CLASSIFYING EEG RECORDINGS OF RHYTHM PERCEPTION

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

ABSTRACT

Electroencephalography (EEG) recordings of rhythm perception might contain enough information to distinguish different rhythm types/genres or even identify the rhythms themselves. In this paper, we present first classification results using deep learning techniques on EEG data recorded within a rhythm perception study in Kigali, Rwanda. 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. During presentation, the participants' brain waves were recorded via 14 EEG channels. We applied stacked denoising autoencoders and convolutional neural networks on the collected data to distinguish African and Western rhythms on a group and individual participant level. Furthermore, we investigated how far these techniques can be used to recognize the individual rhythms.

1. INTRODUCTION

Musical rhythm occurs in all human societies and is related to many phenomena, such as the perception of a regular emphasis (i.e., beat), and the impulse to move one's body. However, the brain mechanisms underlying musical rhythm are not fully understood. Moreover, musical rhythm is a universal human phenomenon, but differs between human cultures, and the influence of culture on the processing of rhythm in the brain is uncharacterized.

In order to study the influence of culture on rhythm processing, we recruited participants in East Africa and Canada to test their ability to perceive and produce rhythms derived from East African and Western music. Besides behavioral tasks, which have already been discussed in [4], the East African participants also underwent electroencephalography (EEG) recording while listening to East African and Western musical rhythms thus enabling us to study the neural mechanisms underlying rhythm perception. We were interested in differences between neuronal entrainment to the periodicities in East African versus Western rhythms for participants from those respective cultures. Entrainment was defined as the magnitudes of steady state evoked potentials (SSEPs) at frequencies related to the metrical structure of rhythms. A similar approach has been used previously to study entrainment to rhythms [17, 18].

But it is also possible to look at the collected EEG data from an information retrieval perspective by asking questions like *How well can we tell from the EEG whether a participant listened to an East African or Western rhythm?* or *Can we even say from a few seconds of EEG data which rhythm somebody listened to?* Note that answering such question does not necessarily require an understanding of the underlying processes. Hence, we have attempted to let a machine figure out how best to represent and classify the EEG recordings employing recently developed deep learning techniques. In the following, we will review related work in Section 2, describe the data acquisition and pre-processing in Section 3 present our experimental findings in Section 4, and discuss further steps in Section 5.

2. RELATED WORK

Previous research demonstrates that culture influences perception of the metrical structure (the temporal structure of strong and weak positions in rhythms) of musical rhythms in infants [20] and in adults [16]. However, few studies have investigated differences in brain responses underlying the cultural influence on rhythm perception. One study found that participants performed better on a recall task for culturally familiar compared to unfamiliar music, yet found no influence of cultural familiarity on neural activations while listening to the music while undergoing functional magnetic resonance imaging (fMRI) [15].

Many studies have used EEG and magnoencephalography (MEG) to investigate brain responses to auditory rhythms. Oscillatory neural activity in the gamma (20-60 Hz) frequency band is sensitive to accented tones in a rhythmic sequence and anticipates isochronous tones [19]. Oscillations in the beta (20-30 Hz) band increase in anticipation of strong tones in a non-isochronous sequence [5, 6, 10]. Another approach has measured the magnitude of SSEPs (reflecting neural oscillations entrained to the stimulus) while listening to rhythmic sequences [17, 18]. 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].

In the field of music information retrieval (MIR), retrieval based on brain wave recordings is still a very young and unexplored domain. So far, research has mainly focused on emotion recognition from EEG recordings (e.g., [3, 14]). For rhythms, however, Vlek et al. [23] already 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, MIR is still largely under-represented here. 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. [24] 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 hand-selected features.

3. DATA ACQUISITION & PRE-PROCESSING

3.1 Stimuli

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 – each at a different pitch. For each cultural type of rhythm, there were 2 groups of 3 individual rhythms for which rhythms could be overlaid with the others in their group. Because an individual rhythm could be one of two pitches/sounds, this made for a total of 12 rhythmic stimuli from each culture, each used for all tasks. Furthermore, rhythmic stimuli could be one of two tempi: having a minimum inter-onset interval of 180 or 240ms.

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	

3.2 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. 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. All 12 rhythms of each type were presented – all at the same tempo (fast tempo for subjects 1–3 and 7–9, and slow tempo for the others). 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. To this end, we split each 32s-long trial recording into three non-overlapping pieces. The first four seconds were used for the validation dataset. 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 next 24 seconds were used for training and the remaining four seconds for testing.

The data was finally converted into the input format required by the neural networks to be learned. [1] If the network just took the raw EEG data, each waveform was normalized to a maximum amplitude of 1 and then split into equally sized frames matching the size of the network's input layer. No windowing function was applied and the frames overlapped by 75% of their length. If the network was designed to process the frequency spectrum, the processing involved:

1. computing the short-time Fourier transform (STFT) with given window length of 64 samples and 75% overlap,
2. computing the log amplitude,
3. scaling linearly to a maximum of 1 (per sequence),
4. (optionally) cutting of all frequency bins above the number requested by the network,
5. splitting the data into frames matching the network's input dimensionality with a given hop size of 5 to control the overlap.

Here, the number of retained frequency bins and the input length were considered as hyper-parameters.

4. EXPERIMENTS & FINDINGS

All experiments were implemented using Theano [2] and pylearn2 [8]. [2] The computations were run on a dedicated 12-core workstation with two Nvidia graphics cards – a Tesla C2075 and a Quadro 2000.

As the first retrieval task, we focused on recognizing whether a participant had listened to an East African or Western rhythm (Section 4.1). This binary classification task is most likely much easier than the second task – trying to predict one out of 24 rhythms (Section 4.2). Unfortunately, due to the block design of the study, it was not possible to train a classifier for the tempo. Trying to do so would yield a classifier that "cheated" by just recognizing the inter-individual differences because every participant only listened to stimuli of the same tempo.

As the classes were perfectly balanced for both tasks, we chose the *accuracy*, i.e., the percentage of correctly classified instances, as evaluation measure. Accuracy can be measured on several levels. The network predicts a class label for each input frame. Each frame is a segment from the time sequence of a single EEG channel. Finally, for each trial, several channels were recorded. Hence, it is natural to also measure accuracy also at the sequence (i.e, channel) and trial level. There are many ways to aggregate frame label predictions into a prediction for a channel or a trial. We tested the following three ways to compute a score for each class:

- **plain:** sum of all 0-or-1 outputs per class
- **fuzzy:** sum of all raw output activations per class
- **probabilistic:** sum of log output activations per class

While the latter approach which gathers the log likelihoods from all frames worked best for a softmax output layer, it usually performed worse than the fuzzy approach for the DLSVM output layer with its hinge loss (see below). The plain approach worked best when the frame accuracy was close to the chance level for the binary classification task. Hence, we chose the plain aggregation scheme whenever the frame accuracy was below 52% on the validation set and otherwise the fuzzy approach.

We expected significant inter-individual differences and therefore made learning good individual models for the participants our priority. We then tested configuration that worked well for individuals on three groups – all participants as well as one group for each tempo, containing 6 and 7 subjects respectively.

4.1 Classification into African and Western Rhythms

4.1.1 Multi-Layer Perceptron with Pre-Trained Layers

Motivated by the existing deep learning approaches for EEG data (cf. Section 2), we choose to pre-train a MLP as an autoencoder for individual channel-seconds – or similar fixed-length chunks – drawn from all subjects. In particular, we trained a stacked denoising autoencoder (SDA) as proposed in [22] where each individual input was set to 0 with a *corruption probability* of 0.2.

We tested several structural configurations, varying the input sample rate (400Hz or down-sampled to 100Hz), the number of layers, and the number of neurons in each layer. The quality of the different models was measured as the mean squared reconstruction error (MSRE). Table 2 gives an overview of the reconstruction quality for selected configurations. All SDAs were trained with tied weights, i.e., the weight matrix of each decoder layer equals the transpose of the respective encoder layer's weight matrix. Each layer was trained with stochastic gradient descent (SGD) on mini-batches of 100 examples for a maximum of 100 epochs with an initial learning rate of 0.05 and exponential decay.

In order to turn a pre-trained SDA into a multilayer perceptron (MLP) for classification, we replaced the decoder part of the SDA with a DLSVM layer as proposed in [21]. [3] This special kind of output layer for classification uses the hinge

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

[2] The code to run the experiments is publicly available as supplementary material of this paper at http://dx.doi.org/10.6084/m9.figshare.1108287

[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. MSRE and classification accuracy for selected SDA (top, A-F) and CNN (bottom, G-I) configurations.

	neural network configuration	MSRE		MLP Classification Accuracy (for frames, channels and trials) in %											
id	(sample rate, input format, hidden layer sizes)	train	test	indiv. subjects			fast (1–3, 7–9)			slow (4–6, 10–13)			all (1–13)		
A	100Hz, 100 samples, 50-25-10 (SDA for subject 2)	4.35	4.17	61.1	65.5	72.4	58.7	60.6	61.1	53.7	56.0	59.5	53.5	56.6	60.3
B	100Hz, 100 samples, 50-25-10	3.19	3.07	58.1	62.0	66.7	58.1	60.7	61.1	53.5	57.7	57.1	52.1	53.5	54.5
C	100Hz, 100 samples, 50-25	1.00	0.96	61.7	65.9	71.2	58.6	62.3	63.2	54.4	56.4	57.1	53.4	54.8	56.4
D	400Hz, 100 samples, 50-25-10	0.54	0.53	51.7	58.9	62.2	50.3	50.6	50.0	50.0	51.8	51.2	50.1	50.2	50.0
E	400Hz, 100 samples, 50-25	0.36	0.34	60.8	65.9	71.8	56.3	58.6	66.0	52.0	55.0	56.0	49.9	50.1	56.1
F	400Hz, 80 samples, 50-25-10	0.33	0.32	52.0	59.9	62.5	52.3	53.9	54.9	50.5	53.5	55.4	50.2	51.0	50.3
G	100Hz, 100 samples, 2 conv. layers			62.0	63.9	67.6	57.1	57.9	59.7	49.9	50.2	50.0	51.7	52.8	52.9
H	100Hz, 200 samples, 2 conv. layers			64.0	64.8	67.9	58.2	58.5	61.1	49.5	49.6	50.6	50.9	50.2	50.6
I	400Hz, 1s freq. spectrum (33 bins), 2 conv. layers			69.5	70.8	74.7	58.1	58.0	59.0	53.8	54.5	53.0	53.7	53.9	52.6
J	400Hz, 2s freq. spectrum (33 bins), 2 conv. layers			72.2	72.6	77.6	57.6	57.5	60.4	52.9	52.9	54.8	53.1	53.5	52.3

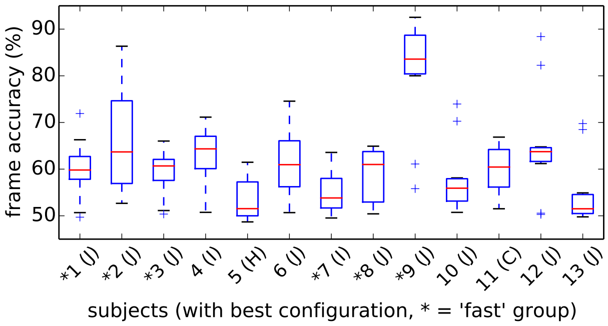


Figure 1. Boxplot of the frame-level accuracy for each individual subject aggregated over all configurations. [5]

loss as cost function and replaces the commonly applied softmax. We observed much smoother learning curves and a slightly increased accuracy when using this cost function for optimization together with rectification as non-linearity in the hidden layers. For training, we used SGD with dropout regularization [9] and momentum, a high initial learning rate of 0.1 and exponential decay over each epoch. After training for 100 epochs on minibatches of size 100, we selected the network that maximized the accuracy on the validation dataset. We found that the dropout regularization worked really well and largely avoided over-fitting to the training data. In some cases, even a better performance on the test data could be observed. The obtained mean accuracies for the selected SDA configurations are also shown in Table 2 for MLPs trained for individual subjects as well as for the three groups. As Figure 1 illustrates, there were significant individual differences between the subjects. Whilst learning good classifiers appeared to be easy for subject 9, it was much harder for subjects 5 and 13. As expected, the performance for the groups was inferior. Best results were obtained for the "fast" group, which comprised only 6 subjects including 2 and 9 who were amongst the easiest to classify.

We found that two factors had a strong impact on the MSRE: the amount of (lossy) *compression* through the autoencoder's bottleneck and the amount of *information* the network processes at a time. Configurations A, B and D had the highest compression ratio (10:1). C and E lacked the third autoencoder layer and thus only compressed at 4:1 and with a lower resulting MSRE. F had exactly twice the compression ratio as C and E. While the difference in the MSRE was remarkable between F and C, it was much less so between F and E – and even compared to D. This could be explained by the four times higher sample rate of D–F. Whilst A–E processed the same amount of *samples* at a time, the input for A–C contained much more information as they were looking at 1s of the signal in contrast to only 250ms. Judging from the MSRE, the longer time span appears to be harder to compress. This makes sense as EEG usually contains most information in the lower frequencies and higher sampling rates do not necessarily mean more content. Furthermore, with growing size of the input frames, the variety of observable signal patterns increases and they become harder to approximate. Figure 2 illustrates the difference between two reconstructions of the same 4s raw EEG input segment using configurations B and D. In this specific example, the MSRE for B is ten times as high compared to D and the loss of detail in the reconstruction is clearly visible. However, D can only see 250ms of the signal at a time whereas B processes one channel-second.

Configuration A had the highest MSRE as it was only trained on data from subject 2 but needed to process all other subjects as well. Very surprisingly, the respective MLP produced much better predictions than B, which had identical structure. It is not clear what caused this effect. One explanation could be that the data from subject 2 was cleaner than for other participants as it also led to one amongst the best individual classification accuracies. [6] This could have led to more suitable features learned by the SDA. In general, the two-hidden-layer models worked better than the three-hidden-layer ones. Possibly, the compression caused by the third hidden layer was just too much. Apart from this, it was hard to make out a clear "winner" between A, C and E. There seemed to be a trade-off between the accuracy of the reconstruction (by choosing a smaller window size and/or higher sampling rate) and learning more suitable features

[5] Boxes show the 25th to 75th percentiles with a mark for the median within, whiskers span to furthest values within the 1.5 interquartile range, remaining outliers are shown as crossbars.

[6] Most of the model/learning parameters were selected by training just on subject 2.

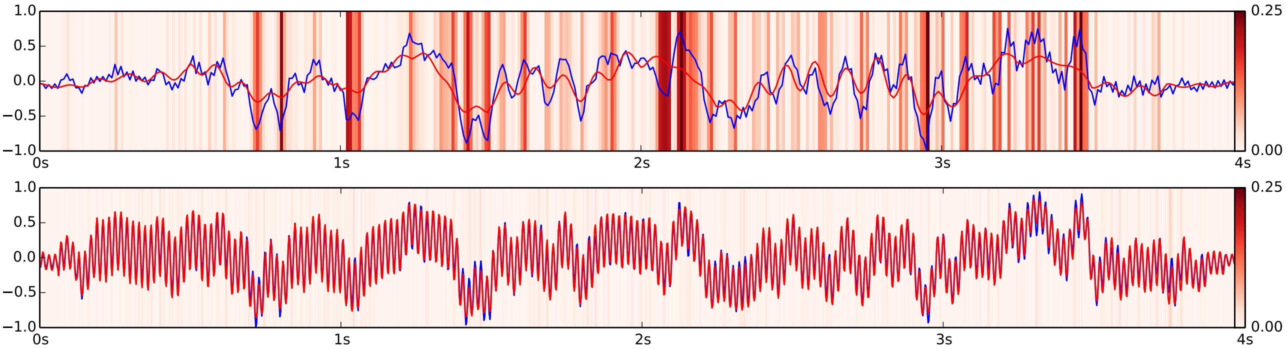


Figure 2. Input (blue) and its reconstruction (red) for the same 4s sequence from the test data. The background color indicates the squared sample error. Top: Configuration B (100Hz) with MSRE 6.43. Bottom: Configuration D (400Hz) with MSRE 0.64. (The bottom signals shows more higher-frequency information due to the four-times higher sampling rate.)

Table 3. Structural parameters of the CNN configurations.

	input	convolutional layer 1				convolutional layer 2			
id	dim.	shape	patterns	pool	stride	shape	patterns	pool	stride
G	100x1	15x1	10	7	1	70x1	10	7	1
H	200x1	25x1	10	7	1	151x1	10	7	1
I	22x33	1x33	20	5	1	9x1	10	5	1
J	47x33	1x33	20	5	1	9x1	10	5	1

for recognizing the rhythm type at a larger time scale. This led us to try a different approach using convolutional neural networks (CNNs) as, e.g., described in [11].

4.1.2 Convolutional Neural Network

We decided on a general layout consisting of two convolutional layers where the first layer was supposed to pick up beat-related patterns while the second would learn to recognize higher-level structures. Again, a DLSVM layer was used for the output and the rectifier non-linearity in the hidden layers. The structural parameters are listed in Table 3. As pooling operation, the maximum was applied. Configurations G and H processed the same raw input as A–F whereas I and J took the frequency spectrum as input (using all 33 bins). All networks were trained for 20 epochs using SGD with a momentum of 0.5 and an exponential decaying learning rate initialized at 0.1.

The obtained accuracy values are listed in Table 2 (bottom). Whilst G and H produced results comparable to A–F, the spectrum-based CNNs, I and J, clearly outperformed all other configurations for the individual subjects. For all but subjects 5 and 11, they showed the highest frame-level accuracy (c.f. Figure 1). For subjects 2, 9 and 12, the trial classification accuracy was even higher than 90% (not shown).

4.1.3 Cross-Trial Classification

In order to rule out the possibility that the classifiers just recognized the individual trials – and not the rhythms – by coincidental idiosyncrasies and artifacts unrelated to rhythm perception, we additionally conducted a cross-trial classification experiment. Here, we only considered all subjects with frame-level accuracies above 80% in the earlier experiments – i.e., subjects 2, 9 and 12. We formed 144 rhythm pairs by combining each East African with each Western rhythm from the fast stimuli (for subjects 2 and 9) and the slow ones (for subject 12) respectively. For each pair, we trained a classifier with configuration J using all but the two rhythms of the pair.[7] Due to the amount of computation required, we trained only for 3 epochs each. With the learned classifiers, the mean frame-level accuracy over all 144 rhythm pairs was 82.6%, 84.5% and 79.3% for subject 2, 9 and 12 respectively. These value were only slightly below those shown in Figure 1, which we considered very remarkable after only 3 training epochs.

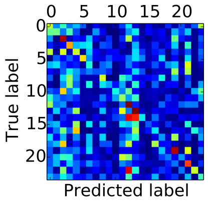
4.2 Identifying Individual Rhythms

Recognizing the correct rhythm amongst 24 candidates was a much harder task than the previous one – especially as all candidates had the same meter and tempo. The chance level for 24 evenly balanced classes was only 4.17%. We used again configuration J as our best known solution so far and trained an individual classifier for each subject. As Figure 3 shows, the accuracy on the 2s input frames was at least twice the chance level. Considering that these results were obtained without any parameter tuning, there is probably still much room for improvements. Especially, similarities amongst the stimuli should be considered as well.

5. CONCLUSIONS AND OUTLOOK

We obtained encouraging first results for classifying chunks of 1-2s recorded from a single EEG channel into East African or Western rhythms using convolutional neural networks (CNNs) and multilayer perceptrons (MLPs) pre-trained as stacked denoising autoencoders (SDAs). As it turned out, some configurations of the SDA (D and F) were especially suited to recognize unwanted artifacts like spikes in the waveforms through the reconstruction error. This could be elaborated in the future to automatically discard bad segments during preprocessing. Further, the classification accuracy for individual rhythms was significantly above chance level and encourages more research in this direction. As this has been an initial and by no means exhaustive exploration of the model- and leaning parameter space, there seems to be a lot more potential – especially in CNNs processing the frequency spectrum – and

[7] Deviating from the description given in Section 3.3, we used the first 4s of each recording for validation and the remaining 28s for training as the test set consisted of full 32s from separate recordings in this special case.



subject	1	2	3	4	5	6	7	8	9	10	11	12	13	mean
accuracy	15.8%	9.9%	12.0%	21.4%	10.3%	13.9%	16.2%	11.0%	11.0%	10.3%	9.2%	17.4%	8.3%	12.8%
precision @3	31.5%	29.9%	26.5%	48.2%	28.3%	27.4%	41.2%	27.8%	28.5%	33.2%	24.7%	39.9%	20.7%	31.4%
mean reciprocal rank	0.31	0.27	0.27	0.42	0.26	0.28	0.36	0.27	0.28	0.30	0.25	0.36	0.23	0.30

Figure 3. Confusion matrix for all subjects (left) and per-subject performance (right) for predicting the rhythm (24 classes).

we will continue to look for better designs than those considered here. We are also planning to create publicly available data sets and benchmarks to attract more attention to these challenging tasks from the machine learning and information retrieval communities.

As expected, individual differences were very high. For some participants, we were able to obtain accuracies above 90%, but for others, it was already hard to reach even 60%. We hope that by studying the models learned by the classifiers, we may shed some light on the underlying processes and gain more understanding on why these differences occur and where they originate. Also, our results still come with a grain of salt: We were able to rule out side effects on a trial level by successfully replicating accuracies across trials. But due to the study's block design, there remains still the chance that unwanted external factors interfered with one of the two blocks while being absent during the other one. Here, the analysis of the learned models could help to strengthen our confidence in the results.

The study is currently being repeated with North America participants and we are curious to see whether we can replicate our findings. Furthermore, 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.

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.

6. REFERENCES

[1] G.F. 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 *Proc. of the Python for Scientific Computing Conference (SciPy)*, 2010.

[3] R. Cabredo, R.S. Legaspi, P.S. Inventado, and M. Numao. An emotion model for music using brain waves. In *ISMIR*, pages 265–270, 2012.

[4] D.J. Cameron, J. Bentley, and J.A. Grahn. Cross-cultural influences on rhythm processing: Reproduction, discrimination, and beat tapping. *Frontiers in Human Neuroscience*, to appear.

[5] T. Fujioka, L.J. Trainor, E.W. 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.J. Trainor, E.W. 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.J. 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.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R.R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

[10] J.R. Iversen, B.H. Repp, and A.D. 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.E. 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] Y.-P. Lin, T.-P. Jung, and J.-H. Chen. EEG dynamics during music appreciation. In *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual Int. Conf. of the IEEE*, pages 5316–5319, 2009.

[15] S.J. Morrison, S.M. Demorest, E.H. Aylward, S.C. Cramer, and K.R. Maravilla. Fmri investigation of cross-cultural music comprehension. *Neuroimage*, 20(1):378–384, 2003.

[16] S.J. Morrison, S.M. Demorest, and L.A. Stambaugh. Enculturation effects in music cognition the role of age and music complexity. *Journal of Research in Music Education*, 56(2):118–129, 2008.

[17] 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.

[18] 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.

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

[20] G. Soley and E.E. Hannon. Infants prefer the musical meter of their own culture: a cross-cultural comparison. *Developmental psychology*, 46(1):286, 2010.

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

[22] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *The Journal of Machine Learning Research*, 11:3371–3408, Dec 2010.

[23] R.J. Vlek, R.S. Schaefer, C.C.A.M. Gielen, J.D.R. Farquhar, and P. Desain. Shared mechanisms in perception and imagery of auditory accents. *Clinical Neurophysiology*, 122(8):1526–1532, Aug 2011.

[24] D.F. Wulsin, J.R. Gupta, R. Mani, J.A. 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.