interferences from the surroundings. For instance, in this experiment, the power supply dominated the frequency band around 50Hz. All these issues have led to the common practice to invest a lot of effort into pre-processing EEG data, often even manually rejecting single frames or channels. In contrast to this, we decided to put only little manual work into cleaning the data and just removed obviously bad channels, thus leaving the main work to the deep learning techniques. After bad channel removal, 12 channels remained for subjects 1–5 and 13 for subjects 6–13.

We followed the common practice in machine learning to partition the data into *training*, *validation* (or model selection) and *test* sets. For testing, we took the first T seconds recorded directly after the auditory presentation of each trial had finished. T was chosen such that it matched the length of one bar of the rhythm presented in the trial, i.e. 2160ms for the fast tempo (subjects 1–3 and 7–9) and 2880ms for the other. The first T seconds from the 32s-long trial recordings were used for validation. The rationale behind this was that we expected that the participants would need a few seconds in the beginning of each trial to get used to the new rhythm. Thus, the data would be less suited for training but might still be good enough to estimate the model accuracy on unseen data. The remaining $32-T$ seconds of each trial recording were used for training.

The data was finally converted into the input format required by the neural networks to be learned.[1] In our previous experiments [21], we found that CNNs processing the power spectrum of the raw EEG signal performed best for classifying EEG of rhythm perception. We decided to continue with this approach but chose a different windows size of 96 samples with a hop size of 24 (25% of the window size) for the spectrum computation. At the sample rate of 400Hz, this corresponds to 240ms windows and 60ms hops which equals one fourth or one third of the beat length in the slow and fast rhythms respectively. With this choice, we hoped to be able to pick up beat-related effects but also to have a window size big enough for a sufficient frequency resolution in the spectrum. Including the zero-frequency band, this resulted in 49 frequency bins up to 200Hz with a resolution of 4.17Hz. All other pre-processing steps were left unchanged such that the whole process involved:

1. computing the short-time Fourier transform (STFT) with a window size of 96 samples and a hop size of 24,
2. computing the log amplitude,
3. scaling linearly to a maximum of 1 (per sequence),
4. splitting the data into frames matching the network's input dimensionality with a given hop size of 1.

4. EXPERIMENTS

CNNs as for instance described in [11], have a variety of structural parameters which need to be chosen carefully. In general, CNNs are artificial neural networks (ANNs) with one or more convolutional layers. In such layers, linear convolution operations are applied for local segments of the input followed by a non-linear transformation and a pooling operation over neighboring segments. The *kernel* for each convolution operation is described by a weight matrix of a certain *shape*. Multiple kernels can be applied in parallel within the same layer whereby each corresponds to a different output *channel* of the layer. The *stride* parameter controls how much the kernels should advance on the input data between successive applications. Here, we fixed this parameter at 1 resulting in a maximal overlap of consecutive input segments. The *pooling*

[1]Most of the processing was implemented through the *librosa* library available at https://github.com/bmcfee/librosa/.

parameter controls how many values of neighboring segments are aggregated using the *max* operation. Generally, each convolutional layer can be considered as a trainable local feature detector. Forming hierarchies of such detectors allows to recognize more complex, higher-level patterns in the inputs. In our previous work, we successfully applied CNNs with two convolutional layers to classify the perceived rhythms into types (East African vs. Western) as well as to identify individual rhythms in a pilot experiment [21]. However, we were only able to test a small number of structural configurations, leaving a considerable potential for further improvement.

Here, we took a systematic approach for finding good structural CNN parameters. To this end, we applied a Bayesian optimization technique for hyper-parameter selection in machine learning algorithms, which has recently been described by Snoek et al. [19] and has been implemented in the Spearmint library.[2] The basic idea is to treat the learning algorithm's generalization performance as a sample from a Gaussian process and select the next parameter configuration to test based on the *expected improvement*. The authors showed that this way, the number of experiment runs to minimize a given objective can be significantly reduced while surpassing the performance of parameters chosen by human experts.

We followed the common practice to optimize the performance on the validation set. Because the 24 classes we would like to predict were perfectly balanced, we chose the *accuracy*, i.e., the percentage of correctly classified instances, as primary evaluation measure. As the Bayesian optimization aims to minimize an objective, we let our learner report the *misclassification rate* instead which is one minus the accuracy. Furthermore, ranking the 24 classes by their corresponding network output values, we also computed the *precision at rank 3 (prec.@3)* and the *mean reciprocal rank (MRR)* – two commonly used information retrieval measures. The former corresponds to the accuracy considering the top three classes in the ranking instead of just the first one. The latter is computed as:

$$MRR = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{1}{rank_i} \quad (1)$$

where D is the set of test instances and $rank_i$ is the rank of the correct class for instance i. The value range is (0,1] where the best value, 1, is obtained if the correct class is always ranked first.

We tested several base configurations as listed in Table 2. For each base configuration, Table 3 shows the hyper-parameters and their value ranges considered for optimization. Like in our previous work, all configurations used a DLSVM output layer as proposed in [23].[3] This special kind of output layer for classification uses the hinge loss as cost function and replaces the commonly applied softmax. The convolutional layers apply the rectifier non-linearity $f(x) = max(0,x)$ which does not saturate like sigmoid functions and thus facilitates faster learning. The input length in the time dimension was adapted to match the bar length T. For the fast rhythms, it was 33 and for the slow ones 45. In both cases, 49 frequency bins were used.

[2] https://github.com/JasperSnoek/spearmint

[3] We used the experimental implementation for pylearn2 provided by Kyle Kastner at https://github.com/kastnerkyle/pylearn2/blob/svm_layer/pylearn2/models/mlp.py

Table 2: CNN base configurations.

id	description
A	1 convolutional layer
B	2 convolutional layers, layer 1 with reduced value range
C	2 convolutional layers, layer 1 with reduced value range
D	2 convolutional layers, no limitation

Table 3: Hyper-parameter value ranges [min, max] considered for optimization for configurations B–D.

	layer	structural parameters			initial
	number	kernel shape	#channels	pooling	learning rate
A	1	[1, 33|45]x49	[1, 30]	[1, 10]	[10^{-5}, 0.01]
B	1	[1, 1]x49	[28, 32]	[8, 12]	
	2	[1, 38]x1	[1, 20]	[1, 10]	[10^{-5}, 0.01]
C	1	[21, 25]x49	[13, 17]	[3, 7]	
	2	[1, 22]x1	[1, 20]	[1, 10]	[10^{-5}, 0.01]
D	1	[1, 45]x49	[1, 30]	[1, 10]	
	2	[1, 45]x1	[1, 20]	[1, 10]	[0.01, 0.01]

For all configurations, models were trained for 20 epochs using stochastic gradient descent (SGD) with exponential decay of the learning rate after each epoch and a constant momentum of 0.5. Furthermore, we applied dropout regularization [9] which largely avoided over-fitting to the training data. These learning parameters were all determined to work well in combination during our earlier experiments with this data set.

We started with a CNN with a single convolutional layer (configuration A), training an individual model for each subject. For the subject with the highest accuracy (subject 4), we then also trained CNNs with a second convolutional layer. Here, the structural parameters for the first layer were either limited to values close to the best solution for configuration A (configurations B and C) or without limitation except for the fixed learning rate parameter (configuration D). The rationale behind limiting the parameter value range was that this would leave more processing capacity for exploring the parameter space of the second layer.

All experiments were implemented using Theano [2] and pylearn2 [8].The computations were run on a dedicated 12-core workstation with two Nvidia graphics cards – a Tesla C2075 and a Quadro 2000. With this hardware configuration, training a CNN with a single convolutional layer for 20 epochs took about 20 minutes and up to ten processes could be run in parallel without notable performance degradation. For training CNNs with two convolutional layers, about half as many could be processed which also took twice as long for the same number of epochs.

5. RESULTS

For configuration A with a single convolutional layer, Table 4 lists the best individual parameter combination found for each participant and the obtained performance

values. Figure 1 visualizes the confusion matrices for the individual subjects and their mean. As in our experiments on classifying the EEG data recorded during perception, the overall best accuracy of 31.6% was obtained for subject 4. In our earlier experiments, it had been only 21.4%. Overall, we could also observe an increase of the accuracy for all participants, almost doubling the mean value from 12.8% to 22.9%. This is especially remarkable as the cross-condition classification task – training on data recorded during perception but classifying data recorded after perception – can be considered much harder. The improvement by over 10% can most likely be credited to the Bayesian hyper-parameter optimization, but the changes in the frequency resolution of the spectrum and matching the input size to the bar length may have also played a role. For more than half of the subjects, the accuracy was over 25% and the precision at rank 3 over 50%. For comparison, the accuracy for random guessing would be only 4.17%.

Taking a closer look at hyper-parameter combinations that worked well for configuration A, we identified two distinct approaches. Either patterns were learned for single 1x49 input frames that roughly correspond to single beats or for large kernel shapes covering 2s or more. We used the best combination for each of the two approaches found for subject 4 as guideline for the first convolutional layer in configuration B and C – i.e., B followed the beat-based approach whereas C followed the other one. Indeed, the Bayesian optimizer already found good combinations for these configurations within less than 25 runs. However, as can be seen in Table 5, these models did not outperform the simpler model from A in terms of accuracy. Only an improvement in the precision at rank 3 could be observed.

An improvement was finally achieved using combination D which did not restrict the structural parameters of the first convolutional layer. Whilst it was only a difference of 3.5% for the accuracy, the effect was much more significant if the top-3 classes were considered. Here, the value increased over 12% to 81.6%. However, more than 350 experiment runs had to be completed before this solution was found. Figure 2 (left) shows the obtained confusion matrix. Remarkably, almost none of the East African rhythm was classified as Western (upper right quadrant). For the East African rhythms, confusion was mostly amongst neighbors (i.e., similar rhythms; upper left quadrant) whereas for the Western rhythms, it looked much different.

If we considered the rhythm sequence pairings [a,b] and [b,a] as the same class, the accuracy would increase to 46.2% (chance level 8.33% for 12 classes) for subject 4 and configuration D. Classifying one of the four groups of three shown in Table 1, we obtained 58.3% (chance level 25%). Considering the problem of distinguishing East African from Western rhythms which we had addressed in our previous experiments, the accuracy was 84% – more than 12% higher than what we had obtained earlier for the easier perception classification task. Aggregating classifications obtained from different channels of the same trial, the accuracies for the 24-classes and 12-classes problem increased slightly to 37.5% and 50.0% respectively. The other values did not change.

6. CONCLUSIONS AND OUTLOOK

With our experiments presented in this paper, using convolutional neural networks that look only at a short segment of the signal from a single EEG channel (corresponding to the length of a single bar of a two-bar stimulus), we have obtained encouraging first results for identifying perceived rhythms from EEG recorded after their auditory presentation. Distinguishing the rhythm stimuli used in this study is not easy as a listener. They are all presented in the same tempo (per subject) and comprise two 12/8 bars. Consequently, none of the participants scored more than 83% in the behavioral rhythm discrimination test. Considering this and the rather sub-par data quality of the EEG recordings, the accuracies obtained for some of the participants are remarkable. They were much higher than anticipated and a considerable improvement to our previous pilot experiments with an easier classification task.

As expected, individual differences were very high with accuracy values ranging from 14.6% to 31.6%. A reason for this could be the highly variable data quality of the recordings but also individual differences between the participants. We will therefore include the results from the behavioral part of the study into future analysis, looking for possible correlations with the classification accuracy. We also plan to analyze the learned models as they might provide some insight into the important underlying patterns within the EEG signals and their corresponding neural processes.

The results reported here still need to be taken with a grain of salt. Because of the study design, there is only one trial per stimulus for each subject. Thus, there is the chance that the neural networks learned to identify the individual trials and not the stimuli based on artifacts in the recordings that only occurred sporadically throughout the experiment. Or there could have been brain processes unrelated to rhythm perception and imagination that were only present during some of the trials. Re-arranging the labels within the confusion matrices such that they correspond to the order of the stimuli presentation (Figure 3 and Figure 2, left) shows some confusion between successive trials (vertical lines along the diagonal) which supports this hypothesis. Repeating the experiment with multiple trials per stimulus for each subject should give more insights into this matter.

The study is currently being repeated with North America participants and we are curious to see whether we can replicate our findings. In particular, we hope to further improve the classification accuracy through higher data quality of the new EEG recordings under lab conditions. Furthermore, encouraged by our results, we want to extend our focus by also considering more complex and richer stimuli such as audio recordings of rhythms with realistic instrumentation instead of artificial sine tones.

7. ACKNOWLEDGMENTS

This work was supported by a fellowship within the Postdoc-Program of the German Academic Exchange Service (DAAD), by the Natural Sciences and Engineering Research Council of Canada (NSERC), through the Western International Research Award R4911A07, and by an AUCC Students for Development Award.

8. REFERENCES

[1] G. Barz. *Music in East Africa: experiencing music, expressing culture*. Oxford University Press, 2004.

[2] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio. Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for*

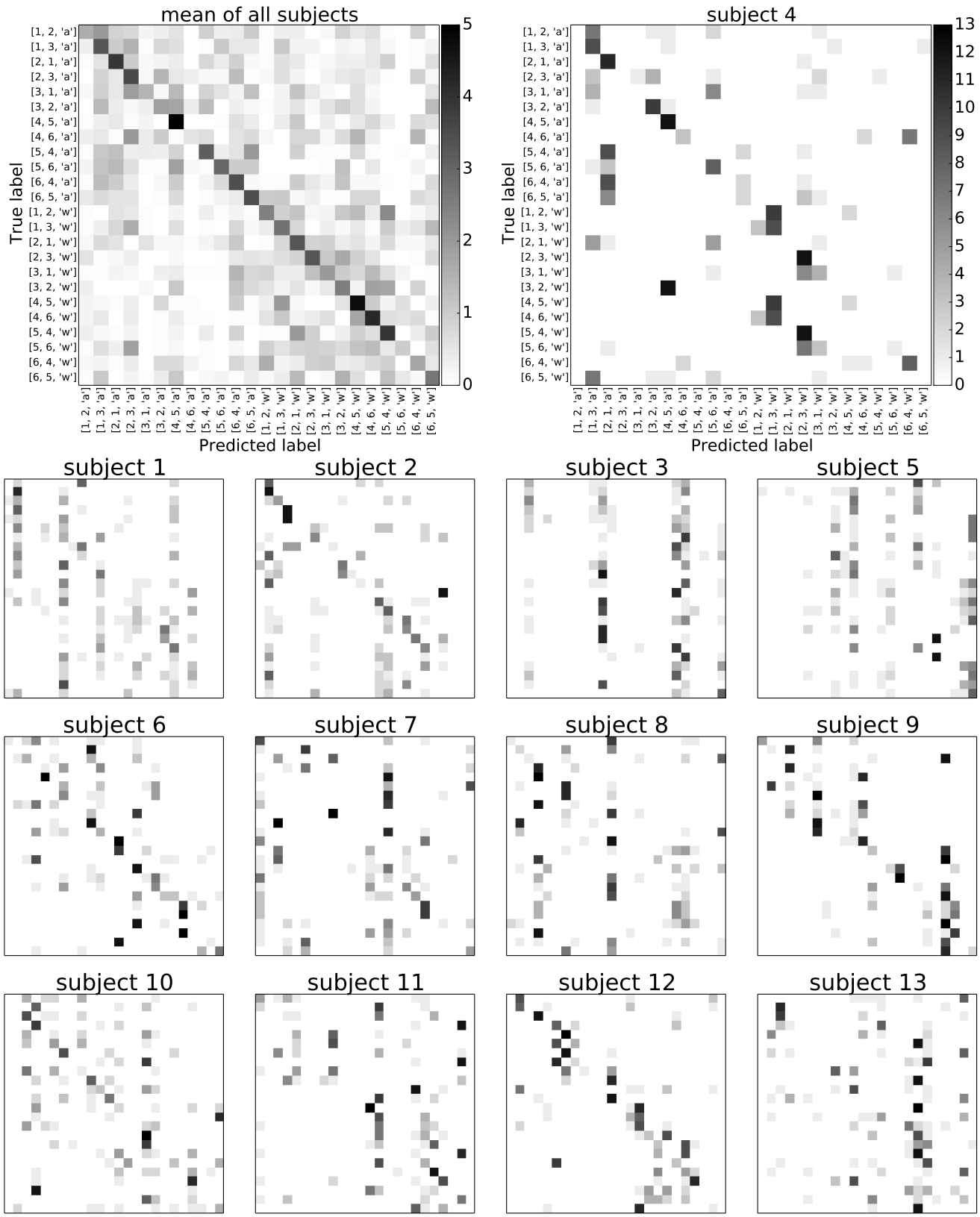

Figure 1: Confusion matrices for the best found models for configuration A. The small matrices for individual subjects use the same color scale and axes as the matrix for subject 4. Labels contain the ids of the high-pitched and low-pitched rhythm sequence (c.f. Table 1) and the rhythm type ('a' for African, 'w' for Western). Labels are arranged such that most similar rhythms are close together.

Table 4: Best hyper-parameter values of configuration A for each subject and their measured performance.

	input shape	kernel shape	#channels	pooling	initial learning rate	accuracy	prec.@3	MRR
value range	33|45x49	[1, 33|45]x49	[1, 30]	[1, 10]	[0.00001, 0.01]			
subject 1	33x49	1x49	18	6	0.00680	15.6%	32.6%	0.31
subject 2	33x49	15x49	23	1	0.00316	27.8%	46.2%	0.42
subject 3	33x49	13x49	30	5	0.00457	17.4%	34.4%	0.31
subject 4	45x49	23x49	15	5	0.00501	31.6%	63.2%	0.52
subject 5	45x49	27x49	1	1	0.00379	14.6%	27.4%	0.29
subject 6	45x49	45x49	30	1	0.00454	29.8%	52.6%	0.48
subject 7	33x49	1x49	15	10	0.00467	21.2%	42.0%	0.38
subject 8	33x49	14x49	30	7	0.00659	16.7%	38.1%	0.33
subject 9	33x49	22x49	16	1	0.00521	28.2%	50.0%	0.45
subject 10	45x49	43x49	30	1	0.00231	27.2%	55.8%	0.46
subject 11	45x49	37x49	30	1	0.00376	22.8%	51.6%	0.41
subject 12	45x49	45x49	19	1	0.00861	26.3%	65.7%	0.49
subject 13	45x49	38x49	14	1	0.00545	19.2%	43.9%	0.38
mean						22.9%	46.4%	0.40

Table 5: Best hyper-parameter values of configurations A–D for subject 4 and their measured performance.

	input shape	layer number	kernel shape	#channels	pooling	initial learning rate	accuracy	prec.@3	MRR
A	45x49	1	23x49	15	5	0.00501	31.6%	63.2%	0.52
B	45x49	1	1x49	29	9				
		2	1x1	16	2	0.00780	28.8%	73.3%	0.50
C	45x49	1	21x49	17	3				
		2	1x1	20	10	0.00709	30.9%	69.4%	0.50
D	45x49	1	26x49	20	1				
		2	6x1	15	5	0.01	34.4%	81.6%	0.55

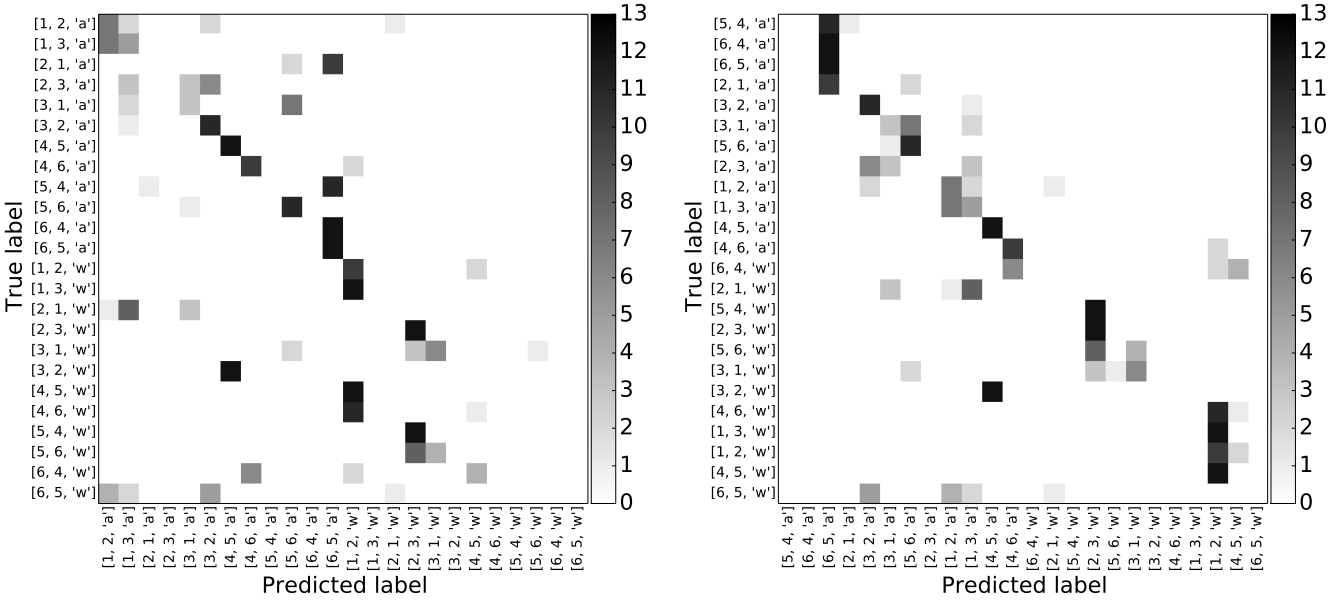

Figure 2: Confusion matrices for the best found model for configuration D. Labels contain the ids of the high-pitched and low-pitched rhythm sequence (c.f. Table 1) and the rhythm type ('a' for African, 'w' for Western). Left: Labels are arranged such that most similar rhythms are close together. Right: Labels are arranged according to the trial order for subject 4.

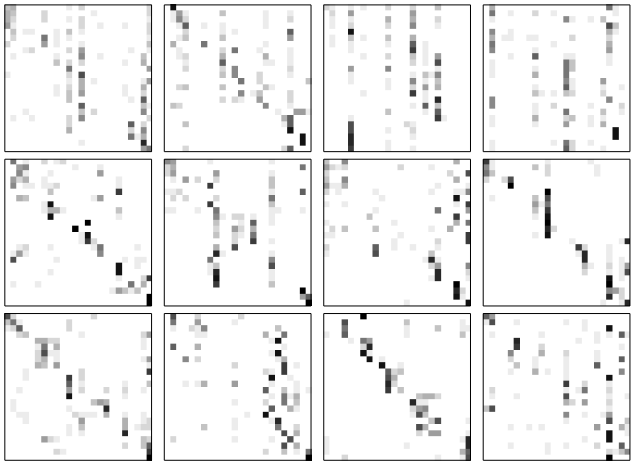

Figure 3: Confusion matrices for the best found models for configuration A for all subjects except subject 4 arranged and with the same color scale as in Figure 1. Labels are arranged according to the (randomized) trial order for each subject. Vertical lines along the diagonal indicate confusion between successive stimuli.

Scientific Computing Conference (SciPy), pages 1–7, 2010.

[3] P. Desain. What rhythm do i have in mind? detection of imagined temporal patterns from single trial ERP. In *Proceedings of International Conference on Music Perception and Cognition (ICMPC)*, page 209, 2004.

[4] P. Fraisse. Rhythm and tempo. In D. Deutsch, editor, *The Psychology of Music*, pages 149–180. Academic Press, 1982.

[5] T. Fujioka, L. Trainor, E. Large, and B. Ross. Beta and gamma rhythms in human auditory cortex during musical beat processing. *Annals of the New York Academy of Sciences*, 1169(1):89–92, 2009.

[6] T. Fujioka, L. Trainor, E. Large, and B. Ross. Internalized timing of isochronous sounds is represented in neuromagnetic beta oscillations. *The Journal of Neuroscience*, 32(5):1791–1802, 2012.

[7] E. Geiser, E. Ziegler, L. Jancke, and M. Meyer. Early electrophysiological correlates of meter and rhythm processing in music perception. *Cortex*, 45(1):93–102, 2009.

[8] I. Goodfellow, D. Warde-Farley, P. Lamblin, V. Dumoulin, M. Mirza, R. Pascanu, J. Bergstra, F. Bastien, and Y. Bengio. Pylearn2: a machine learning research library. *arXiv preprint arXiv:1308.4214*, 2013.

[9] G. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

[10] J. Iversen, B. Repp, and A. Patel. Top-down control of rhythm perception modulates early auditory responses. *Annals of the New York Academy of Sciences*, 1169(1):58–73, 2009.

[11] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1105, 2012.

[12] O. Ladinig, H. Honing, G. Háden, and I. Winkler. Probing attentive and preattentive emergent meter in adult listeners without extensive music training. *Music Perception*, 26(4):377–386, 2009.

[13] M. Längkvist, L. Karlsson, and M. Loutfi. Sleep stage classification using unsupervised feature learning. *Advances in Artificial Neural Systems*, 2012:5:5–5:5, Jan 2012.

[14] D. Moelants, O. Cornelis, and M. Leman. Exploring african tone scales. In *10th International Society for Music Information Retrieval Conference (ISMIR'09)*, pages 489–494, 2009.

[15] S. Nozaradan, I. Peretz, M. Missal, and A. Mouraux. Tagging the neuronal entrainment to beat and meter. *The Journal of Neuroscience*, 31(28):10234–10240, 2011.

[16] S. Nozaradan, I. Peretz, and A. Mouraux. Selective neuronal entrainment to the beat and meter embedded in a musical rhythm. *The Journal of Neuroscience*, 32(49):17572–17581, 2012.

[17] R. Schaefer. *Measuring the mind's ear: EEG of music imagery*. PhD thesis, Radboud University Nijmegen, Nijmegen, The Netherlands, 2011.

[18] R. Schaefer, Y. Blokland, J. Farquhar, and P. Desain. Single trial classification of perceived and imagined music from EEG. *Perception*, 1(S2):S3, 2009.

[19] J. Snoek, H. Larochelle, and R. P. Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2951–2959, 2012.

[20] J. Snyder and E. Large. Gamma-band activity reflects the metric structure of rhythmic tone sequences. *Cognitive brain research*, 24(1):117–126, 2005.

[21] S. Stober, D. J. Cameron, and J. A. Grahn. Classifying EEG recordings of rhythm perception. In *15th International Society for Music Information Retrieval Conference (ISMIR'14)*, 2014.

[22] S. Stober and J. Thompson. Music imagery information retrieval: Bringing the song on your mind back to your ears. In *13th International Conference on Music Information Retrieval (ISMIR'12) - Late-Breaking & Demo Papers*, 2012.

[23] Y. Tang. Deep Learning using Linear Support Vector Machines. *arXiv preprint arXiv:1306.0239*, 2013.

[24] G. Tzanetakis, A. Kapur, and M. Benning. Query-by-beat-boxing: Music retrieval for the dj. In *5th International Conference on Music Information Retrieval (ISMIR'04)*, pages 170–177, 2004.

[25] R. Vlek, R. Schaefer, C. Gielen, J. Farquhar, and P. Desain. Shared mechanisms in perception and imagery of auditory accents. *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, 122(8):1526–1532, Aug 2011.

[26] U. Will and E. Berg. Brain wave synchronization and entrainment to periodic acoustic stimuli. *Neuroscience Letters*, 424(1):55–60, Aug 2007.

[27] D. Wulsin, J. Gupta, R. Mani, J. Blanco, and B. Litt. Modeling electroencephalography waveforms with semi-supervised deep belief nets: fast classification and anomaly measurement. *Journal of Neural Engineering*, 8(3):036015, Jun 2011.