Decoding Imagined Auditory Pitch Phenomena with an Autoencoder Based Temporal Convolutional Architecture

Sean Paulsen Lloyd May Michael Casey

Dartmouth College

DARTMOUTH

Overview

What is Brain Decoding?

- Brain decoding is the problem of **classifying the stimulus** that evoked given brain activity
- Stimulus decoding of functional magnetic resonance imaging (fMRI) data with machine learning models has provided new insights about neural representational spaces and task-related dynamics

What do we want to decode?

- We sought to decode the pitch-class (relative position of a note in a scale) of audio both **heard** and **imagined** during scanning
- However, the power of machine learning models is heavily dependent on adequate dataset sizes, and fMRI studies in particular suffer from notorious **data poverty**

Why music?

- Music's well-defined structure and the wealth of previous results about the neural representation of that structure are thus an appealing foundation for approaching brain decoding

Decoding Tasks

- "**Heard task**:" Classifier is trained and tested on neural activity evoked by heard auditory stimuli
- "**Imagined task**:" Classifier is trained and tested on neural activity while imagining particular pitches
- "**Cross-decoding task**:" Models trained on "heard" data but tested on "imagined" data, to explore overlap between the two processes

Inspiration

- Rather than using a deep model for classification, we used a deep model to augment our dataset and thereby enhance the downstream performance of a more basic support vector machine (SVM) classifier
- *Firat et al.* (2015) demonstrated this approach by improving downstream classifier performance of decoding visual stimuli

Inspiration cont'd

- More details further on, but broadly speaking they used a deep model to learn neural activation patterns latent in their **unlabelled** fMRI data, which are normally discarded, and used the knowledge of these patterns to transform the **labelled** data into a form more easily learned by the downstream classifier
- Observe, performance on data of interest was significantly improved by exploiting data that is almost universally ignored and thrown away. Certainly an inspiring solution, and one might conclude that fMRI data poverty is at least partially **self-inflicted**
- We hypothesized that these benefits would extend to the auditory domain, and thus their work guided the design of our pipeline
- We refer to the result of the final transformation as the "**encoded dataset**" throughout both our paper and these slides

Scanning

- 18 participants
- Each participant's fMRI scan consisted of 8 runs of 21 musical trials and was randomly assigned either the key of E Major or F Major, which was not known by the participant
- Each trial began with an arpeggio in the assigned key for the participant to internally establish a tonal context, followed by a cue-sequence of ascending notes in their assigned major scale

Scanning cont'd

- After a randomized time interval, the participant either **heard** the next ascending note in the scale (data for the "**heard task**"), or was instructed to **imagine** the next ascending note (data for the "**imagined task**")
- Runs alternated between entirely **heard task** and entirely **imagined task**
- Previous literature on imagined and heard tonal pitch-classes directed us to twenty regions of interest in the frontal, temporal, and parietal lobes according to the Desikan-Killiany (D-K) atlas, which are consistent with the scales of relevant previous literature

Classification

- For each pair of participant and ROI, and for each of the heard, imagined, and cross-decoding tasks, we trained a multi-class SVM classifier by inputting the preprocessed voxel data corresponding to the participant **hearing** or **imagining** the next note in the sequence, and outputting a prediction of the pitch class label of that note
- The above process was repeated for each corresponding **encoded dataset** to compare performance

What Kind of Patterns Are We Looking For?

- The canonical hemodynamic response function is observed across 12 seconds, which equates to 6 of our measured timesteps
- Thus we look for the HRF and other latent patterns of activation across 6 TRs in our data

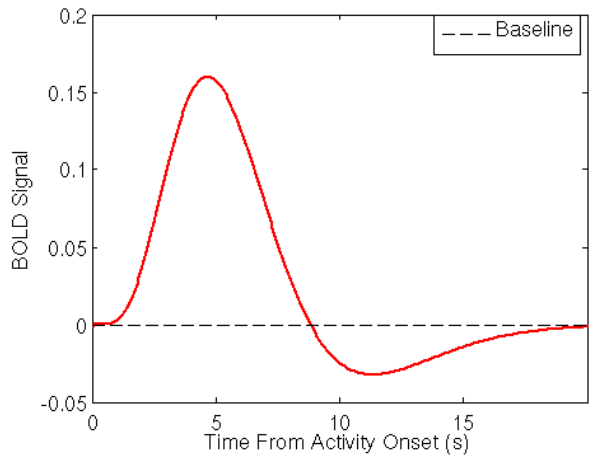


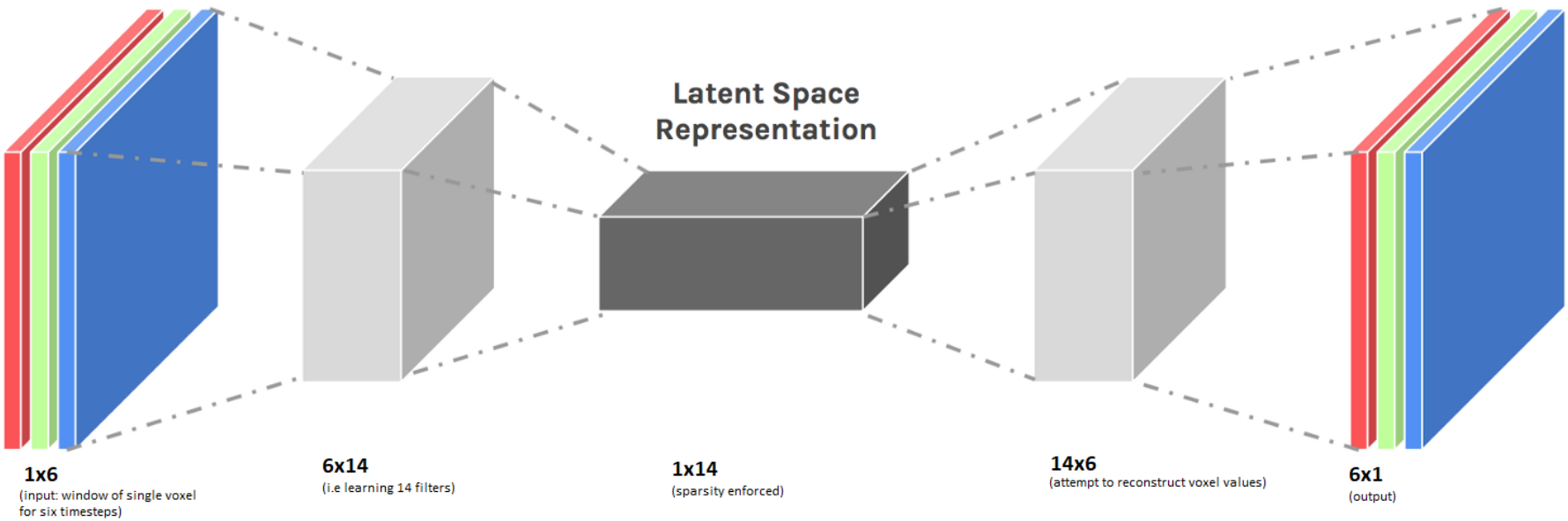
Figure: Image courtesy of https://theclevermachine.wordpress.com/tag/finite-impulse-response-model

Windows of Data

- So we want to consider each voxel separately, and for 6 timesteps at a time, but how should we sample this from the full VxT matrix of voxel data?
- Evenly spaced intervals will potentially miss patterns
- For example, if a voxel had values like "...11122233344..." then the sampling "...[111222][333444]..." would miss that 3 comes after 2, which is potentially even more important than 2 coming after 1
- Sampling every possible 6-TR window, i.e "...[111222][112223][1222333]..." from the preceding example would include every possible pattern but we believed that such significant redundancy could introduce unpredictable biases into the training
- Our solution, then, was to consider each possible 6-TR window at each voxel, but only include it in the training set for that autoencoder with probability $1/6$
- This allowed us to mitigate redundancy while including the potential to find patterns across any 6-TR window

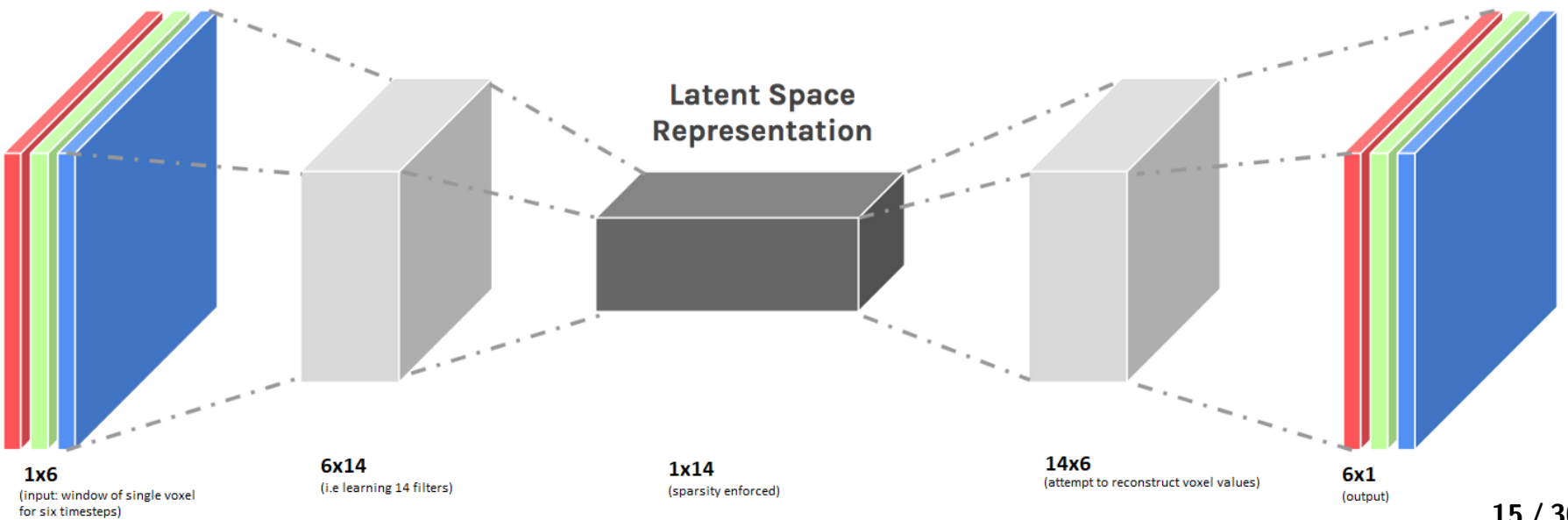
Sparse Autoencoders

- We implemented a **sparse autoencoder** model to perform unsupervised learning of the latent temporal neural activation patterns in our 6-TR windows of voxel data without the need for handcrafted feature



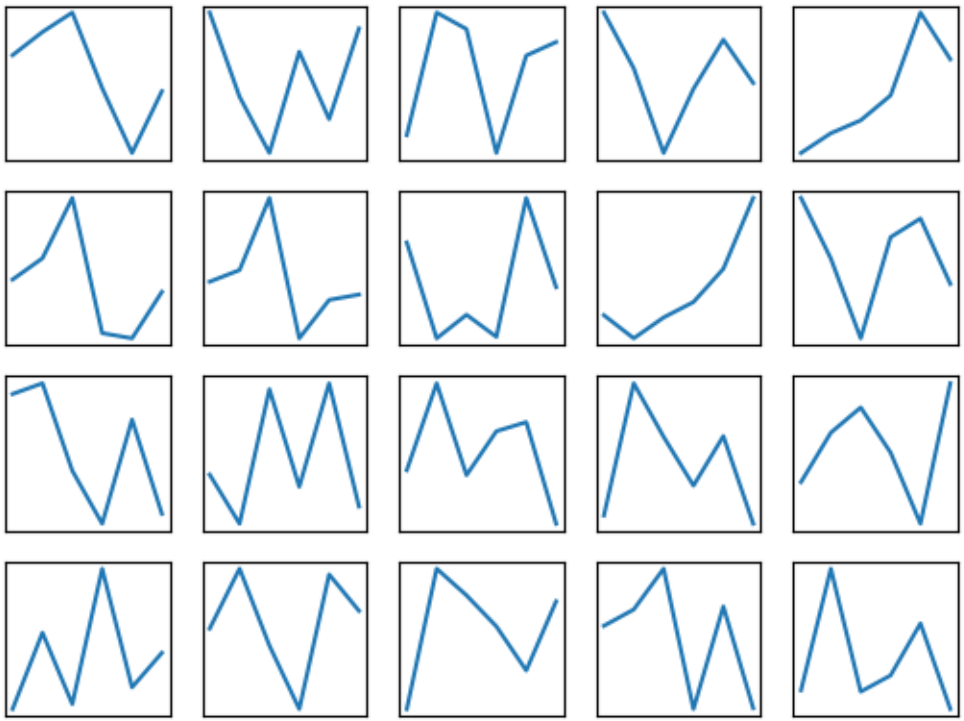
Sparse Autoencoders cont'd

- We hypothesized that by training this **sparse autoencoder**, each of the 14 neurons in the encoding layer, themselves each a 6x1 vector of weights, would learn to capture a sort of neural-activation basis vector, meaning the latent space representation is the expression of the input in terms of this learned basis



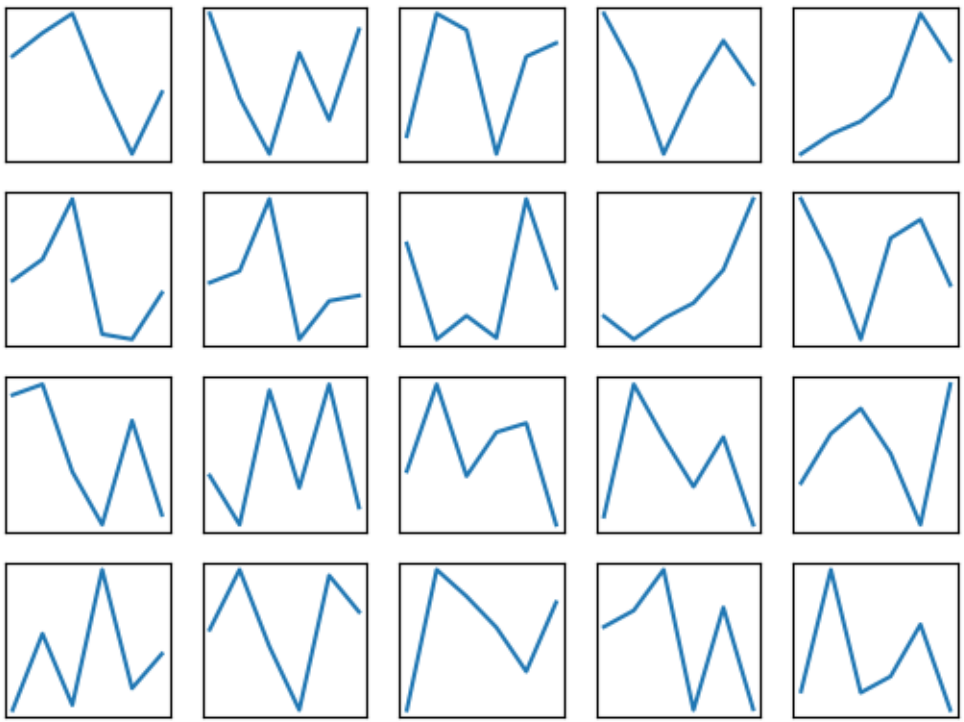
Visualizing Learned Filters

- We plotted a random sample of fully trained 6x1 neurons as timeseries to visualize them as patterns of latent neural activity



Learned HRF

- The hemodynamic response function appears to have been approximated by several of them, which is reassuring

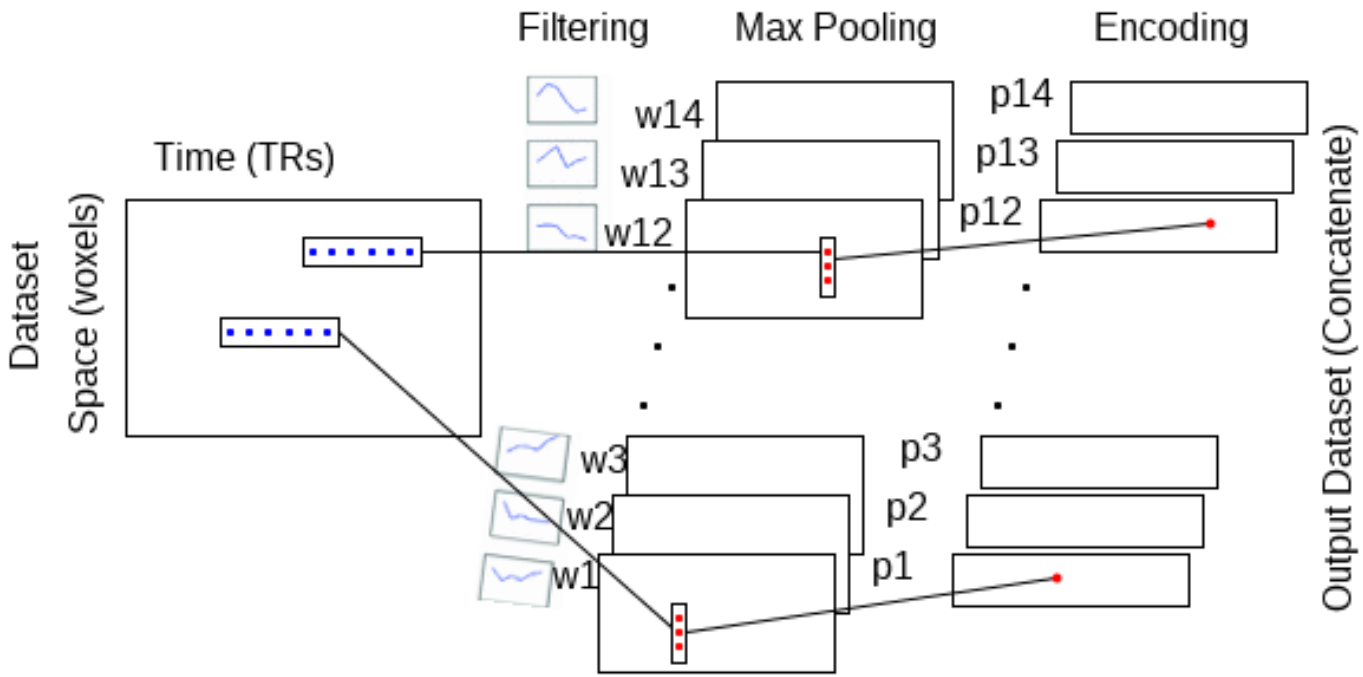


Filtering

- Each of the 14 learned patterns of neural activation could then be used as filters on the *labelled* data (recall that they were learned from *unlabelled* data)
- Simply put, now that we know what we're looking for, let's filter the original voxel data to emphasis those things

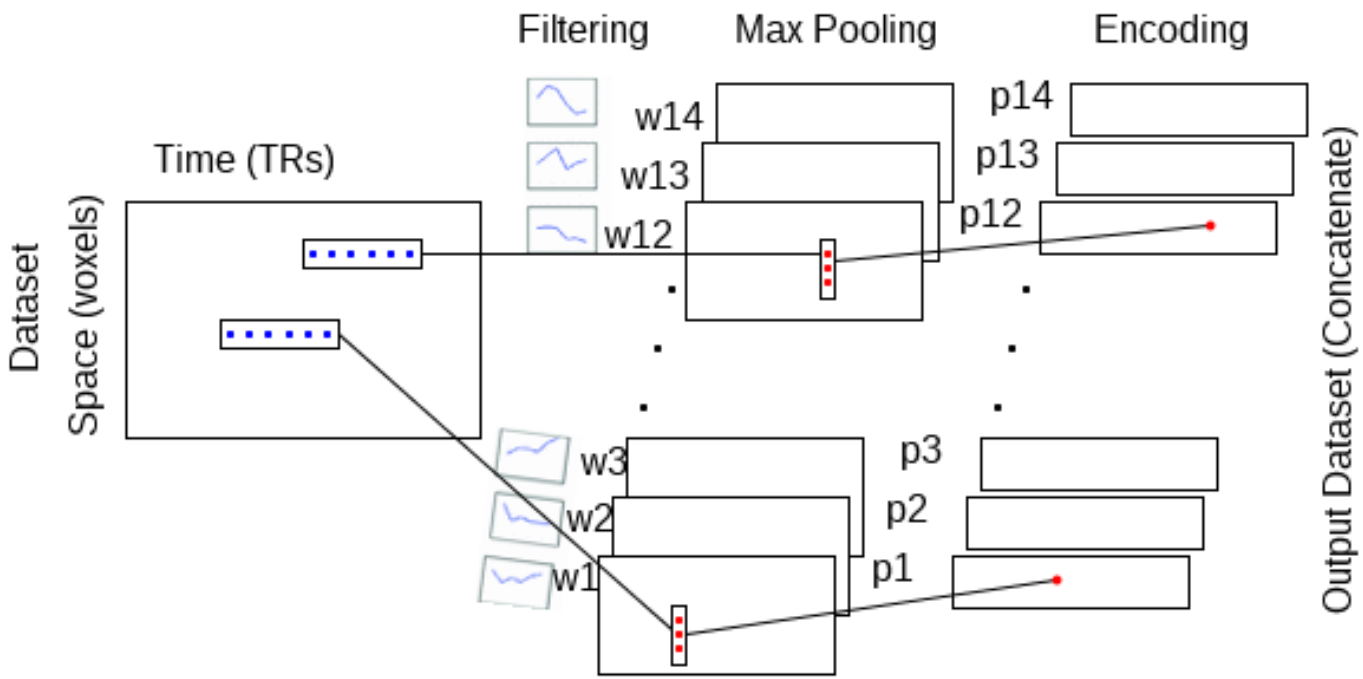
Temporal CNN

- We used a temporal convolutional architecture implemented from scratch using the Keras library to accomplish this



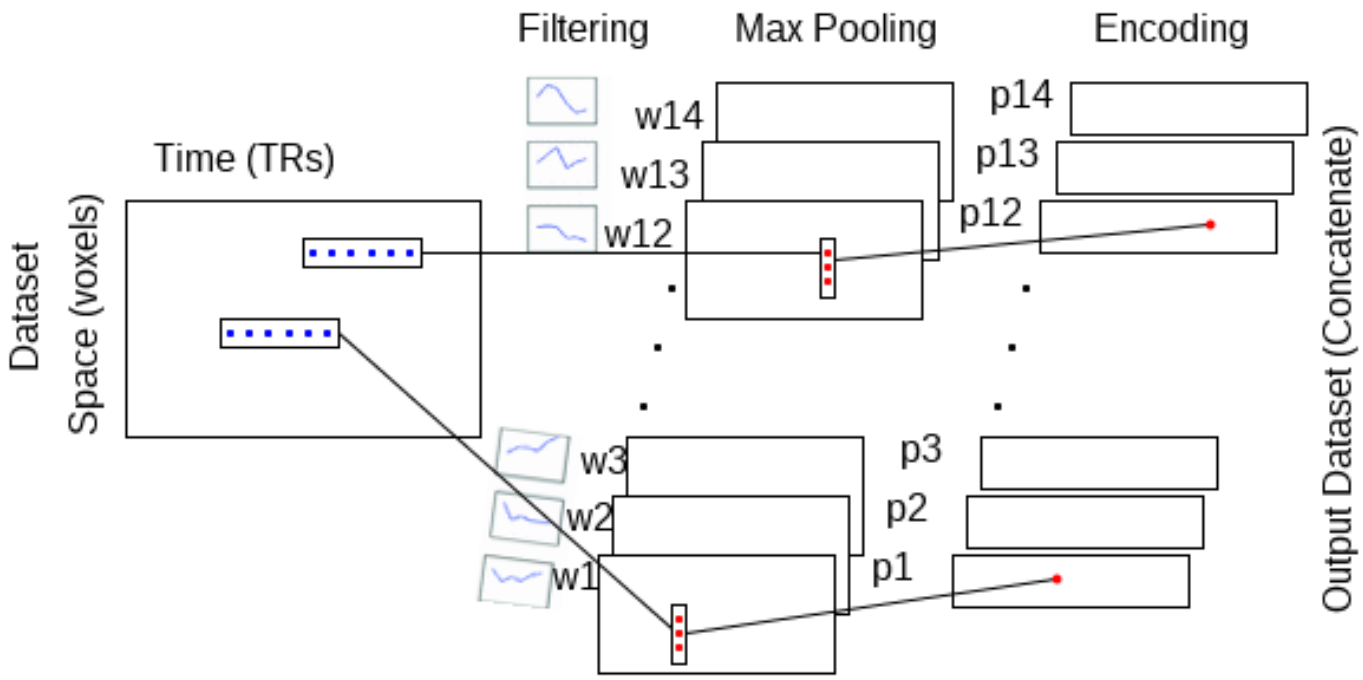
Response Matrices

- The "filtering" step involves full temporal convolution using the 14 learned activation patterns as filters
- This results in 14 **response matrices**, each of which is simply the result of temporal convolution using one of the filters



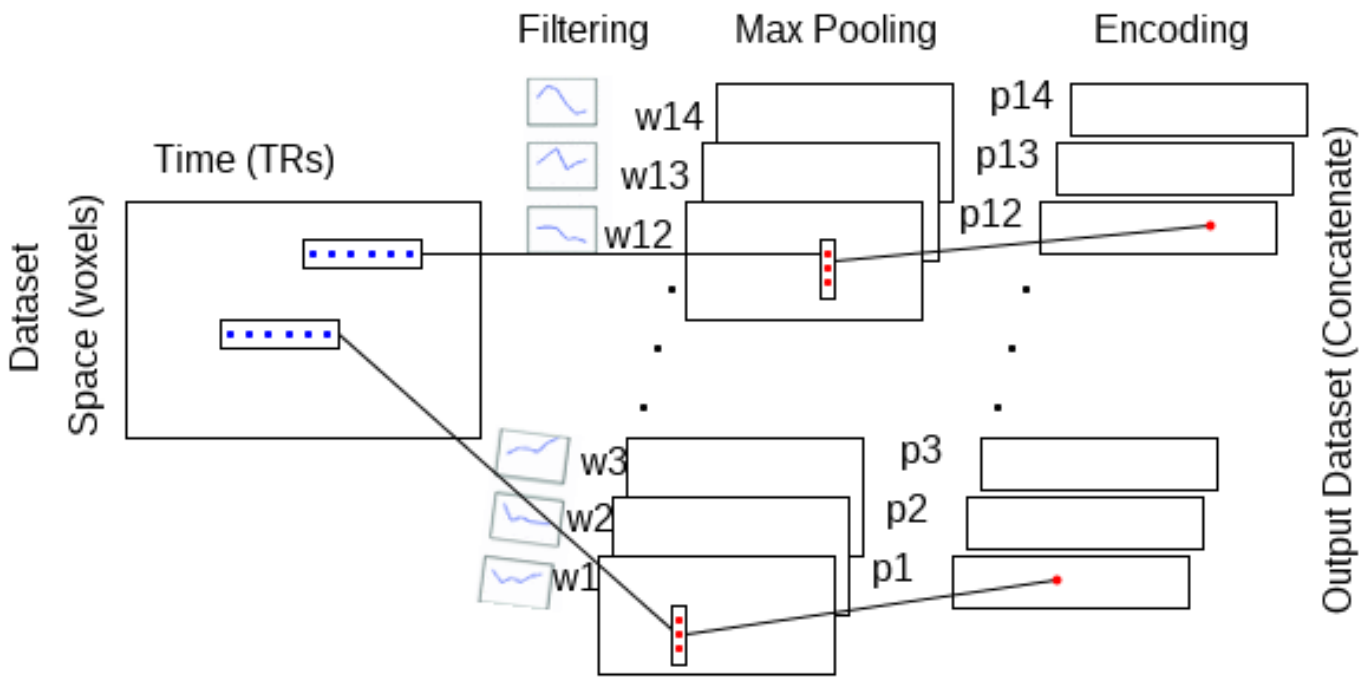
Encoded Datasets

- Max pooling is performed on each response matrix, but the values must first be mapped back to 3-D space since nearby values in the flattened input need not be nearby geographically
- This required quite a bit of care, which is detailed in the paper
- The upshot is that we max-pooled in 3-D space (i.e brain space) with a kernel size of 8 and then flattened



Encoded Datasets

- The 14 resulting flattened vectors were concatenated to create the **encoded dataset** for that participant/ROI pair



Training

- Finally, we trained separate linear SVM classifiers on the *labelled* preprocessed VxT voxel data and the **encoded dataset** resulting from inputting that data to our pipeline, for each participant and ROI, with the pitch class label of the **heard** or **imagined** sound corresponding to each timestep as the target
- We calculated group level significance for each of the **heard**, **imagined**, and **cross-decoding tasks** and each ROI using a t-test between per-participant prediction mean accuracies and null decoding model mean accuracies, detailed extensively in the paper

Significance Testing

		Encoded Dataset		Unencoded Dataset	
Region of Interest	**Task**	**WPC Accuracy Mean (Min, Max) baseline = 0.1429**	**FDR-corrected pvals (threshold = 0.05) (20 ROIs)**	**WPC Accuracy Mean (Min, Max) baseline = 0.1429**	**FDR-corrected pvals (threshold = 0.05) (20 ROIs)**
Left Heschl's Gyrus	H	0.1642 (0.1250, 0.1964)	**0.0039**	0.1523 (0.0833, 0.2262)	0.5808
Right Superior Temporal Sulcus	H	0.1394 (0.0833, 0.2143)	0.8554	0.1754 (0.1190, 0.2560)	**0.0071**
Left Inferior Frontal Gyrus (Orbitalis)	I	0.1625 (0.0893, 0.2381)	**0.0368**	0.1485 (0.0952, 0.2024)	0.6475
Left Precentral Gyrus	I	0.1607 (0.1190, 0.2262)	**0.0368**	0.1530 (0.0952, 0.2202)	0.5228
Left Superior Temporal Gyrus	I	0.1684 (0.1190, 0.2202)	**0.0087**	0.1586 (0.1131, 0.2143)	0.2326
Left Supramarginal Gyrus	I	0.1642 (0.0952, 0.2440)	**0.0355**	0.1502 (0.0893, 0.2083)	0.6291
Left Insula	I	0.1649 (0.1310, 0.2440)	**0.0163**	0.1478 (0.0833, 0.2024)	0.6475
Right Superior Temporal Sulcus	I	0.1604 (0.1131, 0.2083)	**0.0180**	0.1499 (0.0893, 0.2083)	0.6291
Right Inferior Frontal Gyrus (Triangularis)	I	0.1688 (0.0952, 0.2500)	**0.0368**	0.1642 (0.1131, 0.2202)	0.0996
Right Precentral Gyrus	I	0.1719 (0.1071, 0.2321)	**0.0103**	0.1569 (0.0952, 0.2560)	0.5228
Right Superior Temporal Gyrus	I	0.1726 (0.1190, 0.2560)	**0.0124**	0.1453 (0.1012, 0.1845)	0.8149
Right Supramarginal Gyrus	I	0.1656 (0.1250, 0.2440)	**0.0251**	0.1586 (0.0774, 0.2440)	0.5228
Right Insula	I	0.1649 (0.1190, 0.2381)	**0.0124**	0.1506 (0.0893, 0.2024)	0.6291
Right Superior Temporal Gyrus	X	0.1628 (0.1310, 0.1964)	**0.0157**	0.1492 (0.1071, 0.2083)	0.6445
Right Rostral-Middle Frontal Gyrus	X	0.1509 (0.1250, 0.1905)	0.3619	0.1642 (0.1131, 0.2143)	**0.0202**

- Observe one of our critical results, that thirteen of the fifteen successful regions required the **encoded dataset** to obtain statistical significance

Significance Testing cont'd

		Encoded Dataset		Unencoded Dataset	
Region of Interest	**Task**	**WPC Accuracy Mean (Min, Max) baseline = 0.1429**	**FDR-corrected pvals (threshold = 0.05) (20 ROIs)**	**WPC Accuracy Mean (Min, Max) baseline = 0.1429**	**FDR-corrected pvals (threshold = 0.05) (20 ROIs)**
Left Heschl's Gyrus	H	0.1642 (0.1250, 0.1964)	**0.0039**	0.1523 (0.0833, 0.2262)	0.5808
Right Superior Temporal Sulcus	H	0.1394 (0.0833, 0.2143)	0.8554	0.1754 (0.1190, 0.2560)	**0.0071**
Left Inferior Frontal Gyrus (Orbitalis)	I	0.1625 (0.0893, 0.2381)	**0.0368**	0.1485 (0.0952, 0.2024)	0.6475
Left Precentral Gyrus	I	0.1607 (0.1190, 0.2262)	**0.0368**	0.1530 (0.0952, 0.2202)	0.5228
Left Superior Temporal Gyrus	I	0.1684 (0.1190, 0.2202)	**0.0087**	0.1586 (0.1131, 0.2143)	0.2326
Left Supramarginal Gyrus	I	0.1642 (0.0952, 0.2440)	**0.0355**	0.1502 (0.0893, 0.2083)	0.6291
Left Insula	I	0.1649 (0.1310, 0.2440)	**0.0163**	0.1478 (0.0833, 0.2024)	0.6475
Right Superior Temporal Sulcus	I	0.1604 (0.1131, 0.2083)	**0.0180**	0.1499 (0.0893, 0.2083)	0.6291
Right Inferior Frontal Gyrus (Triangularis)	I	0.1688 (0.0952, 0.2500)	**0.0368**	0.1642 (0.1131, 0.2202)	0.0996
Right Precentral Gyrus	I	0.1719 (0.1071, 0.2321)	**0.0103**	0.1569 (0.0952, 0.2560)	0.5228
Right Superior Temporal Gyrus	I	0.1726 (0.1190, 0.2560)	**0.0124**	0.1453 (0.1012, 0.1845)	0.8149
Right Supramarginal Gyrus	I	0.1656 (0.1250, 0.2440)	**0.0251**	0.1586 (0.0774, 0.2440)	0.5228
Right Insula	I	0.1649 (0.1190, 0.2381)	**0.0124**	0.1506 (0.0893, 0.2024)	0.6291
Right Superior Temporal Gyrus	X	0.1628 (0.1310, 0.1964)	**0.0157**	0.1492 (0.1071, 0.2083)	0.6445
Right Rostral-Middle Frontal Gyrus	X	0.1509 (0.1250, 0.1905)	0.3619	0.1642 (0.1131, 0.2143)	**0.0202**

- Eleven of the fifteen significant results were for the **imagined task**, and indeed all of these regions required the **encoded dataset** for significance

Significance Testing cont'd

		Encoded Dataset		Unencoded Dataset	
Region of Interest	**Task**	**WPC Accuracy Mean (Min, Max) baseline = 0.1429**	**FDR-corrected pvals (threshold = 0.05) (20 ROIs)**	**WPC Accuracy Mean (Min, Max) baseline = 0.1429**	**FDR-corrected pvals (threshold = 0.05) (20 ROIs)**
Left Heschl's Gyrus	H	0.1642 (0.1250, 0.1964)	**0.0039**	0.1523 (0.0833, 0.2262)	0.5808
Right Superior Temporal Sulcus	H	0.1394 (0.0833, 0.2143)	0.8554	0.1754 (0.1190, 0.2560)	**0.0071**
Left Inferior Frontal Gyrus (Orbitalis)	I	0.1625 (0.0893, 0.2381)	**0.0368**	0.1485 (0.0952, 0.2024)	0.6475
Left Precentral Gyrus	I	0.1607 (0.1190, 0.2262)	**0.0368**	0.1530 (0.0952, 0.2202)	0.5228
Left Superior Temporal Gyrus	I	0.1684 (0.1190, 0.2202)	**0.0087**	0.1586 (0.1131, 0.2143)	0.2326
Left Supramarginal Gyrus	I	0.1642 (0.0952, 0.2440)	**0.0355**	0.1502 (0.0893, 0.2083)	0.6291
Left Insula	I	0.1649 (0.1310, 0.2440)	**0.0163**	0.1478 (0.0833, 0.2024)	0.6475
Right Superior Temporal Sulcus	I	0.1604 (0.1131, 0.2083)	**0.0180**	0.1499 (0.0893, 0.2083)	0.6291
Right Inferior Frontal Gyrus (Triangularis)	I	0.1688 (0.0952, 0.2500)	**0.0368**	0.1642 (0.1131, 0.2202)	0.0996
Right Precentral Gyrus	I	0.1719 (0.1071, 0.2321)	**0.0103**	0.1569 (0.0952, 0.2560)	0.5228
Right Superior Temporal Gyrus	I	0.1726 (0.1190, 0.2560)	**0.0124**	0.1453 (0.1012, 0.1845)	0.8149
Right Supramarginal Gyrus	I	0.1656 (0.1250, 0.2440)	**0.0251**	0.1586 (0.0774, 0.2440)	0.5228
Right Insula	I	0.1649 (0.1190, 0.2381)	**0.0124**	0.1506 (0.0893, 0.2024)	0.6291
Right Superior Temporal Gyrus	X	0.1628 (0.1310, 0.1964)	**0.0157**	0.1492 (0.1071, 0.2083)	0.6445
Right Rostral-Middle Frontal Gyrus	X	0.1509 (0.1250, 0.1905)	0.3619	0.1642 (0.1131, 0.2143)	**0.0202**

- **Imagining** sound is a more involved process than **hearing**, giving the models more information to learn from, to which we attribute the greater efficacy on the **imagined task**

First Goal Achieved?

- Our first goal was to learn auditory neural activation patterns latent in 6-TR windows of *unlabelled* fMRI data with **sparse autoencoders**
- Several of the plotted neurons are good approximations of the hemodynamic response function, which we expected to be learned by one of the neurons in most of the autoencoders
- Further, none of the patterns are dominated by a single timestep, and the peaks of activity are fairly well distributed across the timesteps, which was the intent of our training data collection method
- These considerations, along with the success of our brain decoding classifiers, provide evidence that each neuron learned a latent auditory neural activation pattern, accomplishing our first goal

Second Goal Achieved?

- Our second goal was to generate a collection of **encoded datasets** by transforming the unencoded voxel data in terms of neural activation patterns latent in *unlabelled* data
- We accomplished this by learning the patterns with a **sparse autoencoder**, and then mapping to our **encoded datasets** with a temporal convolutional architecture

Third Goal Achieved?

- Our third goal was to train a machine learning classifier to predict the **pitch class** labels of **heard** and **imagined** pitches, trained and tested on fMRI data of twenty selected regions of interest
- We accomplished this by training multi-class support vector machines (SVM) with linear kernels on each set of preprocessed voxel data and their corresponding **encoded datasets**
- As shown in the previous table, the statistical significance of outperforming chance relied almost entirely on the **encoded datasets**, indicating that our pipeline reveals fundamental, learnable attributes of auditory imagery that would otherwise remain undetected by machine learning models trained without our pipeline
- Moreover, the significant results on the **cross-decoding task** provide a critical novel result- statistically significant evidence of geographical overlap between heard and imagined sound

Several Immediate Directions for Future Work

- First is toward an end-to-end architecture for this task, rather than a disconnected training session to obtain the **encoded datasets**
- Second is toward decoding/cross-decoding the other information in our fMRI protocol, such as the timbre (clarinet or trumpet) of the **heard** or **imagined** sound
- Third is toward the generalization of our pipeline to other fMRI datasets with auditory tasks
- Fourth is a deeper dive on the ROIs with significant **cross-decoding** results, as these results did not quite match our expectations