Decoding Music in the Human Brain using EEG Data

Chris Foster *, Dhanush Dharmaretnam †, Haoyan Xu ‡, Alona Fyshe §, and George Tzanetakis¶
Department of Computer Science, University of Victoria
Victoria, Canada
Email: *chrisfosterelli@uvic.ca, †dhanushd@uvic.ca, ‡exu@uvic.ca, §afyshe@uvic.ca, ¶gtzan@ieee.org

***Abstract*—Semantic vectors, or language embeddings, are used in computational linguistics to represent language for a variety of machine related tasks including translation, speech to text, and natural language understanding. These semantic vectors have also been extensively studied in correlation with human brain data, showing evidence that the representation of language in the human brain can be modeled through these vectors with high correlation. Further, various attempts have been made to study how the human brain represents and understands music. For example, it has been shown that EEG data of subjects listening to music can be used for tempo detection and singer gender recognition. We propose studying the relationship between the EEG data of subjects listening to audio and the audio feature vectors modeled after the semantic vectors in computational linguistics. This could provide new insight into how the brain processes and understands music.**

I. Introduction

Tasks that apply machine learning to brain data generally involve some sort of classification paradigm, in which a model is trained on many examples of brain data in response to particular stimuli and later used to predict the same stimuli. A classic example of this is the P300 speller [1], in which subjects can use their mind to interact with a computer. In the P300 speller setup, a subject views a digital keyboard on a computer screen. The keyboard is displayed in a grid-like fashion. The program then flashes horizontal and vertical rows of the grid while the subject wears an EEG headset and focuses on the letter they wish to spell. When the row or column of the letter they are focusing on flashes, the subject's brain elicits a well-known event-related potential (ERP) response. However, this ERP can often be masked by noise inherent to EEG data collection. It is possible to detect the ERP in a real-time, noise-resilient manner using machine learning. This requires training a classifier with many examples of a P300 response.

In addition, machine learning can also be applied to brain data in a more generative manner. Prior work by Mitchell et al. [2] showed that the brain response for a given stimulus can be predicted, even for a stimulus that the model had not seen before. This prediction is performed using word vectors, a concept borrowed from the field of natural language processing. Word vectors are single points in a high dimensional space that are designed to capture the semantics of a word. They are typically generated by processing a large text corpus to extract collocation information. Instead of learning the association between a label and brain data, the model learns the association between the semantic word vectors and brain data. This allows the model to predict the brain response to a given stimulus. Our aim is to explore whether or not a similar correlation exists between brain data and semantic music vectors. We follow a similar machine learning methodology to previous work connecting EEG data and natural language understanding. This type of approach can provide new insights into how the brain processes and understands music.

II. Related Work

As discussed, Mitchell et al. [2] showed that fMRI data can be correlated to semantic word vectors in the brain. This same technique was shown by Sudre et al. [3] to work in MEG, which has the additive benefit of tracking neural activity with high time resolution. Work has also been done using EEG by Murphy et al. [4] to apply machine learning techniques to semantic analysis in the brain. Using a similar methodology to Mitchell et al. they were able to predict EEG activity for unseen words and differentiate between two categories of words with 63% accuracy. Xu et al. [5] used the previous fMRI data and MEG data to perform the 2 vs 2 test with popular semantic word vectors, and further revealed the correlation between brain data and word vectors.

Music Imagery Information Retrieval (MIIR) is a sub-field of Music Information Retrieval (MIR) which studies brain activities which are recorded during various music-related tasks such as a person listening or imagining a particular piece of music [6]. EEG has been used on a wide variety of MIR tasks such as predicting emotions, tempo estimation, and audio fingerprinting. Hsu et al. [7] used EEG recording of subjects listening to music to perform music emotion recognition and Stober et al. [8] proved that EEG recordings could be used to classify and understand rhythm perception. The OpenMIIR dataset [6] is widely used for various MIR tasks, and recently, this dataset was used to perform tempo estimation [6].

Wang et al. [9] proposed the music2vec model to learn a probability distribution of music pieces, with the underlying idea that similar music pieces have similar concepts. Their model adopted the Skip-Gram model [10] and was trained on the music listening records of users. Another interesting model is chord2vec [11], which utilizes a sequence-to-sequence modeling approach based on a multilayer Long Short-Term Memory (LSTM) network with two layers of 512 hidden units each, to predict chords which are used in similar contexts.

There are also techniques [12] where acoustic features such as the chromogram, tempogram, mel spectrogram, mel spectrogram cepstral coefficients (MFCCs), contrast, and Tonnetz are extracted and then given as the input to a convolutional neural network (CNN). The activations of the fully connected layer before the output layer are taken as the latent embedding for the input music. Other work generated music feature vectors from symbolic music notation, and used deep auto encoders to learn latent semantic representations for speech signals [13]. Learned music embeddings are also widely used for music recommendation tasks [14]. Our study evaluates music vectors generated from various sources and correlates them with music EEG data to test the hypothesis that the music vectors represent music features in a similar fashion to how the human brain imaging represents music.

III. Dataset

The brain data we used for this experiment is the OpenMIIR dataset [6], which contains 64-channel EEG data of 120 music fragment exposures ranging from 7 – 16 seconds in length at 512Hz. There are 12 music tracks categorized into 4 lyrical songs, 4 songs with the lyrics removed, and 4 non-lyrical longs. The dataset was collected by measuring the response of 10 subjects aged 19 – 36. The brain data has been preprocessed both manually and automatically. Manual preprocessing consists of a standard visual channel inspection, which allows removing bad channels (due to a poor connection or malfunctioning electrode) and interpolating their values based on the surrounding known-good channels. Filters are used to reduce the frequency range to 0.5 – 30Hz, baseline correction is used to adjust for signal drift, and independent component analysis is used to remove cyclic artifacts (such as eye blinks).

IV. Methodology

Our work involved three phases: 1) generating music vectors, 2) extracting music features from the EEG datasets, and 3) performing correlation tests between music vectors and brain data. Music vectors can be generated from either raw audio or symbolic data. The quality of the generated music vectors can be evaluated by performing K-means or agglomerative clustering. The intuition is that similar music pieces should be clustered together in the vector space. Once the vectors are extracted we can correlate them with EEG data. To study the correlation between music vectors and EEG music vectors we use representational similarity analysis (RSA) [15] and the linear models approach [3].

A. Representational Similarity Analysis

Representational Similarity Analysis (RSA) provides a means to measure correlations between two representational models of the same subject. RSA uses the Representational Dissimilarity Matrix (RDM) to compare different representational models. We create an RDM containing a cell for each audio pair in the EEG dataset. The cell holds the correlation distance (1 - correlation coefficient) of the EEG data for that audio pair. In other words, an RDM is a pairwise correlation distance matrix of the audio EEG data. The more similar an audio pair is, the more similar their EEG data should be, hence, the closer their correlation distance should be. Another RDM of the same audio pairs is created for the audio vectors in a similar fashion. This provides us with two correlation matrices we can later evaluate to determine the correlation between the EEG data and audio vectors. If the RDMs of the audio vectors and the EEG data have high correlation, then it shows they both capture the similarity between songs. This analysis has a risk of failure if any of the audio vector or EEG data fails to capture the similarity between songs, but is a simple approach.

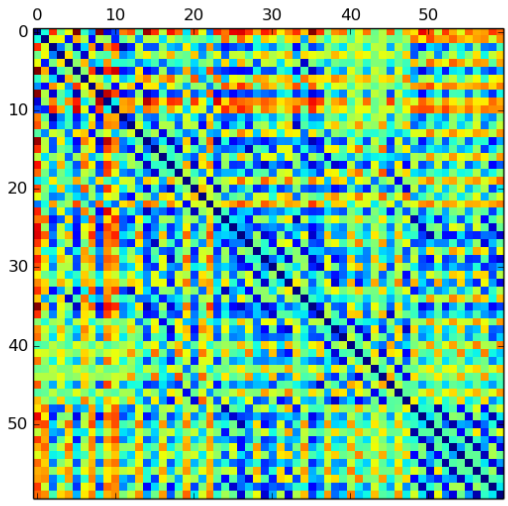

Fig. 1. Visualization of an RDM example

B. Linear Models Approach

The linear models approach utilizes a set of linear models that each predict a semantic feature in the music vector. Following Mitchel et al. [2] and Sudre et al. [3], the individual EEG exposures are averaged for each song to remove noise, generating a set of training data. In our case, we also average across all subjects leaving us with a total of 12 samples. This is one EEG exposure per audio clip. At training time, we train n linear ridge regression models (where n is the length of the audio feature vector used) to each predict their associated element of the feature vector based on the averaged EEG data matching that particular audio clip. At evaluation time, we provide the full EEG exposure to each linear model and each model predicts its respective element of the feature vector. Collectively, this model takes EEG data as input and produces an audio vector in response that is believed to be associated with the underlying music.

C. 2 vs 2 Test

The 2 vs 2 test is a correlation test that simplifies a vector comparison into a binary classification task. This test can be

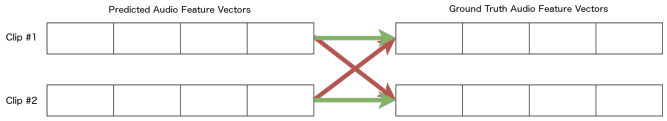

Fig. 2. The 2 vs 2 test allows us to reduce the model to a binary decision task, where a successful comparison occurs when the sum of the distances between the vectors for the correctly aligned songs (green lines) is smaller than the sum of the distances for the incorrectly aligned songs (red lines). Here we label the predicted and ground truth vectors used in the linear models approach.

used as a tool for proving that two data sources are correlated. We utilize the 2 vs 2 test for the RSA model and the linear models approaches to evaluate whether they could capture correlation between the EEG data and the music vectors.

The models are evaluated in a "leave two out" fashion, in which we take every possible combination of two songs from the total set. For the linear models, we train the model on the remaining data and test it on the two held out samples. The two predicted audio vectors are then compared against their ground truths vectors using the 2 vs 2 comparison. For RSA, we take the two rows corresponding to the two songs from the audio vector matrix and EEG exposure matrix. The four vectors are compared in a similar fashion using the 2 vs 2 comparison. To perform the 2 vs 2 comparison, we check if the sum of the distance between the two correctly matched pairs is smaller than the sum of the distance between the two incorrectly matched pairs as in:

$$d(y_i, \hat{y}_i) + d(y_j, \hat{y}_j) < d(y_i, \hat{y}_j) + d(y_j, \hat{y}_i) \qquad (1)$$

If true, the comparison is considered successful. The overall 2 vs 2 accuracy is the percentage of successful 2 vs 2 comparisons. Chance accuracy, or the accuracy we expect to see if the EEG data is not correlated with the audio vectors, will be near 50%. Statistical significance is then validated using a permutation test.

D. EEG Processing

OpenMIIR provides one raw EEG data file per test participant in FIF format. Each file contains all EEG recordings of the participant, which consists of 240 trials (12 stimuli * 4 conditions * 5 blocks). The 12 stimuli are ordered randomly per condition and block [8]. The data has 68 channels with sampling frequency of 512 Hz. The random order of trials made the extraction of the EEG data somewhat challenging, and more preprocessing was involved than anticipated originally. The preprocessing scripts uses the *deep thought* library developed by Sebastian Stober, which utilizes the MNE library and independent component analysis (ICA) (https://github.com/sstober/openmiir-rl-2016) [16].

E. Audio Feature Extraction

Experiments [17] have been done recently to correlate music features such as the RMS and spectral flux with the EEG dataset (NMED-H) [18] using a technique called Canonical Correlation Analysis. Therefore we decided to explore more

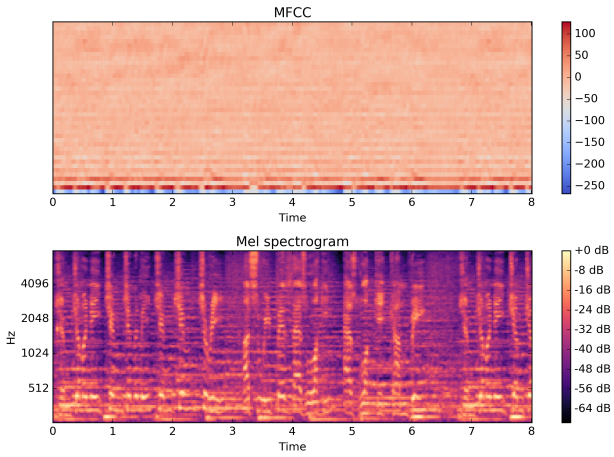

Fig. 3. The MFCC (top) and mel spectrogram (bottom) plot generated for the song "Chim Chim Cher-ee" with 40 MFCCs using the librosa library

musical features such as the MFCCs, spectral centroid, root mean square energy (RMSE), tonal centroid features, and constant-Q chromogram. These features along with other features such as the mel spectrogram, chromogram, and tempogram (a rhythm feature) are generated using the *librosa* library [19] for the same set of songs used by the EEG dataset experiment [6]. The MFCC and mel spectrogram plot for the song "Chim Chim Cher-ee" is shown under Figure 3. The generated features are averaged and pearson correlation matrices are generated for each of the features. These are then used for the RSA and linear models approaches.

F. Tag Feature Extraction

In addition to experimenting with various audio features, we also decided to explore the use of tag prediction models as features. Our goal is to capture the overall semantics of an audio piece, for which tag predictions may work as a reasonable high-level approximation. Many trained models exist that can map a music piece to a vector of probabilistic estimations for each tag, which we treat as our "feature vector". This also allowed us to begin development of the correlation models against a baseline vector that is easy to generate, even if we do not necessarily anticipate high accuracy with them.

We downloaded a convolutional neural network model for this task developed by Choi et al. [20] using Keras and Tensorflow. Pre-trained weights were also available for this model. We adapted the Keras-based code to generate tag feature vectors for each of the songs available in the EEG dataset using the CNN model and weights.

V. Results

The following section details the results of our experiments to find correlation between semantic music features and the brain activity dataset. Additionally, while working on this project we had the opportunity to explore other classifiers related to the dataset.

A. Correlation Experiments

We generated our audio feature music vectors using the *librosa* library in Python. We extracted music features such as the MFCCs, RMSE, spectral centroid, chroma STFT, spectral roll-off, tempogram, harmonics, and beats then attempted to correlate them with the EEG music vectors using the RSA and linear models 2 vs 2 tests. We also generated tag feature music vectors using the Keras/Tensorflow libraries in Python. We detected no correlation between the EEG data and the music vectors for all features with the exception of the MFCC and tempogram features from the audio feature music vectors. The RSA analysis scored **0.63** out of 1.0 for the normalized tempogram features of the songs extracted using librosa. Similarly, we performed the RSA with 100 MFCCs extracted using *librosa* for each song which resulted in a score of **0.62**. The linear models approach did not detect correlation in any of the cases.

However, we can not confirm the presence of any correlation without a permutation test. Therefore, we performed 1,000 iterations of the RSA test after performing randomization of the EEG music vector rows during each iteration. The histogram for the permutation test results on the tempogram features, which creates our null distribution which we can use for significance testing, is shown under Figure 4. A similar permutation test was conducted for the MFCC features as well. The permutation test confirmed we detected statistically significant weak correlation between the music EEG vectors and their corresponding MFCC / tempogram features with $p < 0.01$. These weak correlation scores are based on the average across all the 9 participants. However, we are hesitant to conclude our original hypothesis is correct because it does not seem to generalize well for more semantically-based features that we extracted from the songs or to our other linear models approach. It is also interesting that there is a big deviation in the RSA test scores across participants.

B. Performing Song Identification from EEG Data

Another task that we attempted to perform is song identification from the raw EEG data alone. There were 12 songs in the dataset and each song was repeated five times per participant. The dataset used for song identification therefore has 60 samples per each song (a total of 540). We split this data into train and test with a ratio of 80:20. A logistic regression model was trained using the sklearn library to perform song identification. The linear model was hyper-tuned and we achieved a classification accuracy of 0.287 using increased regularization during training ($C = 0.001$). The F1 score was found to be 0.286 and the confusion matrix for this classification task is shown under Figure 5. This shows that there is enough information in the EEG data to distinguish between songs across all 9 participants. On the analysis of the confusion matrix, we can clearly see that most of the classification errors are between the same pairs of songs which differ only in the presence or absence of lyrics. For example "ChimChimLyrics" is often confused with "ChimChimNoLyrics".

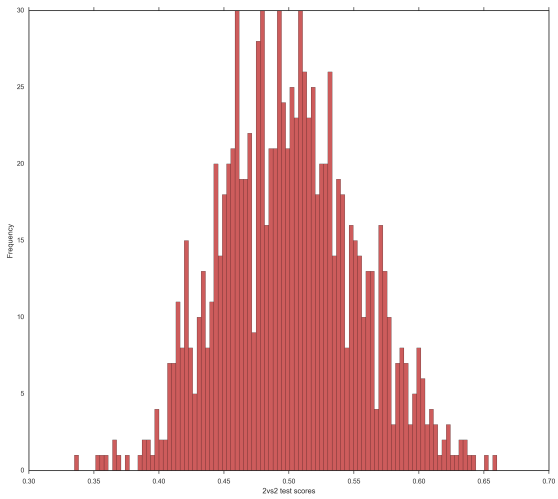

Fig. 4. The histogram of the permutation test for the RSA analysis between song tempogram feature extracted using *librosa* library and the music EEG vectors after randomizing the rows during each iteration of the permutation test.

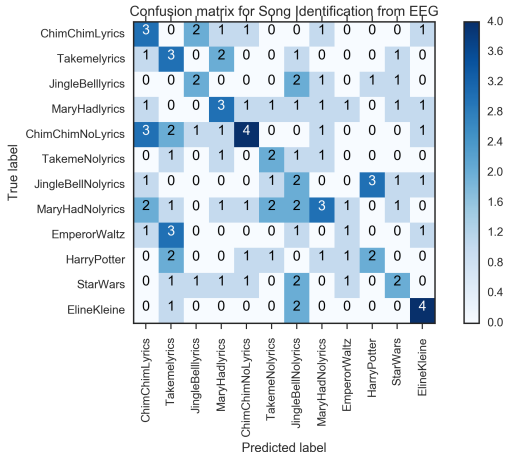

Fig. 5. The confusion matrix generated for the song prediction from raw EEG data using logistic regression.

We hypothesize that EEG data does not encode enough information to help us identify presence or absence of lyrics in music. This could explain why songs with lyrics were often confused with the same song without lyrics, as indicated in the confusion matrix. In order to validate our hypothesis, we trained another logistic regression model to identify whether songs did or did not have lyrics based on the raw EEG data. There were 240 songs with lyrics and 300 songs without lyrics across 5 trials across all 9 participants. The data was split into train and test in the ratio 80:20. The classification accuracy was 0.67 which is within the chance accuracy for this task. This implies that there was not enough information in the raw EEG data available for the classifier to create a decision boundary which separates the songs with lyrics from the songs without

lyrics.

VI. Discussion

Our results show that we have been able to detect some statistically significant correlation, but are not able to confidently show it in all cases. The following is a discussion on the results, including our new classifier.

A. Potential Dataset Issues

The dataset by Stober was useful for its focus on music and its quality EEG data collection. After some experimentation with the dataset throughout the course of this project, we've determined it has several features that make it difficult to adapt for our task and may result in challenges to finding correlations. One of the primary issues is the quantity of data available in the study. The dataset spans twelve songs with ten subjects. Because one subject's data was discarded, this is only a total of 540 trials. Generally, in the equivalent language literature the number of stimuli is several times higher. This is an understandable limitation of applying machine learning models to brain data studies, given that brain data is expensive to collect even in a more affordable methodology such as EEG, but poses a problem for this research. These twelve songs may not provide sufficient coverage of the vector space for the model to learn a useful mapping to differentiate between two stimuli. This is further compounded when considering that four of the tracks are non-lyrical versions of four other tracks, which reduces the amount of information gained from those and leaves only eight fully original tracks.

Subjects were exposed a total of five times for each stimulus. In the equivalent language literature twenty exposures are typically used. EEG data can be challenging to work with due to the noise involved with capturing, and this is typically mitigated by averaging across a number of exposures. With less exposures, the averaging becomes less effective at removing noise. Further, these exposures were captured over different sessions which makes averaging less effective. Between sessions a subject's state-of-mind, level of fatigue, attention, and other attributes change which can effect brain data. Further, a removal and replacement of the cap is likely to produce readings with different electrode connection quality and slightly altered placement. In our experiments, the raw brain data was often not correlated between identical subjects and identical tracks for different sessions which supports our hypothesis that this makes higher level tasks difficult.

B. Correlation Task Discussion

We identified weak yet statistically significant correlation between EEG music vectors and music features, specifically the tempogram and MFCC features. We confirmed this correlation using the permutation test. Our findings are in agreement with Kaneshiro et al. [18] who found weak correlation in their EEG responses dataset based on Hindi songs and music features such as beats, tempo, and zero crossing error. Hindi songs generally have a higher rate of rhythm, beats, and tempo as compared to English carols and classical songs which may explain why our experiments did not find weak correlation with the zcr, beats, or other features. Both the RSA and the linear models 2 vs 2 tests search for weak correlation between two datasets. All the previous experiments, including our experiment, have shown some correlation with tempo of the songs with the EEG datasets. This suggests that tempo may be the easiest feature to detect from EEG.

We were not able to detect correlation with the linear models even in cases which we could with the RSA model, which we hypothesize is related to the size of the available training data as discussed. The RSA 2 vs 2 model does not require training and does not have a similar limitation (although it does benefit from increased data as well). This issue may also be related to the averaging method used, specifically the linear models averages across subjects while the RSA model averages only within subjects. As mentioned, the variance between individual exposures / subjects can lead to issues with this averaging.

C. Classification Task Discussion

Our logistic regression classifier improved on the state of the art in song prediction from raw EEG data. Earlier experiments by Stober [6] extracted and learned music features from EEG data using auto encoders. However, we found that we could match this performance without any expensive or complicated feature learning by just tuning the hyper parameters of a more simple classifier. We attempted to learn the music features directly from raw audio features using the mel spectrogram, however we did not find any correlation with the learned features from raw audio and the EEG data. It may be worthwhile to explore additional methods of feature extraction from EEG data in future work.

VII. Future Work

While working on this project we have found the key difficulty to be the dataset. While it is a good dataset, it is likely not optimal for this task. As such, we have identified a number of possible areas for future work with regards to data collection.

Losorelli et all. [18] have recently created a new tempo focused EEG dataset called NMED-T. The dataset includes EEG recordings of twenty participants listening to ten full-length songs. The songs are all 4.5 minutes to 5 minutes in length and contain vocals. Compared to the OpenMIIR dataset used in the experiments for this paper, the NMED-T dataset features full length songs in a larger variety of genres, which may be more effective compared to when the task is performed on less than ten seconds of recording per song. Although language literature suggests shorter and more sudden stimuli may be more useful, this dataset forms one option for expanding our methodology to other brain data sets.

We are planning a new experiment which would be better tuned to this task. To begin, we would select a series of audio tracks from a variety of genres that provided good coverage of the semantic music space. Sections would be manually extracted from each track to meet a standardized length. While the previous experiment had tracks ranging in length, in our

task it makes sense to record identical length for all tracks since the input EEG data needs to be the same shape between tracks. We would ideally also increase the participant count and exposure count. Increasing the exposure count will allow us to average across trials and reduce noise, while increasing the subject count will reduce the likelihood of our 2 vs 2 accuracy being far from chance value. A higher subject count also shows generalization better and can be used for reducing noise further when averaged across subjects.

The collection approach taken by the original dataset follows good EEG collection practices, and we would follow a similar fashion including full 64 sensor recording and the preprocessing procedure used by Stober. However, we would perform the experiment on subjects in a single-session manner, instead of separating collection over multiple exposures. It is important to balance subject fatigue with these requirements. If the session is too long, subjects will not wish to continue or have difficulty paying attention. A rough estimate of 40 tracks with 10 second recordings repeated 10 times per subject gives a session length of slightly over one hour, which is reasonable and leaves room for breaks, queues, or extensions if necessary.

We have additionally discussed the idea of performing this experiment with sound-based dataset of tracks, as opposed to a music-based dataset of tracks. An example stimuli would be "the sound of door knocks". These events should be significantly shorter than a piece of musical work, which makes it easier to collect a larger variety of data. We also hypothesize that this dataset would contain a larger semantic variety, and be easier to detect in brain data. Comparisons could be made between EEG recording of participants listening to an event, imagining the event, and perceiving the word which represents the event. It would be a novel dataset to create and experiment on.

VIII. Conclusion

We have studied the relationship between the EEG data of subjects listening to audio and the audio feature vectors modeled after the semantic vectors in computational linguistics. We found statistically significant weak correlation between the MFCC and tempogram music features with the music EEG vectors using the RSA 2 vs 2 analysis. We also found that we could distinguish between the songs from the EEG data using a simple linear classifier, specifically logistic regression. Our linear model achieved state of the art accuracy in song identification from just the raw EEG as compared to previous work which achieved the similar results using complex music feature extraction of the EEG data via auto encoder techniques. We also argue that EEG does not capture complex music semantics related to lyrics and further suggest future experiments with better data collection strategies for this task. We believe that the results of our experiments are interesting, and provide new insight into how the brain processes and understands music.

References

[1] R. K. Chaurasiya, N. D. Londhe, and S. Ghosh, "An efficient p300 speller system for brain-computer interface," in *Signal Processing, Computing and Control (ISPCC), 2015 International Conference on*. IEEE, 2015, pp. 57–62.

[2] T. M. Mitchell, S. V. Shinkareva, A. Carlson, K.-M. Chang, V. L. Malave, R. A. Mason, and M. A. Just, "Predicting human brain activity associated with the meanings of nouns," *science*, vol. 320, no. 5880, pp. 1191–1195, 2008.

[3] G. Sudre, D. Pomerleau, M. Palatucci, L. Wehbe, A. Fyshe, R. Salmelin, and T. Mitchell, "Tracking neural coding of perceptual and semantic features of concrete nouns," *NeuroImage*, vol. 62, pp. 451–463, 2012.

[4] B. Murphy, M. Baroni, and M. Poesio, "EEG responds to conceptual stimuli and corpus semantics," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2009, pp. 619–627.

[5] H. Xu, B. Murphy, and A. Fyshe, "Brainbench: A brain-image test suite for distributional semantic models," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 2017–2021.

[6] S. Stober, "Toward studying music cognition with information retrieval techniques: Lessons learned from the openmiir initiative," *Frontiers in psychology*, vol. 8, 2017.

[7] J.-L. Hsu, Y.-L. Zhen, T.-C. Lin, and Y.-S. Chiu, "Affective content analysis of music emotion through eeg," *Multimedia Systems*, pp. 1–16, 2017.

[8] S. Stober, D. J. Cameron, and J. A. Grahn, "Classifying eeg recordings of rhythm perception." in *ISMIR*, 2014, pp. 649–654.

[9] D. Wang, S. Deng, S. Liu, and G. Xu, "Improving music recommendation using distributed representation," in *Proceedings of the 25th International Conference Companion on World Wide Web*, ser. WWW '16 Companion. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2016, pp. 125–126. [Online]. Available: https://doi.org/10.1145/2872518.2889399

[10] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[11] S. Madjiheurem, L. Qu, and C. Walder, "Chord2vec: Learning musical chord embeddings," in *Proceedings of the Constructive Machine Learning Workshop at 30th Conference on Neural Information Processing Systems (NIPS2016), Barcelona, Spain*, 2016.

[12] Z. Xing, E. Baik, Y. Jiao, N. Kulkarni, C. Li, G. Muralidhar, M. Parandehgheibi, E. Reed, A. Singhal, F. Xiao *et al.*, "Modeling of the latent embedding of music using deep neural network," *arXiv preprint arXiv:1705.05229*, 2017.

[13] M. Bretan, S. Oore, D. Eck, and L. Heck, "Learning and evaluating musical features with deep autoencoders," *arXiv preprint arXiv:1706.04486*, 2017.

[14] D. Wang, S. Deng, X. Zhang, and G. Xu, "Learning music embedding with metadata for context aware recommendation," in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. ACM, 2016, pp. 249–253.

[15] N. Kriegeskorte, M. Mur, and P. Bandettini, "Representational similarity analysis–connecting the branches of systems neuroscience," *Frontiers in systems neuroscience*, vol. 2, 2008.

[16] S. Stober, "Learning discriminative features from electroencephalography recordings by encoding similarity constraints," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 6175–6179.

[17] N. G. B. K. J. Berger and J. P. Dmochowski, "Decoding neurally relevant musical features using canonical correlation analysis."

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

[19] 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.

[20] K. Choi, G. Fazekas, and M. Sandler, "Automatic tagging using deep convolutional neural networks," *arXiv preprint arXiv:1606.00298*, 2016.