SHARED GENERATIVE REPRESENTATION OF AUDITORY CONCEPTS AND EEG TO RECONSTRUCT PERCEIVED AND IMAGINED MUSIC

André Ofner **Sebastian Stober**
Research Focus Cognitive Sciences, University of Potsdam, Germany
{ofner, sstober}@uni-potsdam.de

ABSTRACT

Retrieving music information from brain activity is a challenging and still largely unexplored research problem. In this paper we investigate the possibility to reconstruct perceived and imagined musical stimuli from electroencephalography (EEG) recordings based on two datasets. One dataset contains multichannel EEG of subjects listening to and imagining rhythmical patterns presented both as sine wave tones and short looped spoken utterances. These utterances leverage the well-known speech-to-song illusory transformation which results in very catchy and easy to reproduce motifs. A second dataset provides EEG recordings for the perception of 10 full length songs. Using a multi-view deep generative model we demonstrate the feasibility of learning a shared latent representation of brain activity and auditory concepts, such as rhythmical motifs appearing across different instrumentations. Introspection of the model trained on the rhythm dataset reveals disentangled rhythmical and timbral features within and across subjects. The model allows continuous interpolation between representations of different observed variants of the presented stimuli. By decoding the learned embeddings we were able to reconstruct both perceived and imagined music. Stimulus complexity and the choice of training data shows strong effect on the reconstruction quality.

1. INTRODUCTION

Studying the human brain's response to music gained a lot of attention in recent years. Many studies in the field rely on electroencephalography (EEG) recordings, as they provide better temporal resolution than other techniques, such as functional magnetic resonance imaging (fMRI). Previous research suggests that a listener's brain response is modulated in correlation to the perceived auditory stimuli on many different levels and that these modulations can be detected within EEG. One of these effects is the correlation between the frequency and magnitude of neural oscillation patterns, which are modulated by accents and rhythmical patterns in music [3, 20, 21]. Other studies indicate that tracking auditory attention towards a specific sound source in EEG recordings is possible [1, 30].

EEG data has been used to research event-related potentials (ERPs) as a repeatable and distinguishable response to aspects of perceived music. The characteristic brain activity patterns underlying ERPs can be specific, for example, to the structure of musical events, such as note onsets or rhythm and pitch patterns [19, 24]. Other ERPs are related to the timbre of sound and can be modulated even by differences within timbre, such as changes in harmonics [17, 25]. While many ERP components show similar activation across subjects, studies suggest that some are caused by more fine-grained aspects of music, especially within trained musicians [25]. These brain activity patterns extend over the temporal, spatial and frequency domain of the EEG signal.

Motivated by the existence of such features, EEG recordings have been used in several music information retrieval studies based on EEG, such as perceived rhythm or tempo classification [28]. First attempts have been made to reconstruct the loudness envelope of perceived and imagined musical stimuli, but with unsatisfying accuracy [22, 26, 27]. Some of these studies use deep neural networks for classification and regression and the achieved results hint at their usefulness in exploring the complex brain signal. However, the power of employed networks is restricted by size and their general application exclusively to EEG signal denoising or classification. Outside from research on music cognition, recent studies have shown the possibility to use generative models to reconstruct perceived visual stimuli both from fMRI and EEG recordings [4, 10]. Generative models learn to encode a meaningful internal latent representation of a given signal. In addition, they contain a decoding part to either reconstruct the input or another signal that is extractable from the internal latent variable. A recent study has demonstrated the possibility to learn such shared latent embeddings for EEG recordings of music perception and use them as a continuous semantic space representation of the audio [23].

This suggests that a more elaborate generative model could learn a shared encoding of music and brain signals, leading to a conjoint representation of those auditory concepts that are perceived and processed by the brain. As previous research suggests, these concepts span a spectrum of complexity, starting on the level of the subject-specific manifestation and meaning of specific ERP responses to high-level semantic or emotional meaning of music. Therefore, they provide the necessary information to reconstruct the musical stimuli as they are perceived or imagined. Based on this motivation, we propose a generative multi-view model that makes use of deep neural networks to encode and decode spatio-temporal brain signal using a latent embedding. This embedding is simultaneously used to reconstruct and classify music presented and imagined during EEG recording. In this paper we introduce our view

on auditory concepts and the suggested method. We describe two datasets of EEG recordings during music perception and imagination that are used for training and evaluation. Furthermore, we perform model introspection to demonstrate the possibility of interpolating between musically meaningful points within the learned latent space. Finally, we suggest possible ways to extend the framework to include multi-modal processing and learning high level musical concepts.

2. AUDITORY CONCEPTS

Our approach relies on three assumptions for auditory concepts:

1. Coupled auditory and conceptual processing
2. Shared neural representation of music perception and imagination
3. Hierarchical structure of music

Firstly, we assume that there is a tight coupling between auditory and conceptual processing [12]. Several studies suggest that auditory stimuli are processed in a conceptual system that is shared with other modalities, such as visual perception [31]. Furthermore, music processing is based on concepts inherent to the auditory stimuli as well as on external factors, such as visual and social environment or musical training [7]. Secondly, following the ideas of embodied cognition, we assume that the human conceptual system is essentially grounded in perception and that through its interplay with action and cognitive states, music perception at least partially shares conceptual and neural representation with musical imagination [11]. Previous research suggests that auditory concept formation can be traced back to specific ERPs and that the magnitude of some ERP component can be controlled by the presence of an auditory concept in the listeners mind [29]. Thirdly, we follow the idea that music is essentially hierarchical in structure and that auditory concepts equivalently exist on a spectrum of abstraction levels, reflecting and augmenting this structure. They can range from concepts related to single sounds or rhythm to concepts within the emotional or aesthetic processing of music. Together with the previous two assumptions this means that basic elements of perceptual musical processing, such as ERPs related to note onset expectancy, are influenced by their integration into conceptual processing. Music cognition and concept formation can be highly subjective, stimulus-driven as well as context-dependent, e.g. on visual and social aspects of a performance [18]. For these reasons, we hypothesize that a simultaneous retrieval of auditory concepts from multiple sources aids the reconstruction of the processed stimuli while further deepening our understanding of music cognition.

3. RELATED WORK

Various approaches exist to learning a shared embedding from two or more datasets. One method is Canonical Correlation Analysis (CCA) [8]. CCA is non-probabilistic and enables the extraction of linear components to optimize the correlations between two multivariate datasets. CCA in combination with convolutional neural networks has recently been used by Raposo et al. to learn a shared semantic space between audio and EEG signal [23]. Based on CCA, Fujiwara et al. have introduced Bayesian Canonical Correlation Analysis (BCCA), a probabilistic interpretation of CCA [5]. However, BCCA still contains linear observation models, while EEG data is very complex and noisy and requires non-linear computation. To surpass this limitation, Deep Canonically Correlated Autoencoders (DCCAEs) were proposed by Wang et al. [32]. DCCAEs maximize the correlation between the latent embeddings of two separate autoencoders, but do not enable cross-reconstruction between their inputs. While this problem is solved by correlational neural networks (CorrNets), the unregularized latent embeddings of both DCCAE and CorrNet are prone to overfitting, especially in combination with the representational power of non-linear observation models [2]. For these reasons, we follow the suggestion of Wang et al. to use a deep, generative and probabilistic latent variable interpretation of CCA called Deep Variational Canonical Correlation Analysis (VCCA) [32]. A similar approach tailored specifically to a missing view reconstruction for visual stimuli in fMRI data has successfully been demonstrated recently [4]. Here, we show that we can derive a general multi-view generative model capable of joint EEG and stimulus processing that allows multi-modal learning from physiological data as well as directly from the stimuli. To our knowledge, no comparable framework for EEG-based audio stimulus reconstruction or for shared auditory concept learning exists.

4. DATASETS AND PREPROCESSING

We use two datasets, the OpenMIIR speech and the Naturalistic Music EEG Dataset - Tempo (NMED-T) dataset. They are similar in experimental setup but differ in focus and size.

4.1 OpenMIIR speech dataset

One dataset contains EEG of subjects listening to and imagining four rhythmical patterns presented both as sine wave tones and short looped spoken utterances. It stems from the Open Music Imagery Information Retrieval (OpenMIIR) initiative [28] and features four different catchy and easy to reproduce motifs superimposed on a constant metronome click. We refer to it as "OpenMIIR speech dataset". The trials are annotated for containing either speech or sine wave tones and can be used to train and evaluate model performance for the perception and imagination of the same rhythmical trials within two timbres. The metronome clicks serve as cues that are present during perception as well as imagination. The main intention behind this dataset is to reduce stimulus complexity as far as possible while still retaining enough musical structure for building and evaluating models. This dataset contains data from seven subjects with normal hearing and no history of brain injury. It was recorded with 64 EEG channels, horizontal and vertical Electrooculography (EOG) channels sampled at 512 Hz. All perception stimuli have equal tempo and duration of 12 s. Presentation was done in randomized order after 2 s of metronome clicks. They were immediately followed by another 12 s of metronome cues. Participants were asked to imagine the perceived stimulus directly after presentation using these subsequent cue clicks. The concatenated

perception-imagination trials sum up to 26 s of recorded EEG data for each trial. As each trial was presented 6 times, this sums up to a total of 96 presented trials. In total, the dataset contains about 2500 s (42 min) of EEG recordings per subject. We performed common-practice preprocessing steps using the MNE-python toolbox by Gramfort et al. including manual bad channel removal and interpolation after visual inspection [6]. All EEG data was bandpass filtered between 0.5 and 50 Hz. Extended Infomax Independent Component Analysis (ICA) was used to remove EEG artifacts using the EOG signal.

4.2 NMED-T dataset

The NMED-T dataset provides EEG recordings for the perception of 10 naturalistic full length songs. The songs are in Western musical tradition, have durations between 4:30 and 5:00 min in length and contain vocals. They are real-world musical works with pronounced rhythmical properties. 125 channel EEG at 1 kHz sampling rate was recorded for all of the 20 subjects with normal hearing and no history of brain injury. We used the preprocessed version of the dataset, which features EEG down-sampled to 125 Hz and bandpass filtered between 0.3 and 50 Hz. Ocular and cardiac artifacts were removed using the additional EOG channels with ICA after manual bad channel removal. A more detailed description of the preprocessed dataset can be found in [15].

Subjects in both experiments were not required to have musical training, nor did they execute a particular task during listening or imagination. All EEG channels were normalized to zero mean and range [-1, 1]. For training, EEG data was split into excerpts of 1 s length, resulting in 512 samples (OpenMIIR) and 125 samples (NMED-T) length.

We computed Mel spectrograms of audio targets at full sample-rate of 44100 Hz using the librosa library [16] with 64 frequency bands between 0 and 2000 Hz, FFT window size of 2048 and hop length of 1024. Furthermore, we generated loudness envelopes for each stimulus using Hilbert transform of the scipy library at the full sample rate [9]. We then down-sampled the Mel spectrograms and loudness envelopes to the sample rates of the EEG (512 Hz for OpenMIIR and 125 Hz for NMED-T) before splitting into excerpts of 1 s length.

5. LEARNING SHARED REPRESENTATIONS OF AUDIO AND BRAIN SIGNAL

We propose an adaptation of VCCA as proposed by Wang et al. [32] to perform multi-view learning on audio and EEG signal by defining EEG and audio to be two views that can be generated independently from a shared latent embedding z:

$$p(audio,eeg,z) = p(z)p(audio|z)p(eeg|z). \quad (1)$$

As we are essentially interested in the auditory information within EEG signal, we formulate a default model with a single encoder, which processes EEG. Here, z is a learnable space of auditory concepts which are contained implicitly both in the audio and the EEG signal and which generate significant parts of both views. Following the VCCA principle, we project both audio and EEG signal into the shared space z. By declaring the prior $p(z)$, $p(audio \mid eeg)$, and $p(eeg \mid z)$ to be Gaussian, we ensure that the projections $E[z \mid audio]$ and $E[z \mid eeg]$ of the maximum likelihood solution are in the same space as the projections through CCA. As we deal with the reconstruction of complex EEG data, we parametrize the mean of $p_\Theta(eeg \mid z)$ with deep neural networks (DNNs) and apply the same procedure for the mean of $p_\Theta(audio \mid z)$. The approximate posterior $q_\phi(z \mid eeg)$ is optimized by a third DNN. Training the VCCA model is done in analogy to Variational Autoencoders (VAEs) with variational inference by sampling from $q_\phi(z \mid eeg)$. Optimizing the lower bound of the log likelihood $L(eeg,audio;\theta,\phi)$ with stochastic backpropagation is done by optimizing the reconstruction loss of audio and EEG decoder and the Kullback-Leibler (KL) divergence between the learned $q_\phi(z \mid eeg)$ and $p(z)$ using the reparameterization trick [14].

5.1 Multimodal data and additional views

This model can be extended to arbitrary amount of decoders to reconstruct multiple views, as long as they are dependent mainly of a shared latent variable. Here, we use several decoders to reconstruct different aspects of the audio signal: Mel spectrograms of the audio stimuli, their loudness envelope as well as an additional decoder to classify the trial types. Based on our retrieval intention, here we focus on the learned embedding and the reconstructed Mel spectrograms. We use the remaining decoders to enhance the training quality. Similarly, we can add additional encoders, if they represent data based on the latent variable, by making use of additional private latent variables introduced with the VCCA model. They store only the view-specific aspects of additional input, e.g. from other biological modalities, such as fMRI, audio or EEG signal during imagination. Figure 1 shows an example of the modified VCCA architecture with one EEG decoder and two audio decoders. Here, we test the model with a single EEG encoder and multiple decoders.

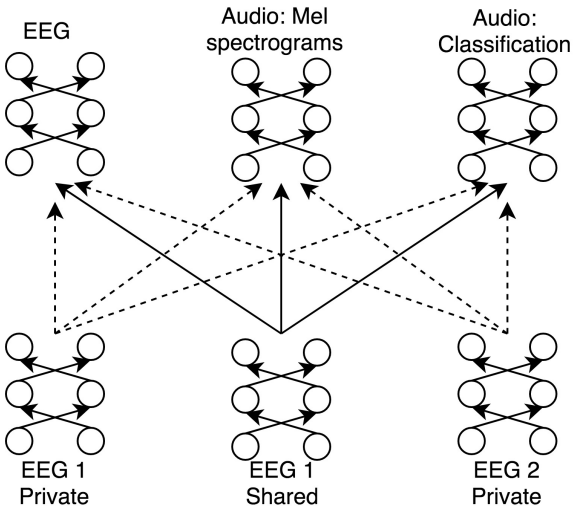

Figure 1. VCCA architecture for shared auditory concept and EEG representation learning. Latent variables parametrized by optional private encoders are indicated with dashed lines.

5.2 EEG encoder architectures

Both NMED-T and OpenMIIR speech EEG encoders featured 4 convolutional layers with filter numbers linearly ascending from 64 to 512 per layer. Convolution was performed on two dimensional inputs. Each column of the input represented the same linear concatenation of EEG channels for a single sample within the inputs of 1 s length. This resulted in inputs of size 512*64 for the OpenMIIR speech and 125*125 channels for the NMED-T inputs. The kernel size was set to [2x2] for all layers. Here and for all further kernel dimensions, we define the first index to be within the channel domain (or frequency for spectrograms) and the second within the temporal domain. Each convolutional layer was followed by 30 % dropout.

5.3 EEG and audio decoder architectures

We used similar EEG decoder architectures for both datasets. The OpenMIIR speech EEG decoder featured 6 hidden deconvolution layers with three layers of 16 and another three layers of 32 filters. The kernel size was set uniformly to [2x16] with stride 2 except for a [2x1] kernel in the third layer with stride 1. A final dense output layer consisted of 512*64 units. The decoder for the NMED-T dataset followed the same deconvolution architecture, except for kernels with dimension of [4x16] and [4x1] instead of [2x16] and [2x1]. A final dense layer consisted of 125*125 units. Both OpenMIIR speech and NMED-T decoders for Mel spectrograms consisted of four layers: Two deconvolution layers of 32 filters and two layers with 64 filters. As the length of Mel spectrograms mirrors those of the EEG excerpts, but in combination with a frequency resolution of 64 bins, the final dense layer featured 512*64 and 125*64 units respectively. The kernel dimensions were set to [4x8] uniformly, except for the fourth deconvolution layer of the OpenMIIR speech decoder, with a [2x8] kernel. The decoder for loudness envelope reconstruction consisted of a bidirectional LSTM layer with 128 hidden units, followed by a dense layer of size equal to the length of the audio excerpt. Finally, the decoder used for classification of the OpenMIIR speech dataset consisted of two hidden dense layers with 32 filters and a dense output layer of 1 unit. All internal units used Rectified Linear Unit (ReLU) activations, all output units had sigmoid activation. The size of the latent embedding was 128 units.

5.4 VCCA training and prediction

The extended VCCA model was trained both intra-subject and cross-subject in an end-to-end fashion purely on the perception trials using Adam optimization with a constant learning rate of 0.0001 [13]. For both datasets we used 60 % of available perception trials for training and another 20 % for validation. The remaining 20 % and the imagination trials were used for testing. All trials were shuffled randomly before training. For tests on imagination data, we evaluated both imagination trials whose corresponding perception trials were included in the training as well as entirely unknown trials. All models were trained up to saturation of the Mel spectrogram reconstruction loss, between 1000-2000 epochs. Reconstruction loss was computed as the mean squared error between reconstructions and targets.

5.5 Introspection

After training we inspected the learned latent space by linearly interpolating between multiple existing EEG inputs extracted either from the training or testing dataset. This way, we received embeddings for the given inputs as well as a fixed number of embeddings that connect them in the learned projection space. We then used the model to reconstruct the Mel spectrogram and EEG signal for the embeddings.

6. QUALITATIVE ANALYSIS OF MUSICAL STIMULUS RECONSTRUCTION

6.1 Perceived stimulus reconstruction

We were able to use the modified VCCA model to reconstruct the Mel spectrograms of perceived audio within both datasets at various levels of accuracy. Figure 2 shows exemplary reconstructions of speech and sine wave tone patterns for intra-subject training and testing on both trial types of the OpenMIIR speech dataset. The reconstructions are characterized by rhythmical and timbral alignment with the target. In some cases we noticed erroneous temporal shifts of the whole predicted rhythmical pattern within a reconstructed excerpt. Additional tests with smaller window sizes lead to a decrease in amount and size of such errors, while increasing the amount of false positive predictions of both sine wave and speech patterns. In some cases speech and sine wave patterns were mixed up, but still with correct temporal alignment of note onset positions between target and predictions. Figure 3 shows reconstructions after training on all subjects of the OpenMIIR speech dataset. Multi-subject training lead to results with improved temporal alignment of targets and predictions. Here, in more cases the two timbres (sine wave and speech pattern) were confused. This indicates that the correct prediction of the timbre is more subject-specific than the temporal and rhythmical aspects. Increasing the amount of training data for both trials enhanced the overall reconstruction quality, training only on the speech trials still lead to temporally meaningful reconstructions of the sine wave tone patterns. We found the stimulus reconstruction quality to be best when including 4 subjects for cross-subject training and testing.

Increasing the amount of dropout within the EEG decoder (up to 40 %) turned out to be crucial for reconstructions of comparable quality for trials in subjects that were excluded entirely from the training procedure. Training with randomized window start positions and using overlapping overlapping windows proved to enhance the reconstruction quality. This suggests that Mel spectrogram reconstruction quality for this dataset is limited by the amount of available training data.

Compared to the OpenMIIR dataset, the NMED-T dataset provided more training data with increased target complexity. The reconstructions showed different characteristic in visual inspection. Often times, the timbre reconstruction dominated the reconstruction of temporal aspects, especially in parts that featured multiple instruments or singing voice. In fewer cases, but within all songs, the onsets of percussion, speech or other sounds were reconstructed. For all trained models, timbre reconstruction was visible after around 500 epochs, while temporal aspects were learned at later stages. Figure 4 provides examples for reconstructed excerpts of the perceived full-length

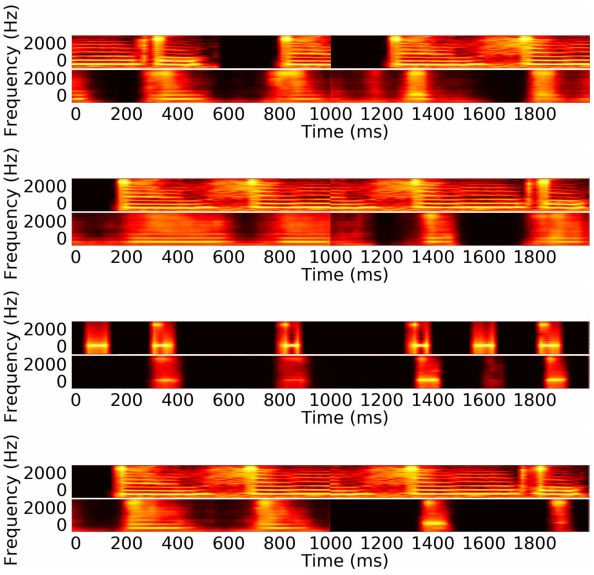

Figure 2. Mel spectrogram reconstructions of perceived rhythmical trials for the VCCA model trained on subject 'P13' of the OpenMIIR speech dataset. Target stimuli are presented above their reconstructions.

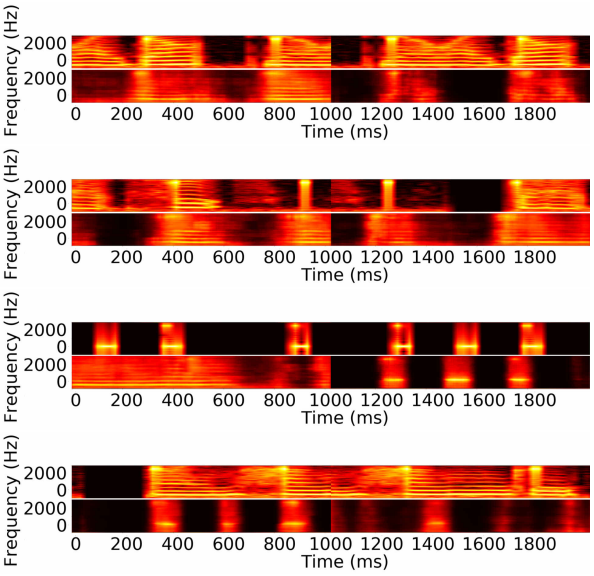

Figure 3. Mel spectrogram reconstructions of perceived rhythmical trials for the VCCA model trained on all subjects of the OpenMIIR speech dataset. Target stimuli are presented above their reconstructions.

songs contained in the NMED-T dataset. We found no substantial difference in the quality of reconstructions within subjects included into training and those from subjects excluded during training. This might be due to the small amount and long duration of 10 stimuli in combination with a single presentation per stimulus. Increasing the dropout rate after each convolutional layer in the EEG encoder over 30 % increased the models tendency to reconstruct temporal aspects, such as percussion onsets. Training sets with a larger amount of subjects generally

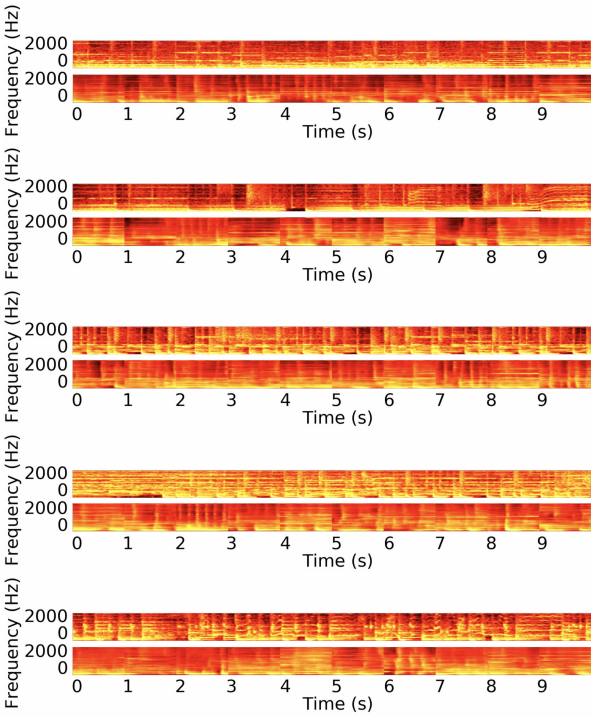

Figure 4. Excerpts of reconstructed Mel spectrograms from the NMED-T dataset. The target stimuli are shown above their reconstructions. The two top rows are based on training on all subjects. The three bottom rows are based on training on 10 subjects and testing on subjects that were excluded during training.

improved reconstruction quality. Furthermore, the introduction of overlapping EEG input windows increased the amount of reconstructed temporal features. Models trained for more than 2000 epochs showed more sparse reconstruction within the frequency domain. This indicates that adding more data and increasing training length can further increase the reconstruction quality for naturalistic music. Often times, the size of temporal misalignments was equal at all positions within reconstructed excerpts. This indicates that the reconstruction quality is dependent of the window size. Future work could test this assumption by simultaneously training on EEG or audio excerpts of various sizes within different encoders of the VCCA model. This would furthermore allow the representation of the latent concepts to include contexts of various size. For example, in the audio domain, such contexts could range from single note onsets to changes in song structure.

6.2 Imagined stimulus reconstruction

VCCA models trained on perceptual OpenMIIR speech data could be applied to imagination trial reconstruction. The reconstructed stimuli showed the same typical rhythmical patterns and could be divided into speech and sine wave predictions. However, the correct rhythmical predictions were less often visible and more blurry. It is important to note that the imagination was performed superimposed on a constant metronome click. This means, that only the difference

between the rhythmical structure and timbre was based on pure imaginative processes, while there were still perceptual cues for temporal alignment. Models trained on multi-subject perceptual data showed less blurry reconstructions. Adding private encoders with imagination based EEG signal did not cause a visible increase in reconstruction quality.

7. QUALITATIVE ANALYSIS OF LEARNED AUDITORY CONCEPTS

We found musically meaningful representations of the OpenMIIR speech stimuli in the latent space of models trained intra-subject as well as cross-subject. Both EEG signal from training and testing subsets could be used to produce continuous interpolation. Processing EEG inputs from both testing and training data sets and using the target audio stimuli as validation, we found continuous representation across the temporal, rhythmical and timbral domain. For any given input, we could change the temporal position of the rhythmical pattern as well as the timbre (within speech and sine wave tones). Furthermore, the latent space enabled interpolation between metronome clicks and the sine wave tones of increased loudness. However, this difference was found to a lesser degree with data of the test set. Figure 5 (a) shows an example for the interpolation between 3 embeddings based on EEG inputs of the OpenMIIR speech training data set. Here, interpolation between a syncopated and non-syncopated part of the rhythm was done while simultaneously shifting the temporal position of the rhythmical pattern within the reconstructed excerpt. The non-syncopated excerpt was further interpolated into its representation with speech signal. Figure 5 (b) shows topographic projections of the brain activity reconstructed for each embedding that was computed in Subfigure (a). For the sake of clarity we show six topographic plots out of the total amount of 512 per embedding. Qualitative comparison of the EEG signal with the original inputs indicated that overfitting the EEG data is not possible when we stop training when the audio reconstruction loss is saturated. For other use cases, higher quality EEG reconstructions could be achieved with different training procedures, such as unsupervised EEG reconstruction pretraining. Models with smaller latent embeddings sizes (e.g. 8 units) did still produce meaningful and continuous interpolations, but with more blurring across the temporal and frequency domains. The model forces EEG and audio to be shared even in these smaller latent spaces. The neuroscientific meaningfulness of the EEG reconstructions might further be validated in future work, for example with shared fMRI representation in private encoders.

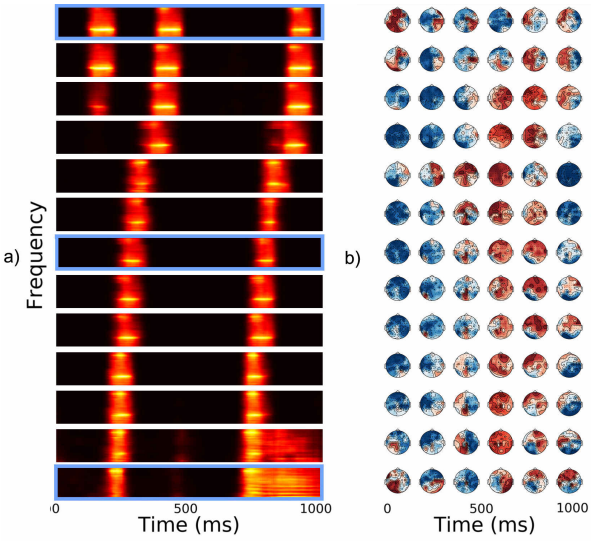

Figure 5. (a) Reconstructed Mel spectrograms after interpolation in the learned latent space learned for Subject 'P13' of the OpenMIIR speech dataset. Embeddings that correspond to real EEG inputs are framed. (b) Topographic visualization of the reconstructed temporal brain activity. Each row represents the brain activity reconstructed for the embedding in the same row of Subfigure (a).

8. CONCLUSIONS

In this paper, we presented the application of a multi-view generative model for shared auditory concept learning and musical stimulus reconstruction from EEG signals. We showed that the model can learn representations of simple rhythm and timbre related concepts that are shared in audio and EEG data. Furthermore, we could see first successes in approaching naturalistic music and imagined stimulus reconstruction. The presented framework is designed to be expandable to additional modalities, such as fMRI data, or additional reconstruction targets, such as emotional aspects of music cognition. In combination with the ability to perform introspection on the shared representation of stimuli and electrophysiological responses, the model can be an aid for future EEG based music information retrieval and research in music cognition.

9. ACKNOWLEDGMENTS

The OpenMIIR speech dataset was kindly shared by the Music and Neuroscience Lab at Western University in London, Ontario. The authors especially would like to thank Avital Sternin who recorded the data. This research has been funded by the Federal Ministry of Education and Research of Germany (BMBF).

10. REFERENCES

[1] A. Aroudi, B. Mirkovic, M. De Vos, and S. Doclo. Auditory attention decoding with eeg recordings using noisy acoustic reference signals. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 694–698. IEEE, 2016.

[2] S. Chandar, M. M. Khapra, H. Larochelle, and B. Ravindran. Correlational neural networks. *Neural computation*, 28(2):257–285, 2016.

[3] L. K. Cirelli, D. Bosnyak, F. C. Manning, C. Spinelli, C. Marie, T. Fujioka, A. Ghahremani, and L. J. Trainor. Beat-induced fluctuations in auditory cortical beta-band activity: using eeg to measure age-related changes. *Frontiers in psychology*, 5:742, 2014.

[4] C. Du, C. Du, and H. He. Sharing deep generative representation for perceived image reconstruction from human brain activity. In *Neural Networks (IJCNN), 2017 International Joint Conference on*, pages 1049–1056. IEEE, 2017.

[5] Y. Fujiwara, Y. Miyawaki, and Y. Kamitani. Modular encoding and decoding models derived from bayesian canonical correlation analysis. *Neural computation*, 25(4):979–1005, 2013.

[6] A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, et al. Meg and eeg data analysis with mne-python. *Frontiers in neuroscience*, 7:267, 2013.

[7] K. Hoenig, C. Müller, B. Herrnberger, E.-J. Sim, M. Spitzer, G. Ehret, and M. Kiefer. Neuroplasticity of semantic representations for musical instruments in professional musicians. *NeuroImage*, 56(3):1714–1725, 2011.

[8] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

[9] E. Jones, T. Oliphant, and P. Peterson. Scipy: open source scientific tools for python. 2014.

[10] I. Kavasidis, S. Palazzo, C. Spampinato, D. Giordano, and M. Shah. Brain2image: Converting brain signals into images. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 1809–1817. ACM, 2017.

[11] M. Kiefer and L. W. Barsalou. 15 grounding the human conceptual system in perception, action, and internal states. *Action science: Foundations of an emerging discipline*, page 381, 2013.

[12] M. Kiefer, E.-J. Sim, B. Herrnberger, J. Grothe, and K. Hoenig. The sound of concepts: four markers for a link between auditory and conceptual brain systems. *Journal of Neuroscience*, 28(47):12224–12230, 2008.

[13] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[14] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[15] S. Losorelli, D. T. Nguyen, J. P. Dmochowski, and B. Kaneshiro. Nmed-t: A tempo-focused dataset of cortical and behavioral responses to naturalistic music. ISMIR, 2017.

[16] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, pages 18–25, 2015.

[17] M. Meyer, S. Baumann, and L. Jancke. Electrical brain imaging reveals spatio-temporal dynamics of timbre perception in humans. *Neuroimage*, 32(4):1510–1523, 2006.

[18] N. Moran. Social implications arise in embodied music cognition research which can counter musicological individualism. *Frontiers in psychology*, 5:676, 2014.

[19] R. Näätänen and T. Picton. The n1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. *Psychophysiology*, 24(4):375–425, 1987.

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

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

[22] A. Ofner. Reconstruction of perceived and imagined music from eeg recordings with deep neural networks. In *Proceedings of the MEi:CogSci Conference 2017*, page 53.

[23] F. Raposo, D. M. de Matos, R. Ribeiro, S. Tang, and Y. Yu. Towards deep modeling of music semantics using eeg regularizers. *arXiv preprint arXiv:1712.05197*, 2017.

[24] R. S. Schaefer, P. Desain, and P. Suppes. Structural decomposition of eeg signatures of melodic processing. *Biological psychology*, 82(3):253–259, 2009.

[25] A. Shahin, L. E. Roberts, C. Pantev, L. J. Trainor, and B. Ross. Modulation of p2 auditory-evoked responses by the spectral complexity of musical sounds. *Neuroreport*, 16(16):1781–1785, 2005.

[26] A. Sternin, S. Stober, J. Grahn, and A. Owen. Tempo estimation from the eeg signal during perception and imagination of music. In *1st International Workshop on Brain-Computer Music Interfacing/11th International*

Symposium on Computer Music Multidisciplinary Research (BCMI/CMMR15), 2015.

[27] S. Stober, D. J. Cameron, and J. A. Grahn. Using convolutional neural networks to recognize rhythm stimuli from electroencephalography recordings. In *Advances in neural information processing systems*, pages 1449–1457, 2014.

[28] 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*, pages 763–769, 2015.

[29] D. Stuss, A. Toga, J. Hutchison, and T. Picton. Feedback evoked potentials during an auditory concept formation task. In *Progress in brain research*, volume 54, pages 403–409. Elsevier, 1980.

[30] M. S. Treder, H. Purwins, D. Miklody, I. Sturm, and B. Blankertz. Decoding auditory attention to instruments in polyphonic music using single-trial eeg classification. *Journal of neural engineering*, 11(2):026009, 2014.

[31] R. Vigo, M. Barcus, Y. Zhang, and C. Doan. On the learnability of auditory concepts. *The Journal of the Acoustical Society of America*, 134(5):4064–4064, 2013.

[32] W. Wang, X. Yan, H. Lee, and K. Livescu. Deep variational canonical correlation analysis. *arXiv preprint arXiv:1610.03454*, 2016.