MODELING PERCEPTION WITH HIERARCHICAL PREDICTION: AUDITORY SEGMENTATION WITH DEEP PREDICTIVE CODING LOCATES CANDIDATE EVOKED POTENTIALS IN EEG

André Ofner and Sebastian Stober
Otto von Guericke University, Magdeburg, Germany
ofner@ovgu.de stober@ovgu.de

ABSTRACT

The human response to music combines low-level expectations that are driven by the perceptual characteristics of audio with high-level expectations from the context and the listener's expertise. This paper discusses surprisal based music representation learning with a hierarchical predictive neural network. In order to inspect the cognitive validity of the network's predictions along their time-scales, we use the network's prediction error to segment electroencephalograms (EEG) based on the audio signal. For this, we investigate the unsupervised segmentation of audio and EEG into events using the NMED-T dataset on passive natural music listening. The conducted exploratory analysis of EEG at locations connected to peaks in prediction error in the network allowed to visualize auditory evoked potentials connected to local and global musical structures. This indicates the potential of unsupervised predictive learning with deep neural networks as means to retrieve musical structure from audio and as a basis to uncover the corresponding cognitive processes in the human brain.

1. INTRODUCTION

Studying the human perception of music has received increased interest in Music Information Retrieval (MIR). As humans solve tasks such as beat tracking, genre identification or musical prediction with ease, many MIR methods rely on computational models inspired by human perception. At the same time, studying the brain's response to auditory stimuli is still limited by the lack of resources that map complex musical stimuli to neural processes. Studies in cognitive neuroscience and brain computer interfacing (BCI) on auditory evoked brain states require labor intensive manual preparation and often focus on isolating particular brain responses using sparse stimuli presented individually [1, 2]. While datasets on brain states evoked by natural music exist, they often lack fine-grained annotations of the event structure and corresponding neural activity [3–5]. This entails a demand for efficient and unsupervised mapping techniques between natural music and evoked brain states. Furthermore, there is a need for biologically plausible and multi-modal models for such mapping, as induced brain states are a mixture of stimulus-derived and subjective, cognitive or contextual factors.

We address these challenges with predictive coding, one of the most dominant theoretical frameworks of human perception [6, 7]. Predictive coding offers a comprehensive description of how humans parse and predict sounds and map auditory stimuli to musically meaningful and hierarchically organized units [8]. In predictive coding, the neural response to music is shaped by hierarchically organized expectations [7]. This hierarchy of expectations connects predictions about low-level auditory features to more global context, such as the listener's musical expertise or levels of entrainment during listening [8]. The underlying dependencies between expectancy and uncertainty in predictive coding are particularly interesting in the context of music perception, as music perception can be described as continuously resolving uncertainty and forming new expectations [9–11]. This is in line with evidence on the predictive nature of human music perception, especially within studies on unexpected stimulus deviations and the influence of the listener's expectancy on attention and perceptual precision [9, 11].

Predictive coding offers an efficient algorithmic motif that allows unsupervised learning. Learning in predictive coding systems can be seen as a hierarchy of predictive modules that form predictions over various temporal scales. These predictions can either be about future states in the stimulus domain or about the future of internal states of the systems and are often cast in the context of Bayesian (i.e. probabilistic) inference [12]. In this hierarchical generative model of perception, long-term expectations from temporally stable aspects of music, such as genre or tempo form top-down predictions about the activity of layers closer to the actual auditory information [8]. By propagating the deviations between predictions and observations, the generative model and with that the model of the processed stimulus is updated [7].

Here, we connect predictive coding as a algorithmic motif for unsupervised stimulus representation with deep neural networks and recurrent variational inference in order to segment natural music into units that are musically meaningful. Following the assumption that hierarchical

predictive coding of music explains a substantial amount of evoked brain states, we analyse the retrieved musical structure in terms of the induced neural activity in electroencephalographic signal (EEG). Using the Naturalistic Music EEG Dataset—Tempo (NMED-T) of passive listening to natural music we the demonstrate that the approach allows to locate and visualize event related potentials (ERPs) on local and global scale [3].

2. RELATED WORK

Within the field of MIR, the capacity of predictive coding algorithms to compress and represent auditory information on the sensory level has been exploited for various tasks such as speech re-synthesis or audio compression since many years [13,14]. The human brain, however, augments such low-level sensory representations with a hierarchy of more abstract, semantic predictions from other brain areas [7]. This aspect of hierarchical predictive learning has found traction in the domain of deep neural networks (DNNs), but so far has been applied mostly to images and video processing [15, 16]. Furthermore, most popular implementations of deep predictive coding often only rely on non-linear transformation of the sensory error and not yet abstract away from pure sensory prediction. Autoregressive modeling of audio has seen tremendous progress in recent years, with a plethora of models performing tasks such as sample level audio prediction or speech synthesis, often with impressive results [17–19]. However, such autoregressive models are computationally expensive and sample-level models still tend to struggle with incorporating more abstract and long-term musical features.

2.1 Auditory evoked potentials and musical structure

Recent years have shown a variety of approaches to studying the human brain's response to auditory stimuli, especially with functional magnetic resonance imaging (fMRI) and electroencephalography (EEG). EEG is especially adequate in the context of music due to its higher temporal resolution. A multitude of auditory features, such as loudness, frequency, tempo and rhythm have been traced in EEG recordings of brain activity during music perception [20–23]. Next to these stimulus-derived aspects, recorded brain activity has further been analysed with respect to more contextual aspects of music perception, such as the listener's attention, which is modulated by aspects such as expertise or engagement [24]. Two extensively researched aspects of the neural response underlying perception potentials are event-related potentials (ERPs) and steady-state evoked potentials (SSEPs) [25, 26]. ERPs and SSEPs differ mainly in their temporal scope: While ERPs are aligned to a single loation (typically the onset of a particular event), SSEPs show frequency alignment to stimulus periodicity over longer time frames [27]. For ERPs, the brain response aligned to the event type of interest is analysed after averaging large amounts of trials [28]. Auditory event-related potentials (AEPs) are modulated by aspects such as rhythm, pitch, timbre or the duration of musical events, all of which play an important role in human audio segmentation [25,29–33]. Many of these evoked potentials have been explained in the context of predictive coding as a mixture of bottom-up and top-down mechanisms that are modulated both contextual expectations and the auditory stimulus itself [34, 35]. Similar to ERPs, SSEPs are inspected after averaging over many trials, but don't require zero valued phase offset between stimulus and response. Instead, SSEPs characterize periodic mappings between auditory features and brain response, such as phase locking to perceived frequencies or loudness envelopes. Both ERPs and SSEPs can be related to predictive cognitive processes aiming at structuring the incoming sensory signal into meaningful events in a hierarchical fashion [35, 36].

3. A HIERARCHICAL PREDICTIVE CODING MODEL FOR MUSIC

Predictive coding describes hierarchical predictions of sensory states and hidden states of the network across various time-scales. Sample based predictions about audio requires a model with high temporal resolution that captures the causal dependencies between adjacent samples. Thus, a desired predictive coding model for audio connects low-dimensional predictions over many time-steps with fine-grained predictions at the sensory level. Transforming audio features to high-level representations is a complex task, which is often solved with the non-linear processing found in DNNs. We approach these requirements with a recurrent DNN that generates autoregressive predictions based on long short-term memory (LSTM) [37]. Instead of predicting individual frames, we process mel spectrogram representations of audio. The reduced temporal resolution of spectrograms helps reducing the computational complexity while still capturing fine-grained auditory information. As spectrograms extend into time and frequency, we employ convolutional neural networks (CNNs) to extract features from the spectrograms.

3.1 Autoregressive predictive coding

In order to enable hierarchical predictions across multiple time-scales, we stack multiple LSTM layers and allow each layer to predict the future states of the next lower layer. In line with Bayesian views on brain function and research on the effectiveness of probabilistic recurrent modelling, we express the current state in each layer as Gaussian prior distributions, parameterized by mean and variance parameters [12, 38]. While the lowest layer predicts future audio signal, the network's hidden layers predict future states of the lower layer's representations. More specifically, we first sample the prior distribution of each layer and transform the resulting activation with a convolutional decoder network. The decoder of the lowest layer parameterizes the prediction of expected next spectrogram input window. The decoders in hidden layers output predictions about the mean and variance parameters of the next lower layer. In contrast to the related class of recurrent variational autoencoders (VAE), we do not employ an

encoder network that directly transfers observations to a posterior distribution [39]. Instead, the network processes only the deviation e_t between predicted p_t and observed values o_t with a error encoder network:

$$e_t = p_t - o_t \tag{1}$$

3.2 Variational inference with deep neural networks

With the previously introduced decoder, error encoder and recurrence networks, the model can be trained to perform variational inference by constructing a variational bound on the data log-likelihood. More specifically, the model is trained to maximize the evidence for its current inferred state, the network's "belief" about what causes an observation. Mathematically speaking, maximizing the model evidence can be expressed as minimizing the complexity of the model's generative model while providing maximally accurate predictions for future audio inputs. The model thus reduces the complexity of states with respect to observations o_t and states s_t at a discrete time step t:

$$Complexity = E_{q(s_{t-1}|o_{\leq t-1})}[KL[q(s_t|o_{\leq t})\|p(s_t|s_{t-1})]] \tag{2}$$

Simultaneously, the accuracy of predicted observations maximized:

$$Accuracy = E_{q(s_t|o_{\leq t})}[\ln p(o_t|s_t)] \tag{3}$$

The observations in the lowest (sensory) layer refer to the observed audio, while observations in the hidden layers refer to the observed state posteriors in terms of mean and variance. The model optimizes both terms simultaneously for all layers. For this the approximate state posteriors $q(s_{1:T}|e_{1:T}) = \prod_{t=1}^{T} q(s_t|e_{t-1})$ are inferred by filtering past prediction errors $\{e_t\}_{t=1}^{T}$. By selecting a pair of adjacent layers and minimizing the accuracy and complexity term between them, this structure allows to form predictions that are consistent between layers, i.e. show small or no top-down prediction error. Such a design prevents error propagation across many layers in a single step. This is a potential drawback and could be improved in future iterations. Throughout all experiments, we used a network with three predictive coding layers. We model $q(s_t|e_{t-1})$ as diagonal Gaussian for all layers with mean and variance parameterized by a convolutional neural network (CNN) with two layers of 64 and 128 units each. The convolutional layers were followed by a dense network of 1024, 512 and 256 units respectively.

3.3 Deterministic transitions with a probabilistic step

The LSTM states of the network are conditioned on its previous states $\{h_t\}_{t=1}^{T}$, top-down predictions $\{e\_td_t\}_{t=1}^{T}$ from the next higher layer as well as the prediction error of the last outgoing prediction, the bottom-up prediction error $\{e\_bu_t\}_{t=1}^{T}$. At each time-step, the bottom-up prediction error is forced to pass a sampling step when updating the prior to the posterior distribution. This means that any incoming sensory information $\{e\_bu_t\}$ must pass

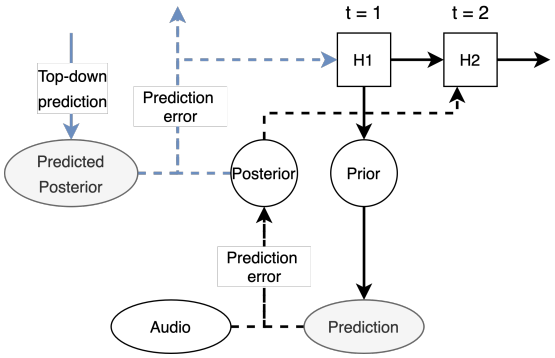


Figure 1. Transitions in a single layer of the predictive coding network. The black pathways show transitions in the lowest layer of the network. Predictions about future audio are conditioned on past states and prediction errors. The blue pathways indicate top-down posterior predictions, which allow to predict the states of lower layers in the network in terms of mean and variance parameters. Multi-step predictions can be generated by updating the recurrent states without sampling new observations.

a stochastic step before being integrated into the deterministic memory states $\{h_t\}$. Figure 1 shows an overview of the transitions in a single layer and the connection to the top-down predictive pathway.

3.4 Model training

The model is trained using a timestep-wise variational evidence lower bound (ELBO) that combines the previously introduced complexity (2) and accuracy (1) terms. Similarly to the objective function in recurrent VAEs [40], the model maximes the ELBO for the approximate posteriors in each layer by accumulating evidence over past timesteps:

$$ELBO(q) = \sum_{t=1}^{T} \Big(o_{q(s_t|o_{\leq t})}[\ln p(o_t|s_t)] - o_{q(s_{t-1}|o_{\leq t-1})}[KL[q(s_t|o_{\leq t})\|p(s_t|s_{t-1})]] \Big) \tag{4}$$

This structure can be viewed as a hierarchical Kalman filter, making the connection to predictive coding as a Bayesian update scheme or generalized Kalman filtering apparent. We used ReLU activations for all CNNs and hyperbolic tangent activations for the decoder's output layer [41]. In each layer, the prediction error was computed with respect to positive and negative prediction error. Each layer was then ReLU activated before propagation to the encoder networks. For all presented experiments, we trained the model to convergence of the input layer reconstruction loss. For this, we used the Adam optimizer with a learning rate of 10^{-3} [42]. The KL divergence terms for each layer were scaled proportionally to the prediction errors. Furthermore, we weighted the reconstruction losses by 2:1:1 for the employed three layer model.

4. THE NMED-T AND FMA DATASETS

We used the Naturalistic Music EEG Dataset—Tempo (NMED-T) for the evaluations in all presented experiments [3]. NMED-T features EEG recordings from 10 commercially available music pieces, with durations between 270 and 300 seconds, spanning 55 to 150 BPM in various genres. 20 participants were allowed to freely and passively listen to the music, without any additional cognitive load. We used the provided preprocessed version of the EEG data at a sampling rate of 125 Hz. For all presented ERP experiments, we re-referenced the EEG data to the average of all 125 EEG channels and filtered out background noise using a Savitzky-Golay filter before averaging the evoked responses. For network training, we resorted to the "small" partition of the Free Music Archive (FMA) dataset, featuring 8000 songs with 30 seconds duration [43]. We computed magnitude spectrograms for all ten provided audio files of the NMED-T dataset and the FMA audio files before mapping to the mel scale, resulting in mel spectrograms at 125 Hz, equal to the EEG sampling rate. All audio processing steps were done with the librosa library [44]. We tested different mel spectrogram lengths as inputs to the lowest network layer and found lengths between 50 and 150 ms to be the sweet spot with low computation time and without quick overfitting.

5. EXPERIMENTS

For all following experiments, network training was done first on the FMA dataset followed by a evaluation phase using the NMED-T stimuli. After training on the FMA audio, we froze the network weights and processed the NMED-T audio to generate predictions and corresponding prediction errors. For each processed NMED-T audio stimulus we extracted both positive (PPE) and negative (NPE) valued prediction errors. In this context, PPEs refer to areas where the model predictions are lower than the observed threshold, while NPEs refer to predictions that are higher than the actual values. Predictions were computed in a single pass over each song, i.e. without repeated inference of the current musical context. However, such "active learning" or "active inference" schemes could be explored in the future.

5.1 Deriving segmentation boundaries from prediction errors

In order to inspect the effect of predictive coding at the audio level, we first deactivated the recurrent parts of the lowest layer, forcing the model to express next states as a function of previous observation and the top-down prediction. For model evaluation, we extracted positive and negative prediction errors from each layer of the network. In all layers, we a applied a magnitude threshold to pick peaks from the continuous error response, followed by a peak-picking step that ignores repeated error peaks in a sliding window of fixed size. Both magnitude and window size could be learned by the network itself, leaving the room for more complex and self-supervised segmentation techniques. All presented experiments use the mean of positive and negative prediction errors, if not further specified. Figure 2 shows two examples for input and predicted audio as well as the corresponding prediction errors and selected peaks. The examples illustrate that autoregressive predictive coding decorrelates large parts of the processed audio in the first layer, by reducing the redundancies in the signal using non-linear weighted predictions based on the past values. This is in line with the spatial and temporal whitening effects described by Rao et al. in the context of center–surround receptive fields in the retina [7]. For the following experiments, we use these sensory predictions to derive segmentation boundaries and explore temporally aligned ERPs in the brain.

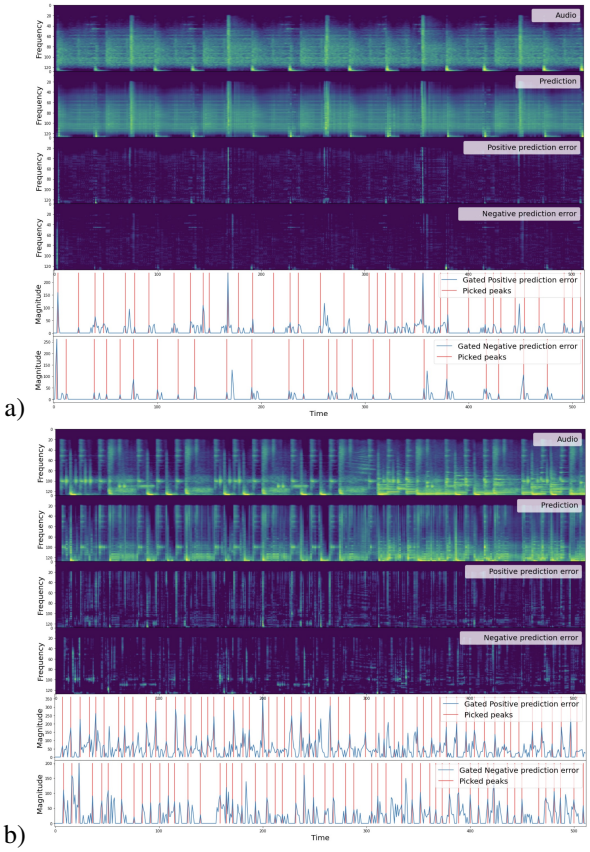


Figure 2. Predicted audio and positive and negative prediction errors in the first predictive coding layer for songs with a) 55 and b) 108 BPM. The model generates local predictions about inputs in a sliding window of 50 ms size. This autoregressive and non-linear process removes temporal redundancy in the residual error response. The bottom rows show the thresholded prediction error and picked peaks.

Increasing the weight of the prediction errors in the hidden layer decreased the error magnitudes. This is expected, as the network now learns to include more global temporal context over multiple steps of the lower layer. Ideally, the network learns to predict the rhythmic and timbral structure perfectly and successfully suppresses the prediction error in the first and second layer. If the recurrent parts are

active in the lowest layer, the long-term temporal dependencies can be memorized in the first layer additionally. In our experiments we found that (with a fixed weighting of the prediction errors between layers) deactivating the recurrence in the lowest layer is essential to learning predictive representations in the hidden layers. As visible in Figure 2, the tempo of the song as well as the rhythmic density have influence on the effectiveness of input decorrelation in the lowest layer.

5.2 Grand average ERP

To inspect the possibility to detect ERP events based on the sensory surprise, we extracted the prediction errors from the lowest predictive coding layer and averaged the EEG signal over all trials in all songs and subjects. We were able to derive a total of 242960 trials within 10 songs and 20 subjects using the proposed method. This equates 22140 to 28740 trials per song and between 1108 and 1437 unique event locations per song.

Figure 3 a) shows the grand average ERP for all ten songs in the NMED-T dataset at locations of prediction errors peaks. In comparison to the tempi reported in the orginal NMED-T paper, we sorted the songs between 83 and 151 BPM using beat tracking in the librosa library. The difference between our tempo measures and the ones in the original paper can be explained as being multiples of each other, e.g. 110 BPM being a multiple of 55 BPM. The averaged ERP shows an activity peak for positively correlated channels at the predicted event location, followed by a negative peak around 60 ms after onset. The grand average ERP further shows two smaller peaks around 120 and 170 ms after onset, indicating the presence of surrounding onsets with variable latency. The reduced magnitude of these delayed peaks can be explained by the differences in tempo between songs.Specifically, the difference in peak size between activity close to the predicted onsets and those with greater temporal distance indicates a separation between tempo-independent components (close to the prediction error peak) and attenuated tempo-dependent components. Figure 3 b) shows the grand average ERP in five positively activated channels, sorted by the prediction error magnitude. The magnitude of the first evoked peak after stimulus onset grows proportional with the error magnitude for large error values. For smaller prediction error values, the response shows larger latency. Peaks with similar latency of the evoked activity have magnitudes proportional to the prediction error magnitude. This fits with the assumption that the grand average ERP shows temporally variable peaks induced by differences in tempo.

5.3 Evaluating song-level segmentation with low frequency EEG

Next to inspecting the predicted ERP responses with the local predictions of the input layer, we want to inspect the possibility to segment stimuli on the song level with the model. For this, we repeat the unsupervised training of the previous experiments, but weight the prediction errors in all layers equally after pretraining for 100000 updates.

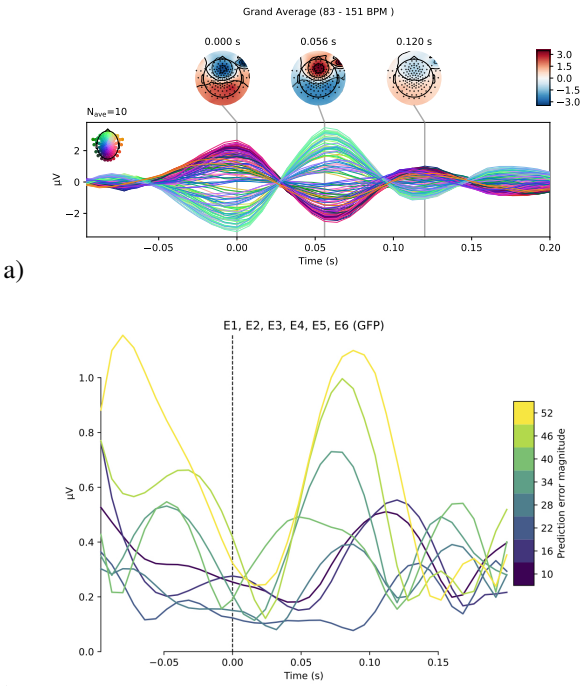
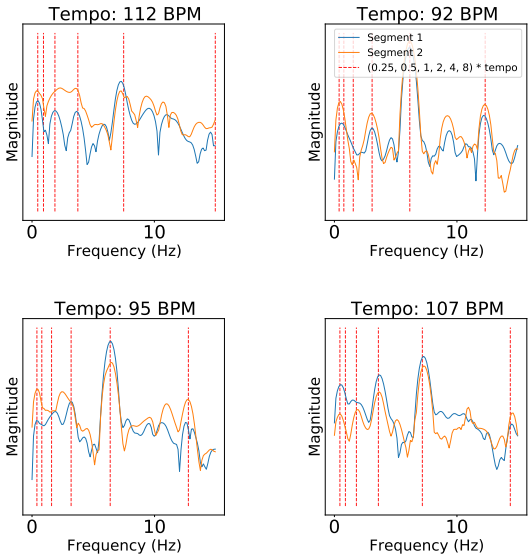


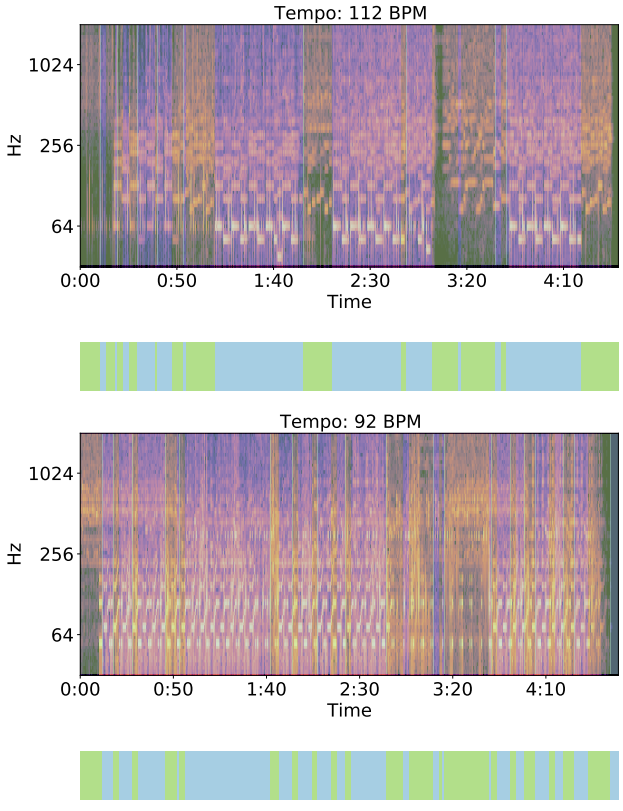
Figure 3. a) Grand average ERP for all songs in the NMED-T dataset at locations of prediction errors peaks generated by the predictive coding network. b) Grand average ERP in five positively correlated channels for trials sorted after the prediction error magnitude of the predictive coding network.

This approach puts more focus on the temporal consistency of the predictions in the hidden layers. Furthermore, we train with multi-step predictions of length 8, i.e. prediction errors are generated with respect to 8 future states at a time. This follows the assumption that both multi-step predictions and increased weighting of the hidden layer prediction errors increase the network's tendency towards more global predictions. In order to evaluate the ability to retrieve meaningful musical structure with the network's predictions, we extracted the timings of prediction errors from the lowest predictive coding layer and used them as starting points of EEG epochs, subsequently averaging the EEG signal over all epochs within each segmented class. Following previous work that illustrates differences in beat processing with SSEPs, we inspect averages of low frequency EEG to detect changes in beat processing or entrainment between the segmented classes derived from prediction errors in the predictive coding network [3]. For this, we use the same epoched data derived from locations computed in the previous experiments but average over all epochs within the bounds of each segmented class.

Here, we want to inspect whether changes in network prediction triggered by peaks in prediction errors show changes that are detectable with SSEPs. To generate binary segmentation, we threshold the prediction error like in the previous steps with a fixed value for each song and switch between segmentation masks when the positive error surpasses the negative error and vice versa. Both



a) SSEPs in low frequency EEG within the segments derived from gated prediction errors of the predictive coding network. Indicated with dashed lines are multiples of the song tempo, ranging from 1 to 16 Hz. Differences between the peaks in the power spectrum of both segments indicate different rhythmic processing between the two segmented classes.



b) Corresponding audio and temporal binary segmentation of two songs derived from gated prediction errors after training the proposed network for unsupervised multi-step prediction. Only the lowest 6 octaves are included for illustration purposes.

Figure 4. Segmented audio and evoked SSEPs in low frequency EEG of the NMED-T dataset.

down-sampling the inputs or increasing to length of multi-step predictions leads to more coarse grained segmentation. We found that using hop lengths up to as much as 16000 successive frames during spectrogram computation were suited to generate song-level segments while simultaneously reducing computation time. Intuitively speaking, the changes in hidden prediction error magnitude reflect the "mid-level" surprise of the network, as the pure sensory surprise is largely minimized in the input layer and the residual errors are further propagated. Future iterations of the model could use learnable error thresholds for improved and self-supervised segmentation. To help visualize the effect of segmentation we reduced spatial EEG dimensionality using Principal Components Analysis (PCA) before averaging the data and analyzed only the first component. Figure 4 a) shows the induced SSEPs in the magnitude of low-frequency EEG for selected songs. Audio and segmentation boundaries for two of these songs are displayed in Figure 4 b). Visible are peaks in the low frequency EEG components within all segmented parts that are aligned with multiples of the song tempo. In most processed songs the noticeable magnitude shifts go along with a stable distribution of the frequencies of evoked peaks, indicating rhythmic differences between the annotated segments which are embedded into the same global tempo.

6. DISCUSSION

This paper explored deep predictive coding for unsupervised audio representation learning inspired by human cognition. We compared the network's prediction errors with evoked potentials in EEG. For this, we related the hierarchical predictions of the model on ten naturalistic musical pieces to onset-aligned evoked potentials captured in EEG. We derived locations for individual musical events from the sensory surprise and inspected steady-state evoked potentials that capture rhythmic differences in the segmented songs. The employed model combines deterministic sequential predictions with probabilistic representations. While the deterministic parts allow to learn regularities over time-scales, the probabilistic elements lessens overfitting and helped shortening training duration. While sensory-level predictions can be employed for local event annotations, the predictions and prediction errors in hidden layers target higher levels of temporal abstraction.

Our results indicate the usefulness of predictive coding for the retrieval of events across the local and global structure of musical works. The model allows to approach audio segmentation jointly with structuring recorded brain activity, forming a basis for retrieval of information about cognitive processes in music perception. This offers an appealing method for researching auditory evoked potentials, as it eases the mapping between stimulus characteristics and connected evoked potentials across time-scales. Future improvements could enhance the capacity of the model, e.g. by allowing the model to segment inputs based on learned error gating.

7. REFERENCES

[1] P. Paavilainen, C. Kaukinen, O. Koskinen, J. Kylmälä, and L. Rehn, "Mismatch negativity (MMN) elicited by abstract regularity violations in two concurrent auditory streams," *Heliyon*, vol. 4, no. 4, 2018.

[2] I. Nambu, M. Ebisawa, M. Kogure, S. Yano, H. Hokari, and Y. Wada, "Estimating the intended sound direction of the user: toward an auditory brain-computer interface using out-of-head sound localization," *PloS one*, vol. 8, no. 2, 2013.

[3] 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.

[4] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.

[5] S. Stober, A. Sternin, A. M. Owen, and J. A. Grahn, "Towards music imagery information retrieval: Introducing the openmiir dataset of EEG recordings from music perception and imagination." in *ISMIR*, 2015, pp. 763–769.

[6] L. Aitchison and M. Lengyel, "With or without you: predictive coding and bayesian inference in the brain," *Current opinion in neurobiology*, vol. 46, pp. 219–227, 2017.

[7] R. P. Rao and D. H. Ballard, "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects," *Nature neuroscience*, vol. 2, no. 1, pp. 79–87, 1999.

[8] P. Vuust and M. A. Witek, "Rhythmic complexity and predictive coding: a novel approach to modeling rhythm and meter perception in music," *Frontiers in psychology*, vol. 5, p. 1111, 2014.

[9] S. L. Denham and I. Winkler, "Predictive coding in auditory perception: challenges and unresolved questions," *European Journal of Neuroscience*, vol. 51, no. 5, pp. 1151–1160, 2020.

[10] M. Heilbron and M. Chait, "Great expectations: is there evidence for predictive coding in auditory cortex?" *Neuroscience*, vol. 389, pp. 54–73, 2018.

[11] S. Koelsch, P. Vuust, and K. Friston, "Predictive processes and the peculiar case of music," *Trends in Cognitive Sciences*, vol. 23, no. 1, pp. 63–77, 2019.

[12] K. Friston and S. Kiebel, "Predictive coding under the free-energy principle," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 364, no. 1521, pp. 1211–1221, 2009.

[13] B. S. Atal and M. R. Schroeder, "Adaptive predictive coding of speech signals," *Bell System Technical Journal*, vol. 49, no. 8, pp. 1973–1986, 1970.

[14] G. Schuller and A. Hännä, "Low delay audio compression using predictive coding," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2. IEEE, 2002, pp. II–1853.

[15] W. Lotter, G. Kreiman, and D. Cox, "Deep predictive coding networks for video prediction and unsupervised learning," *arXiv:1605.08104*, 2016.

[16] K. Han, H. Wen, Y. Zhang, D. Fu, E. Culurciello, and Z. Liu, "Deep predictive coding network with local recurrent processing for object recognition," in *Advances in Neural Information Processing Systems*, 2018, pp. 9201–9213.

[17] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv:1609.03499*, 2016.

[18] A. v. d. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. v. d. Driessche, E. Lockhart, L. C. Cobo, F. Stimberg *et al.*, "Parallel wavenet: Fast high-fidelity speech synthesis," *arXiv:1711.10433*, 2017.

[19] Y.-A. Chung, W.-N. Hsu, H. Tang, and J. Glass, "An unsupervised autoregressive model for speech representation learning," *arXiv:1904.03240*, 2019.

[20] S. Nozaradan, I. Peretz, M. Missal, and A. Mouraux, "Tagging the neuronal entrainment to beat and meter," *Journal of Neuroscience*, vol. 31, no. 28, pp. 10 234–10 240, 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*, vol. 32, no. 49, pp. 17 572–17 581, 2012.

[22] 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*, vol. 5, p. 742, 2014.

[23] 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*, vol. 11, no. 2, p. 026009, 2014.

[24] 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*. IEEE, 2016, pp. 694–698.

[25] G. Plourde, "Auditory evoked potentials," *Best Practice & Research Clinical Anaesthesiology*, vol. 20, no. 1, pp. 129–139, 2006.

[26] S. Morgan, J. Hansen, and S. Hillyard, "Selective attention to stimulus location modulates the steady-state visual evoked potential," *Proceedings of the National Academy of Sciences*, vol. 93, no. 10, pp. 4770–4774, 1996.

[27] P. M. Picciotti, S. Giannantonio, G. Paludetti, and G. Conti, "Steady state auditory evoked potentials in normal hearing subjects: evaluation of threshold and testing time," *Orl*, vol. 74, no. 6, pp. 310–314, 2012.

[28] T.-P. Jung, S. Makeig, M. Westerfield, J. Townsend, E. Courchesne, and T. J. Sejnowski, "Analyzing and visualizing single-trial event-related potentials," in *Advances in neural information processing systems*, 1999, pp. 118–124.

[29] T. W. Picton, *Human auditory evoked potentials*. Plural Publishing, 2010.

[30] 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*, vol. 24, no. 4, pp. 375–425, 1987.

[31] R. S. Schaefer, P. Desain, and P. Suppes, "Structural decomposition of EEG signatures of melodic processing," *Biological psychology*, vol. 82, no. 3, pp. 253–259, 2009.

[32] M. Meyer, S. Baumann, and L. Jancke, "Electrical brain imaging reveals spatio-temporal dynamics of timbre perception in humans," *Neuroimage*, vol. 32, no. 4, pp. 1510–1523, 2006.

[33] 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*, vol. 16, no. 16, pp. 1781–1785, 2005.

[34] T. Baldeweg, "ERP repetition effects and mismatch negativity generation: a predictive coding perspective," *Journal of Psychophysiology*, vol. 21, no. 3-4, pp. 204–213, 2007.

[35] I. Winkler and I. Czigler, "Evidence from auditory and visual event-related potential (erp) studies of deviance detection (mmn and vmmn) linking predictive coding theories and perceptual object representations," *International Journal of Psychophysiology*, vol. 83, no. 2, pp. 132–143, 2012.

[36] S. Nozaradan, I. Peretz, and P. E. Keller, "Individual differences in rhythmic cortical entrainment correlate with predictive behavior in sensorimotor synchronization," *Scientific Reports*, vol. 6, p. 20612, 2016.

[37] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[38] D. Hafner, T. Lillicrap, I. Fischer, R. Villegas, D. Ha, H. Lee, and J. Davidson, "Learning latent dynamics for planning from pixels," *arXiv:1811.04551*, 2018.

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

[40] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio, "A recurrent latent variable model for sequential data," in *Advances in neural information processing systems*, 2015, pp. 2980–2988.

[41] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

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

[43] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "FMA: A dataset for music analysis," *arXiv:1612.01840*, 2016.

[44] 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*, 2015, pp. 18–25.