Interpretable Deep Generative Modeling of EEG*

Jack Lynch†
North Carolina State University

(BME 512 001)
(Dated: May 7, 2020)

EEG is one of the most prevalent paradigms of brain-signal measurement. A maturing understanding of relevant ECG features, including event-related potentials and neural oscillations, has produced a concurrent maturation of digital-signal processing techniques for extracting these features. Recent advances in deep generative modeling have enabled the integration of such strictly interpretable DSP into powerful deep architectures. This work will attempt to leverage these recent advances to model EEG data using interpretable, modular DSP components.

Purpose: Grading by Dr. Lalush.

I. INTRODUCTION

A. Electroencephalography (EEG)

Electroencephalography (EEG) is the cheapest and most widely-available non-invasive method of measuring electrical brain activity [1]. EEG utilizes electrodes placed at various points on the skull to measure voltage fluctuations related to neuronal ionic current [2]. Typically, analysis includes the study of event-related potentials (ERPs)—for example, in response to an input stimulus— or spectral content (identifying specific kinds of neural oscillations). EEG features a high temporal resolution [3] but a low spatial resolution, in part due to the low-pass characteristic of tissue between the brain and scalp [4]. Other limitations of the modality include an inability to measure "interior" neuronal activity and a low signal-to-noise ratio.

1. Event-Related Potentials (ERPs)

The study of event-related potentials consists of decomposing them into component voltage deflections with positive or negative parity and a given duration. Components are labeled by parity and duration—e.g., an N100 component has negative polarity and occurs 100ms after its associated stimululs. ERP's have been shown reliable indicators of different cognitive processes and have proven informative as features for downstream analysis, such as input stimulus reconstruction [5].

2. Neural Oscillations

In EEG, neural oscillations are only observed in aggregate, through spectral content. Analysis of neural oscillations benefits from an understanding of their aggregate activity patterns. Neural oscillations are believed to serve diverse funcitonal roles, from motor pattern generation [6] to such high-level cognitive faculties as perception [7].

B. Deep Generative Modeling

Generative modeling attempts to model a data distribution for purposes of sampling and analysis. The field of deep learning has seen the emergence of multiple dominant families of generative model, including variational autoencoders (VAEs), generative adversarial networks (GANs), and (in 1D) recurrent neural networks (RNNS).

1. Variational Autoencoders (VAEs)

Variational autoencoders (VAEs) are a variational extension of the deep autoencoder model. Instead of learning an efficient compression of a dataset, VAEs parametrize a latent probability distribution to match the data distribution. They employ an encoder network to reduce input data samples into distribution parameters and a decoder network to convert samples from the latent distribution into higher-dimensional samples like those from the target dataset. Once trained, a VAE enables the sampling of new data similar to that from the training distribution. [8].

2. Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) are a family of deep generative model involving competing networks. GAN training has been formulated traditionally as a minmax game in which the generator network learns to produce samples similar to samples from a dataset distribution, while the discriminator network learns to distinguish generated from authentic samples. This is equivalent to minimizing JS divergence between the modeled and actual dataset distributions [9]. Other formulations

* I resisted the urge to produce a stupid acronym ("DEEG"? "DIEEG"?), though it seems to be a reflex in deep-learning papers.

† jmlynch3@ncsu.edu

have been proposed, minimizing other distribution distance metrics like the Wasserstein earth-mover's distance [10] and the Pearson 2 divergence [11].

Generally, GANs suffer from instability during training and hyperparameter-dependent failures to converge; most novel GAN architectures seek to address these issues. However, GANs often greatly outperform other generative approaches (variational autoencoders, Boltzmann machines), especially at high resolutions [12]. GANs are typically used to generate images but have been shown effective at modeling time series as well [13]. In such cases, the primary building blocks can be convolutional or autoregressive.

Like many deep-learning image networks, GANs can be difficult to interpret [14, 15]. Recent work has attempted to allay this by forcing the generator to work with interpretable differentiable modules [16]. This approach yields other benefits, such as the ability to incorporate engineered features and domain knowledge into the training process.

3. Recurrent Neural Networks (RNNs)

Recurrent neural networks are distinguished by their use of previous inputs to the model. A traditional neural network can be thought of as a special case of an RNN where a single input produces a single output (disregarding batching, of course); but RNNs can also produce a single output for a sequence of inputs (classification), a single output from a sequence of inputs (generation), or a sequence of outputs from a sequence of inputs (translation/transfer) [17]. The most prominent iterations of RNNs, long short-term memory and gated recurrent unit networks (LSTMs and GRUs), accommodate long-term signal dependencies in different ways [18].

C. Deep Generative Modeling of EEG Signals

Though deep models have been used in the past for discriminative EEG modeling [19], generative approaches are comparatively rare. Recently, GANs have been used to model single-channel EEG signals associated with common tasks [20]. In this work a convolutional GAN was used to promote interpretability, though (at the time) no interpretation was made.

It seems intuitive that a generative model equipped with interpretable building blocks, especially those with biological relevance, might produce a more interpretable and effective model of the relevant EEG data. Such is the purpose of this project.[1]

[1] I had originally planned to do this with a GAN—the method is agnostic w/r/t the generative paradigm adopted—but realized, per your feedback and further research, that an adversarial loss was not necessary, in that an RNN would suffice. RNNs are much stabler and thus far more desirable.

II. METHODS

A. Deep Digital Signal Processing

The incorporation of traditional signal processing directly into gradient-driven deep learning is a relatively recent integration of the two families of methods. I follow the method of [16], training an encoder to embed an input signal into two familiar features—fundamental frequency f_0 and approximate "loudness" A—and training a GRU-RNN decoder to convert these latent features into parametrizations for basic synthesis tools: an additive synthesizer and a windowed FIR filter.

The additive synthesizer consists of a sum of sinusoids:

$$x(n) = \sum_{k=1}^{K} A_k(n) \sin\left(\phi_k(n)\right)$$

where the time-varying component sinusoid amplitudes $A_k(n)$ are multiples of the latent embedding feature $A(n)$ (loudness):

$$A_k(n) = A(n)c_k(n)$$

with normalized distribution over harmonics $c(n)$ ($\sum_k c_k(n) = 1$).

Similarly, the time-varying component sinusoid phases are multiples of the latent embedding feature $f_0(n)$ (the fundamental frequency), via

$$\phi_k(n) = 2\pi \sum_{m=0}^{n} f_k(m) + \phi_{0,k},$$
$$f_k(n) = k f_0(n)$$

The long and short of the above is that the additive synthesizer is handily parametrized entirely by the two latent features the encoder produces.

The output of the additive synth is summed with noise filtered through the aforementioned FIR filter. This result is then fed through a reverb stage via multiplication in the frequency domain.[2] This produces a synthetic reconstruction of the input signal; a multi-scale spectrogram loss guides the training process.

B. Data

1. EEG Correlates of Genetic Predisposition to Alcoholism

The dataset used [21] consists of 256 Hz, 64-channel, 1-second EEG recordings of alcoholic and non-alcoholic

[2] Of course, the need for reverb is not nearly as intuitive without an acoustic setting. In adversarial contexts, this could prove useful as a form of instance smoothing; but it seems lacking in strictly biological motivation.

subjects shown visual stimuli from [22]. Each subject underwent over one hundred trials and repeated each trial multiple times; however, some trial data was missing from certain subjects. A sample of the data is shown in Figure 1.

Here I limited my focus to a single channel (33: AF8), for simplicity and computational feasibility during training. I trained the aforementioned autoencoder to model alcoholic EEG signals, then attempted to "translate" signal characteristics between distributions (e.g., from a "sober" EEG signal to an "alcoholic" one). In such an application, the inherent interpretability of the basic synthesis tools used might prove illuminating.

The full set of one-second trial recordings were concatenated per-subject and divided into smaller chunks for processing.[3]

III. RESULTS

1. Training the Autoencoder

The autoencoder proved difficult to train, reaching a loss plateau early in training, as shown in Figure 2.

Though this magnitude loss corresponds to good performance on more conventional auditory datasets, it did not for this dataset: while the spectrogram loss did produce similar overall spectra (as shown in Figre 3), it underfitted severely in terms of timelike similarity, as shown in Figure 4.

It is unclear why the autoencoder failed to fully converge on the dataset. It's possible it would have progressed to a more sophisticated minimum eventually; if not, perhaps a thorough hyperparameter search would help. Though I had imagined the additive synthesizer a capable representative tool, it's possible the embedding features chosen constituted too aggressive a bottleneck. More on this in the conclusion.

It's also possible an L1 loss is not ideal for this dataset. I've heard that GANs prefer cosine proximity loss when dealing with spectrograms, but I'm not sure why that would be.

2. Signal Characteristic Transfer

As the autoencoder underfitted, it was of course difficult to transfer signal characteristics as planned. Attempting to "translate" from a control sample to an alcoholic one was inconclusive, as shown in Figure 6. The spectral results are shown in Figure 7.

In the very least, it's clear the translated *spectrum* is more similar to the true alcoholic spectrum—note the absence of lobes—but the time domain remains problematic.

IV. CONCLUSION

Overall, more time/compute is needed to determine the severity of the weaknesses observed. I'm confident it's a matter of hyperparameter-tuning. For this I would have to optimize the data pipeline: preprocessing the many minutes of EEG data for use with TensorFlow takes far, far longer than the training itself (and, more damningly, happens on the CPU).

Much could be done to make the modeling approach more targeted, including the implementation of new DSP modules tailored specifically to the EEG modality. In fact, I spoke recently with the authors of [16] about adding more EEG-specific blocks to their model, and will be in touch in the coming weeks about it.

Further work could also extend this approach to more channels, or perhaps operate upon a learned single-channel embedding of the 64-channel whole. Additionally, this approach could be made conditional and restricted to specific stimuli, to further isolate the influence of addictive correlates on the signals.

I'll keep at it—

[1] P. M. Vespa, V. Nenov, and M. R. Nuwer, Continuous EEG monitoring in the intensive care unit: Early findings and clinical efficacy, Journal of Clinical Neurophysiology **16 (1)**, 1 (1999).

[2] E. Niedermeyer and F. L. da Silva, *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields* (Lippincott Williams Wilkins, 2004).

[3] M. Hämäläinen, R. Hari, R. J. Ilmoniemi, J. Knuutila, and O. V. Lounasmaa, Magnetoencephalography-theory, instrumentation, and applications to noninvasive studies of the working human brain, Reviews of Modern Physics **65 (2)**, 413 (1993).

[4] C. Ramon, W. J. Freeman, M. Holmes, A. Ishimaru, J. Haueisen, P. H. Schimpf, and E. Rezvanian, Similarities between simulated spatial spectra of scalp EEG, MEG and structural MRI, Brain topography **22 (3)**, 191 (2009).

[5] D. Nemrodov, M. Niemeier, A. Patel, and A. Nestor, The neural dynamics of facial identity processing: Insights from EEG-based pattern analysis and image reconstruction, eNeuro **5 (1)**, ENEURO.035.17.2018 (2018).

[6] E. Marder and D. Bucher, Central pattern generators and

[3] I concatenated the trial recordings to produce larger overall data points. This was required by a bug in Google's codebase, which is still in significant flux—major and relevant changes were made to it during the course of this project, which you can imagine was not ideal—but this concatenation is not ultimately obfuscatory to the training process, as the trials are actually contiguous in time.

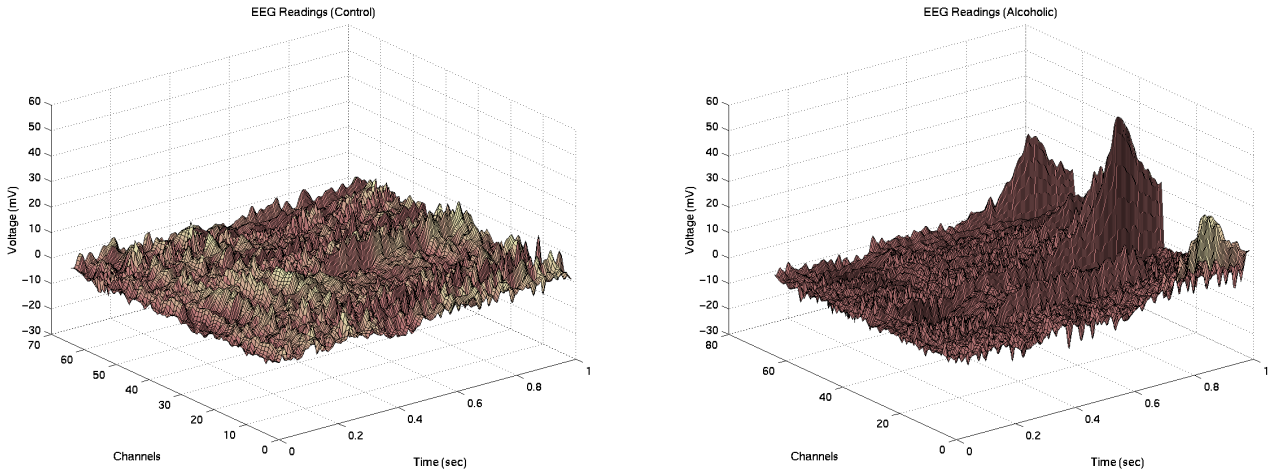

FIG. 1. Figures from [21] illustrating differences in EEG activity between control (left) and alcoholic (right) subjects. Available here and here.

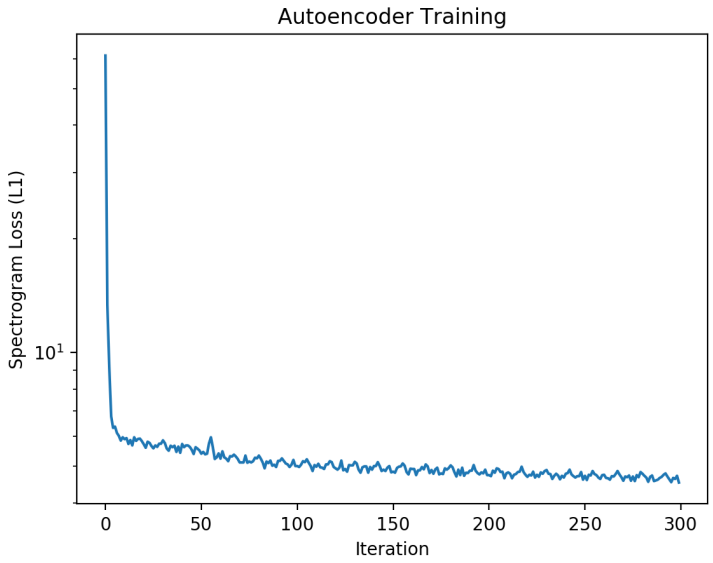

FIG. 2. Spectrogram loss during training. Loss equilibrated early.

the control of rhythmic movements, Current Biology **11 (23)**, R986 (2001).
[7] C. M. Gray, P. König, A. K. Engel, and W. Singer, Oscillatory responses in cat visual cortex exhibit intercolumnar synchronization which reflects global stimulus properties, Nature **338 (6213)**, 334 (1989).
[8] D. P. Kingma and M. Welling, An introduction to variational autoencoders, Foundations and Trends® in Machine Learning **12**, 307–392 (2019).
[9] I. Goodfellow *et al.*, Generative adversarial networks, arXiv:1406.2661 [stat.ML] (2014).
[10] M. Arjovsky *et al.*, Wasserstein GAN, arXiv:1701.07875 [stat.ML] (2017).
[11] X. Mao *et al.*, Least squares generative adversarial networks, arXiv:1611.04076 [cs.CV] (2017).
[12] T. Karras *et al.*, Progressive growing of GANs for improved quality, stability, and variation, arXiv:1710.10196 [cs.NE] (2017).
[13] C. Donahue *et al.*, Adversarial audio synthesis, arXiv:1802.04208 [cs.SD] (2018).
[14] C. Olah *et al.*, The building blocks of interpretability, Distill (2018).
[15] S. Carter *et al.*, Activation atlas, Distill (2019).
[16] J. Engel, L. H. Hantrakul, C. Gu, and A. Roberts, Ddsp: Differentiable digital signal processing, in *International Conference on Learning Representations* (2020).
[17] A. Amidi and S. Amidi, Recurrent neural networks cheatsheet, Stanford CSC 230 .
[18] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, Empirical evaluation of gated recurrent neural networks on se-

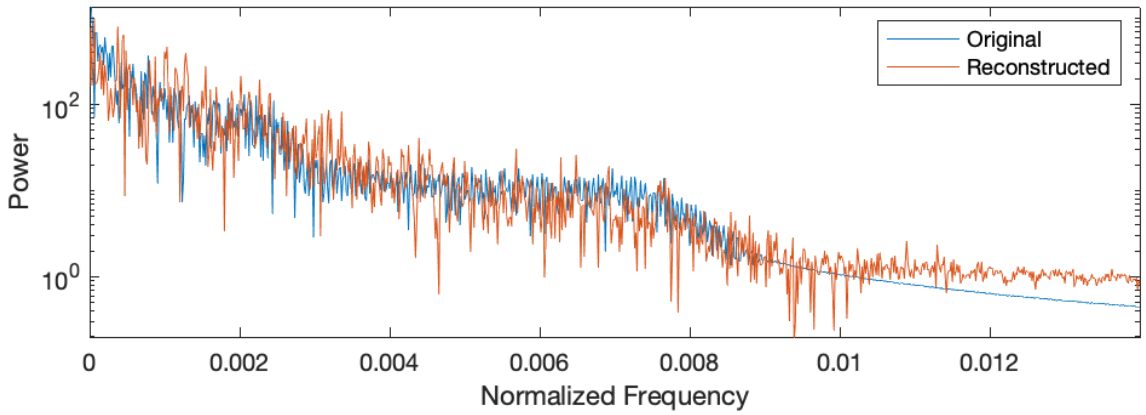

FIG. 3. Original and reconstructed spectra.

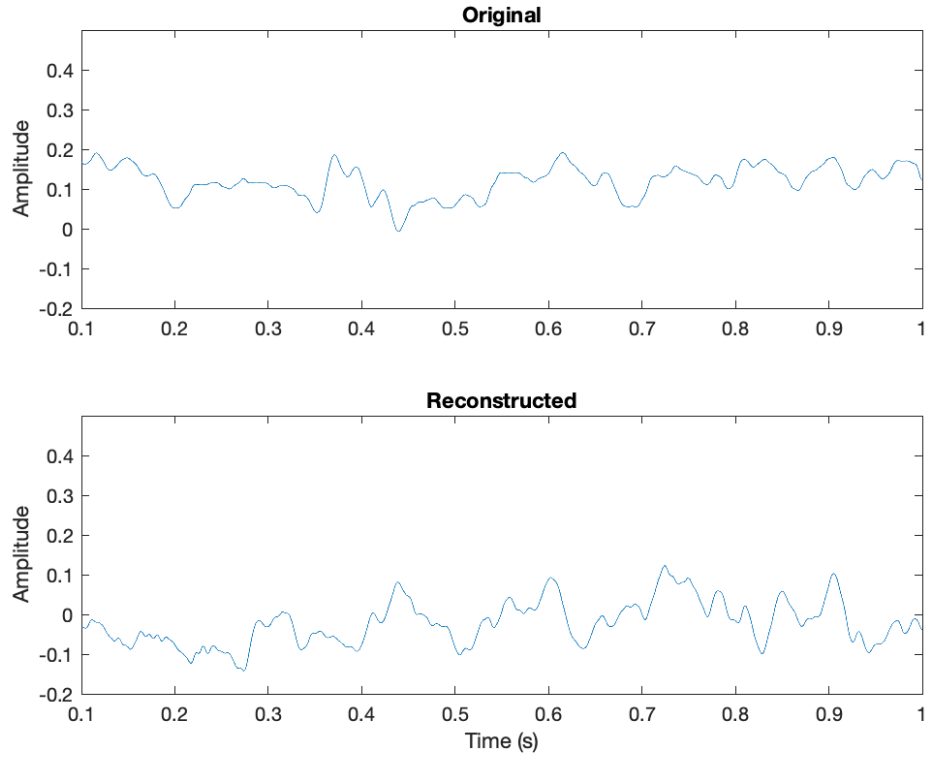

FIG. 4. Original and reconstructed signals.

quence modeling (2014), arXiv:1412.3555.

[19] C. Alexander *et al.*, Deep learning for electroencephalogram (EEG) classification tasks: a review, Journal of Neural Engineering **16**, 031001 (2019).

[20] K. G. Hartmann *et al.*, EEG-GAN: Generative adversarial networks for electroencephalograhic (EEG) brain signals, arxiv:1806.01875 [eess.SP] (2018).

[21] H. Begleiter, Eeg database data set, UCI Machine Learning Repository (1999).

[22] J. G. Snodgrass and M. Vanderwart, A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity, Journal of Experimental Psychology: Human Learning and Memory **6(2)**, 174 (1980).

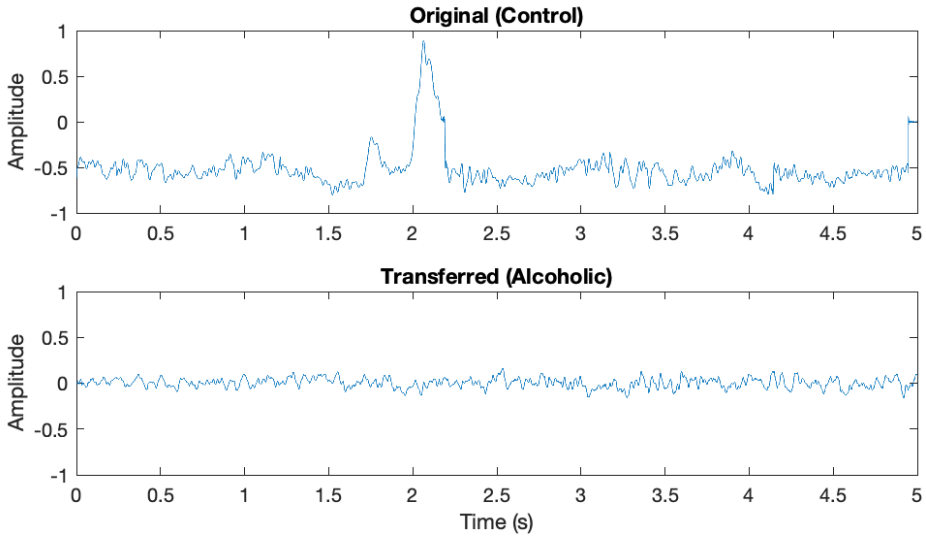

FIG. 5. Original and "translated" signals.

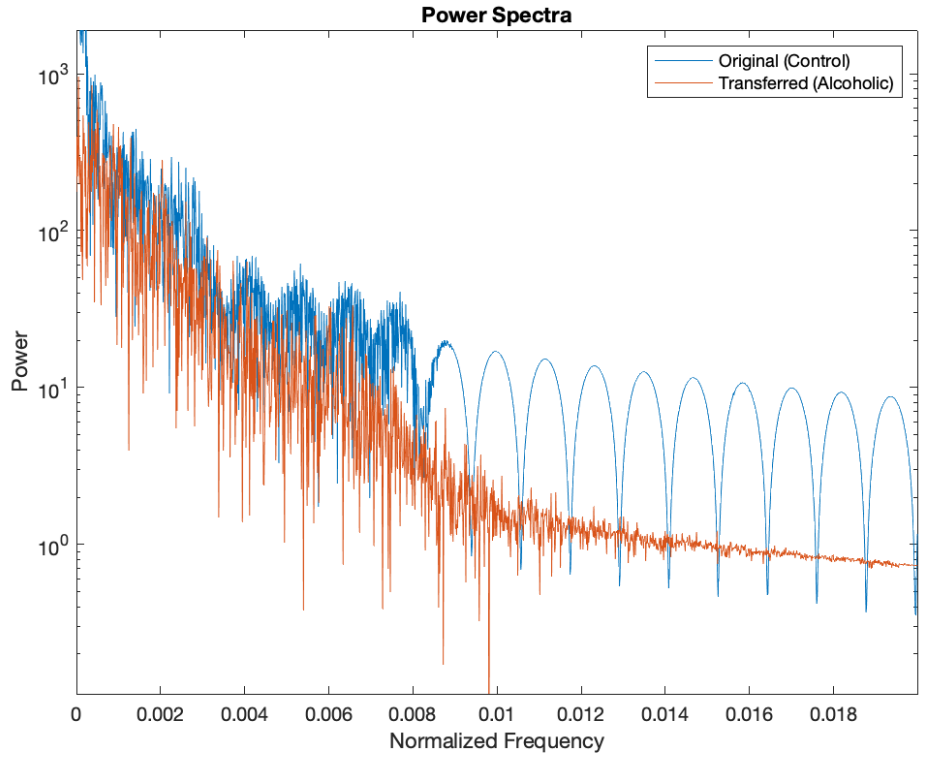

FIG. 6. Original and "translated" spectra.