Approximation fMRI data from the audio time series

Anastasia German
MIPT
Moscow, Russia
german.aia@phystech.edu

Daniil Dorin
MIPT
Moscow, Russia
dorin.dd.contact@gmail.com

Andrey Grabovoy
MIPT
Moscow, Russia
grabovoy.av@phystech.edu

Abstract

Nowadays understanding how the brain perceives and processes external stimuli is essential for advancing neuroscience and improving diagnostic tools. In this article, we explore the relationship between human perception of the outside world and fMRI scanner readings. The analysis focuses on the correlation between a sequence of fMRI images and an auditory signal. A method for predicting fMRI readings based on the auditory sequence is proposed. The task of predicting complex, non-linear time series, influenced by exogenous factors and exhibiting multiple periodicities, is approached through the application of the Granger causality test. TODO

Key words: fMRI, Linear model, Audio time series, Correlation analysis

1 Introduction

The human brain is perhaps the most complex biological system, composed of over 100 billion neurons and an even greater number of glial cells organized into the cerebrum, cerebellum, and brainstem, whose intricate networks underlie perception, cognition, and motor control . Dysfunctions in these networks manifest as a wide spectrum of neurological and psychiatric disorders. Functional magnetic resonance imaging (FMRI)[1] has become the leading non-invasive tool for studying brain function by measuring hemodynamic reactions. It uses the blood oxygen level-dependent (BOLD) signal[2], which reflects variations in oxygenated blood levels, serves as a key indicator of neural activity. fMRI has been widely applied in neuroscience research and medicine, including the study of brain function in conditions such as autism and Alzheimer's disease[3], as well as in predicting and potentially treating disorders like traumatic brain injuries.

Every second our brain process different type of information, including audio. When sound enters the ear, it causes tiny hair cells in the inner ear to vibrate[4]. These cells convert the mechanical movement into electrical signals, which travel along the auditory nerve to the brainstem—the first point where the brain begins to process sound. From there, the signals pass through several key brain regions that help identify aspects like the direction, timing, and type of sound. Eventually, the information reaches the thalamus, a central relay station that sends it to the auditory cortex, located in a part of the brain called Heschl's gyrus. This area is arranged by pitch, meaning different regions respond to different sound frequencies. Nearby regions further analyze complex features like speech patterns, musical tones, and rhythm. With the help of fMRI, scientists can observe how these brain areas respond to different types of sound in real time[5][6], providing valuable insight into how we perceive and make sense of the acoustic world around us.

Preprint. Under review.

Accurately processing auditory signals presents a significant challenge. In this study, we address it by utilizing Mel-Frequency Cepstral Coefficients (MFCCs)[7][8].This method takes short parts of an audio signal and turn them into values that represent how humans naturally hear pitch and tone. These features have been effectively used to predict patterns in brain scans (like fMRI) when people hear speech or music[9]. More advanced techniques, such as spectro-temporal modulation, help improve how well these sounds are matched to brain responses.

This study aims to understand how changes in sound over time relate to brain activity measured by fMRI. We do this by comparing sound features (MFCCs) with signals from specific brain areas, while also taking into account the delay between brain activity and the blood flow response. We use an open multimodal dataset[10] comprising fMRI recordings from 30 participants (ages 7–47) who viewed a short audiovisual film. The film stimulus is richly annotated with detailed speech and video event markers, facilitating precise alignment of MFCC features with the preprocessed fMRI time series. Auditory features will be combined with a standard model of the brain's blood flow response[11] (the hemodynamic response function) and then compared with fMRI signals. This will help us find which areas respond more to audio. By analyzing both individual brain voxels and larger regions, we aim to understand how sounds influence brain activity, and whether it is possible to predict brain responses based on changes in the sound features.

2 Related Work

When working with fMRI data, researchers may encounter various methodological and technical challenges.One of them in auditory fMRI modeling is the brain's intrinsic "dark energy"[12] — the default mode network's continuous activity that consumes the majority of neural energy and overlaps with stimulus-driven signals.Techniques used to reduce noise in fMRI data —like removing signals related to head movement or breathing —can sometimes also remove meaningful brain activity patterns. This highlights the importance of using careful and well-balanced denoising methods[13].Spontaneous changes in fMRI signals (BOLD fluctuations) vary across different brain regions and are linked to how those regions are physically connected[14]. This means that general noise models may not accurately reflect local differences in brain activity over time. New methods, such as topological analysis using cubical persistence, help identify stable features in the data that can group similar brain activity patterns, even during complex, real-world tasks[15].

The integration of functional magnetic resonance imaging (fMRI) with advanced machine learning techniques has significantly advanced the decoding of auditory stimuli and the reconstruction of neural signals. Recent studies have explored the potential of deep learning models to predict semantic information from brain activity[16][17]. Using Deep neural networks (DNNs) optimized for auditory tasks have demonstrated remarkable success in replicating human auditory behavior and predicting cortical activation patterns. These task-optimized DNNs reveal a hierarchical processing structure in the auditory cortex, mirroring the brain's organization for processing speech and music[18]. But not only complex neural networks are used in this field. Due to the requirements of careful handling of inherent variability and noise in working with fMRI data, researchers have developed Bayesian general linear models. They allows for variability in both the noise level (heteroscedasticity) and the temporal dependencies (autoregressive structures) across different brain areas. These models adapt to the unique characteristics of each voxel, improving the detection of genuine brain activity signals[19]. In our work we aim to use the simpliest linear model, which does not require large computing power and is well interpreted.

3 Problem statement

It is required to propose a method for predicting FMRI readings based on the auditable sound series. We will denote the frequency of occurrence of FMRI images $\mu \in \mathbb{R}$. The sequence of images is set

$$\mathbf{S} = [\mathbf{s}_1, \dots, \mathbf{s}_{\mu t}], \quad \mathbf{s}_\ell \in \mathbb{R}^{X \times Y \times Z}, \tag{1}$$

where X, Y and Z — dimensions of voxel image.

The sampling frequency $\nu \in \mathbb{R}$ and the duration $t \in \mathbb{R}$ of the audio sequence are set. A a time-discrete signal is defined

$$\mathbf{P} = [p_1, \dots, p_{\nu t}], \quad p_\ell \in \mathbb{R}, \tag{2}$$

The task is to construct a mapping that accounts for the delay Δt between the fMRI scan and the audio stream, as well as previous tomographic readings. Formally, we need to find such a mapping $\mathbf{f}$ that

$$\mathbf{f}(p_1, \dots, p_{k_\ell - \nu\Delta t}; \mathbf{s}_1, \dots, \mathbf{s}_{\ell-1}) = \mathbf{s}_\ell, \ \ell = 1, \dots, \mu t, \tag{3}$$

For the ℓ-th fMRI scan, the index k_ℓ of the corresponding signal is set by the formula

$$k_\ell = t\nu = \frac{\ell}{\mu}\nu. \tag{4}$$

4 Description of method

The audio stream embeddings will be represented by Mel-frequency cepstral coefficients (MFCCs) [8]. To compute MFCCs, the continuous audio signal is first divided into short, overlapping segments called frames. Each frame is then transformed using the short-time Fourier transform (STFT), which converts the signal from the time domain to the frequency domain. The resulting frequency spectrum is passed through a series of triangular filters spaced along the mel scale, which models how humans perceive pitch. These filters produce a set of energy values for different frequency bands. These energies are then compressed using a logarithmic function to simulate how humans perceive loudness. Finally, a discrete cosine transform (DCT) is applied to these log-energy values to reduce redundancy and produce the final MFCC features. So, for each signal instance, we have a $\mathbf{d}$ - dimensional vector:

$$\mathbf{x}_\ell = [x_1^\ell, \dots, x_d^\ell]^\mathsf{T} \in \mathbb{R}^d, \ \ell = 1, \dots, \frac{\nu t}{h}. \tag{5}$$

where $\mathbf{x}_\ell$ - is the vector related to l-frame.

We aim to reconstruct the function $\mathbf{f}$ under the Markov property assumption:

$$\mathbf{f}(\mathbf{x}_{k_\ell - \nu\Delta t - g}, \dots, \mathbf{x}_{k_\ell - \nu\Delta t}) = \mathbf{s}_\ell - \mathbf{s}_{\ell-1} = \boldsymbol{\delta}_\ell, \quad \ell = 2, \dots, \mu t, \tag{6}$$

where $\boldsymbol{\delta_\ell} = [s_{ijk}^\ell - s_{ijk}^{\ell-1}] = [\delta_{ijk}^\ell] \in \mathbb{R}^{X \times Y \times Z}$ — difference between two consecutive scans.

Taking (4) into account, the total number of (signal, scan) pairs is $N = \mu(t - \Delta t)$. Thus, for each voxel, we have the following dataset:

$$\mathfrak{D}_{ijk} = \{(\mathbf{x}_\ell, \delta_{ijk}^\ell) \mid \ell = 2, \dots, N\}.$$

This reduces to a standard regression problem:

$$y_{ijk} : \mathbb{R}^d \to \mathbb{R}. \tag{7}$$

Treating each voxel independently, let $\mathbf{Y}_{ijk} \in \mathbb{R}^N$ represent the voxel time series and $\mathbf{X} \in \mathbb{R}^{N \times d}$ the feature matrix. The assumed relationship is:

$$\mathbf{Y}_{ijk} = \mathbf{X}\theta + \varepsilon, \tag{8}$$

where $\theta \in \mathbb{R}^d$ - are model coefficients, and $\varepsilon \sim N(0, \Sigma)$ - represents noise.

The goal is to find parameters $\widehat{\theta}$ that maximize the likelihood function given hyperparameters Δt and d, where d – is the MFCC dimension:

$$L_X(\theta) = \prod_{v=1}^{N} p_\theta(Y_{ijk}^v) \longrightarrow max_\theta \tag{9}$$

5 Experiments

5.1 Experimental Setup

We conducted a comprehensive empirical study to validate theoretical predictions . The primary objective of this experiment is to quantify the correlation between neural activity, captured via fMRI and acoustic features, represented as time series while optimizing the hyperparameter Δt for individualized models. A linear regression framework was employed to establish this relationship.

Audio features were extracted using Mel-Frequency Cepstral Coefficients (MFCCs) with a fixed sampling frequency of 44.1 kHz. The MFCC feature dimension was standardized to $d = 15$

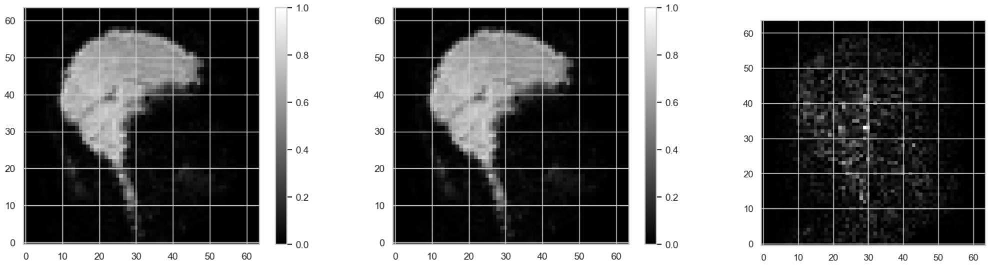
The study used a neuroimaging dataset including MRI scans of 30 participants aged 7 to 47 years. The mean age of participants is 22 years. People were exposed to identical audiovisual stimuli that contained rich auditory information: both dialogues and music were present in them. The detailed description of the dataset is given in the table 2

Table 1: Description of dataset

Name	Notation	Value
Length of the movie	t	390 s
Scan frequency of fMRI	μ	1.64 Hz
Audio frequency	ν	44.1 Hz
fMRI dimensions	X, Y, Z	$[0,1]^{40\times64\times64}$
Audio series	X_1, Y_1, Z_1	$[-1,1]^{1\times390\times44100}$

5.2 Results of Experiments

In this experiment data subsets from three representative participants — 7th, 22nd and 31st were selected for detailed analysis.

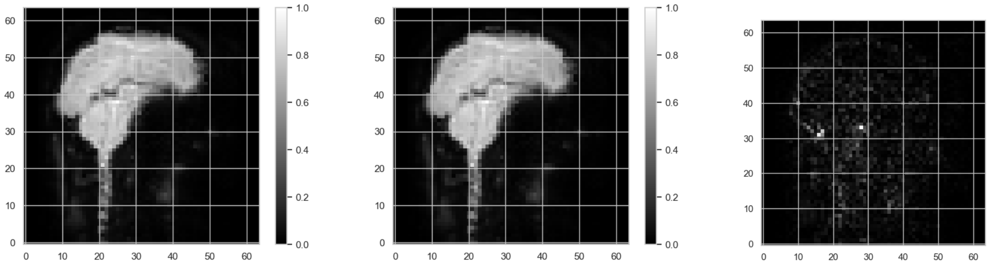


(a) Ground truth

(b) Predicted

(c) Difference

Figure 1: The figure presents slices from the test sample, displaying both the original and reconstructed fMRI images alongside their differences. The pictures were taken by the 7th participant.

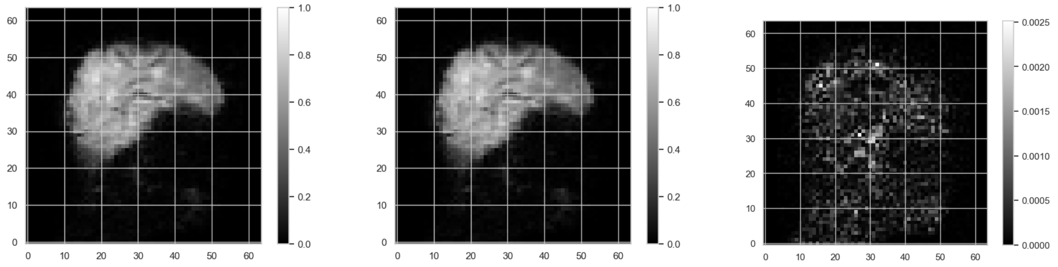


(a) Ground truth

(b) Predicted

(c) Difference

Figure 2: Results of 22nd participant



(a) Ground truth (b) Predicted (c) Difference

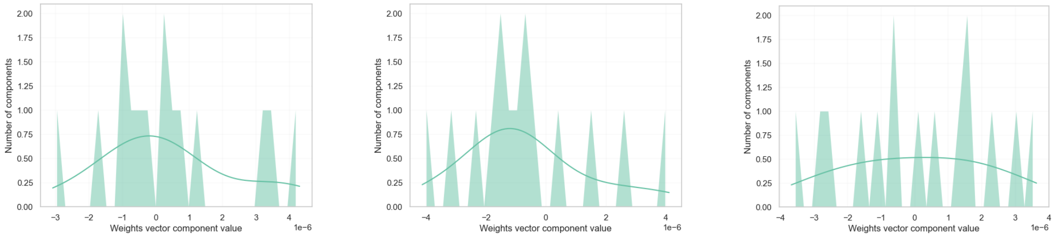
Figure 3: Results of 31st participant

Table 2: The table shows the Mean squared error (MSE) of each participant. Since all voxels have been normalized to a segment$[0, 1]$, the algorithm shows fairly good results.

	7th participant	22nd participant	31st participant
MSE	$6.39 \cdot 10^{-5}$	$7.46 \cdot 10^{-5}$	$9.00 \cdot 10^{-5}$

The voxel values were transformed using the following algorithm. First, the maximum and minimum values of all voxels are calculated. Secondly, the minimum value is subtracted from the value of each voxel, and then the resulting value is divided by the difference between the maximum and minimum value. Thus, the voxel values were normalized to the interval$[0, 1]$. thus the errors reported in 2 demonstrate the algorithm's robust performance.

5.3 Weight analysis



(a) 7th participant (b) 22nd participant (c) 31st participant

Figure 4: Distribution of the components of the weight vector. The model's weight distribution shows no concentration around any single value, reflecting a non-degenerate spread. This implies that the model has been well-trained and captures a broad spectrum of features.

To assess the applicability of our model, we performed an analysis of its weight distribution across multiple participants 4. The experiment was conducted using data from the 7th, 22nd, and 31st participants. The findings indicated that the model's weights are uniformly distributed along the entire axis, with no prominent peak identified. This pattern suggests that the model captures a broad spectrum of features, rather than emphasizing any single feature, thereby reflecting its capacity to generalize across diverse data characteristics.

5.4 Estimation of Δt

During the training the best hyperparameter Δt was analysed. Fistly, all areas of the brain were taken into account in the study, but then it was corrected and shrinked to the area, which is responsible for the perception of audio data. After the correction only one global minimum on each graphic was observed 5 6 7.

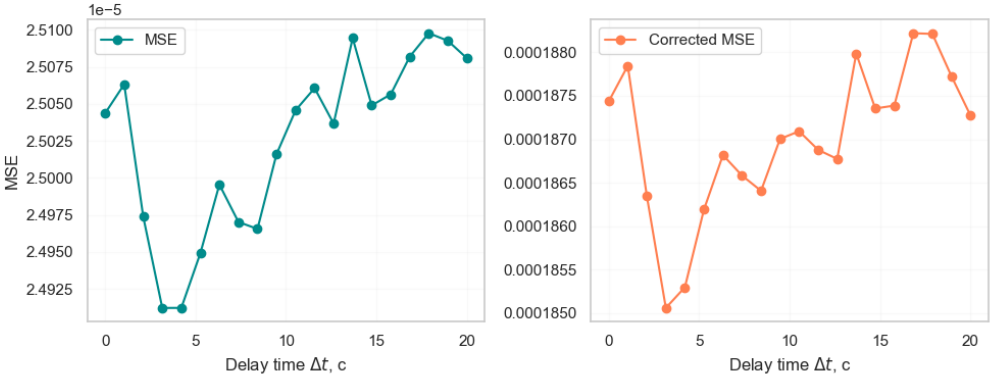


Figure 5: The left graphic illustrates the Mean Squared Error (MSE) as a function of delay time, where 2 global minimum are observed.This absence of a clear minimum suggests that the MSE calculation across the entire fMRI image is influenced by noise, potentially masking the true underlying signal. In order to correct this the MSE was recalculated by focusing on the area, which is responsible for audio perception.The results presented were derived from the data of the 31st participant

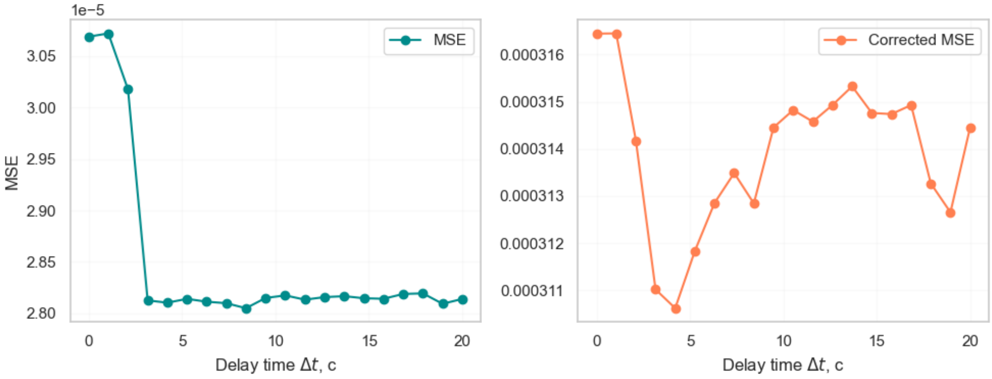


Figure 6: In the analysis of the 22nd participant's data, the uncorrected Mean Squared Error (MSE) curve exhibited multiple local minima, suggesting the presence of noise. After applying a correction only one global minimum was observed.

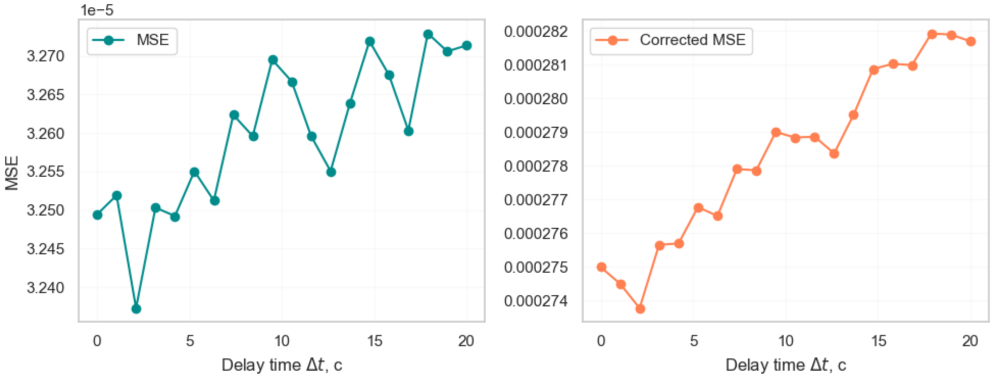


Figure 7: Both images display a single global minimum, but after corrections a noticeable reduction in noise on the graph were shown. The results presented were derived from the data of the 31st participant

6 Discussion

In this study, we used a linear modeling approach to approximate FMRI data based on time series of audio signals to figure out the relationship between auditory stimuli and corresponding neural responses.Despite the good results, there are disadvantages of the work.

Firstly, the sample size comprised only 30 participants, which may limit the generalizability and statistical power of our findings. Small sample sizes in fMRI research have been associated with reduced replicability and increased variability in results, potentially leading to both false positives and negatives.

Secondly, while linear models offer clarity in interpretation, they may not capture the complex, non-linear relationships inherent in neural data. The brain's response to auditory stimuli involves intricate interactions that linear models might oversimplify, potentially overlooking subtle yet significant patterns.

Future research could benefit from incorporating larger, more diverse participant samples to enhance the robustness of findings. Additionally, exploring nonlinear modeling techniques, such as deep learning approaches, may provide a more comprehensive understanding of the neural dynamics associated with auditory processing.

7 Conclusion

This study introduced a linear modeling approach to approximate fMRI time series data based on auditory stimuli. The results demonstrated a significant correlation between the audio signals and the BOLD responses, indicating that linear models can effectively capture aspects of auditory processing in the brain.

Analysis of the model's weight coefficients revealed non-degenerate values, suggesting that the model successfully identified a diverse range of auditory features influencing neural activity. Furthermore, the estimation of the hemodynamic delay parameter, Δt, provided insights into the temporal dynamics of the BOLD response, aligning with previous findings on hemodynamic variability in fMRI studies.

References

[1] H. White. A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity. doi: 10.2307/1912934.

[2] Catherine A Cooper, Marc R Wilkins, Keith L Williams, and Nicolle H Packer. Bold—a biological o-linked glycan database. *ELECTROPHORESIS: An International Journal*, 20(18): 3589–3598, 1999.

[3] Rose Dawn Bharath. Functional mri: Genesis, state of the art and the sequel. *doi: 10.4103/0971-3026.130684.*

[4] Robert Fettiplace. Hair cell transduction, tuning, and synaptic transmission in the mammalian cochlea. *Comprehensive Physiology*, 7(4):1197–1227, 2011.

[5] Michelle Moerel, Essa Yacoub, Omer Faruk Gulban, Agustin Lage-Castellanos, and Federico De Martino. Using high spatial resolution fmri to understand representation in the auditory network. *Progress in neurobiology*, 207:101887, 2021.

[6] Colin Humphries, Einat Liebenthal, and Jeffrey R Binder. Tonotopic organization of human auditory cortex. *Neuroimage*, 50(3):1202–1211, 2010.

[7] Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366, 1980.

[8] Zrar Kh Abdul and Abdulbasit K Al-Talabani. Mel frequency cepstral coefficient and its applications: A review. *IEEE Access*, 10:122136–122158, 2022.

[9] Corneliu Toader, Calin Petru Tataru, Ioan-Alexandru Florian, Razvan-Adrian Covache-Busuioc, Bogdan-Gabriel Bratu, Luca Andrei Glavan, Andrei Bordeianu, David-Ioan Dumitrascu, and Alexandru Vlad Ciurea. Cognitive crescendo: how music shapes the brain's structure and function. *Brain sciences*, 13(10):1390, 2023.

[10] Julia Berezutskaya, Mariska J Vansteensel, Erik J Aarnoutse, Zachary V Freudenburg, Giovanni Piantoni, Mariana P Branco, and Nick F Ramsey. Open multimodal ieeg-fmri dataset from naturalistic stimulation with a short audiovisual film. *Scientific Data*, 9(1):91, 2022.

[11] Peter A Bandettini, Eric C Wong, R Scott Hinks, Ronald S Tikofsky, and James S Hyde. Time course epi of human brain function during task activation. *Magnetic resonance in medicine*, 25 (2):390–397, 1992.

[12] Marcus E Raichle. The brain's dark energy. *Science*, 314(5803):1249–1250, 2006.

[13] Molly G Bright and Kevin Murphy. Is fmri "noise" really noise? resting state nuisance regressors remove variance with network structure. *Neuroimage*, 114:158–169, 2015.

[14] John Fallon, Phillip GD Ward, Linden Parkes, Stuart Oldham, Aurina Arnatkevičiūtė, Alex Fornito, and Ben D Fulcher. Timescales of spontaneous fmri fluctuations relate to structural connectivity in the brain. *Network neuroscience*, 4(3):788–806, 2020.

[15] Bastian Rieck, Tristan Yates, Christian Bock, Karsten Borgwardt, Guy Wolf, Nicholas Turk-Browne, and Smita Krishnaswamy. Uncovering the topology of time-varying fmri data using cubical persistence. *Advances in neural information processing systems*, 33:6900–6912, 2020.

[16] Mingqian Zhao and Baolin Liu. An fmri-based auditory decoding framework combined with convolutional neural network for predicting the semantics of real-life sounds from brain activity. *Applied Intelligence*, 55(2):1–12, 2025.

[17] Yun Liang, Ke Bo, Sreenivasan Meyyappan, and Mingzhou Ding. Decoding fmri data with support vector machines and deep neural networks. *Journal of Neuroscience Methods*, 401: 110004, 2024.

[18] Alexander JE Kell, Daniel LK Yamins, Erica N Shook, Sam V Norman-Haignere, and Josh H McDermott. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644, 2018.

[19] Anders Eklund, Martin A Lindquist, and Mattias Villani. A bayesian heteroscedastic glm with application to fmri data with motion spikes. *NeuroImage*, 155:354–369, 2017.