Reconstructing Speech from fMRI-Recorded Brain Activity

A Thesis Submitted

to the University of Texas at Austin

for the Turing Scholars program

by

Ashay Swadi

to

Department of Computer Science

College of Natural Sciences

University of Texas at Austin

May 2025

ABSTRACT

Brain decoding is the process of reconstructing stimuli by interpreting brain activity. In particular, given brain signals recorded using functional magnetic resonance imaging (fMRI), we want to determine the language that incited these brain responses. Previous research has successfully demonstrated reconstruction of word sequences from these fMRI recordings, but this does not capture the more refined aspects of speech, such as intonation, emotion, and intent. In this project, our aim is to improve the performance of brain decoding by reconstructing audio segments that recover the acoustic properties of perceived speech.

1 Introduction

Previous research in the field of brain decoding has shown that speech can be extracted from intracranial brain recordings [1], which has enabled people who have lost the ability to speak to communicate effectively again [16, 10]. However, these invasive approaches have critical limitations; they require surgical implantation of electrodes in the brain, which carries significant medical risks, high costs, and ethical concerns, making them impractical for widespread use among individuals with speech impairments. Non-invasive recordings have been proven to capture linguistic information effectively [8, 4, 2, 3], offering a feasible approach to decoding brain activity in a more practical way. If we can show that decoding using non-invasive techniques is possible, this would provide a more accessible approach to restoring communication abilities.

In this work, we propose a non-invasive decoding approach based on functional magnetic resonance imaging (fMRI) data. fMRI data is gathered by measuring blood-oxygen-level-dependent (BOLD) signals in the brain as an individual listens to audio; this serves as an indicator of a subject's neural activity in response to auditory stimuli. Using fMRI data for decoding presents a key challenge. While BOLD signals provide an effective indirect measurement of how different parts of the brain respond to stimuli, they have a low temporal resolution - the signals themselves rise and fall over approximately 10 seconds [9], a much longer period of time compared to the time in which a stimulus might cause neural activity. This means that even a single brain image can contain information about several different time points in a stimulus. Recent studies have been able to handle the low temporal resolution issue of fMRI data effectively; with the emergence of new encoding models that are able

to accurately predict BOLD signals in the brain in response to stimuli, it has become possible to map how semantic information is represented over time [8].

Brain decoding has been shown to work well when reconstructing text from fMRI-recorded brain responses [13]. However, decoding spoken audio presents a different challenge: unlike text, which is made up of discrete words, speech is a *continuous* signal that carries both meaning and sound over time. Because of this fundamental difference, decoding speech from brain activity requires a different strategy than decoding discrete word sequences. Another paper has shown the feasibility of brain decoding when reconstructing movie clips from fMRI data [12]. Like speech, visual stimuli are continuous signals - therefore, a similar approach to vision decoding is also applicable to speech decoding. The decoding algorithm presented in this paper mimics that of the algorithm presented for the purposes of visual scene decoding.

In this paper, we introduce a Bayesian decoder that produces a reconstruction of natural speech from fMRI-recorded brain responses. We define reconstruction as the production of the speech stimulus with the highest posterior probability of inciting the recorded brain response [11]. The purpose of the speech decoder is to capture both words *and* paralinguistics, such as emotion, intonation, and intent of speech, to add precision to decoded communication.

1.1 Bayesian Decoding

A Bayesian decoder is based on Bayes theorem, which is defined as:

$$p(s \mid r) \propto p(s)\, p(r \mid s)$$

Here, $p(s \mid r)$ is the posterior distribution, or the probability that the stimulus s occurs given the response r. $p(s)$ is the distribution of the sampled prior, which comes from a set of stimuli. $p(r \mid s)$ is the likelihood of observing response r given stimulus s. Given an input response r, decoding aims to find the stimulus s that maximizes the posterior probability.

The key to Bayesian decoding lies in the likelihood $p(r \mid s)$. In practice, this likelihood is provided by an encoding model, which predicts how stimuli map onto brain responses. An encoding model is mathematically defined as a conditional probability that a recorded response r was caused by a stimulus s [11]. By reversing this relationship, Bayesian decoding leverages the encoding model to evaluate which stimulus is most consistent with an observed response. Importantly, the Bayesian decoding framework is generalizable - it can be applied using any encoding model that defines a mapping from a stimulus to brain responses.

Previous work has used self-supervised deep neural networks to create an encoding model that predicts how the brain will respond to natural speech [14]. These speech encoding models enable us to perform Bayesian decoding on reconstructed natural speech, which has not been done previously.

2 Encoding

As mentioned above, we want to adapt the encoding model for decoding purposes. Previous encoding models have typically relied on contextual information from surrounding time points to improve predictive performance. However, this reliance on context across fMRI time point (TR) boundaries is not suitable for decoding, since

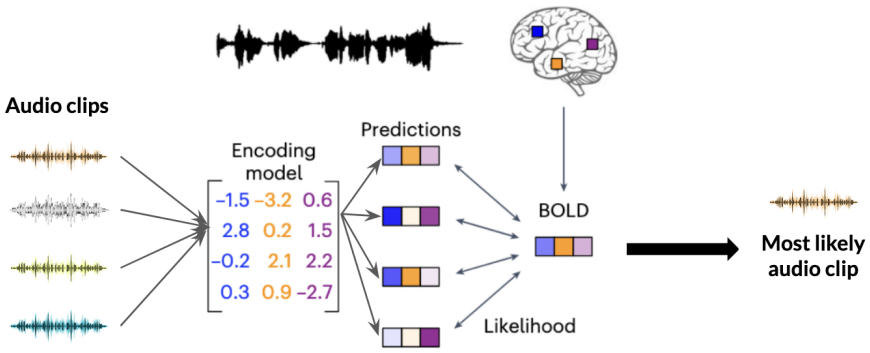


Figure 1: Example flow of decoding using a speech encoding model. Audio clips from a database are passed through an encoding model to predict corresponding brain responses (represented by activations of different colored voxels). These predicted responses are then compared to an observed brain response (fMRI-recorded BOLD signals). The audio clip whose predicted brain response most closely matches the observed response is identified as the most likely stimulus.

decoding requires reconstructing each segment independently. To address this, we train a new encoding model that predicts brain responses on a TR-by-TR basis, using only the features from a single 2-second audio segment aligned with each TR. This new encoding model is better suited for decoding, as it forces the encoding model to learn how to represent a single segment of audio without information. The model was trained on 106 stories retrieved from *The Moth Radio Hour*, which consist of naturalistic stimuli that are highly representative of commonly heard speech [14].

To ensure that removing the contextual element of the speech encoding model does not impact the accuracy of its predictions, we evaluate the performance of the no-context model. In the context of a speech encoding model, performance is defined as the correlation across time between the model's prediction and the recorded brain response. We then compared the performance of the no-context model to a speech encoding model that uses 2 seconds of context and compared their performances. In

Figure 2, we see that the performance of the model without context is very comparable to the performance of the model that uses context. The performance was measured on 3 test stories, all of which were narrative stories also taken from *The Moth Radio Hour*.

An important note when comparing performance between different encoding models is that not all subdivisions of the fMRI data, known as voxels, are important to consider. Because we are trying to solve an auditory task, it follows that we should primarily consider the performance of the voxels involved in auditory processing. This includes the auditory cortex, Broca's area, and the superior ventral premotor speech area (sPMv) regions of the brain [4, 7].

Because BOLD responses are delayed from the onset of a stimulus, our encoding model predicts the brain response at each time point using stimulus features from several time points in the past. Most encoding models utilize information from multiple delays, or previous time points, to provide more accurate predictions of voxel responses [12, 8]. However, this makes decoding more complex – a response that is generated from multiple time delays makes it difficult to isolate a specific stimulus that evoked a response. To simplify the decoding process, our model uses a single time delay of 4 seconds; that is, each brain response prediction is based on information from the stimulus that occurred 4 seconds prior. To measure the efficacy of this model, we compared its performance relative to an encoding model trained to use all 4 previously mentioned delays in a stimulus.

Based on the results in Figure 3, we see that using the second delay (4 seconds in the past of the stimulus) results in not only the best encoding model performance in the regions of interest, but also a very comparable performance to the encoding

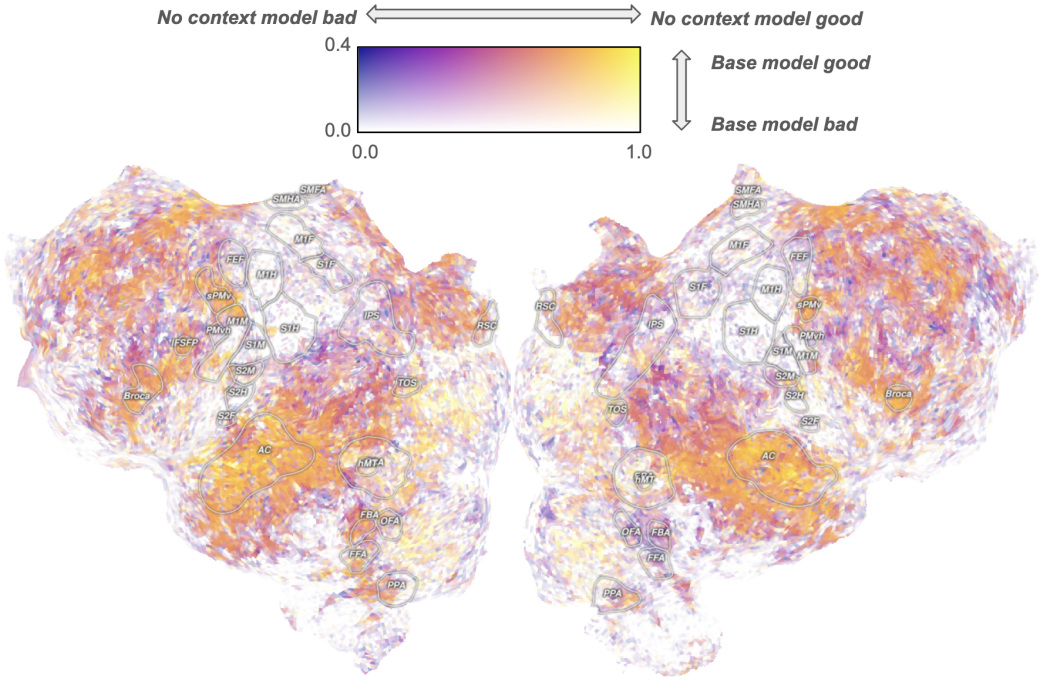


Figure 2: 2-D flatmap comparing performance of an encoding model trained using context and the encoding model not trained on context. The vertical colorbar shows performance of the encoding model trained on context, and the horizontal colorbar shows correlation between the two models. A higher correlation for a specific voxel implies that the two models perform more similarly in that brain region. The results shown are for a single test story, "*Where There's Smoke*".

model that is trained to use four time delays (2s, 4s, 6s, 8s).

2.1 Identification

To estimate the feasibility of the decoding approach, we constructed a test to measure identification accuracy, which mimics the decoding process and measures how accurately the decoding model can associate a BOLD response with the specific speech stimulus that triggered it. For the identification test, we used audio clips from a test

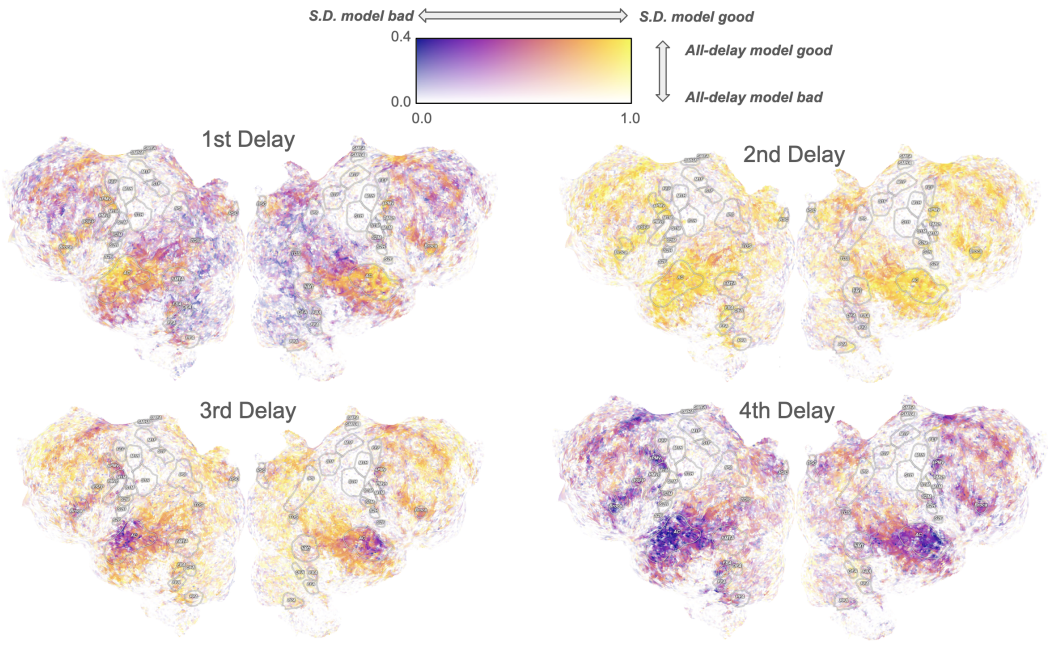
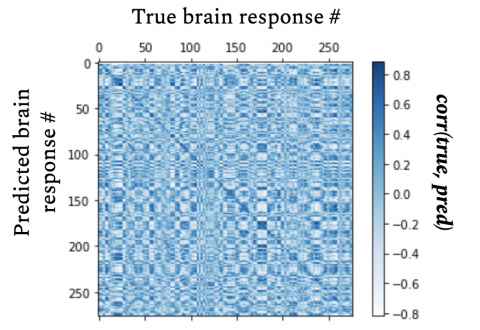


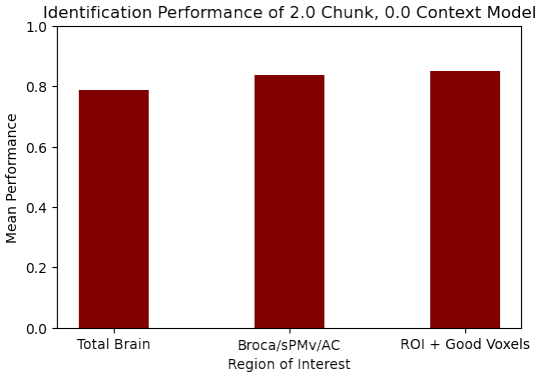
Figure 3: Flatmaps comparing performance of single delay encoding models (each using a different time delay) compared to the performance of all-delay encoding model. The alpha channel shows performance of the model that uses all-delays, and the hue shows performance of the single-delay models relative to the all-delay model.

story for which fMRI-recorded brain responses are known. Identification is considered successful if the brain response recorded at time t matches the predicted brain response at time t.

In Figure 4, we see that we achieve the highest identification performance when only evaluating performance of the voxels at the intersection of the previously mentioned regions of interest (auditory cortex, Broca's area, sPMv) and what we refer to as the "good voxels", which are the 10,000 voxels with the highest encoding performance. This subset of voxels is what we will use to evaluate the effectiveness of decoding. These results were based on fMRI recordings taken while a subject listened to the "*Where There's Smoke*".



(a) Identification accuracy on the masked regions of interest. The test data in our experiment consisted of 276 different BOLD signals evoked by the test story and predicted responses. The blueness of the point in the ith row and jth column represents the correlation between predicted response i and true response j, which were generated and recorded in sequential order (that is, predicted response k should most closely match true response k).



(b) Identification performance. In this experiment, performance is defined as the position of predicted response k to true response k when ranking all predicted responses by correlation to true response k. If predicted response k ranks the highest, it would receive a score of 1.0; if it ranks the lowest, it would receive a score of 0.0. These scores are then averaged over all 276 time points of the original response to compute the overall identification performance.

Figure 4: Identification results.

We also evaluated how the identification accuracy scaled with respect to the size of the set of speech clips. To do this, we expanded the set of speech stimuli considered in the identification test to include additional stimuli from the speech clip database. These clips were 2 second segments taken from the stories used to train the speech encoding model.

The expanded identification results help us evaluate how reconstruction will perform more generally. As shown in Figure 5, identification accuracy naturally declines as the number of distractor clips increases, but it remains well above chance even with many distractors (here, the chance of the correct clip appearing in the top-10

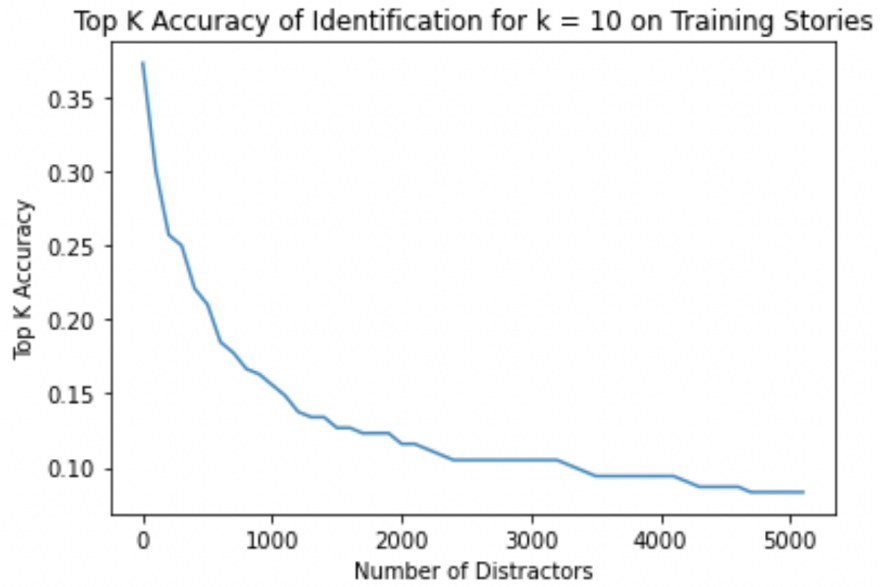


Figure 5: Expanded identification accuracy. The graph shows how top-10 accuracy changes when performing identification as the database of speech segments (number of "distractor clips") increases in size.

ranked clips is $\frac{10}{w+k}$, where k is the number of distractor clips, and w is the number of recorded responses in the test set, which is 276 in our case). This trend demonstrates that our decoding pipeline can generalize beyond a limited test set and gives us confidence that our reconstruction method—relying on similarity between predicted and actual brain responses—can effectively recover the speech content a subject is hearing.

3 Decoding and Reconstruction

Now that we have finalized an encoding model to use in our decoding process, we can discuss the decoding algorithm, which obtains the speech that elicited an input

fMRI-recorded brain response. Using the Bayesian decoding framework (Section 1.1) and the speech encoding model, we can define decoding to be the reconstruction of a stimulus whose predicted brain response most closely correlates to the input brain recording. This allows us to model the decoding algorithm as a search over a database of stimuli based on their predicted brain responses generated by the speech encoding model.

3.1 Setup

First, we must create a database of speech segments. Given that the encoding model is trained to predict brain responses to 2-second long speech segments, the database is comprised of 2-second clip segments from the 106 *The Moth Radio Hour* stories used to train the encoding model. Next, each of these 2-second clips is passed through the no-context encoding model; this results in a new database of predicted responses, where each speech clip has a corresponding predicted brain response.

3.2 Reconstruction

Now that we have constructed a database of speech clips and their associated predicted brain responses, we have the information necessary to perform Bayesian decoding. To start, we input an fMRI-recorded brain response taken from a subject who is listening to a story. These brain responses are taken from a held out story that the encoding model was not trained on (in this case, we used *Where There's Smoke* again). This input recorded response is then compared to the database of predicted brain responses that was previously generated for the various sound clips.

By computing the correlation between the recorded brain response and each predicted response, we can rank the corresponding natural speech clips based on how well their predicted responses match the actual fMRI data. The top ranked clips are then selected and their corresponding stimuli are combined to produce a final "decoded" sound that approximates what the subject originally heard.

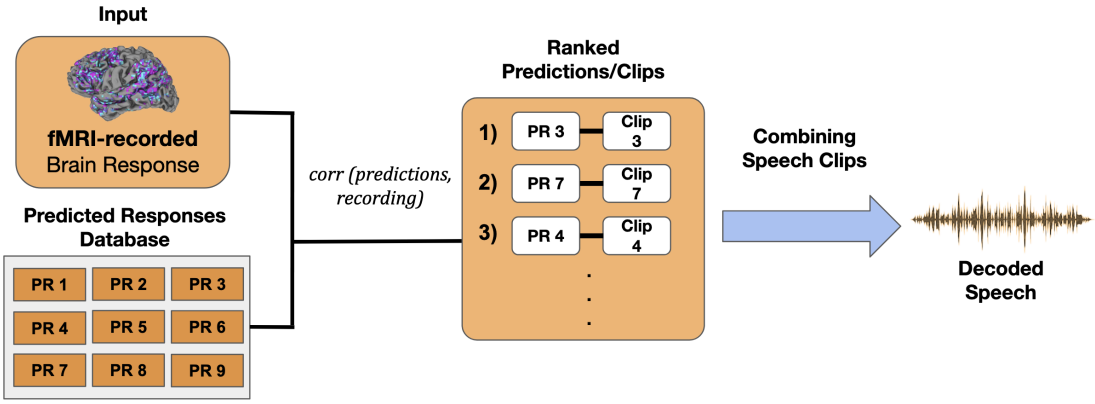


Figure 6: Reconstruction pipeline based on Bayesian decoding. The predicted responses in the database are generated by the speech encoding model, each of which also has a corresponding speech clip used by the encoding model to create the predictions. The predictions are compared to the input brain response based on correlation across voxels.

3.3 Evaluation Metrics

In order to measure the effectiveness and precision of reconstruction, it is vital to have an evaluation metric that represents a way to compare the decoded audio and the natural speech. For audio comparison, there are a number of ways to do this. In this section, we will evaluate several different metrics and their effectiveness.

The simplest way to compare the similarity of two speech segments is by scoring the similarity between their respective waveforms, which capture amplitude changes

over time. However, this method is highly sensitive to differences in pitch, speaking rate, and background noise, making it unreliable for capturing perceptual or semantic similarities (which are important when considering speech similarity).

There are three other metrics that we studied closer to help us reduce this sensitivity, each of which captures a different aspect of reconstructed speech quality. The first is **latent space similarity**, which measures how similar two audio segments are in a compressed latent space. This can help mask fine-grained details and introduce a more generic representation of a speech segment. The second is **power-weighted coherence**, a metric that quantifies the similarity of two signals across the most powerful frequency bands. This helps in making the similarity metric invariant to noise and slight shifts. The third is **spectrogram similarity**, which compares the time-frequency patterns of two audio segments. Because this metric is sensitive to both the timing and spectral content of speech, it is useful for detecting whether key acoustic features have been preserved. Using multiple metrics allows for a more comprehensive evaluation of decoding quality, as it helps measure the performance of decoding based on different acoustic properties.

3.4 P-Testing

The evaluation metrics discussed above have differing scales and variances, making them not very interpretable as standalone values. As such, determining the accuracy of a reconstructed clip based on a metric is only useful if we know how the reconstructed clip scores relative to a null distribution. This motivates the design of a p-test that will compute whether the observed similarity score between the reconstructed clip and the original stimulus is significantly greater than what would be

expected by chance.

To construct this p-test, we generate a null distribution of similarity scores by comparing the reconstructed clip to a large set of clips. To mimic the process of reconstruction, the clips in the null distribution are generated in a similar fashion to the reconstructed clip; this process is different based on how the decoded clip was reconstructed (we will cover a few different approaches in the following sections). The empirical p-value is then computed as the proportion of scores in the null distribution that are greater than or equal to the observed score for the reconstructed clip.

3.5 MAP and AHP Reconstruction

Using the sampled prior of speech clips and their respective predicted brain responses, we are able to assign a score to each speech clip in the prior. This score is based on the correlation between the predicted brain response and the input fMRI-recorded response. The clip with the highest score represents the maximum a posteriori (or MAP) reconstruction for the given BOLD signal.

A clear setback of MAP reconstruction is that the performance of decoding is severely limited by the speech clips in the prior. Because of this, the current decoding approach does not have the ability to generate new audio outside of the prior. To improve on this approach, we can construct a combined or averaged version of the speech clips from the prior with the highest score. This new approach, called averaged high posterior (or AHP), offers a way to construct new speech clips that might be more similar to the stimulus that incited the input fMRI response. The following sections discuss different approaches to AHP reconstruction that we tested.

3.6 Using Codecs in Reconstruction

The first (and simplest) approach we tried was averaging the waveforms of the top ranked clips. However, as mentioned earlier, this often led to poor results due to the fine-grained nature of waveform representations of audio.

One solution to this is to use a richer representation of the speech clips. We utilize audio codecs to do this, which are neural network models that compress audio signals into compact latent representations and reconstruct them with high fidelity. Audio codecs are typically implemented as autoencoders, with an encoder that converts the raw audio into a latent representation and a decoder that converts the latent form back into the original audio. The latent representations that lie in this architecture capture key acoustic and semantic features, which makes them a well-suited medium for combining speech segments during reconstruction. For each of the top ranked clips considered for the final decoded audio clip, we extract their latent representations from the codec by passing them through only the encoding side. These latent representations are then combined using a weighted average based on the correlations between the predicted response for the audio clip and the input fMRI-recorded brain response. Finally, the combined latent representation is passed through the decoder stage of the codec, giving us our final audio output.

We considered two different codecs to use for combining the top ranked audio clips. The first was Encodec, a state-of-the-art, high-fidelity audio codec developed by Facebook. It consists of an encoder-decoder architecture and a quantized latent space which serves as a meaningful way to combine sound segments [5]. The second was Mimi, a powerful codec that was trained to perform well specifically on speech due to its ability to capture acoustic *and* semantic information [6]. Both of these

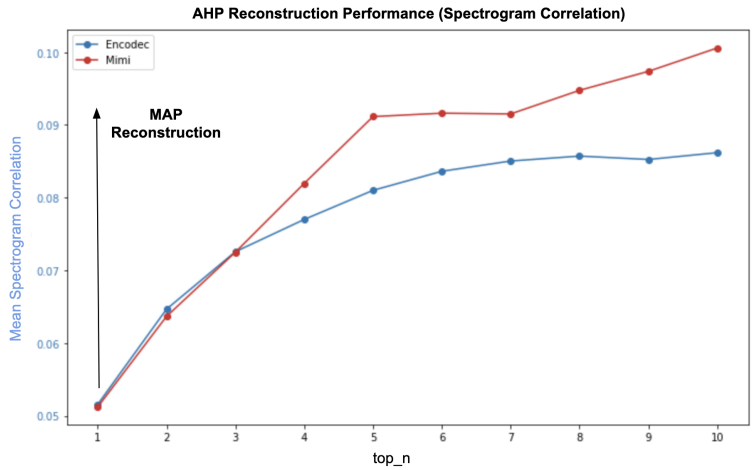
models could prove to be useful for speech decoding since we want the output of the decoding algorithm to sound as close to natural speech as possible.

From Figure 7, we see that the average p-values are approximately similar across different values of *top_n*, where *top_n* is the number of top-ranked clips selected based on their based on their posterior probabilities under the Bayesian decoding model. This shows that reconstruction using latent representations from both models results in similar performance in relation to a randomly sampled distribution of clips. However, the mean spectrogram correlation is, on average, higher for Mimi. This motivates us to use Mimi over Encodec for this approach and others when reconstructing a speech clip.

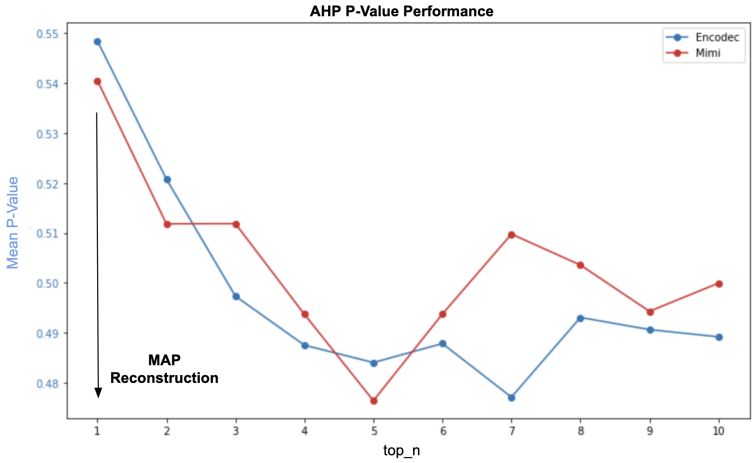
3.7 Reconstruction with a Neural Network

One setback of the codec latent space averaging approach is that it assumes the reconstructed clip can be represented as a linear combination of the top-ranked speech clips, which limits the ability of the decoder to capture more complex relationships between acoustic features and output more accurate reconstruction. In contrast, a neural network that is trained to learn the relevant audio features across different speech clips offers a more non-linear approach to combining the clips that have the highest likelihood of relating to the stimulus. This motivates the design and implementation of a neural network that can learn to combine audio clips effectively rather than relying on a weighted averaging.

The neural network we implemented uses a Transformer encoder [15]. Regarding parameters of the model, the latent dimensions were extracted from the latent representations of the 2-second speech clips were in the shape of (512, 17), where the



(a) Comparison of AHP reconstruction performance when using Mimi vs using Encodec as the codec in which latent spaces are combined. The graph shows mean spectrogram correlation as *top_n*, or the top n ranked clips that are used in the reconstructed clip, changes.



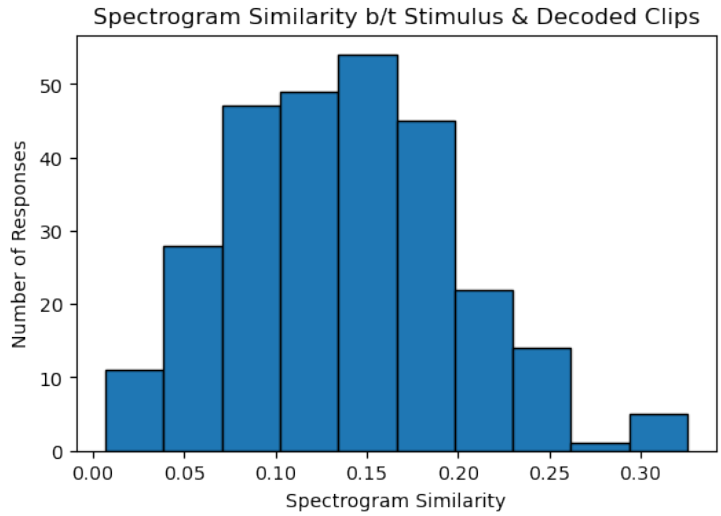
(b) Comparison of p-values in AHP reconstruction when using Mimi vs using Encodec. The graph shows mean p-value as *top_n* changes.

Figure 7: AHP reconstruction results for Mimi and Encodec.

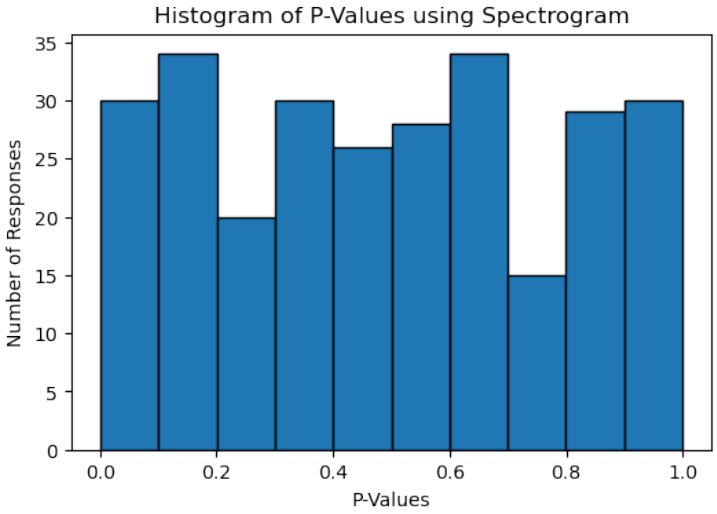
first dimension represents the dimension of the Mimi latent vectors, and the second dimension represents the number of time steps. We also used a hidden state dimension of 768 for both the transformer layer and to project the model inputs down to a lower dimension that the transformer could handle. Finally, we gave the model 5 input clips, which was based on the fact that AHP reconstruction when using Mimi as the codec yielded the best results when we used the 5 top-ranked clips (see Figure 7).

In our decoding setup, the top-ranked clips identified during reconstruction are not exact matches to the original stimulus, but rather partial matches that capture different acoustic facets, such as pitch, tempo, or spectral content. To simulate this, during the training stage of the network, the input speech clips are actually perturbations to original stimuli, and the neural network learns to recover the original audio segment from the multiple perturbed inputs (this is the output of the network). The model was trained in this way based on what we hope to extract from reconstruction, which is that the top-ranked clips should be similar to the initial stimulus based on some set of acoustic features. Note that both the inputs and output were in the form of Mimi-based latent representations, and this was how we computed the loss in the network.

Figure 8 shows reconstruction results when combining the top-ranked clips using the neural network described above. The average spectrogram similarity is 0.138, which is approximately a 53% improvement on reconstruction using averaged codec latent representations. This suggests that the neural network can more effectively learn combinations of features that relate to the stimulus speech audio from the input clips. The average p-value is 0.49, which is comparable to that of codec-based averaging, indicating that although the reconstructions result in higher spectrogram



(a) AHP reconstruction performance when using a neural network trained to combine the top-ranked audio clips.



(b) Histogram of p-values in AHP reconstruction when using the neural network in reconstruction.

Figure 8: AHP reconstruction results for the transformer-based neural network. Note that there is a score and p-value for each of the 276 TRs, as we reconstruct a speech segment for the recording at each TR.

similarity, they are still not entirely distinguishable from selecting a clip at random. Future improvements to the neural network can be made by making the perturbations more diverse and more significant, forcing the model to learn more abstracted variations of the original stimulus. We may also want to try using Encodec latent representations rather than Mimi to see if any improvements to statistical significance can be made.

4 Conclusion

In this project, we introduced a Bayesian decoding framework capable of reconstructing speech from non-invasive fMRI-recorded brain responses. By adapting an encoding model trained to predict how the brain will respond to natural speech stimuli, we were able to construct a prior of natural speech clips and perform decoding by correlating predicted brain responses to actual fMRI signals.

We also implemented Averaged High Posterior (AHP) reconstruction using latent representations from audio codecs to increase the range of decoding. This method allowed us to synthesize new audio outputs by combining the top-ranked speech segments in the latent space. These results indicate the potential of codec-guided decoding as a promising direction for improving the naturalness and accuracy of reconstructed speech.

While our current approach is constrained by the expressiveness of the prior database and the resolution of fMRI data, this work lays a foundation for future methods that combine neural decoding with generative models. These advances could ultimately allow us to recover detailed acoustic features, including intonation and

emotion, from non-invasively gathered brain activity, offering a powerful new means for restoring communication in individuals who have lost the ability to communicate.

Future works may include expanding the diversity and size of the speech clip database to improve reconstruction coverage. Additionally, for the most part, this paper focused on using spectrograms as the evaluation metric. In the future, we would like to do thorough reconstruction analyses on each of the metrics to see if certain metrics are more sensitive to specific aspects of reconstruction quality. Finally, we would like to refine the neural network and experiment with different architectures to see if any of them yield higher quality reconstructed outputs.

References

[1] Gopala K Anumanchipalli, Josh Chartier, and Edward F Chang. "Speech synthesis from neural decoding of spoken sentences". In: *Nature* 568.7753 (2019), pp. 493–498.

[2] Michael P Broderick et al. "Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech". In: *Current Biology* 28.5 (2018), pp. 803–809.

[3] Charlotte Caucheteux and Jean-Rémi King. "Brains and algorithms partially converge in natural language processing". In: *Communications biology* 5.1 (2022), p. 134.

[4] Wendy A De Heer et al. "The hierarchical cortical organization of human speech processing". In: *Journal of Neuroscience* 37.27 (2017), pp. 6539–6557.

[5] Alexandre Défossez et al. “High fidelity neural audio compression”. In: *arXiv preprint arXiv:2210.13438* (2022).

[6] Alexandre Défossez et al. “Moshi: a speech-text foundation model for real-time dialogue”. In: *arXiv preprint arXiv:2410.00037* (2024).

[7] Gregory Hickok and David Poeppel. “The cortical organization of speech processing”. In: *Nature reviews neuroscience* 8.5 (2007), pp. 393–402.

[8] Alexander G Huth et al. “Natural speech reveals the semantic maps that tile human cerebral cortex”. In: *Nature* 532.7600 (2016), pp. 453–458.

[9] Nikos K Logothetis. “The underpinnings of the BOLD functional magnetic resonance imaging signal”. In: *Journal of Neuroscience* 23.10 (2003), pp. 3963–3971.

[10] David A Moses et al. “Neuroprosthesis for decoding speech in a paralyzed person with anarthria”. In: *New England Journal of Medicine* 385.3 (2021), pp. 217–227.

[11] Thomas Naselaris et al. “Bayesian reconstruction of natural images from human brain activity”. In: *Neuron* 63.6 (2009), pp. 902–915.

[12] Shinji Nishimoto et al. “Reconstructing visual experiences from brain activity evoked by natural movies”. In: *Current biology* 21.19 (2011), pp. 1641–1646.

[13] Jerry Tang et al. “Semantic reconstruction of continuous language from non-invasive brain recordings”. In: *Nature Neuroscience* 26.5 (2023), pp. 858–866.

[14] Aditya R Vaidya, Shailee Jain, and Alexander G Huth. “Self-supervised models of audio effectively explain human cortical responses to speech”. In: *arXiv preprint arXiv:2205.14252* (2022).

[15] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).

[16] Francis R Willett et al. “High-performance brain-to-text communication via handwriting”. In: *Nature* 593.7858 (2021), pp. 249–254.