Music Reconstruction: Translating Thoughts Into Music

Christian de Abreu[a,b] **and Alejandro Villasmil**[a,c]

[a]University of Pennsylvania Department of Bioengineering (BSE); [b]University of Pennsylvania Department of Bioengineering (MSE); [c]University of Pennsylvania Department of Mechanical Engineering (MSE)

Decoding electrocortical signals has been studied extensively in the past decades because of its multiple applications in seizure detection, artificial limb movement, and image reconstruction. With ever growing availability of electroencephalography (EEG) data, experts in machine learning been entering the field, applying statistical models to enhance the state of the art in neural decoding. Although most researchers today use EEG data and deep learning for image reconstruction, and only to a limited accuracy, there is scarce knowledge in the field of audio reconstruction from EEG. Given the nature of this problem was considered to be of lower dimension than an image, perhaps a simpler algorithm would be able to perform this decoding. Using the NMIIR dataset, we employed a simple neural network architecture to translate EEG data from subjects listening to different songs into a reconstructed and replayable audio waveform. Through a principle component analysis, we selected a combination of the 64 input channels available from the dataset to train the network. Overall, we were able to play the songs outputted by the network, and even though the sounds were qualitatively unlike the original audio, waveform similarities were present, meaning further investigation in this field could lead to more accurate audio reconstructions.

Music | EEG | Reconstruction | Neural Networks | Brain

Electrical signals extracted from the brain can provide meaningful information about the stimuli that humans perceive. Prior research has proven just this with the novel finding that there is a significant correlation between the audio one is able to perceive and the electrical activity of the brain [1]. Furthermore, prior research has also proven that is possible to reconstruct stimuli perceived by human beings [2]. One instance of this can be seen in a publication by Nemrodov et al [3], where spatio temporal EEG information was used in conjunction with a predictive algorithm to reconstruct images that were shown to subjects. These two findings point to the conclusion that audio reconstruction from EEG data is within the realm of what algorithms today can learn, and what high quality data is available publicly.

This potential implementation could have tremendous impact in terms of its application to neurotechnology. One application could perhaps be in its use to communicate with people who are not able to communicate verbally. In other words, audio reconstruction from thought would give mute people the ability to one day, carry on complete conversations. Law enforcement could also make use of this technology to potentially aid them in recognizing suspects from eye witnesses. Instead of relying on the caricature capabilities of an artist, which is a very subjective art form both the eye witness standpoint as well as interpretation by the officer, an algorithm would be able to work side by side with the officer to enhance the accuracy of that reconstruction. In the field of brain computer interfaces, decoding EEG data is the foundation of much of today's technology. From electrode size to invasiveness, there is much to improve on the hardware side of brain activity recordings. While those technologies are being developed, algorithms are more easily iterated on, and in our investigation, we will use EEG data to reconstruct music perceived from several patients.

Data Collection and Preprocessing

The data used for this research was acquired via electroencephalography (EEG) recordings. This is the most commonly used and least invasive method to record the electrical activity of the

Each author contributed equally to the work presented here

No conflicts of interest

[2] To whom correspondence should be addressed. E-mail: dechrist@seas.upenn.edu or avil@seas.upenn.edu

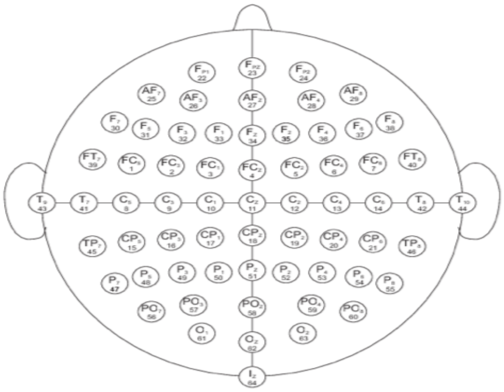

Fig. 1. 64 Channel EEG Layout used for data acquisition

cortex. A net of several electrodes, ranging from 20 to 64, is placed on the scalp of the subject and this is essentially able to record the electrical field being produced by billions of neurons firing synchronously. An electrocorticogram (ECoG) can also be used as another means to gather more reliable and less noisy electrical data, however, this method is substantially more invasive as it requires the electrodes to be placed directly on the cortical surface of the brain.

The specific dataset used for this research project was obtained from the OpenMIIR dataset [4], a public domain dataset of EEG recordings for music imagery information retrieval. This data was recorded using a 64 channel EEG. A map layout of this data retrieval structure can be seen in *Figure 1.* One important thing that was kept in mind was, as is the case with all analog to digital signal processing, the sampling frequency of the EEG data. If the sampling frequency happens to be too low this can lead to aliasing of the data and can result in poor signal integrity. If the sampling frequency is too high, this can lead to more demanding computations and can slow down training significantly. However, it is always better to oversample than to undersample because downsampling method techniques are typically lose less data than interpolating/upsampling.

One downside of using EEG data as our primary dataset is its susceptibility to noise and a variety of undesirable artifacts. These can be attributed to either physiologic or extra-physiologic causes [5]. For instance, a physiological artifact could be due to muscle movements on or near the subject's head. Given that EEG electrodes are simply measuring electrical fields, the numerous action potentials occurring during muscle movements (e.g. cheek, jaw, etc.), could easily interfere with the brain activity that is being recorded. Additionally, if the patient sweats during a trial, the conductive nature of human sweat can lead to inductive changes at the electrodes and in turn lead to erroneous measurements. Some artifacts may also come about due to extra-physiologic reasons, reasons not having to do with the subject. One primary contributor to this would be the common 60Hz AC noise that is present anywhere there is an electrical power line. It is also a possibility that individual electrodes are inherently faulty.

In the case where a relatively integrous signal was corrupted signal by some sort of high frequency noise (e.g. muscle movement, 60 Hz AC noise) we decided to break the signal into its frequency components by taking a simple Fast

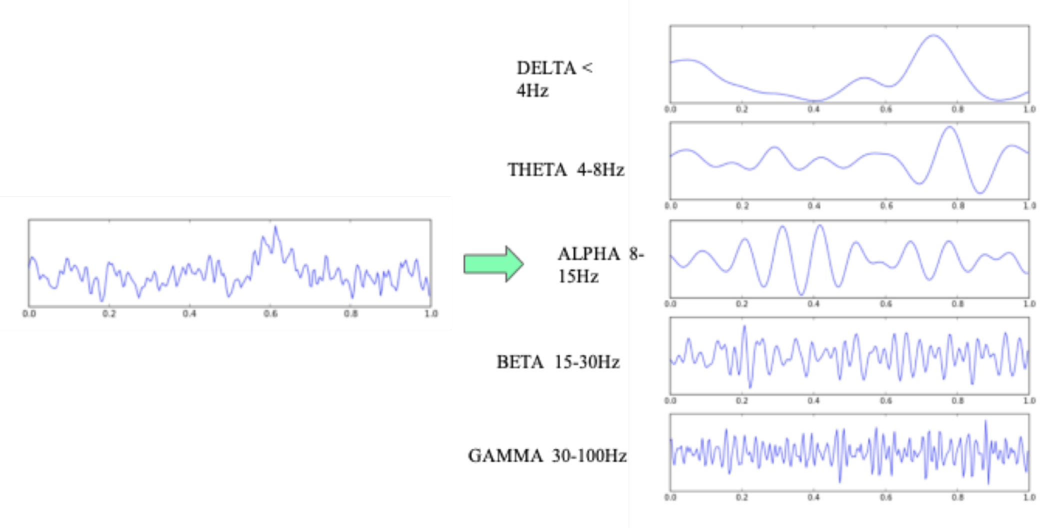

Fig. 2. Frequency component breakdown into respective bands

Fourier Transform. Thereafter, we were able to bin these frequencies into 5 distinct bands, as can be seen in Figure 2 above. The undesirable high frequency oscillations mentioned above can be seen in the gamma band that houses the component of the signal corresponding to frequencies between 30 and 100 Hz.

In order to mitigate the contaminating effects of these high frequency oscillations we decided to feed our EEG data through a bandpass filter with cutoff frequencies of 0.5 Hz and 30 Hz. In the case of a faulty electrode in which the entire time series EEG data was defective and abnormal we decided it would be best to completely get rid of the channel, given that we were dealing with 64 electrodes in total and ridding ourselves of a couple would not affect the training data. We came to the conclusion it would be more beneficial to exclude this data by setting it to a baseline signal of 0 V, given that a future PCA implementation would select the most active channels. Since these faulty channels likely had higher oscillations than usual, they could potentially be included as a principle component and compromise results. Therefore, we set the faulty channels to 0V instead, such that they would contribute no variance.

So in summary, our preprocessing pipeline can be visualized as in Figure 3. At the culmination of preprocessing we are left with the data which will be used to train our model. This data is comprised of grand total of 514 trials, with each trials composed of a single subject, out of 9, and a single song, out of 12. Each trial houses 6.87 seconds of EEG recordings for a single patient responding to an individual song. These 6.87 seconds of time series EEG data were sampled with a sampling frequency of 512 Hz giving us 3,518 samples per trial.

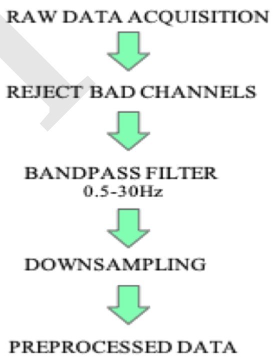

Fig. 3. Preprocessing pipeline used in order to get data ready for training

Training and Architecture

Raw audio data was made available through the NMIIR dataset. Another file in that datset labelled “meta_data” contained information such as subject ID and stimulation ID, which allowed us to match trials with subjects and songs played. At this point in the pipeline, our data was com-

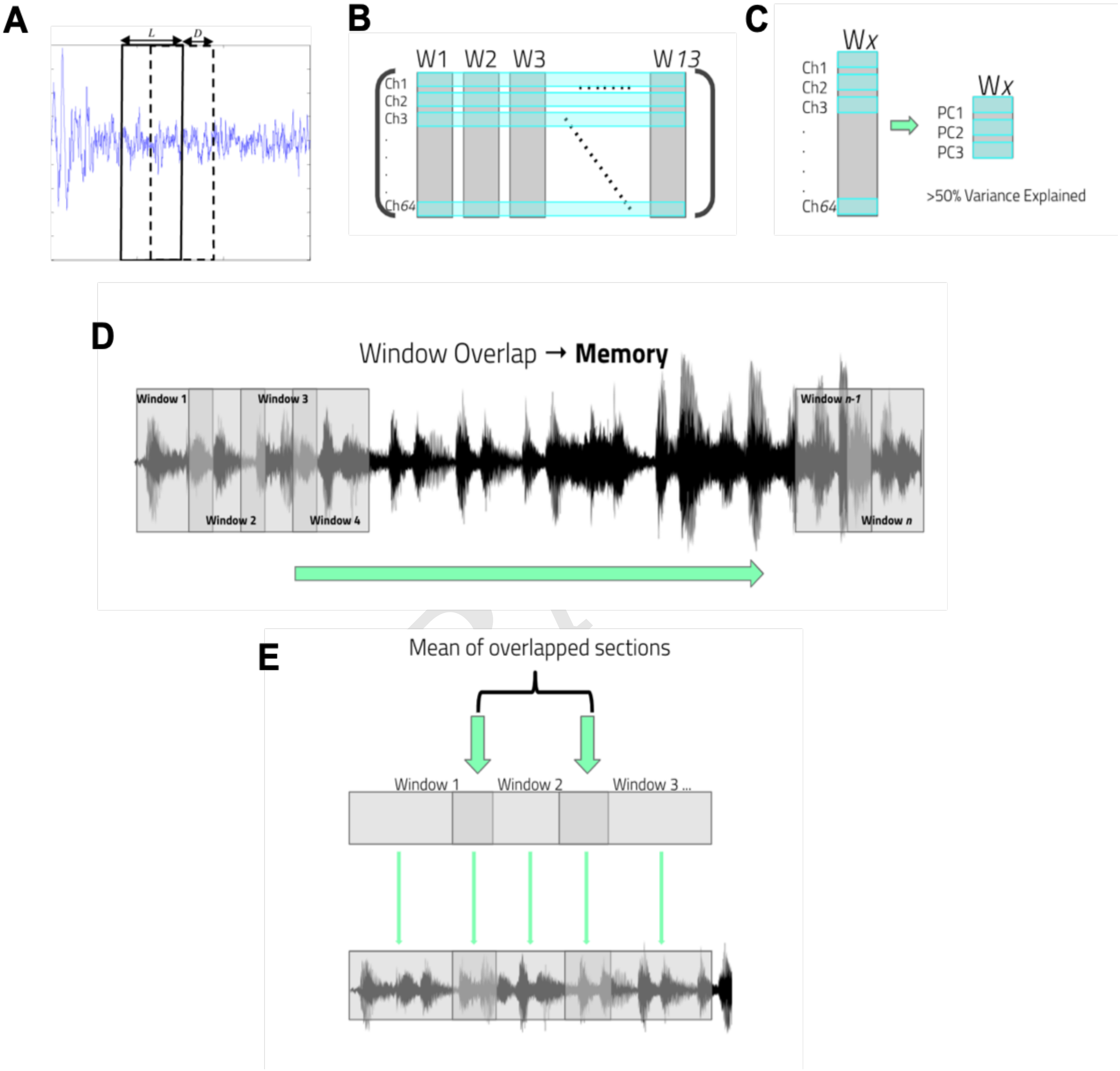

Fig. 4. Algorithm Training Setup and Procedure. (A) Sliding window schematic. By designating a window displacement and window length, one can slide through the time series data to extract features. **(B)** Feature matrix representation. Each column designated as W_x, corresponds to data in a time window, including features and raw signal. Every row corresponds to a row. **(C)** Selection of channels with most information via PCA **(D)** Schematic of how sliding window encodes memory as well as feature reinforcement (E) Reconstruction of original audio file. After averaging intersection of predicted windows, we up-sampled to original audio length.

prised of one audio *WAV* file with sampling frequency 44.1kHZ, and a matrix of EEG recordings with 540 trials, each trial comprised of 64 channels and each channel containing 3158 samples. As is typically done with time series signals such as EEG, one uses a sliding time window [5], similar to how a kernel is used in convolutional neural networks. Figure 4 (A) depicts this method of sliding at an jump of 0.5 seconds (corresponding to 0.5s*512Hz = 256 samples for each window of 0.75 seconds (corresponding to 0.75s*512Hz = 384 samples).

Under this setup, we extracted features for each window for each EEG channel into what we dubbed as "R Matrix". The features extracted for each window can be broken down into six categories. First of all, we extracted the raw EEG signal for each window (and for each channel) and appending it to the R Matrix. Most neural network transformations are black box in the sense that the tuning during backpropagation has one objective which is to minimize a loss function. For our case, the raw data carries the most information for each subject, so including the raw data as a feature was an essential component of the investigation. In order to augment the nonlinear transformation of EEG to audio, we also appended five additional features that have been previously used to classify epilepsy onset among or brain computer interface applications. The first of these features is power per frequency band. In order to decompose the EEG signal into each frequency band, we applied a Fourier transform to the EEG and then binned the corresponding power distribution into the delta, theta, alpha, beta waves. Since we had band-passed the EEG data to remove any wave above gamma (35Hz) and because most of the useful information encoded in EEG lies in those first 4 bands, we appended those to the R Matrix. Next, we introduce the Area and Energy features. Area is nothing more than total area of each window represented by the integral of the signal. Energy on the other hand is the squared value of each sample. How this differs from area is that it captures higher amplitudes and places more importance on those sample. The fifth feature included is the number of zero crossings [6]. A signal that is agitated around the mean is indicative of some kind of increased activity and that is definitely something that we want to include in our decoding procedure. Finally, line length was appended to the R Matrix. Line length is one of the most interesting features as it considers the pairwise distance between samples [7]. The intuition behind is would be to envision the signal as a wire and line length would be how long the wire would be if extended from point to point.

Once the features were extracted, the R Matrix was then transformed into an easily manipulated 3-D matrix whose dimensions number of channels x number of samples per window x number of windows. With this new representation, we can easily train the algorithm using a time window approach, something we denote as "Sequential Batch Learning".

Before beginning sequential batch learning, we wanted to further select channels based on information they provided. Training on useless channels would just incur noise and make computations even more expensive, so we propose a principle component analysis of all 64 channels, reducing them to 3 linear combinations of the channels that represent the highest variance. Since the principle components ordered by variance explained, this allowed us to train a simple neural network with 9 hidden layers and a *tanh* activation sequentially. The structure for this neural network can be seen in Figure 5. Therefore, for each time window, we would reduce the number of channels to 3 principle components, then train the first principle component 1000 times to fit the audio time window, then repeat training 800 times for the second principle component, and finally 600 times for the third principle component.

After this batch training was performed on a specific time window, we would slide the window by 0.5 seconds along the EEG channels. Note that each window is actually 0.75 seconds long, so by sliding by 0.5 seconds, there is an overlap of 0.25 between the *nth* window and the *n+1*

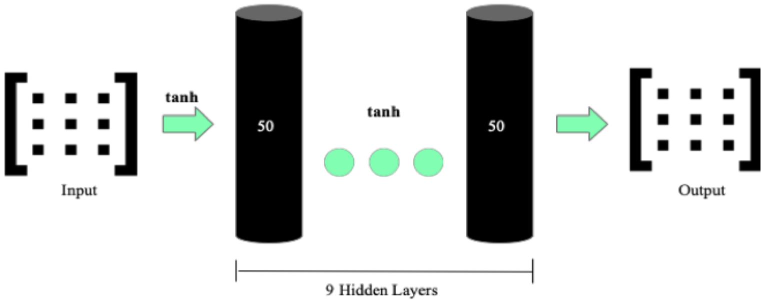

Fig. 5. Neural Network Architecture

window (see Figure 4 (D)). The intuition behind this decision was that the system would then have "memory" of the previous time window. In other words, weights would not be changed in sharp increments. Rather, by ensuring that 33% of the data in each new window had already been seen before, we could either reinforce certain features that were learned or slowly forget those that were specific to a time window and not the global transformation between EEG and audio.

Once predictions were made for each time window, we had to reconstruct the final audio in two steps. Figure 4 (E) shows a schematic of this process. Since each prediction is calculated for a single time window, and each time window overlaps with the one before it and the one after it, a concatenation of each window would produce the reconstructed and complete audio. In the sections with overlap, a numerical mean was calculated and appended to a final audio waveform. With the newly constructed audio, we could replay the songs in python and see how they compared to the original input.

Results

As mentioned above, our final output from the algorithm was essentially a reconstructed version of the originally presented audio. We first wanted to compare the output of the network on the EEG time series used to train the algorithm to see how it would fare with an input that it should be able to predict fairly well, given that it was trained on the same data.

As can be seen in Figure 6 (A) a majority of the signal integrity is lost during the prediction and reconstruction, however, if given a closer look one can see that the peaks and troughs seem to match up relatively well with the audio presented as a stimulus to the subject. Thereafter, we sought out to see how the algorithm would fare in predicting and reconstructing a song with an EEG signal it had not seen before. In Figure 6 (B) one can see the predictions of a testing EEG time series and can notice that the features of the true audio waveform does not match up as well with the prediction.

Generating these past predictions all involved training the algorithm with a single song and a single EEG signal. Given this, we thought it would be worthwhile to see the effect of training on multiple song and multiple subjects. Therefore, we first decided to train the algorithm with 90% of the available songs and their respective EEG signals corresponding to a single patient. The predictions generated from this method can be seen in Figure 6 (C). It is apparent that this prediction carries substantially more of the signal power carried by the original audio waveform, however, upon playback of the reconstruction one can realize it is not close to sounding like the original.

Next, we trained the algorithm with a single song but with EEG data corresponding to 90% of the subjects. As can be seen in Figure 6 (D), we obtained a similar prediction to that of the result when trained on 90% of the songs. The signal integrity was much better than the prediction generated from training on a single song and patient, however, the playback is not close to sounding like the original audio.

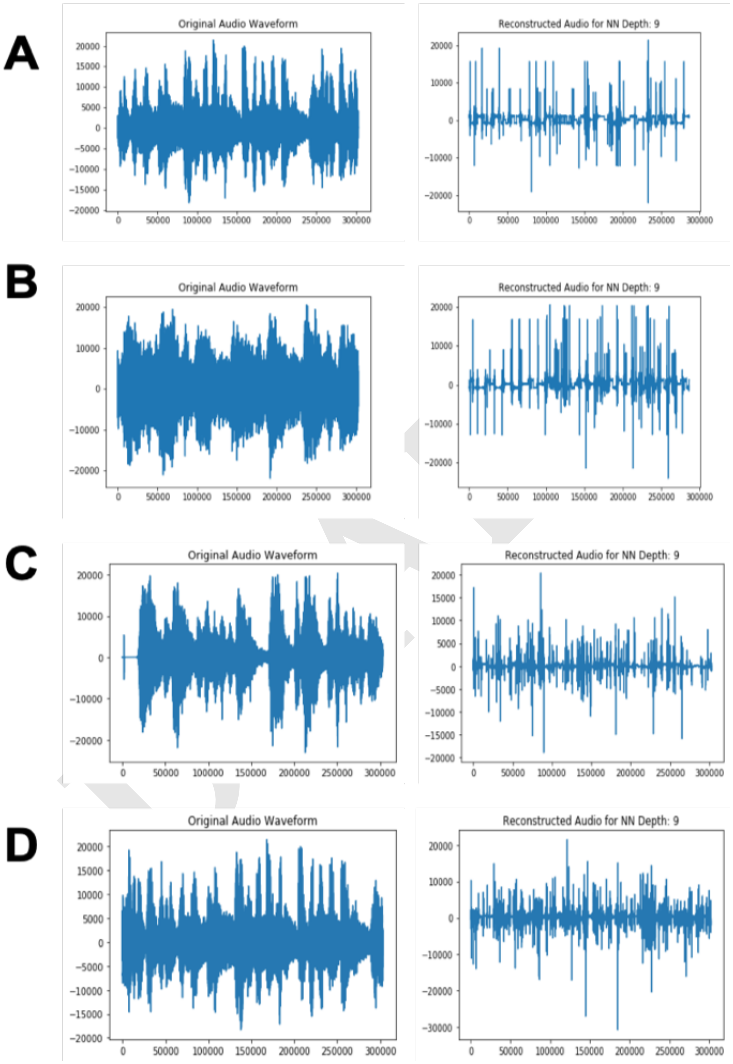

Fig. 6. Audio Reconstructions (A) Original audio waveform and predicted audio for training EEG. (B) Original audio waveform and predicted audio for testing EEG. (C) Original audio waveform and predicted audio after training on 90% of the songs for a single subject. (D) Original audio waveform and predicted audio after training on 90% of the subjects for a single song.

Conclusions and Future Directions

Decoding information from EEG data has been a growing field mostly due to the availability of high quality data as well as multimodal datasets. From recording patients prior to having seizures up until recording subjects while they listen to music, EEG data is in abundance [8]. With the increased resources on neural network architecture, researchers have been able to translate artistic style to pictures [9] and music [10-12]. What we have proposed in our investigation is a first step towards reconstructing music solely form 64 channel EEGs. Although the reconstructed versions of the songs were not qualitatively similar to the original audio waveforms, features of the original were in fact preserved in the final transposition. This leads us to believe that the method we employed here is a first step towards future progress in understanding how our brain processes sound.

We have methodological considerations for future studies. For our investigation to have yielded a more accurate reconstruction of EEG into music, there are a few specific things that need to be addressed. First of all, the epoched EEG limited the amount of training data we had available. When the dataset was preprocessed, we cut the EEG data to the shortest trial in the set, which was 6.87 seconds, for symmetry and to facilitate parsing. On top of that, there were numerous instances where we downsampled both EEG and audio data to minimize computational expenses. Downsampling influences both the amount of data to train on and the granularity of learning. Moreover, when reconstructing the song back to its original sampling rate, a lot of information is lost in the interpolation process. While these improvements are specific to the data processing and training components of our investigation, it is worth mentioning other architectures that could have potentially outperformed ours.

A look at how the R Matrix was constructed and a convolutional neural network comes to mind. In fact, it was the runner up in the final architecture decision making process. However, because we wanted to be in greater control of what we feed the algorithm through the batch sequential learning technique we ultimately employed. Another interesting and alternative learning architecture would have been to use a recurrent neural network, specifically an LSTM. The idea behind the long short term memory system would be best suited for a different purpose than reconstruction. Where we believe it could be interesting to explore the properties of an LSTM would be for song completion. The output of the LSTM would be whatever the architecture learns to be the next sequence of audio after having learned a previous sequence. Albeit there is no concrete reason to investigate this approach, it is both amusing and awesome to see the capabilities of an algorithm to make music.

The future of the field of thought processing is promising. With recent successes in style transfer in image and video, the study of the arts and machine learning is getting more and more traction[10]. A few research teams have begun the music reconstruction project independent of Gerwin Schalk, but we have yet to see any significant results. Even in style transfer, music is more complicated than initially predicted. While the signal is strictly speaking one dimensional and image data is 3 dimensional, what is encoded in the music time series signal is vastly more complicated than a conglomeration of RGB values. From timbre and note quality to instrument and vocals, the multi-modal nature of music makes it very difficult to transfer style between bands for example. However, one only one instrument is present, there has been progress made [11], reconstructing songs of Beethoven in the style of Mozart for example. Still, when it comes to reconstructions from EEG, music has not been a focus in the field. Shen et al have used deep learning to reconstruct images from a human fMRI data is a starting point for decoding what the brain is processing, and had yielded some decent results thus[13]. Other methods being studied now include new decoding architectures such as generative adversarial neural networks (GANs) [14]. The field is still very much open and we hope that our study here sparks interest to those with the skills and competency to build on so that one day, low fidelity but highly available EEG data can be used to reconstruct sounds perceived by human at all levels of complexity.

References

[1] Paul Y (2018) Various epileptic seizure detection techniques using biomedical signals: a review. Brain Inform 5(2):6.

[2] Potes C, Gunduz A, Brunner P, Schalk G (2012) Dynamics of electrocorticographic (ECoG) activity in human temporal and frontal cortical areas during music listening. Neuroimage 61(4):841–848.

[3] Nemrodov D, Niemeier M, Patel A, Nestor A (2018) The Neural Dynamics of Facial Identity Processing: Insights from EEG-Based Pattern Analysis and Image Reconstruction. eNeuro 5(1). doi:10.1523/ENEURO.0358-17.2018.

[4] Stober S, Sternin A, Owen AM, Grahn JA (2015) Towards Music Imagery Information Retrieval: Introducing the OpenMIIR Dataset of EEG Recordings from Music Perception and Imagination. ISMIR, pp 763–769.

[5] Puce A, Hämäläinen MS (2017) A Review of Issues Related to Data Acquisition and Analysis in EEG/MEG Studies. Brain Sci 7(6). doi:10.3390/brainsci7060058.

[6] Elgohary S, Eldawlatly S, Khalil MI (2016) Epileptic seizure prediction using zero-crossings analysis of EEG wavelet detail coefficients. 2016 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), pp 1–6.

[7] Esteller R, Echauz J, Tcheng T, Litt B, Pless B (2001) Line length: an efficient feature for seizure onset detection. 2001 Conference Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp 1707–1710 vol.2.

[8] Yao Y, Plested J, Gedeon T (2018) Deep Feature Learning and Visualization for EEG Recording Using Autoencoders. Neural Information Processing (Springer International Publishing), pp 554–566.

Jing Y, et al. (2017) Neural Style Transfer: A Review. arXiv [csCV]. Available at: http://arxiv.org/abs/1705.04058.

[9] Dai S, Zhang Z, Xia GG (2018) Music Style Transfer: A Position Paper. arXiv [csSD]. Available at: http://arxiv.org/abs/1803.06841.

[10] Mor N, Wolf L, Polyak A, Taigman Y (2018) A Universal Music Translation Network. arXiv [csSD]. Available at: http://arxiv.org/abs/1805.07848.

[11] Sturm I, Blankertz B, Potes C, Schalk G, Curio G (2014) ECoG high gamma activity reveals distinct cortical representations of lyrics passages, harmonic and timbre-related changes in a rock song. Front Hum Neurosci 8:798.

[12] Raposo F, de Matos DM, Ribeiro R, Tang S, Yu Y (2017) Towards Deep Modeling of Music Semantics using EEG Regularizers. arXiv [csIR]. Available at: http://arxiv.org/abs/1712.05197.

[13] Shen G, Dwivedi K, Majima K, Horikawa T, Kamitani Y (2019) End-to-End Deep Image Reconstruction From Human Brain Activity. Front Comput Neurosci 13:21.

[14] Donahue C, McAuley J, Puckette M (2018) Adversarial Audio Synthesis. arXiv [csSD]. Available at: http://arxiv.org/abs/1802.04208.