Sound Reconstruction from Human Brain Activity

Jong-Yun Park

Supervisor: Prof. Dr. Yukiyasu Kamitani

Department of Intelligence Science and Technology
Graduate School of Informatics
Kyoto Univeristy

This dissertation is submitted for the degree of
Doctor of Philosophy

August 2023

Declaration

I, Park Jong-Yun, hereby declare that this thesis, entitled "Sound Reconstruction From Human Brain Activity" is original and my own work.

I declare that:

- This work was done solely while a candidate for the research degree at the Graduate School of Informatics, Kyoto University.
- No part of this work has previously been submitted for a degree at this or any other university.
- References to the work of others have been clearly attributed. Quotations from the work of others have been clearly indicated, and attributed to them.
- In cases where others have contributed to part of this work, such contribution has been clearly acknowledged and distinguished from my own work.

Jong-Yun Park
August 2023

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my advisor, Prof. Yukiyasu Kamitani. Your unwavering patience, insight, and expertise have shaped this research project and fostered my development as a scholar. Your encouragement when times were challenging was a constant source of motivation, and your insightful feedback was instrumental in helping me navigate the complex world of neuroscience research. Thank you for your continued support and for inspiring me to always push the boundaries of my knowledge and abilities.

I extend my appreciation to Prof. Hidehiko Takahashi. Your continued support and care during the degree course were crucial in overcoming numerous hurdles. Your commitment to my development and academic progression has been fundamental in this journey, and for that, I am profoundly grateful.

To my esteemed colleagues, Dr. Mitsuaki Tsukamoto, Dr. Misato Tanaka, Dr. Kei Majima, and Dr. Tomoyasu Horikawa, I owe a special mention. Your dedication, collaborative spirit, and shared wisdom have been a source of invaluable learning and inspiration. I am grateful for the enlightening discussions, constructive criticisms, and ceaseless assistance that marked our collective endeavor.

I am also indebted to Dr. Eizabro Doi, Dr. Shuntaro Aoki, Jun Kai Ho, Fan Cheng, Ken Shirakawa, and Takaya Ido, along with all members of the Kamitani Lab in Kyoto University and ATR. Your insightful feedback, guidance, and constant encouragement have significantly contributed to the success of this project and my development as a researcher.

I want to extend my deepest thanks to our administrators and secretaries, Yukari Kado and Yukiko Masuda. Your diligent work behind the scenes ensured I had the necessary resources and support to concentrate on my research. Your assistance is truly invaluable.

My gratitude extends to the facilities staff of the Kyoto University Institute for the Future of Human Society for providing an excellent research environment and the necessary resources to carry out this project.

Lastly, to my family and friends – your unwavering love, understanding, and support have been my anchor during this academic voyage. A special remembrance of my beloved grandmother who passed away in June; your spirit continues to be a source of strength and inspiration in my journey. The faith you had in me propels me to strive and excel even more.

Thank you all for contributing to this journey in your unique ways. It has been an honor to learn from and work with each one of you.

Abstract

The reconstruction of perceptual experiences from human brain activity has opened up new possibilities for understanding neural representations of sensory experiences. Despite substantial advancements, sound decoding studies have often shied away from reconstructing arbitrary sounds under unrestricted conditions, due to the complexity of temporal sequences in sounds, as well as the limited temporal resolution of neuroimaging tools. Nevertheless, leveraging the insights into the hierarchical nature of brain auditory processing offers a promising direction for reconstructing sounds from brain activity. In essence, the hierarchical processing of auditory features, a characteristic attribute of the human auditory system, paves the way for a more efficient and effective sound reconstruction. Furthermore, the advancements in audio-generative machine learning models offer unprecedented capabilities to translate compressed representations back into high-resolution sounds. In this light, this thesis introduces a novel method for reconstructing sound from functional magnetic resonance imaging (fMRI) responses. The approach combines the decoding of hierarchical auditory features from a DNN model with an audio-generative model. In Chapter 1, the thesis starts with a comprehensive introduction to human auditory processing and a review of the current state of reconstructing perceptual experiences in both vision and audition. Chapter 2 describes the compact representation designed to bridge the gap between sound and neuroimaging. Chapter 3 provides a detailed description of the experimental protocols used throughout this thesis to measure fMRI responses to natural sounds. Chapter 4 involves a feature decoding analysis using various auditory features from the auditory cortex. Chapter 5 presents the application of the proposed method for sound reconstruction from fMRI responses, demonstrating its capability to reconstruct complex spectral-temporal patterns that broadly maintain content and quality similar to the actual sound stimulus. In Chapter 6, a training dataset ablation analysis is conducted to investigate the generalizability of the proposed model. Chapter 7 explores the role of hierarchical auditory areas and DNN features in sound reconstruction. In Chapter 8, the study extends to "cocktail party conditions",

illustrating the potential of the proposed model to reconstruct the subjective content of top-down auditory attention. Finally, Chapter 9 discusses the implications of the proposed model and includes a preliminary analysis of potential applications for reconstructing auditory perceptual experiences.

Table of contents

List of figures

Nomenclature

Acronyms / Abbreviations

AAC	Auditory Association Cortex
AC	Auditory Cortex
CNN	Convolutional Neural Network
DNN	Deep Neural Networks
EAC	Early Auditory Cortex
ECoG	Electrocorticography
EEG	Electroencephalogram
Env.	Environmental sounds
F0	Fundamental frequency
fMRI	Functional Magnetic Resonance Imaging
HNR	Harmonic to noise ratio
MEG	Magnetoencephalography
RNN	Recurrent Neural Network
ROI	Region of interest
SC	Spectral centroid

Chapter 1

Introduction

Sound perception is a vital sense, allowing us to decipher and engage with our acoustic surroundings. The convergence of advancements in neuroimaging and machine learning technologies has opened new vistas of understanding sensory neural representations. This has enabled the reconstruction of perceptual experiences from human brain activity, offering deeper insights into our sensory world. Despite significant progress in neuroscience, most sound reconstruction studies have typically shied away from reconstructing sounds under unrestricted conditions for arbitrary sounds. This is mainly due to the wide diversity and complex temporal sequencing of sounds, as well as the relatively low resolution of neuroimaging modalities. Furthermore, our understanding of how the human auditory system extracts auditory information from perceived sounds remains largely confined within the sphere of peripheral sound processing. In the following section, I will introduce how the human auditory system processes auditory information. The second section will introduce brain decoding methods for the reconstruction of perceptual experiences, with a particular focus on visual and sound reconstruction, a cornerstone of this field. Following this, I will introduce how Deep Neural Networks (DNNs) serve as potent tools for sound decoding and reconstruction. Lastly, in the final section, I will outline the organization of this thesis.

1.1 Human auditory system

1.1.1 From the ear to brain

Auditory perception is a critical function not only for animals but also for humans. It allows us to deduce vital information such as the location of an object, its identity, who is speaking, and what they're saying. Auditory perception kicks off when sound enters the ear, converting air vibration into electrical action potentials, which are then transmitted to the brain cortex via the auditory nerve. Figure 1.1A illustrates the key processes of the ear involved in auditory perception. Initially, sound reaches the eardrum in the form of air vibrations. The sound then triggers the eardrum to vibrate and these vibrations are transmitted via the ossicles (the tiny bones in the middle ear) to the cochlea. Located in the inner ear, the cochlea is a spiral structure filled with a fluid. Integral to the cochlea is the organ of Corti, which is connected to the basilar membrane. The organ of Corti detects the fluid's vibrations in the cochlea, converts these vibrations into electrical signals, and transmits them to the brain via the auditory nerves. After the cochlea converts the vibrations into neural signals, several brain regions, including the midbrain, thalamus, and primary auditory cortex, process and convert the raw acoustic input into neural representations.

1.1.2 Auditory cortex

Brain anatomy and connectivity studies have helped researchers categorize the human auditory cortex into three distinct regions: a core region that receives input from the cochlea via the auditory nerves and thalamus; a belt region that surrounds the core area; and a parabelt region that surrounds the belt area (Kaas and Hackett, 2000; Sweet et al., 2005). Also referred to as the primary auditory cortex (PAC), the core region resides in Heschl's gyrus, nestled within the brain's lateral sulcus. The PAC receives electrical signals from the thalamus's auditory nucleus. Researchers have studied the function and structure of the PAC more extensively than other auditory regions, leading to the proposal of several computational models that explain the PAC's auditory processing.

1.1.3 Primary auditory cortex

Neurons in the PAC have a frequency preference and are particularly sensitive to complex spectra as opposed to pure tones (Moshitch et al., 2006). This frequency preference led

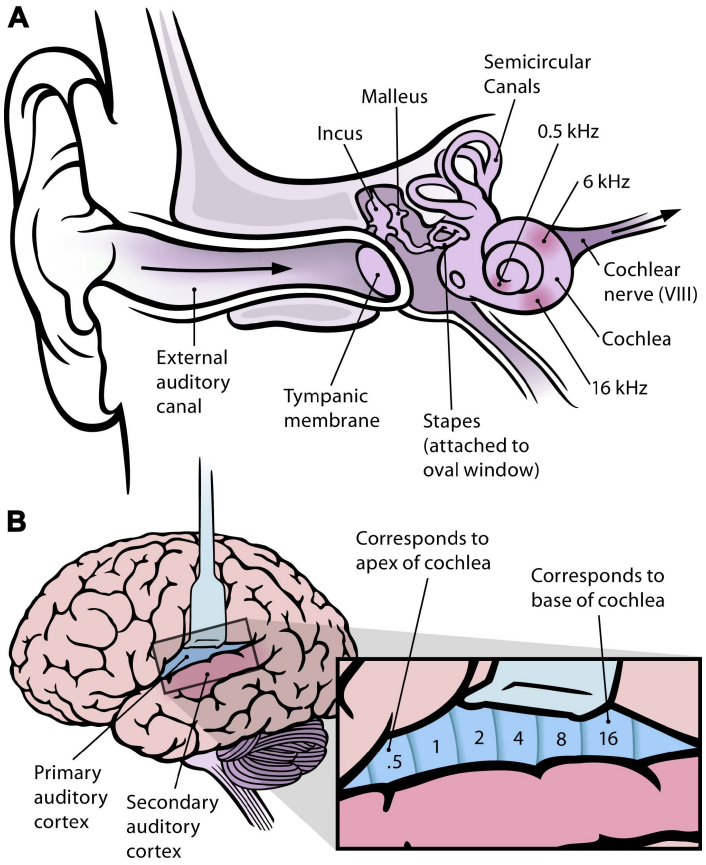

Fig. 1.1 Encoding of sound frequencies in the human ear and brain. (A) Sound waves arrive at the cochlea after passing through the eardrum, causing the basilar membrane and the hair cells situated on it to vibrate. This vibration is converted into electrical potentials by the hair cells, which in turn initiate neural activity. This electric signal is then conveyed to the primary auditory cortex via auditory nerve fibers. (B) The primary auditory cortex comprises a topographic map representing the frequency spectrum of the cochlea, denoted in kilohertz. Figure adapted from (Chittka and Brockmann, 2005), licensed under CC BY.

researchers to discover a tonotopic architecture in the PAC, an organization where neurons are spatially sorted according to their frequency preference (Figure 1.1B). Researchers have detected tonotopic maps from single-unit to fMRI studies in humans (Formisano et al., 2003; Humphries et al., 2010; Moerel et al., 2014) and monkeys (Petkov et al., 2006).

1.1.4 The dual stream

The dual stream hypothesis, a prevalent model for brain processing in both the auditory and visual domains, proposes that sensory input is processed via two distinct pathways. The dorsal pathway, often referred to as the "where pathway," is implicated in recognizing the location of objects or sounds. This pathway spans from the primary sensory areas to the posterior parietal cortex. On the other hand, the ventral pathway, known as the "what pathway," is responsible for recognizing objects or environmental sounds, and it extends from the primary sensory areas to the anterior temporal cortex (Rauschecker and Scott, 2009).

Anatomical research on monkey brains has revealed distinct streams in the anterior and posterior regions of the auditory cortex (Munoz-Lopez et al., 2010). These studies have uncovered an extensive anatomical tract stretching from the anterior belts to the ventrolateral prefrontal cortex (PFC), as well as another from the caudal belt to the dorsolateral PFC. In addition to anatomical studies, single-unit investigations have lent credence to the dual pathway hypothesis. For instance, studies by Tian (2001) revealed that neurons in the caudal belt have a more active response to the spatial location of sound sources compared to the core and anterior belt regions. Furthermore, research by Recanzone and colleagues showed a strong correlation between neural response and sound source localization in the caudal belt, bolstering the validity of the "where" pathway (Recanzone et al., 2000).

The functional specialization of the ventral areas has also been observed in single-unit studies (Lewis and Van Essen, 2000). These studies found that the ventral pathway processes visual, auditory, and somatomotor representations. These findings collectively provide robust support for the dual pathway hypothesis in the auditory domain and highlight the diverse roles of different brain regions in auditory processing.

1.1.5 Computational models of neuronal tuning

Understanding auditory processing has long been a key focus of neuroscience, and computational models have emerged as an essential tool in these endeavors. These models are crucial

because they can simulate and predict how neural behavior responds to complex sound inputs, providing insights into the intricate processes of the auditory system. In particular, computational models of neuronal tuning often center around spectral and temporal modulations, key factors processed by neurons in the auditory cortex.

Spectral modulation, or power variation across the frequency axis, is a fundamental part of auditory perception. Given that sounds such as speech, music, or natural sounds each have distinct frequency representations, neural tuning to spectral modulation is integral to cortical responses (Barbour and Wang, 2003). Typically, computational models for auditory processing aim to convert sound features into these temporal and spectral modulations. These conversions result in representations similar to spectrograms or cochleagrams, facilitated by a filter bank (Chi et al., 2005).

While early studies of human auditory systems have primarily concentrated on initial auditory processes like sound transduction, many of the more complex aspects of auditory perception remain largely uncharted. Recently, there has been a shift toward examining mid-level and high-level representations beyond the auditory core using computational models (McDermott, 2018).

One such study used a natural sounds dataset to measure the hierarchical cortical processes in the auditory area (Norman-Haignere et al., 2015). The researchers recorded fMRI responses from subjects listening to natural sounds and trained a model to predict voxel responses from natural sound stimuli. They introduced "voxel decomposition analysis," revealing the primary components of natural sound and identifying distinct components for the music and speech categories. Interestingly, they found that other components were more responsive to frequency and modulation representations, rather than sound categories. This result supports the concept of hierarchical processing in auditory perception, particularly in music and speech categories.

These explorations into hierarchical auditory processing and understanding neural tuning at different levels underpin further advancements in auditory perception research, highlighting the pivotal role of computational modeling in auditory neuroscience.

1.2 Externalization of perceptual experience

Brain decoding, a prevalent method in neuroscience, stands at the intersection of neuroimaging and machine learning advancements. This technique allows us to interpret information from sensory input reflected in specific brain regions, providing vital insights into the neural representation of mental content. Over the years, numerous studies in the field of vision have successfully leveraged decoding methods to interpret what individuals see (Haxby et al., 2001; Horikawa and Kamitani, 2017a; Kamitani and Tong, 2005), imagine (Andersson et al., 2019; Hassabis et al., 2014; Horikawa and Kamitani, 2017a), and dream (Horikawa and Kamitani, 2017b; Horikawa et al., 2013). Similarly, auditory research has decoded acoustic features (Sankaran et al., 2018), sound category recognition (Zhang et al., 2018), speech recognition (Heelan et al., 2019), and even inner speech (Martin et al., 2018) from brain responses. The rise of deep learning, particularly in the realm of computer vision, has further advanced our understanding of mental content. Deep neural networks (DNNs) have led to the development of increasingly sophisticated decoding models, allowing for the reconstruction of subjective visual experiences (Shen et al., 2019b).

Despite these significant advancements, reconstructing arbitrary sounds from brain activity remains a considerable challenge. This difficulty arises due to the complexity of temporal sequences in sounds and the relatively low resolution offered by neuroimaging modalities. Such hurdles necessitate innovative solutions and careful refinement of our methodologies as I continue to advance in the field of neuroscience. The promise of externalizing perceptual experiences offers exciting opportunities, such as exploring the neural basis of auditory hallucinations or developing more advanced communication methods for individuals with speech impairments. These possibilities underscore the importance and urgency of further research in this area.

1.2.1 Visual reconstruction

Image reconstruction of subjective visual experiences offers substantial insights for understanding the processing of neural representations. Although the reconstruction of sensory stimuli theoretically demands an infinite stimulus set, it is impractical to gather brain activity data for all possible visual stimuli. In an early study, Miyawaki et al. (2008) proposed a visual image reconstruction method that combined multi-scale image patches to predict binary contrasts from fMRI activity (Figure 1.2). The researchers trained a brain decoder

to predict local image contrasts at multiple scales from the fMRI responses in the primary visual cortex. By integrating these locally predicted contrasts, they facilitated unconstrained visual image reconstruction and successfully managed to reconstruct geometric shapes and alphabets. This advancement serves as a testament to the potential of brain decoders and their role in arbitrary sensory stimuli reconstruction.

The advancement of deep learning in the domain of computer vision has significantly augmented the sophistication of decoding models, enabling the reconstruction of subjective mental content. Researchers have increasingly turned their focus towards features processed hierarchically, akin to the functioning of the human brain. This focus stems from the demonstrated parallels in hierarchical processing structures between the human sensory system and deep neural network (DNN) models.

Taking inspiration from this hierarchical representation homology between the brain and DNNs, Shen et al. (2019b) proposed an image reconstruction method. They utilized decoded DNN features from multiple layers of fMRI responses to build a reconstruction algorithm. This algorithm optimized the pixel values of an image to align its DNN features with those decoded from human brain activity across multiple layers (Figure 1.3). Their proposed method demonstrated a reliable capacity to produce reconstructions that resembled a range of viewed natural images. Furthermore, when applied to mental imagery, the same analysis yielded basic reconstructions of the subjective content. These results suggest that integrating hierarchical neural representations provides a novel lens into reconstruction of perceptual visual experiences.

Generative models offer a new avenue for reconstructing visual experiences from human brain activity. In particular, Generative Adversarial Networks (GANs) effectively lend semantically meaningful details to the reconstructions (Goodfellow et al., 2016). A GAN is structured with two neural networks: a generator and a discriminator. The generator learns to transform the input (latent space) to create images resembling the training images. Conversely, the discriminator learns to differentiate between real images from the training set and fabricated images from the generator. The power of GAN lies in its ability to transform random latent features into meaningful images across multiple scales.

Seeliger et al. (2018) implemented a linear model to predict the latent features of the generative model from brain responses. Initially, they trained a GAN model to understand the latent space using an unsupervised method on a large image dataset. After completing

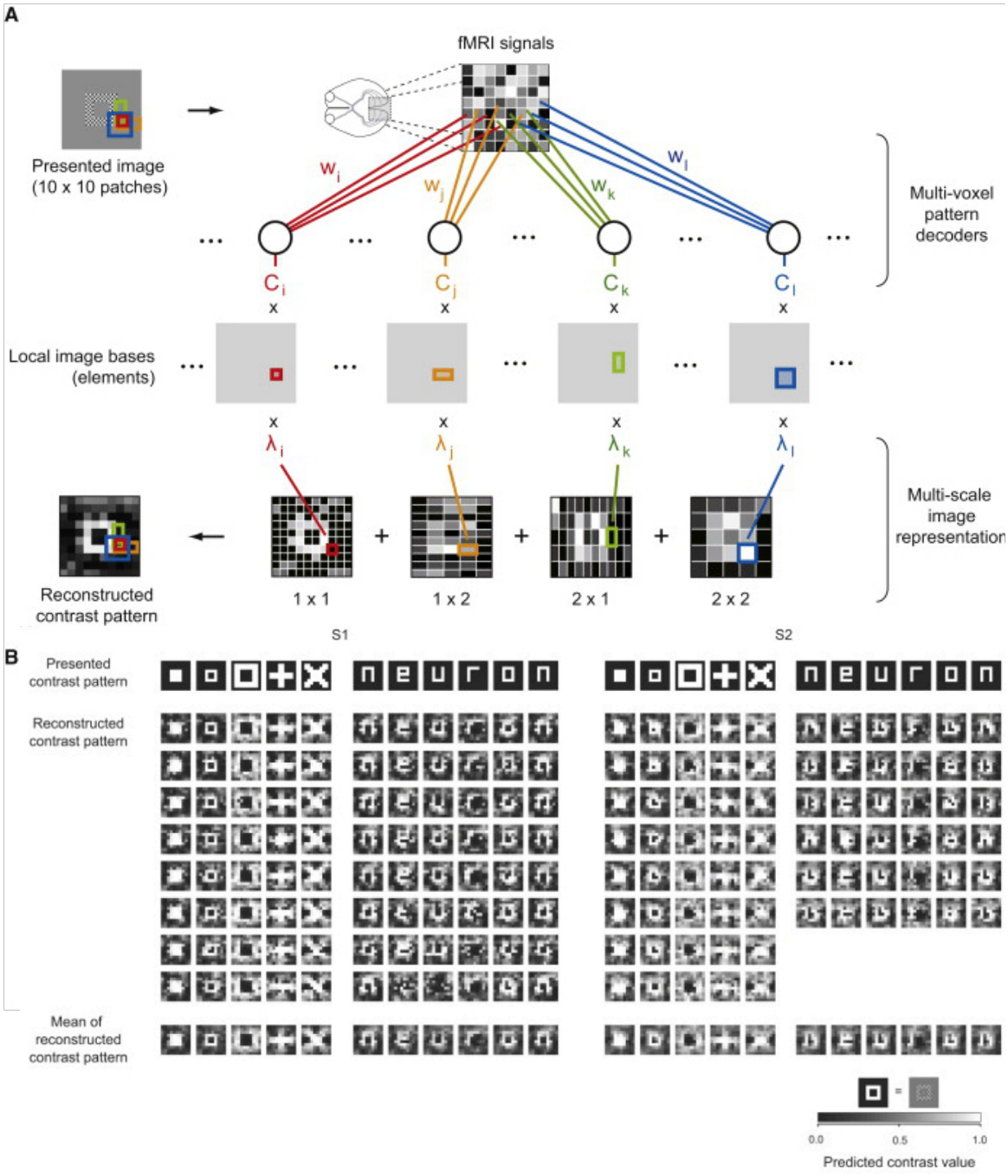

Fig. 1.2 Visual image reconstruction using a combination of multiscale local image decoders. (A) Several brain decoders were trained to predict four local image contrasts at various scales using the fMRI responses from the primary visual cortex. (B) By integrating these locally predicted contrasts, geometric shapes and alphabets were succesfully reconstructed from fMRI responses. Figure adapted from (Miyawaki et al., 2008). Copyright 2008 with permission from Elsevier.

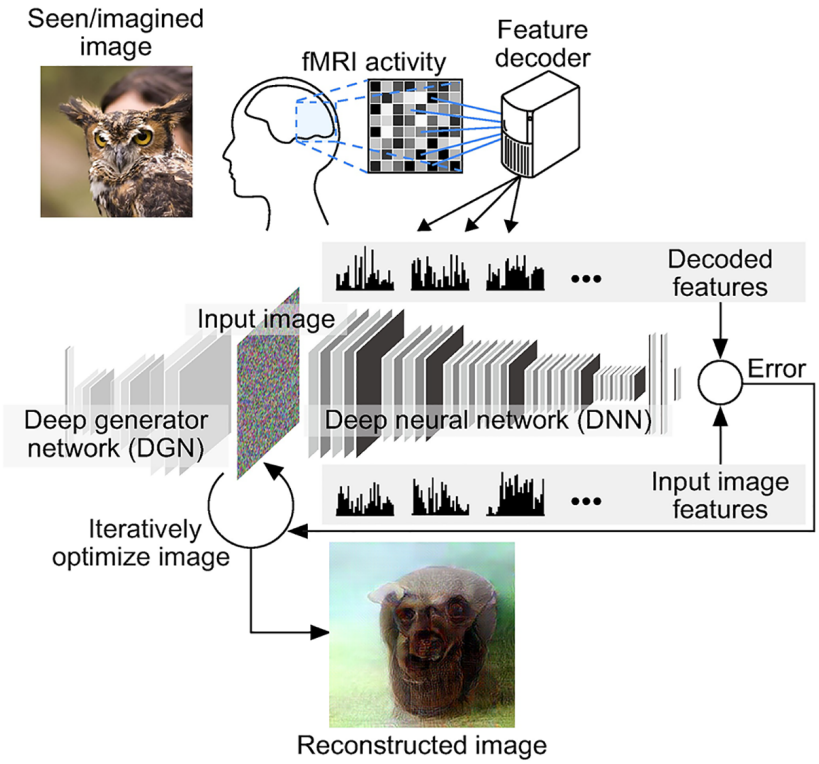

Fig. 1.3 Visual image reconstruction using a hierarchical DNN features. Decoded DNN features across multiple layers of fMRI responses were used to create a reconstruction algorithm. This algorithm was designed to optimize an image's pixel values to make its DNN features coincide with those decoded from multi-layered human brain activity. The demonstrated capability of their method was its consistent ability to produce reconstructions resembling a diverse array of viewed natural images.. Figure adapted from (Shen et al., 2019b), licensed under CC BY.

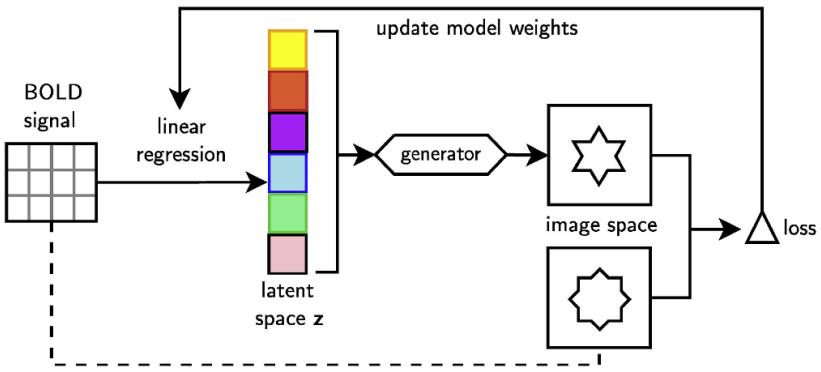

Fig. 1.4 Visual image reconstruction using generative DNN model. Initially, a generator model was trained in an unsupervised manner using a large image dataset, thereby facilitating the understanding of the latent space. Following this, linear models were trained to predict latent features from fMRI responses, aiming to minimize the disparity between the original and reconstructed images. Figure adapted from (Seeliger et al., 2018). Copyright 2018 with permission from Elsevier.

this training, they trained linear models to predict latent features from fMRI responses by minimizing the discrepancy between the true and reconstructed images (Figure 1.4). Applying this approach enabled us to reconstruct both structural and certain semantic features of a subset of the natural images. Moreover, Shen et al. (2019a) introduced a modified GAN model that employed fMRI responses as a direct input and trained the GAN model from scratch. This approach was innovative, as most reconstruction models avoid using fMRI responses as direct inputs due to the limited data size of fMRI studies. To address this issue, they utilized a modified GAN strategy, incorporating a generator, a discriminator, and a comparator. The comparator used in the reconstruction was a pre-trained DNN for image object recognition, and its weight remained fixed during the generator's training. The success achieved in visual reconstruction highlights the potential of DNNs in interpreting and translating complex neural activities.

1.2.2 Sound reconstruction

In parallel, researchers have sought to apply similar principles to the auditory system to unravel the complex neural mechanisms underpinning our auditory experiences . However, unlike visual reconstruction, sound decoding studies typically avoid reconstruction under unconstrained conditions for arbitrary sounds. This is primarily due to the broad diversity and complex temporal sequencing of sounds, coupled with the relatively low resolution of neuroimaging modalities. Traditionally, neuroimaging modalities such as electroencephalog-

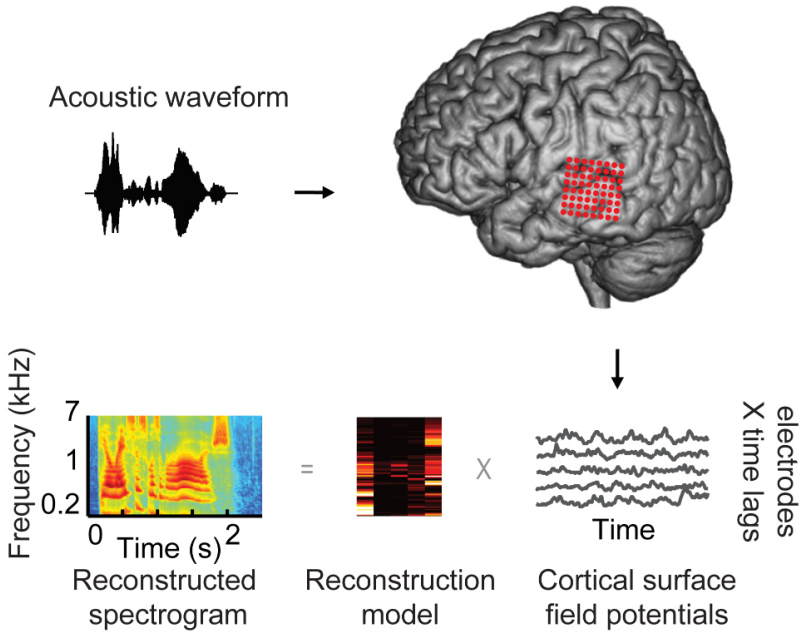

Fig. 1.5 Sound reconstruction using invasive recording for predicting spectrogram. Pasley et al. (2012) collected ECoG signals from the nonprimary auditory cortex of participants who were listening to isolated words. A model was then trained using these signals, with the aim of directly predicting the modulation or spectrogram of the auditory stimuli being perceived. Figure adapted from Pasley et al. (2012), licensed under CC BY.

raphy (EEG) and magnetoencephalography (MEG) have been favored for auditory decoding due to their superior temporal resolution, as they capture real-time electrical activity from the scalp or sensors placed on the head. Pasley et al. (2012) conducted a study where they measured ECoG signals from the nonprimary auditory cortex of subjects as they listened to isolated words. They trained a model to directly predict the modulation or spectrogram of the speech heard from these signals (Figure 1.5. Their findings highlighted the capability of a linear model, based on the auditory spectrogram, to accurately reconstruct slow and intermediate temporal fluctuations. Nevertheless, to reconstruct fast temporal fluctuations, a nonlinear approach to sound representation was required, relying on temporal modulation energy. This decoded representation of speech allowed the identification of individual words from brain activity during single trial record.

As advancements in Deep Neural Networks (DNNs) have significantly enhanced visual reconstruction, they have also greatly improved the sophistication of sound reconstruction. One common technique involves the reconstruction of spectrograms from neural responses. Akbari et al. (2019) employed DNN models with ECoG signals to predict spectrograms or

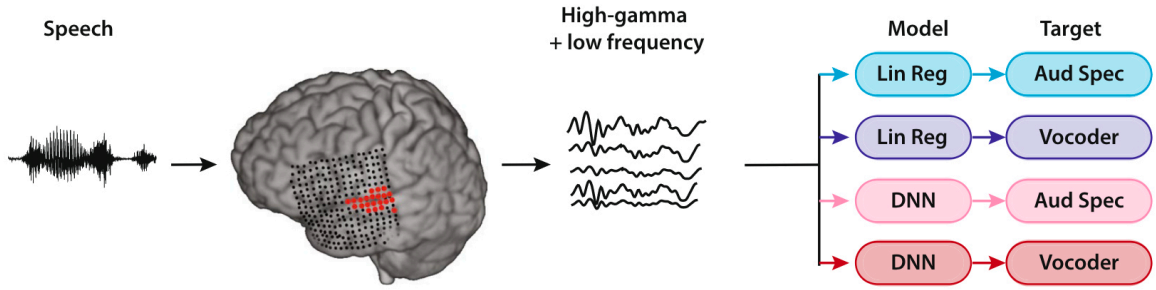

Fig. 1.6 Sound reconstruction based on DNN model for predicting spectrogram from invasive recording. While participants listened to brief, isolated words, such as digits, ECoG signals were collected. Initially, an autoencoder model was trained to extract latent representations from the audio signal. Subsequently, DNN models were trained to predict these latent representations using the collected ECoG signals. Figure adapted from (Akbari et al., 2019), licensed under CC BY.

vocoder features. These could then be transformed back into sound. The team gathered ECoG signals while participants listened to brief, isolated words, such as digits. They first trained an autoencoder model, which learned to internalize the acoustic features of stimuli and subsequently regenerate those features from the latent representations. They then trained DNN models to anticipate these latent representations from the ECoG signals (as depicted in Figure 1.6). In a related study, Wang et al. (2018) proposed a model that modified the architecture of WaveNet—a generative model used for sound synthesis. This adapted model was able to create spectrograms using ECoG time series as inputs (see Figure 1.7). These methods demonstrated intelligible recognition results in both a quantitative and qualitative sense. However, due to the invasive nature of the data collection process and the subsequent limitations on dataset sizes, their utilization has been confined to classifying predefined speech (Chakrabarti et al., 2015; Martin et al., 2018; Moses et al., 2019; Pei et al., 2011) and reconstructing constrained examples such as digits (Akbari et al., 2019) and words (Wang et al., 2018). Furthermore, the intrinsic temporal resolution limitations of fMRI have primarily confined its use to classification approaches (Correia et al., 2015; Formisano et al., 2008).

Contrary to this common practice, recent studies propose that reconstructing unconstrained sounds may be possible without the need for exact temporal alignment between neural recordings and auditory stimuli. This involves leveraging the spatial patterns in fMRI data to compensate for its limited temporal resolution, allowing for the prediction of intricate temporal information. Santoro et al. (2017) developed a computational model that decoded

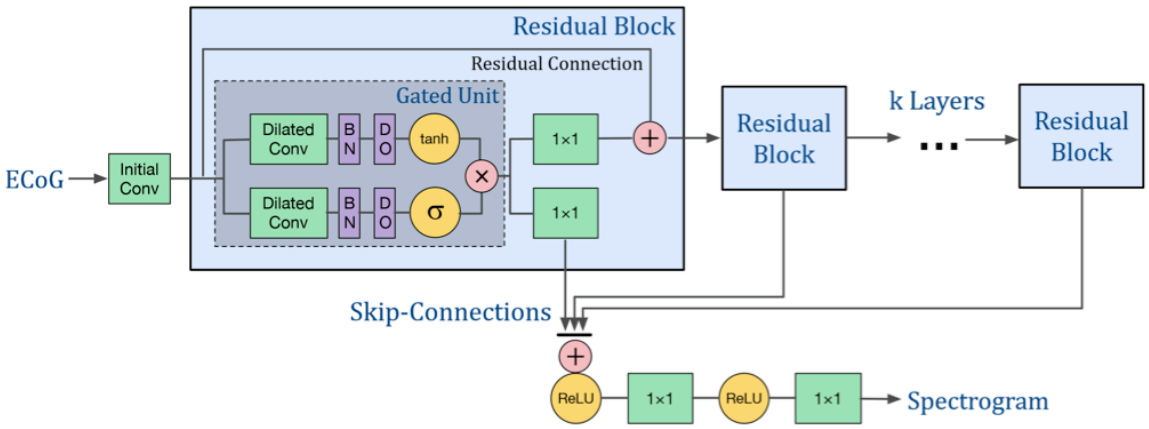

Fig. 1.7 Sound reconstruction based on audio generative model. Wang et al. (2018) proposed a reconstruction model that modified the structure of WaveNet, a audio generative model. This model was capable of generating spectrograms using time series ECoG data as inputs. Figure adapted from (Wang et al., 2018), Copyright 2020 IEEE.

the physical features of natural sounds from high spatial resolution 7T fMRI responses. This model utilized several multivariate decoders to predict spectral-temporal modulation features from fMRI activation patterns. Impressively, these trained decoders were able to predict subtle modulation changes even from fMRI's coarse temporal sampling (2.6 seconds). To facilitate interpretation of the decoded results, the study converted the decoded features back into sounds. Despite the encouraging results, the reconstructed sounds lacked complex spectro-temporal patterns. This resulted in temporally smoothed reconstructions, which were challenging for human listeners to recognize.

1.3 Deep Neural Network for sound reconstruction

Deep Neural Networks (DNN) are a category of machine learning models that mimic the structure and function of the human brain. They consist of multiple layers of interconnected nodes, or "neurons," which allow these models to learn complex patterns and relationships in data. In the field of sound reconstruction, DNNs are particularly potent, as they can be employed to decode brain activity by mirroring the processing features of the human auditory system. DNNs are notable for their exceptional ability to handle large-scale time-series data like sounds, as they can efficiently extract concise representations and reconstruct sound back. These capabilities significantly contribute to the field of neural sound reconstruction, marking an intriguing intersection between neuroscience and artificial intelligence. This crossover

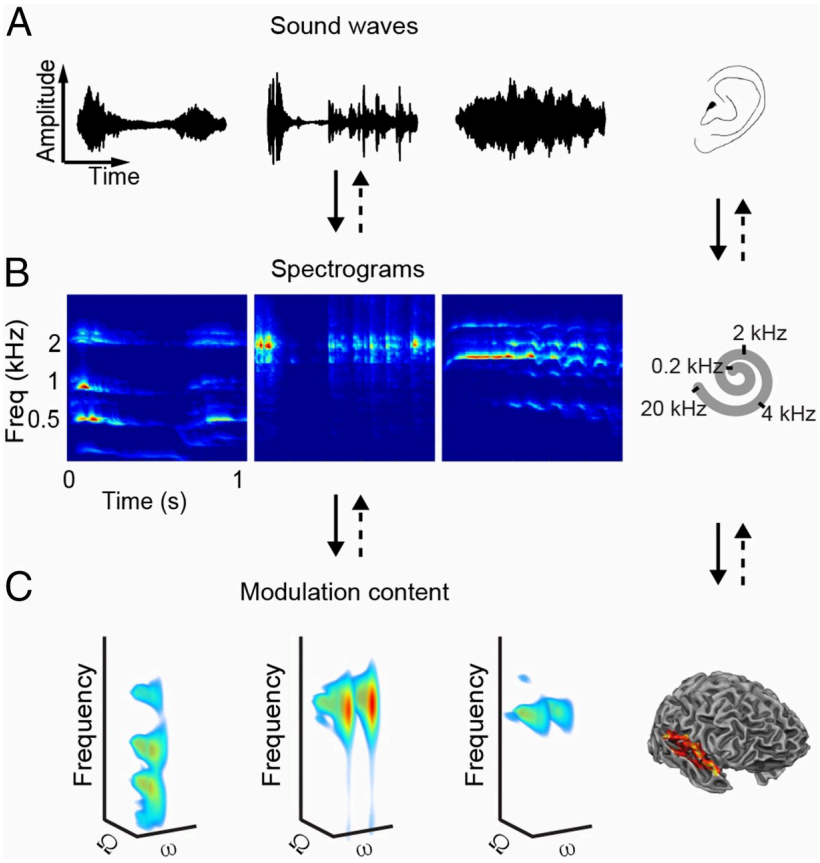

Fig. 1.8 Sound reconstruction using spatial patterns in fMRI voxel responses. (A) Sound waves arrive at the human ear as a waveform. (B) The cochlea breaks down this sound waveform into its component frequencies. (C) The auditory cortex processes the modulations of these spectral-temporal components. Santoro et al. (2017) used a brain decoder to predict these spectro-temporal modulation features from fMRI responses, which were then converted back into a spectrogram and sound waveform, respectively. Figure adapted from (Santoro et al., 2017). Copyright 2018 by the National Academy of Sciences.

holds vast potential to not only enhance our comprehension of how the brain processes sound but also to pave the way for novel applications.

1.3.1 Sound recognition

In this section, I explore the extraordinary role of DNN in sound recognition. The aptitude of DNNs in processing intricate time-series data such as sound, coupled with their capability to form precise representations of this data, significantly fosters advancements in the field of neural sound reconstruction. Early models employed in this field often mirrored the architecture of Convolutional Neural Networks (CNNs), extensively utilized in image object classification (Hershey et al., 2017). Using a nearly identical structure and training tasks, these models were successfully applied to acoustic scene classification, predict and identify a wide array of sounds, thereby significantly outperforming traditional methods (Han et al., 2017; Salamon and Bello, 2017; Stowell et al., 2015). Speech recognition, a significant and widely researched area within sound recognition, is particularly challenging. Human speech is highly variable due to differences in speakers, speaking attributes, and environmental noise. This necessitates mapping variable-length speech signals into words or speech representations (Deng et al., 2013). To handle the sequential nature of these temporal processes, sequential DNN models like Recurrent Neural Networks (RNNs) or Long Short-Term Memory networks (LSTMs) are commonly employed in sound recognition tasks (Graves et al., 2013; Hannun et al., 2014).

Analogous to the human auditory system's approach to processing and interpreting various sounds, DNNs can be tailored to simulate these functions. The network layers can be fine-tuned to mimic the hierarchical processing structure of the human auditory system, which spans from the initial detection of basic sound features to the higher-level interpretation of complex auditory scenes. This capability of DNNs to mimic the processing features of the human auditory system is extremely beneficial when applied to decode brain activity. It not only enhances the performance of the DNN models, but also provides valuable insights into the functioning of the human auditory system.

1.3.2 Audio generation

Audio-generative models, which leverage the capabilities of DNN to create novel sound sequences, are at the forefront of current research. These generative models discern the underlying distribution of training data, enabling them to craft new data instances that

resemble the original. In the realm of sound, these models can generate new sound sequences that preserve the statistical properties of the training sounds. There are various types of generative models employed in the domain of sound, including Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and autoregressive models such as WaveNet. Each of these models possesses unique characteristics, making them suited to different sound generation tasks.

VAEs, for instance, are capable of learning a compressed, or "latent," representation of sound data. This latent representation can be sampled to generate new sounds. VAEs construct a probabilistic mapping of input data into a lower-dimensional latent space, where the complexity of the data distribution is tamed. New data points (sounds) can then be generated by sampling from this simpler, learned distribution and mapping it back to the data space. Recently, a variant of VAEs known as Vector Quantization-VAEs (VQ-VAEs) has been developed, which uses discrete rather than continuous latent representations (Oord et al., 2018). This approach has shown promise in mitigating the problem of “posterior collapse” — a situation where the latent variables are ignored when paired with a powerful autoregressive decoder, such as a WaveNet. This issue often occurs when the decoder can model the data well without the need for information from the latent variables, leading to an underutilization of the latent space. By using discrete representations, VQ-VAEs encourage the model to use the latent space more effectively, thereby enhancing the quality of the generated sounds.

On the other hand, GANs involve a ’game’ between two networks—a generator that creates new sounds and a discriminator that attempts to differentiate the generated sounds from real ones. The interplay between these two networks eventually leads to the generation of highly realistic sounds. Donahue et al. (2019) introduced WaveGAN, a model that applies GANs to the unsupervised synthesis of raw-waveform audio. In WaveGAN, a flattened version of the DCGAN architecture is employed to generate one-dimensional samples as audio waveforms (Radford et al., 2016). Similarly, Kumar et al. (2019) proposed MelGAN, a model designed to generate raw waveforms of high temporal resolution. MelGAN utilizes a non-autoregressive, feed-forward convolutional architecture to capture high-frequency representations of an audio signal. This approach promises a significant avenue for generating complex sounds from spectrogram, highlighting the potential of DNNs in audio generation tasks.

Finally, autoregressive models like WaveNet generate sounds by predicting the next sample in a sound sequence based on the previous samples. WaveNet, in particular, has

shown remarkable results in generating realistic and high-quality speech sounds (Oord et al., 2016). It operates by taking a sequence of audio samples and predict the next sample in the sequence. Over time, this allows WaveNet to generate a complete sound sequence that is typically much more realistic and natural-sounding than sounds generated by other types of models. Each of these models offers unique capabilities that can be harnessed for sound generation, making this a rich and exciting area of research.

1.3.3 Neural coding and sound DNN

Neural coding refers to how the nervous system translates sensory information. When it comes to sound, neural coding encapsulates the intricate series of transformations that convert auditory stimuli into patterns of neural activity. Decoding these transformations is critical for understanding the neural code of sound perception, and this is an area where DNN can offer invaluable insights. Sound recognition models, engineered to emulate the hierarchical processing of the human auditory system, can learn to discern patterns in complex sound data, predict, and categorize various sounds. This capability makes them an excellent tool for studying and modeling the neural coding of sound. DNNs can decipher the complex relationship between a sound and its neural representation by training on extensive datasets of sound stimuli and corresponding neural responses. This learned relationship can then be used to predict the neural responses to new sounds, offering a quantitative model of the neural coding process.

An approach referred to as encoding modeling has successfully predicted neural responses to sounds in various auditory brain regions using this principle. Evidence of a similar hierarchical processing structure in both the human auditory system and DNN models supports this. A comprehensive brain encoding analysis conducted by Kell et al. (2018) predicts human auditory responses from DNN model responses, highlighting the hierarchical homology between the DNN model and fMRI data. They developed a DNN architecture for sound recognition, mirroring the hierarchical processing integral to the human auditory system. The structure of this DNN segregates common layers, responsible for low-level processing akin to early auditory stages, from branching layers that handle task-specific processing, such as speech recognition or music genre classification. Using this trained DNN model and fMRI data for natural sounds, their encoding analysis revealed a clear hierarchical correspondence between the brain and DNN models. It was observed that the responses from the early auditory cortex of the brain were more accurately predicted using DNN features derived from the common layer. Conversely, the responses from non-primary brain regions

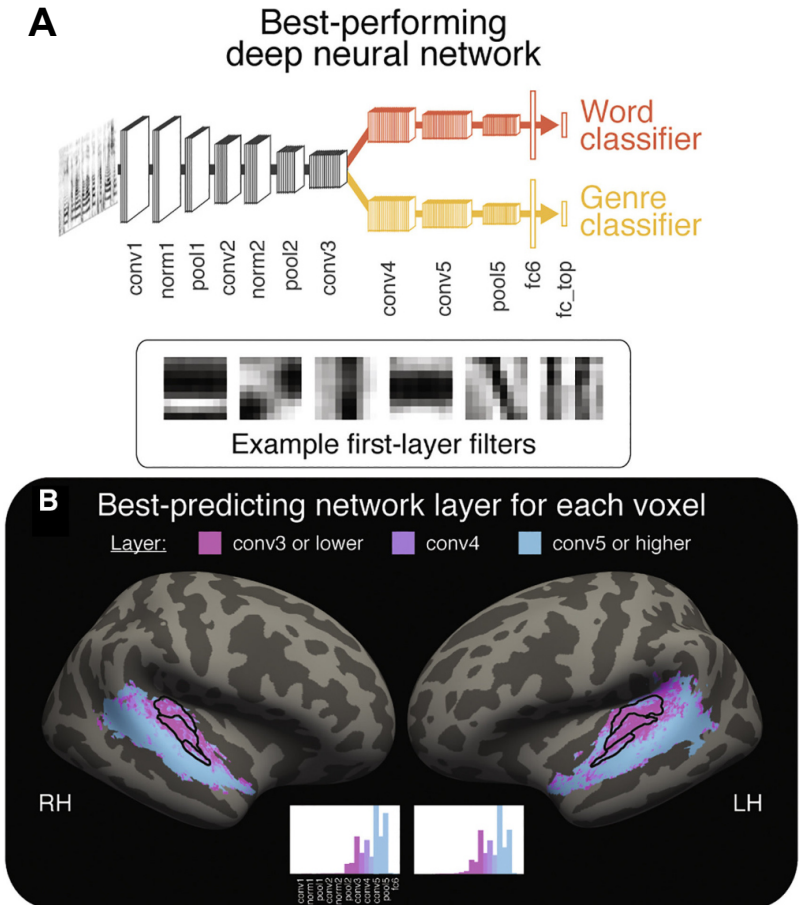

Fig. 1.9 The architecture of the DNN model emulates the hierarchical processing of the human auditory system. (A) An example of a DNN architecture designed for sound recognition mirrors the hierarchical processing integral to the human auditory system. The structure of this DNN segregates common layers, which are responsible for low-level processing akin to early auditory stages, from branching layers that handle task-specific processing, such as speech recognition or music genre classification. (B) The encoding analysis using this hierarchical DNN model alongside fMRI responses revealed a distinct hierarchical correlation between brain and DNN models. The responses from the early auditory cortex in the brain were more accurately predicted using the DNN features derived from the common layer of the model. In contrast, the responses from non-primary brain regions were more accurately predicted using DNN features from the branched layers. The figure is adapted from (Kell et al., 2018). Copyright 2018 with permission from Elsevier.

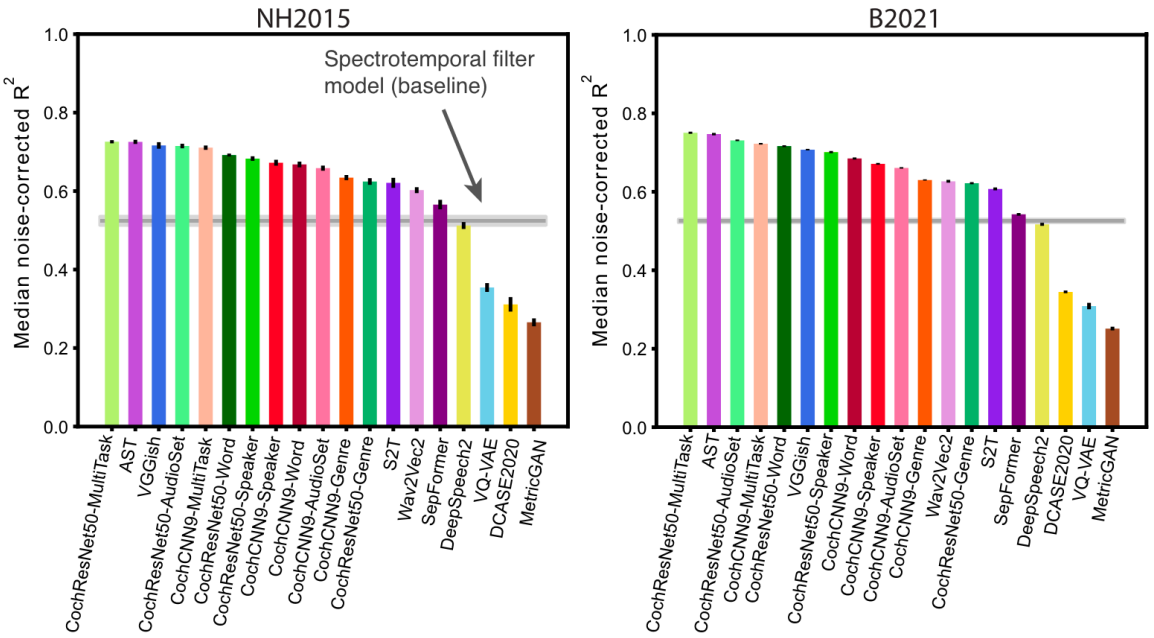

Fig. 1.10 Encoding performance of various sound DNN models compared to a baseline of spectro-temporal modulation features (represented by the grey line). While most models show higher encoding performance and correspondence than the baseline, certain state-of-the-art models, including generative models such as VQVAE and MelGAN, perform below the baseline, indicating a potential need for further refinement and research. Figure adapted from (Tuckute et al., 2023), licensed under CC BY-NC-ND.

were better predicted using DNN features obtained from the branched layer. This suggests a clear hierarchical correspondence between neural auditory processing and the structure of hierarchical DNN models.

However, a subsequent study observed that not all DNNs exhibit homology with the human brain as their encoding performance and hierarchical correspondence can significantly vary depending on the structure and optimization tasks of the DNNs (Tuckute et al., 2023). Especially, most models showed higher encoding performance and correspondence than spectro-temporal modulation, but the latest state-of-the-art models, especially generative models such as VQVAE and MelGAN, showed lower performance than modulation features.

The inverse process, known as decoding approach, can also be performed using DNNs. In this scenario, the models are trained to infer the sound stimulus from the observed neural responses, effectively 'decoding' the neural code. This approach can yield insights into how different sound features are represented in the brain and can also be utilized for neural sound

reconstruction. However, due to limitations in neuroimaging, research on brain decoding using sound DNNs has been relatively scarce.

1.4 Thesis organization

In this dissertation, I demonstrate the potential for sound reconstruction from human brain activity utilizing DNN models that are architected to mimic the human auditory system, alongside state-of-the-art sound generative models. In Chapter 4, I conduct a feature decoding analysis using fMRI responses and a variety of auditory features. Among the auditory features, those that were processed through hierarchical means displayed superior decoding performance, aligning with previous encoding analyses. In Chapter 5, I exploit these DNN features to reconstruct sound. The methodology synergistically combines the decoding of auditory features with an audio-generative model, enabling the disentanglement of temporally compressed information within DNN features. Chapter 6 showcases the generalizability of the approach by reconstructing sound categories not included in the training dataset. In this section, I also introduce how intermediate representations can be interpreted. In Chapter 7, I explore the hierarchical combinations of DNN layers and individual ROIs (Regions of Interest) that contribute to sound reconstruction. Chapter 8 comprises an experiment involving selective auditory attention to one of the overlapping sounds to determine if the reconstructions indeed mirror actual subjective perceptual experiences. Finally, Chapter 9 will summarize and discuss our results, contributions, and potential future applications.

Chapter 2

Bridging temporal gaps in sound domain

2.1 Introduction

The endeavor to decode and reconstruct sound from brain activity requires innovative methods to address the temporal intricacies within the sound domain. Despite previous progress in using spatial patterns from neuroimaging and hierarchically processed auditory features for sound reconstruction, the limited temporal resolution of neuroimaging data remains a considerable obstacle to achieving high-quality sound reconstructions. A key challenge in the realm of sound processing and analysis is transforming high-dimensional audio data into a more manageable, lower-dimensional form. This chapter delves into the approaches and techniques used to perform this crucial dimensionality reduction for effective sound data management.

Various contemporary studies have utilized an array of methods, from traditional techniques like principal component analysis (PCA) to more advanced deep learning architectures. PCA, a widely-used technique, reduces the complexity of high-dimensional data by transforming it into a new coordinate system with orthogonal axes, known as principal components.

Yet, as artificial intelligence and machine learning evolve, more sophisticated techniques have emerged. Autoencoders, a type of artificial neural network, have been used for efficient encoding of input data, aiding in reducing dimensionality in sound data (Hinton and Salakhutdinov, 2006; Vincent et al., 2008). Further advancements, such as Variational Autoencoders

(VAE) and Generative Adversarial Networks (GANs), have also been employed to create lower-dimensional audio data representations (Donahue et al., 2019).

Despite the effectiveness of these techniques, they may still struggle to bridge the significant temporal gap between sound data and neuroimaging. In response to this challenge, Iashin and Rahtu (2021) introduced SpecVQGAN, a method aiming to overcome the temporal limitations of conventional approaches. This method combines the power of GANs in capturing complex data distributions with the efficiency of vector quantization (VQ) in compressing high-dimensional data into a lower-dimensional discrete space.

With SpecVQGAN, we can transform high-dimensional spectrograms into a compact codebook representation, effectively bridging the gap between high-dimensional spectrograms and the limited temporal resolution of neuroimaging data. The implementation and effectiveness of SpecVQGAN in our sound reconstruction framework will be examined in-depth in this chapter, laying the groundwork for a more comprehensive understanding of how the human auditory system hierarchically processes auditory features.

In this chapter, we will highlight the effectiveness of SpecVQGAN and provide insights into the interpretation of codebook representations. The codebook representation used in this chapter was also employed for sound reconstruction in chapter 5.

2.2 Methods

In this chapter, we incorporated the SpecVQGAN model initially established in the work of Iashin and Rahtu (Iashin and Rahtu, 2021). The pre-trained models and corresponding scripts can be accessed at https://iashin.ai/SpecVQGAN. Specifically, we leveraged their pre-trained models for VGGish-ish. Additionally, we trained models for SpecVQGAN with the objective of generating 4-second sound segments. This utilization of established and trained models contributed to the efficiency and effectiveness of our sound reconstruction methodology.

2.2.1 Data processing

For training, test and validation of DNN models, I used the VGGsound dataset (Chen et al., 2020). This public audio-visual databaset, comprising of 200,000 videos extracted from YouTube, offers reliably annotated labels across 309 categories. I omitted any videos with missing or invalid links to ensure data integrity. This resulted in balanced validation and test

datasets, each containing an equal distribution of audio clips across the 309 categories. The training dataset housed 157,000 audio clips, while the validation and test datasets held 19,000 and 15,000 audio clips respectively. All audio clips were trimmed to a uniform duration of 4 seconds for consistency.

All the sound stimuli were processed by involving resampling audio clips at a frequency of 22050 Hz. I then generated log-Mel-spectrograms using a Short-Term Fourier Transform (STFT) with 1024 bins, 256 hop lengths, and 80 Mel band scales, centered on frequencies between 125 and 7600 Hz. The Mel-spectrograms were further processed by cropping them in the time domain from 80x345 to 80x336, thereby ensuring compatibility with subsequent downscaling operations during the training phase.

2.2.2 VGGish-ish classifier

We utilized the VGGish-ish model, a convolutional neural network (CNN) comprising 13 convolutional layers and three fully connected layers. This model was specifically trained for sound recognition tasks using the VGGSound training dataset. Crucially, this model was employed to compute the perceptual loss of SpecVQGAN. The SpecVQGAN model relies on a pretrained classifier model to extract perceptually-rich features (Iashin and Rahtu, 2021). This design choice ensures that the generated representations from SpecVQGAN are perceptually meaningful and align with the characteristics of human auditory perception.

2.2.3 SpecVQGAN

I adopted a SpecVQGAN to achieve two objectives: to extract compact discrete codebook representations and to ensure the reconstruction of high-quality sound from these representations. SpecVQGAN is a variant of the Vector Quantized Variational Autoencoder (Walker et al., 2021) model, which uses vector quantization techniques to convert latent features into discrete units. This approach is instrumental in bypassing the 'posterior collapse' problem, a challenge often encountered when the model's complexity is high or when there are insufficient constraints in the generative model's latent space. The SpecVQGAN model includes an encoder, which is a standard 2D-Conv stack augmented with self-attention layers that operate on an encoded representation. The decoder, on the other hand, mirrors the architecture of the encoder, by upsampling layer that doubles the spatial resolution previous the convolutional kernel with nearest-neighbor interpolation. Following the default parameters from the referenced paper (Iashin and Rahtu, 2021), a 4-second Mel-spectrogram with the shape

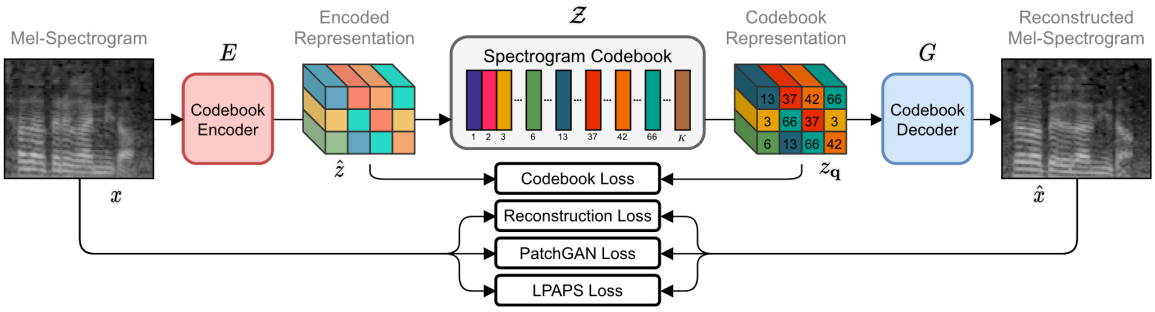

Fig. 2.1 Architecture of SpecVQGAN. A schematic of SpecVQGAN model for training codebook representations from Mel-spectrogram. The SpecVQGAN utilized the Vector Quantization (VQ) operation to extact a compact codebook representation from the Mel-spectrogram through training encoder and decoder. The model was trained to minimize the reconstruction loss as well as reducing the codebook loss, adversarial loss, and perceptual loss (LPAPS). Reprinted from Iashin, V. and Rahtu, E. (2021). Taming visually guided sound generation, licensed under CC BY 4.0.

of (nspectral $\times$ ntemporal = 80 $\times$ 336) produced concise codebook indices in the shape of (nspectral $\times$ ntemporal = 5 $\times$ 21).

2.3 Results

2.3.1 Spectrogram codebook representations

As suggested in the original paper, our application of SpecVQGAN, newly trained with 4-second stimuli, effectively reduces the dimension of Mel-spectrograms while maintaining high fidelity sound upon reconstruction. In the upper panel of Figure 2.4, we compare the reconstructions of various sounds from the VGGSound's test dataset. The results show that not only are detailed spectral patterns superbly reconstructed, but temporal information is also accurately replicated, resulting in sounds highly relevant to the originals.

Specifically, bottom pannel explores the reconstruction of artificial sounds. Even though artificial sounds like pure tones introduced harmonic noise and precisely reconstructed frequency range, the model's ability to accurately reconstruct silence demonstrates the generalizability of the SpecVQGAN model. In spite of the harmonic noise introduced in the reconstruction of artificial sounds like pure tones, the SpecVQGAN model shows its versatility in accurately reconstructing silence. These findings indicate that codebook representations

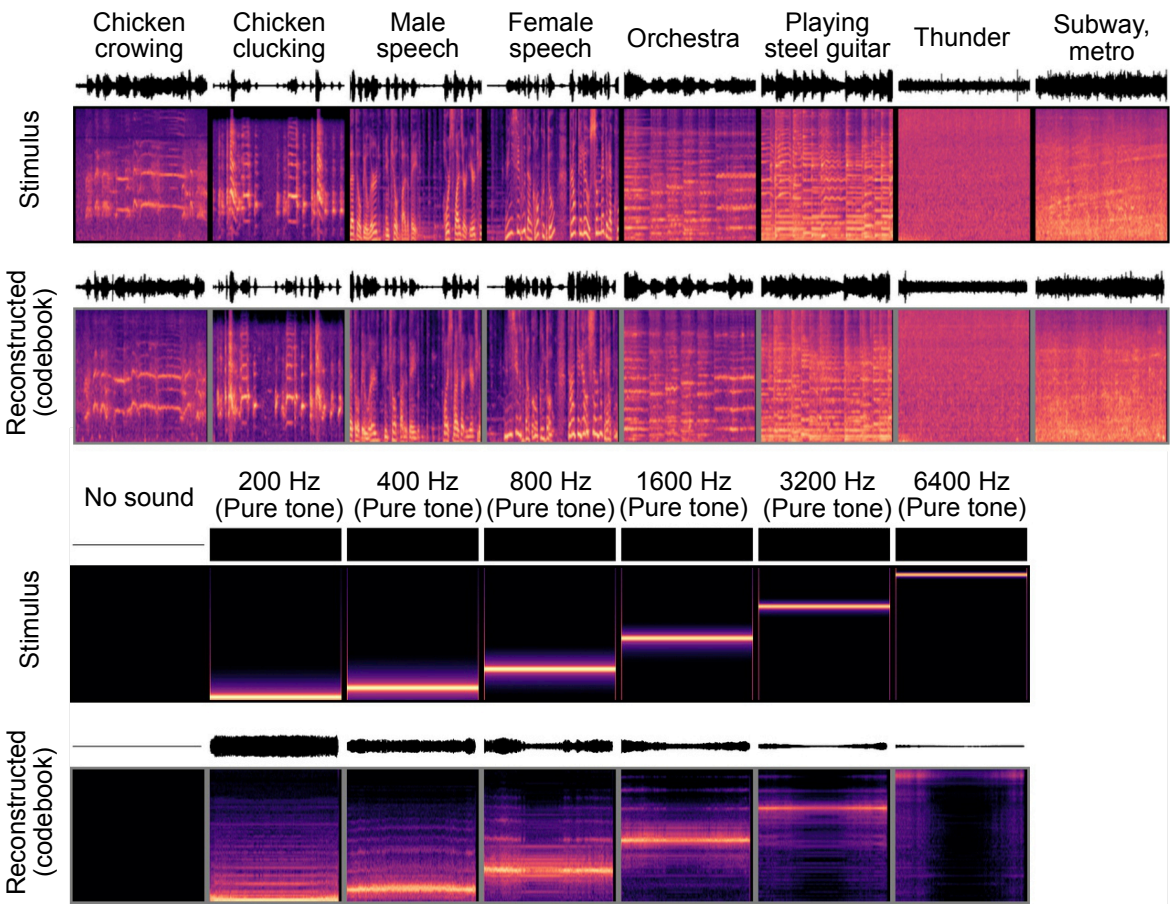

Fig. 2.2 Examples of reconstructed sound using SpecVQGAN. (A) Illustrates reconstructed sound examples obtained from true codebook indices of the VGGSound test dataset. (B) Showcases examples of reconstructed sound derived from artificial sounds.

effectively offer a concise representation of Mel-spectrograms, further underscoring the model's adaptability and effectiveness in sound reconstruction.

I examined the distribution of the codebook indices used in sound generation. Despite setting number of codes (1024) for the training of the SpecVQGAN following the parameters from previous studies, I discovered that only a portion of them (187) are utilized by the trained SpecVQGAN model for the creation of a 4-second sound. This phenomenon, known as index collapse, can occur as the input gets shorter, enabling sound representation with fewer codes and potentially inhibiting the reconstruction of more complex patterns. To investigate this, I examined the histograms of codebook indices using our sound dataset from the experiment.

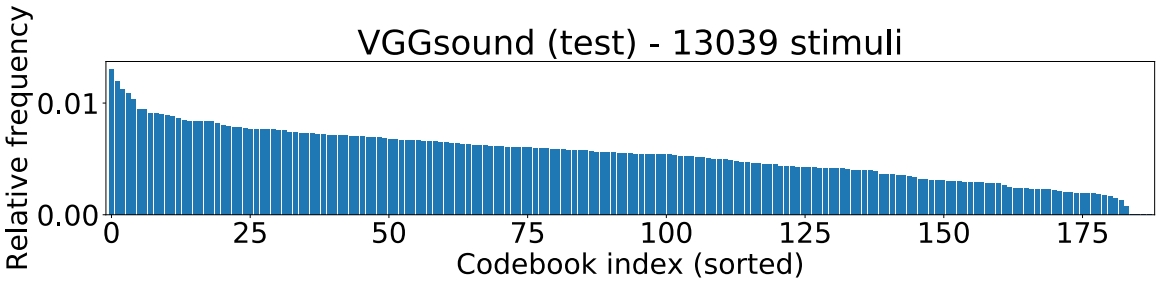

Fig. 2.3 Histogram of the codebook indices used for representing the VGGsound test dataset. Figure depicts a histogram of codebook indices utilized when mapping all the stimuli in the VGGSound test dataset. The codebook indices are sorted based on their relative frequency. The histogram provides insight into the distribution of codebook indices usage, which could help in identifying potential index collapse issues.

Figure 2.3 depicts a histogram of codebook indices used for representing the VGGSound test dataset. Notably, despite the phenomenon of index collapse, the code usage appears to be broadly distributed and relatively balanced. This suggests that the SpecVQGAN model does not over-rely on a specific subset of codes, thus allowing for a diverse range of sound reconstructions. However, given that only 187 out of the 1024 available codes are utilized, future work could explore techniques to enhance the efficiency of code usage, thereby potentially improving the complexity and diversity of the reconstructed sounds.

2.3.2 Interpretation of codebook representations

I utilized the codebook representation as a concise representation of the Mel-spectrogram, which can be transformed back into a Mel-spectrogram. Each codes is calculated from a patch in the Mel-spectrogram and represents a portion of the spectrogram. To explore the patterns that each codes carries within Mel-spectrogram patches, I inputted a 5x21 codebook indices, consisting entirely of the same codebook index, into the codebook decoder 2.4A. Given that I downsample a Mel-spectrogram of size 80x336 into a 5x21 codebook indices, reducing the size by a factor of 16, I examined patterns within patch sizes of 16x16. The resulting Mel-spectrogram was segmented into several patches based on the sampling size. The average of these patches provided a crude pattern, visualized in Figure 2.4B. This figure shows examples of patch patterns for each code in the Mel-spectrogram.

Despite many codes displaying uniform patterns along the temporal dimension, I found that each code possesses diverse spectral and temporal patterns within patches. It is worth noting that in actual sound generation, the patches employed for each codes are larger than

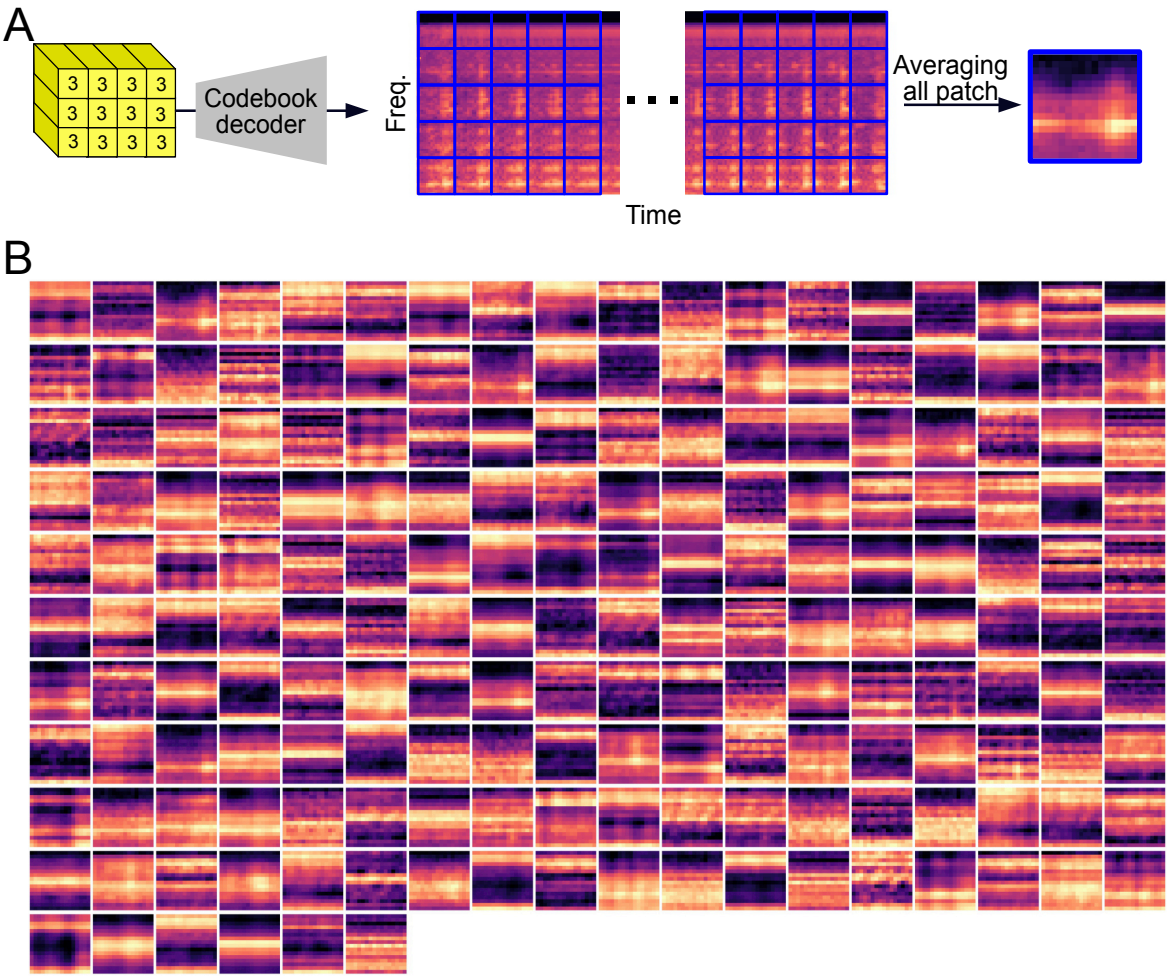

Fig. 2.4 Examples of patch patterns for each code in the Mel-spectrogram. (A) To visualize the patch patterns corresponding to each code within the Mel-spectrogram, I generated a spectrogram using only a single code for sound generation. The resulting spectrogram was then segmented into distinct patches, which were subsequently averaged to match the size of the patch. (B) Examples of patch for each code in Mel-spectrogram by the visualization of its distinct pattern.

the sampling size of 16x16. Each pixel within the Mel-spectrogram is determined by complex operations involving multiple codes.

2.4 Discussion

In this chapter, we utilized SpecVQGAN, an approach outlined in prior studies, which efficiently reduces dimensionality from audio waveforms (Iashin and Rahtu, 2021). This novel technique generates a codebook representation, effectively bridging the chasm between

high-dimensional spectrograms and the limited temporal resolution of neuroimaging data. Not only does the SpecVQGAN perform the task of dimensionality reduction, but it also serves as a perceptually rich prior, significantly aiding in sound reconstruction (Dhariwal et al., 2020; Liu et al., 2021; Zhao et al., 2020). Visualizing patch patterns corresponding to each code highlights the SpecVQGAN's capacity to capture a wide array of spectral and temporal sound characteristics. This diverse repertoire allows our model flexibility in reconstructing a multitude of sounds from concise codebook representations.

However, our trained model is not without potential areas for improvement. The issue of index collapse, wherein a model excessively relies on a small subset of codes, ignoring others and diminishing the diversity of sound generation, presents a challenge. Despite our model's promising capability in reconstructing arbitrary sounds, index collapse could limit the complexity and variety of the produced sounds. Several strategies to counter index collapse could be employed. Using different VQ-VAE architectures or applying methods such as the Exponential Moving Average (EMA) update used in VQ-VAE (Oord et al., 2018) could ensure a more uniform codebook usage, yielding a broader range of sound reconstructions. Additionally, exploration of a larger or dynamic codebook might be beneficial. A dynamic codebook, adaptable to the specific requirements of the sound being reconstructed, could provide more flexibility and effectively address the diverse and intricate nature of sounds.

Chapter 3

Auditory neuroimaging with fMRI

3.1 Introduction

The pursuit of understanding the human brain has been significantly propelled by the advent of neuroimaging techniques. These techniques offer a remarkable opportunity to peer into the human brain and observe how it reacts to different stimuli. These techniques give us a window into the brain, allowing us to visualize how it responds to different stimuli. This chapter delves into the use of neuroimaging techniques in understanding human brain activity, specifically focusing on auditory processing and the neuroimaging experiments of natural sound listening conditions for our reconstruction analysis. Neuroimaging techniques allow us to visualize the intricate patterns of brain activity triggered by diverse sounds, thereby offering novel insights into the complexity of the human auditory system.

Research into auditory processing extensively employs a spectrum of functional neuroimaging techniques, including Electroencephalography (EEG), Electrocorticography (ECoG), Magnetoencephalography (MEG), and functional MRI (fMRI). These techniques grant us dynamic images of brain activity, revealing how the brain responds to sound stimuli. By using these techniques, I can identify specific brain regions engaged in the perception and interpretation of auditory signals. Nevertheless, each technique presents its own set of unique challenges. For instance, while fMRI excels in providing high spatial resolution, it is not as proficient in temporal resolution. Conversely, techniques like EEG offer excellent temporal resolution but fall short in terms of spatial precision. The choice of technique often relies on the specific requirements and aims of the research question.

Traditionally, neuroimaging tools such as EEG, ECoG, and MEG have been favored for decoding auditory information due to their high temporal resolution, which allows for the capture of real-time electrical activity from the scalp or sensors placed on the head. These techniques have delivered impressive results in terms of both quantitative and qualitative recognition. However, the invasive nature of data collection and the subsequent restrictions on dataset sizes have meant that their use has largely been confined to classifying predefined speech (Chakrabarti et al., 2015; Martin et al., 2018; Moses et al., 2019; Pei et al., 2011) and reconstructing limited examples such as digits (Akbari et al., 2019) and words (Wang et al., 2018). The inherent limitations of fMRI in temporal resolution have, for the most part, restricted its usage to classification approaches (Correia et al., 2015; Formisano et al., 2008).

Interestingly, recent studies have started to challenge this conventional practice, suggesting that it might be possible to reconstruct unrestricted sound without requiring an exact alignment in temporal resolution between neural recordings and auditory stimuli. One method to address this involves leveraging the spatial patterns in fMRI to compensate for its limited temporal resolution, which in turn allows for the prediction of detailed temporal information. In this regard, Santoro et al. (2017) developed a computational model to decode the physical features of natural sounds using high spatial resolution 7T fMRI responses. This model used multiple multivariate decoders to predict spectral-temporal modulation features from fMRI activation patterns. Impressively, these trained decoders were capable of predicting subtle modulation changes from the relatively coarse temporal sampling of fMRI (2.6 seconds). To make the decoded results easier to interpret, they were converted back into sounds. Despite these promising results, the reconstructed sounds lacked complex spectro-temporal patterns, which resulted in temporally smoothed reconstructions and posed challenges for human listeners.

In the upcoming section, I will delve into the principles of fMRI to understand the brain's response. I will focus on how these responses are captured and interpreted using fMRI. Following this, I will introduce MRI experiments conducted under natural sound listening conditions that were utilized for our reconstruction analysis. The contents of this chapter is based on the section Materials and methods: *Subjects, stimuli, MRI acquisition, MRI data preprocessing, and Regions of interest* of (Park et al., 2023).

3.2 Basics of fMRI

In this section, I delve into the principles and technology behind functional Magnetic Resonance Imaging (fMRI), a non-invasive neuroimaging technique that has proven to be invaluable in cognitive neuroscience. Magnetic Resonance Imaging (MRI) came into the medical imaging scene in the early 1970s, providing a new approach to visualize the body's internal structures. Nearly two decades later, fMRI was introduced, adding a new dimension to the capabilities of MRI. While traditional MRI offers detailed visualizations of the anatomical structures within the brain, fMRI extends this by identifying areas of brain activation linked with specific cognitive or sensory tasks. By combining high-resolution anatomical data with functional information, fMRI has proven to be a potent tool, bridging the gap between neurophysiology and cognitive neuroscience.

The fundamental workings of fMRI rely on the interactions between biological tissues and magnetic fields. In particular, it uses the Blood Oxygen Level Dependent (BOLD) contrast. The BOLD signal leverages the magnetic properties of deoxyhemoglobin, a substance with paramagnetic properties. This feature allows researchers to infer neural activity from observed changes in blood flow and oxygenation (Ogawa et al., 1990). When a region in the brain is active, it triggers a localized response which involves increased blood flow and oxygen consumption. These changes affect the BOLD signal, which can then be detected by the fMRI scanner. The entire process involves subjecting the individual to strong magnetic fields generated by a superconducting magnet. These fields polarize the subject, causing hydrogen protons to align along the direction of the magnetic field. Short bursts of radiofrequency (RF) pulses disturb this alignment, and as the protons relax back to their equilibrium state, they emit signals that can be detected. These signals are transformed into digital data, which are then processed to create images of the brain. Underpinning the BOLD signal is the hemodynamic response, which is a complex interplay between neural activity, metabolism, blood flow, volume, and oxygenation. As neurons fire, the demand for oxygen increases, leading to local vasodilation, increased blood flow, and a decrease in deoxyhemoglobin concentration. This sequence of events leads to a local increase in the BOLD signal, providing an indirect measure of neural activation.

3.3 Experimental settings

In my experiments involving natural sound listening, I collected fMRI responses while sound stimuli were played through fMRI-compatible headphones (Kiyohara, KAS-3000HK). These headphones ensured a sound pressure level (SPL) within the range of roughly 68-75 dB. The stimuli were continuously delivered, adopting a method that enabled simultaneous presentation of the stimuli and recording of the fMRI responses, without the need for intervening silent periods. Before the scanning process, subjects were allowed to adjust the sound level to their comfort.

3.3.1 Subjects

Our study involved five non-native English-speaking subjects with normal hearing abilities, including one female participant. The average age of the subjects was 27.6 years. One subject (S1) was utilized for exploratory analysis to establish the reconstruction model, while the remaining four subjects served to independently validate the results. Prior to the scanning sessions, all subjects provided their informed consent. The study protocol was approved by the Ethics Committee of the Advanced Telecommunications Research Institute International (approval no: 106) and adhered to the principles outlined in the Declaration of Helsinki.

3.3.2 Sound stimuli

The natural sound fMRI experiments employed a total of 1,250 audio clips, sourced from the VGGsound test dataset (Chen et al., 2020). All stimuli underwent meticulous assessment by human listeners for sound quality. For the subject S1, each audio clip was of 10-second duration, however, only 8 seconds of data was used in the experiment to maintain consistency with other subjects. The training dataset consisted of 1,200 indiscriminately selected audio clips, disregarding category label information in an effort to emulate natural auditory scenes. Each audio clip in the training set potentially encompassed multiple sound categories. In the pilot study, 162 stimuli that created categorization difficulties were replaced with fresh stimuli in the subsequent experiments involving the four subjects. The test dataset was carefully curated, including four representative sound categories as per criteria defined in earlier studies (Norman-Haignere et al., 2015): Human speech (including English), animal sounds, musical instruments, and environmental sounds. Categories unsuitable for longer sound stimuli, such as non-speech vocalizations, were excluded. This led to a compilation

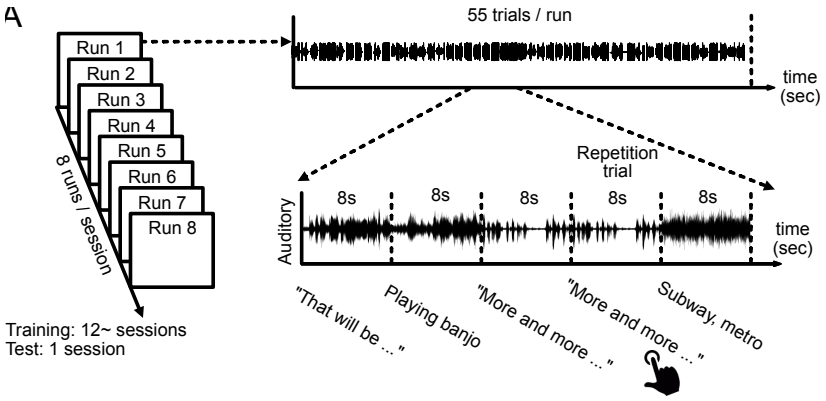

Fig. 3.1 Schematic of the experimental design for the natural sound listening condition. Participants were continuously presented with 8-second sound stimuli without any intervening silent periods. To ensure participants' attention was uniformly distributed across all stimuli, they were instructed to perform a one-back repetition detection task. Participants were required to press a button whenever they detected consecutive trials featuring the same sound.

of 50 audio clips for the test dataset, each distinctly representing a single sound category. All audio clips included in the fMRI dataset were resampled to a frequency of 22050 Hz, center-cropped to an 8-s duration (10-s for S1 but only 8-s of data was used for the analysis), and normalized to ensure equivalent energy levels.

3.3.3 Experimental design

In the natural sound presentation experiment, subjects passively listened to a variety of audio clips of natural sounds. I recorded whole-brain fMRI responses while subjects listened to 1,200 stimuli designated for training and 50 for the test dataset. Each subject underwent a series of scanning sessions spread over approximately three months, with 12-16 training sessions followed by a separate single test session. Every session included 4-8 functional runs, each not exceeding 90 minutes. Each run started with a rest period of 30-s, followed by 55 stimulus presentation blocks of 8-s each (comprising 50 unique sound stimuli and five randomly interspersed behavioral task blocks), and ended with a 10-s rest period. This sequence resulted in a total run duration of 8 minutes. To maintain subject concentration, I incorporated a one-back repetition detection task, where subjects were required to press a button if the subsequent stimulus presented was identical to the previous one. These repetition blocks (five per run) were not included in the analysis. Experiments for training sets were repeated four times and experiments for test sets were repeated eight times.

3.3.4 MRI acquisition

Functional MRI data were collected using a 3.0-Tesla Siemens MAGNETOM Verio scanner at the Kyoto University Institute for the Future of Human Society. An interleaved T2*-weighted gradient-echo echo planar imaging (EPI) sequence was used to collect functional images covering the entire brain (TR = 2000 ms, TE = 44.8 ms, flip angle = 70 degrees, FOV=192 × 192 mm, voxel size=2 × 2 × 2 mm, slice gap = 0 mm, number of slices = 76, multiband factor = 4). Additionally, T1-weighted magnetization-prepared rapid acquisition gradient-echo (MP-RAGE) images of the entire head were obtained to provide high-resolution structural information (TR = 2250 ms, TE = 3.06 ms, TI = 900 ms, flip angle = 9 degrees, FOV = 256 × 256 mm, voxel size=1.0 × 1.0 × 1.0 mm, number of slices = 208).

3.3.5 MRI data preprocessing

The preprocessing of the MRI data was carried out using the pipeline provided by fMRIPrep (Esteban et al., 2019). First, a BOLD reference image was first generated from acquired functional data of eac run using fMRIPrep. The next step was motion correction using mcflirt from FSL (Jenkinson et al., 2012). After motion correction, slice time correction was applied to the data using 3dTshift from AFNI (Cox, 1996). The data were then co-registered to the corresponding T1-weighted image using the boundary-based registration approach implemented by bbregister from FreeSurfer (Fischl, 2012). Finally, the co-registered BOLD time-series were resampled onto their original space using antsApplyTransforms from ANTs (Avants et al., 2008), utilizing Lanczos interpolation for this process.

I adjusted the preprocessed functional data by shifting them forward by 2 seconds to account for the hemodynamic delay. To augment the number of available data samples, I slid a 4-second time window across the original 8-second stimulus at 2-second intervals. For each 4-second sound stimulus, an fMRI sample was created by averaging the three consecutive functional volumes post the stimulus onset (Figure 3.2A). This procedure resulted in three data samples from each original 8-second trial. As a result, a total of 14,400 training samples were obtained (1,200 stimuli × 4 repetitions × 3 samples = 14,400 samples. For the test datasets, I enhanced the Signal-to-Noise Ratio (SNR) by averaging the fMRI responses to identical sound stimuli across multiple repetitions. This approach yielded a total of 150 test samples (50 stimuli × 3 samples = 150 samples) (Figure 3.2B).

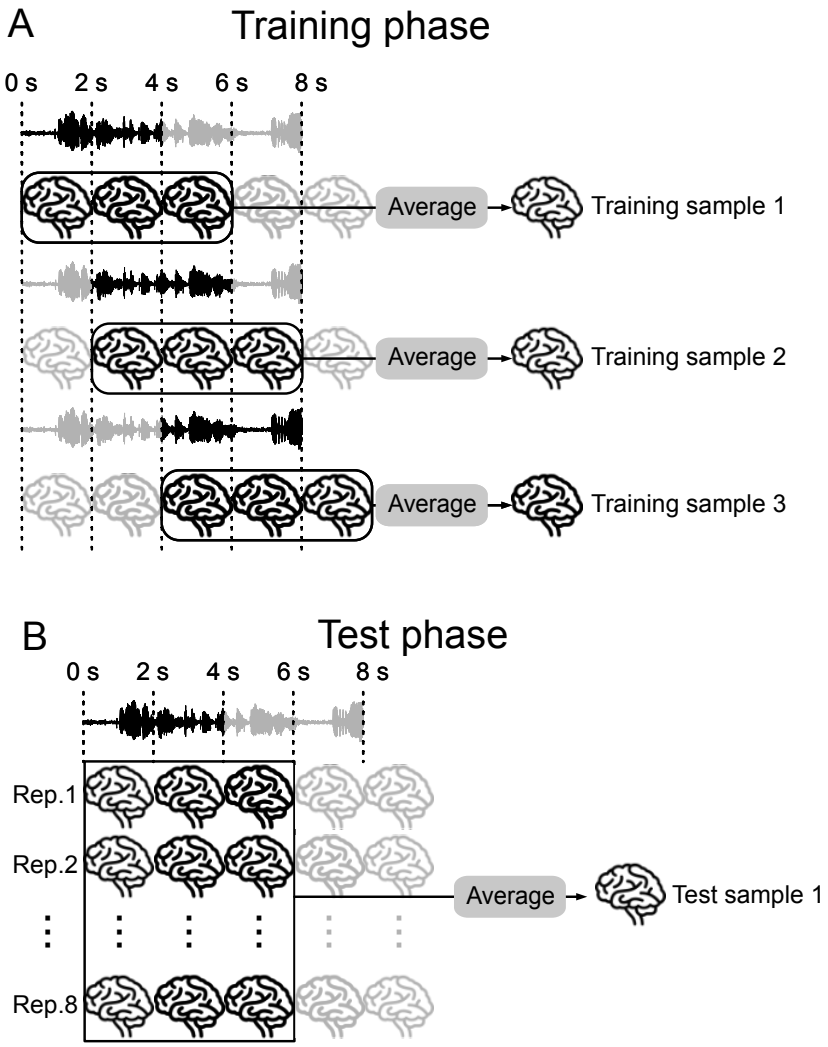

Fig. 3.2 Schematic of calculating fMRI samples from preprocessed data. (A) This panel represents the calculation of fMRI samples for the training set. In the fMRI experiments, each 8-second stimulus was divided into three samples, each with a 4-second time window. The training samples were computed by averaging the three consecutive functional volumes for each 4-second stimulus onset, after adjusting for a 2-second volume shift to account for hemodynamic delay. Since I conducted four fMRI experiments for training, I obtained four training samples for each 4-second stimulus. (B) This panel illustrates the calculation of fMRI samples for the test set. Similar to the training set, I extracted a 4-second stimulus from the 8-second stimuli. To enhance the HNR in the test set, I computed the test samples by averaging all the samples calculated from eight repeated fMRI experiments. Therefore, I obtained a single test sample for each 4-second stimulus.

3.3.6 Region of interest (ROI)

I utilized a multi-modal cortical parcellation developed by the Human Connectome Project (HCP) (Glasser et al., 2016) to delineate Regions of Interest (ROI) within the auditory cortex. I identified thirteen anatomical ROIs in both hemispheres, spanning the early auditory cortex (which included A1, LBelt, MBelt, PBelt, RI), and the auditory association cortex (which included A4, A5, TA2, STGa, STSd anterior, STSv anterior, STSd posterior, and STSv posterior). The combined set of voxels from both the early auditory cortex and the auditory association cortex was collectively referred to as the Auditory Cortex (AC).

3.4 Statistics

All statistical tests were carried out individually, with each subject's results treated as within-subject replications of an experiment (Ince et al., 2022). A 95% confidence interval was used to determine if the mean identification accuracy of the reconstructed sounds across test stimuli exceeded the chance level of 50%. The sample size for the natural sound test experiment (N = 50) was predetermined before the experiment was conducted. This is greater than the sample size required to detect an effect size of Cohen's d = 0.5 at a significance level of 0.05 (N = 27). Although data samples from a 4-s time window were used for decoder training and reconstruction, statistical evaluations were carried out on data points corresponding to 8-s stimulus blocks. This approach was adopted to handle the lack of independence among the three samples derived from an 8-s stimulus block. For the single sound test sample analyses, the identification accuracies of the three samples were averaged to define a single data point for statistical analysis, resulting in 50 data points for each condition and subject.

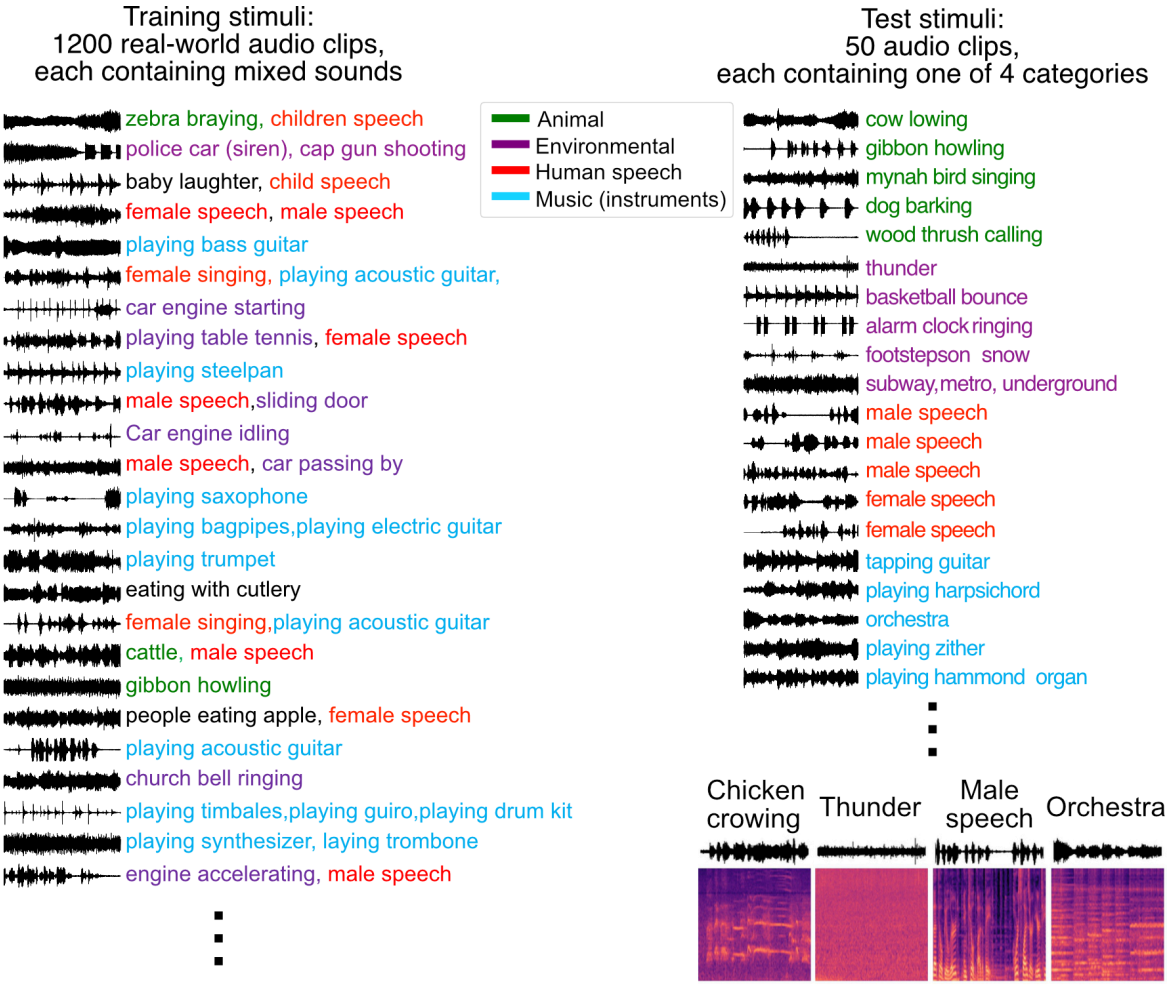

Fig. 3.3 Sound stimuli for fMRI experiments consisted of a diverse range of real-world audio clips. The training dataset comprised of 1,200 clips (left panel), each containing a mixture of different sound categories that represent a wide array of natural auditory scenes. On the other hand, the test dataset consisted of 50 audio clips (right panel), each of which contained only one specific sound category. These test categories included human speech, animal sounds, musical instruments, and environmental sounds. This categorization is in line with the divisions used in previous auditory perception studies.

Chapter 4

Brain decoding of auditory features

4.1 Introduction

The intersection of progress in neuroimaging and machine learning has given rise to the concept of 'brain decoding.' Brain decoding utilizes machine learning algorithms to interpret or decode information derived from observed patterns of brain activity. The focus of these decodings can be on a variety of auditory features, determined by the design of the experiment or the hypothesis under investigation.

A spectrogram is one of the commonly used auditory features, because it mirrors the processes carried out in the cochlea. The spectrogram delivers a representation of the spectrum of frequencies present in a sound as they change over time. Created by mapping the frequency content of the signal across time, the spectrogram offers a time-frequency analysis of the sound. Spectrograms find extensive use in auditory neuroscience, as the cochlea, the auditory segment of the inner ear, performs a kind of real-time spectrogram analysis on incoming sounds. The cochlea filters sound into diverse frequency bands and encodes the intensity of each frequency, effectively crafting a 'live spectrogram' that is then forwarded to the brain for further processing. Consequently, the spectrogram serves as a model of the primary stage of auditory processing in the brain and has found widespread use in Electroencephalography (EEG) and other neuroimaging studies (O'Sullivan et al., 2017; Pasley et al., 2012).

In addition, modulation features serve as hand-engineered models of the auditory system, capturing changes in sound over time and frequency. These features reflect the dynamic nature of auditory perception, where the information carried by a sound varies not only across frequency bands but also changes over time. In the human auditory system, various stages of the auditory pathway are known to process and encode these spectro-temporal modulations (Chi et al., 2005). For instance, the auditory cortex has neurons that respond selectively to specific spectro-temporal modulations, thereby acting as a spectro-temporal filter. This suggests that these modulations are vital to our understanding and interpretation of complex sounds. Previous fMRI-based decoding studies have employed modulation features as brain modeling features (Santoro et al., 2017),

Furthermore, Deep Neural Network (DNN) feature spaces can act as proxies for hierarchical representation. DNN features refer to the abstracted representations learned by DNNs trained to perform various tasks related to sound, such as speech recognition, sound classification, and music genre classification. These features are the activation patterns of the neurons in the network, with each layer in the network encoding different levels of abstraction of the original input sound. At the lower levels, these features might correspond to basic auditory elements, such as spectral and temporal modulations, akin to the early processing stages in the human auditory system. As I move up the network layers, the features become more abstract and task-specific, encoding higher-level perceptual qualities of the sounds, such as phonetic or semantic information. These can be akin to the later stages of auditory processing in the brain, where the initially processed sound signals are interpreted and made sense of. What makes DNN features particularly interesting for auditory neuroscience is their potential to provide a model of hierarchical sound processing in the brain. Just as the brain processes sounds in a hierarchical manner, from simple to complex, so does a DNN, learning to represent sounds at various levels of abstraction. Recent studies have suggested that there might be parallels between the hierarchies in DNNs and those in the brain's auditory system. For instance, a study by Kell et al. (2018) found that the responses of the early auditory cortex in the brain were best predicted by the features from the lower layers of the DNN, while the responses of non-primary auditory regions were best predicted by features from the higher layers of the DNN. This suggests a certain degree of "brain-likeness" in the DNN features, providing further support to the idea of using DNN features for decoding brain activity related to sound perception.

In this chapter, I will elucidate the fMRI-based decoding approach, with a special focus on the decoding of auditory features. I employed mel-spectrogram features. Then, to

capture temporal and spectral changes in sound, I employed spectro-temporal modulation features. Finally, I resorted to the DNN feature space to encapsulate the hierarchical auditory processing. To map brain activity into each of these auditory features, I adopted a feature decoding method proposed by Horikawa and Kamitani (2017a). Post this, I made use of a variety of evaluation metrics to compare decoding performances, and the auditory feature that exhibited the highest decoding performance was identified as the 'brain-like' feature. This feature was then used in subsequent reconstruction analyses. The contents of this chapter is based on the section Materials and methods: *Feature decoding analysis* and Results: *Brain decoding of auditory features* of (Park et al., 2023).

4.2 Methods

4.2.1 Data processing

In this study, I utilized the VGGsound dataset (Chen et al., 2020), a publicly available audio-visual database encompassing 200,000 YouTube video clips, annotated reliably across 309 categories. To maintain the integrity of the data, I discarded any videos with missing or invalid links. I adhered strictly to the data split criteria from prior research that used the VGGsound dataset for training the Deep Neural Network (DNN) model. The audio clips used in my fMRI experiments for natural sounds came from the test set, which was not used in the training of the DNN model to be utilized in the subsequent sound reconstruction chapter. In total, the fMRI experiments incorporated 1,250 audio clips, each lasting 8 seconds. Every stimulus was rigorously evaluated for sound quality by human listeners. During the data preprocessing stage, the audio clips were resampled at a frequency of 22050 Hz. To augment the number of available data samples, I applied a 4-second time window to the original 8-second stimulus and moved it at 2-second intervals. This strategy created three data samples from each original 8-second trial, yielding a grand total of 14,400 training samples (1,200 stimuli × 4 repetitions × 3 samples). For the test datasets, the Signal-to-Noise Ratio (SNR) was enhanced by averaging the fMRI responses to identical sound stimuli across multiple repetitions. As a result, I obtained a total of 150 test samples (50 stimuli × 3 samples) and 144 attention samples (48 stimuli × 3 samples).

4.2.2 Auditory features

Mel-spectrogram

In order to generate the Mel-spectrograms from our audio stimuli, I first resampled the audio clips at a frequency of 22050 Hz. I then used a short-term Fourier transform (STFT) with 1024 bins and a hop length of 256 to produce log-Mel-spectrograms, utilizing 80 Mel band scales that were centered on frequencies ranging from 125 Hz to 7600 Hz. Following this, I carried out additional processing on the generated Mel-spectrograms. This involved center-cropping in the time domain to transform the shape from ($n_{spectral} \times n_{temporal} = 80 \times 445$) to ($n_{spectral} \times n_{temporal} = 80 \times 336$). This adjustment was necessary to ensure compatibility with the subsequent downscaling processes during the training phase. Consequently, each stimulus resulted in a Mel-spectrogram with a shape of ($n_{spectral} \times n_{temporal} = 80 \times 336$).

Spectro-temporal modulation features

As another crucial auditory feature for our decoding analysis, I calculated spectrotemporal modulation features. The method I employed closely followed the steps outlined in the work of Santoro et al. (2017). Initially, I generated audio spectrograms by using a bank of 128 overlapping bandpass filters that were evenly spaced along a logarithmic frequency axis. The output from this filter bank was subjected to several processing stages, which included bandpass filtering, frequency axis differentiation, half-wave rectification, and short-term temporal integration.

Following this, I computed the modulation content of the auditory spectrogram. This computation was done using a bank of 2D modulation-selective filters and performing a complex wavelet decomposition. Consequently, I achieved a representation composed of $\times$ 20 temporal modulation frequencies $\times$ 40 time bins $\times$ 128 frequencies, totaling 614,400 features.

It is worth noting that while the audio spectrogram used for modulation calculation is a time-frequency representation akin to the Mel-spectrogram, it is not identical to it. I made the decision to use the audio spectrogram rather than the Mel-spectrogram to facilitate a more direct comparison with earlier fMRI-based reconstructions.

DNN features

The third type of auditory feature I utilized were DNN features derived from the Mel-spectrogram. To acquire these, I passed the previously computed Mel-spectrograms through a pretrained convolutional neural network (CNN) model known as VGGishish. This model, consisting of 13 convolution layers and three fully connected layers, had been specifically trained for sound recognition tasks using the VGGsound training dataset. I focused on the unit responses from the highest convolution layer, conv5_3, in the VGGishish model, treating these as DNN features. Thus, the Mel-spectrograms were transformed into a set of DNN features with dimensions of ($n_{channels} \times n_{spectral} \times n_{temporal}$). I then reshaped these features to the format ($n_{channels*spectral} \times n_{temporal}$), maintaining the temporal dimension, in order to utilize them as conditioning input in the audio transformer model. Out of all the layers within the VGGishish model, I singled out six representative ones, one from each convolutional and fully-connected layer block, that demonstrated superior decoding performance. As a result, the DNN features I obtained had a shape of ($n_{channels} \times n_{spectral} \times n_{temporal} = 512 \times 5 \times 21$).

Latent features from SpecVQGAN encoder

Additionally, another type of DNN feature I utilized for our analysis was the codebook representation. This type of representation was computed using SpecVQGAN, a variant of an autoencoder, which was specifically trained to produce concise representations, distinct from the sound recognition model, VGGishish. The architecture of SpecVQGAN comprises an encoder and a decoder. The encoder is a standard 2D-Convolutional stack with added self-attention layers that operate on the encoded representation. Conversely, the decoder mirrors the encoder's architecture, except for the presence of an upsampling layer. This upsampling layer doubles the spatial resolution before the convolutional kernel with nearest-neighbor interpolation.

By adhering to the default parameters from the referenced paper, I converted the 4-second Mel-spectrogram in the shape of ($n_{spectral} \times n_{temporal} = 80 \times 336$) into a set of concise codebook indices in the shape of ($n_{spectral} \times n_{temporal} \times n_{dimensionofcodebook} = 5 \times 21$ x 256)). These latent representations calculated from SpecVQGAN encoder, providing a compact and efficient representation of the original Mel-spectrogram, were then used as auditory features in the subsequent stages of our analysis.

4.2.3 Feature decoding analysis

I established a brain decoding framework, inspired by previous works (Horikawa and Kamitani, 2017a, 2022), that enables prediction of auditory features from multi-voxel fMRI responses. In our study, I utilized 14,400 samples from the training dataset for each combination of auditory features and brain areas.

The training phase started with voxel selection, wherein I chose the voxels based on their correlation coefficient with the target features. I selected 500 voxels from the auditory cortex (AC) and 200 from each individual region of interest (ROI). The responses from these selected voxels were normalized using the mean and standard deviation calculated from the training samples. Subsequently, I applied z-score normalization to the stimulus feature values, which utilized the mean and standard deviations derived from the training data. An L2-regularized linear regression model was then used to predict the normalized feature values from the multi-voxel patterns of the normalized fMRI responses.

During the testing phase, each fMRI sample from the test dataset was first normalized using the mean and standard deviation derived from the training dataset. The trained decoders were then applied to these normalized samples to predict the auditory features from 150 fMRI samples. Post-prediction, the decoded features underwent denormalization using the mean and standard deviation of each feature from the training dataset. I acknowledged potential discrepancies between the distribution of actual and decoded features and thus incorporated a posthoc normalization process. This step involved normalization of the decoded feature values by the square root of the number of repetitions. During this process, I ensured that the mean of the posthoc normalized decoded features remained consistent with that of the decoded features calculated from the brain decoder.

To evaluate decoding performance, I employed two metrics: 1) The Pearson correlation coefficient between the actual and decoded auditory features across the test stimuli in each pixel or unit (Figure 4.1), and 2) An identification analysis that assessed the ability of the decoded auditory features to identify the actual stimuli from a set of candidate stimuli.

Pair-wise identification analysis

I examined the identification performance of the decoded features to assess their potential to discern perceived sounds from a set of all test sounds. Specifically, the identification accuracy was determined by evaluating the correlation coefficients between the decoded features and the actual auditory features across all test stimuli. This comparison involved contrasting the correlation between the decoded features and each candidate stimulus with the correlation between the decoded features and the actual stimuli that were presented during the test phase. If the correlation between the decoded and actual stimulus was higher than the correlations between the decoded features and all other candidate stimuli, the identification was considered correct. The cumulative identification accuracy for each decoded feature was then computed as the ratio of the number of correctly identified pairs to the total number of stimuli presented. This provided a metric for how well the decoded features could accurately represent and identify the original auditory stimuli.

4.3 Results

4.3.1 Feature decoding performance

I employed L2-regularized linear regression to predict auditory features from the responses of thirteen anatomically defined ROIs within the early auditory cortex and the auditory association cortex, as delineated from the Human Connectome Project (HCP) Glasser et al. (2016). Our primary focus were A1, LBelt, and PBelt in the early auditory cortex, and A4 and A5 in the auditory association cortex, which follow the ventral pathway. Furthermore, I incorporated the responses from all thirteen ROIs within early auditory and auditory association cortices to delineate an auditory cortex (AC) (Figure 4.1A). I trained the decoder to predict auditory features such as 1) pixel values of the Mel-spectrogram, 2) spectro-temporal modulation features, and 3) DNN features from the sound recognition model, VGGish-ish, using training fMRI samples of natural sound (Figure 4.1B). Post training, I employed the brain decoder to predict the decoded feature values from the fMRI responses in the test dataset.

To assess the decoding performance, I computed the correlation coefficients between the actual and decoded auditory features (Figure 4.1C). Positive correlations were found across all combinations of feature types and ROIs from all subjects, including our pilot study subject

S1 (Figure 4.1D). Delving deeper, most auditory ROIs demonstrated a correlation higher than 0.5 for the Mel-spectrogram and DNN features, and above 0.4 for modulation features. Intriguingly, the decoding performance for the Mel-spectrogram and modulation features gradually decreased as I moved from the A1 to the nonprimary cortex. In contrast, the DNN features retained correlations above 0.5 up to A5, without a substantial drop in decoding performance. This suggests that while spectro-temporal features like spectrograms and modulation features are effectively decoded from the early auditory cortex, their performance begins to wane as I move toward peripheral regions. Conversely, DNN features, while displaying decoding performance similar to spectro-temporal features in the early auditory cortex, contain higher-level information via hierarchical processing akin to the human auditory system. This feature enables DNN features to be decoded even in the nonprimary auditory cortex.

To further assess the capability of our decoded features, I measured their ability to identify specific sounds from all test sounds. I computed identification accuracy by comparing correlation coefficients between decoded and actual auditory features across all test stimuli. This involved measuring the correlation between decoded features and each candidate stimulus, then contrasting it with the correlation between decoded features and the actual stimuli presented. I quantified the identification accuracy of each decoded feature by the number of correctly identified pairs.

As shown in Figure 4.2, the identification accuracy using decoded Mel-spectrograms slightly exceeded chance levels across all auditory ROIs for all subjects. Intriguingly, all subjects could accurately identify sounds using decoded modulation features and decoded DNN features, both in the AC and individual ROIs. Remarkably, DNN features consistently showed superior performance compared to other auditory features across all ROIs, achieving over 80% identification accuracy for each subject. This suggests that the DNN features, derived from a hierarchical sound recognition model, outshine traditional Mel-spectrogram or modulation features in predictive performance, earning them the title of "most brain-like" features due to their impressive decoding capabilities.

4.3.2 Hierarchical correspondence between brain and DNN model

To investigate accordance with previous encoding analysis that utilized hierarchical auditory areas and features, I conducted a comparative assessment of the decoding performance between the individual layers of the DNN and auditory regions. As illustrated in Figure 4.3,

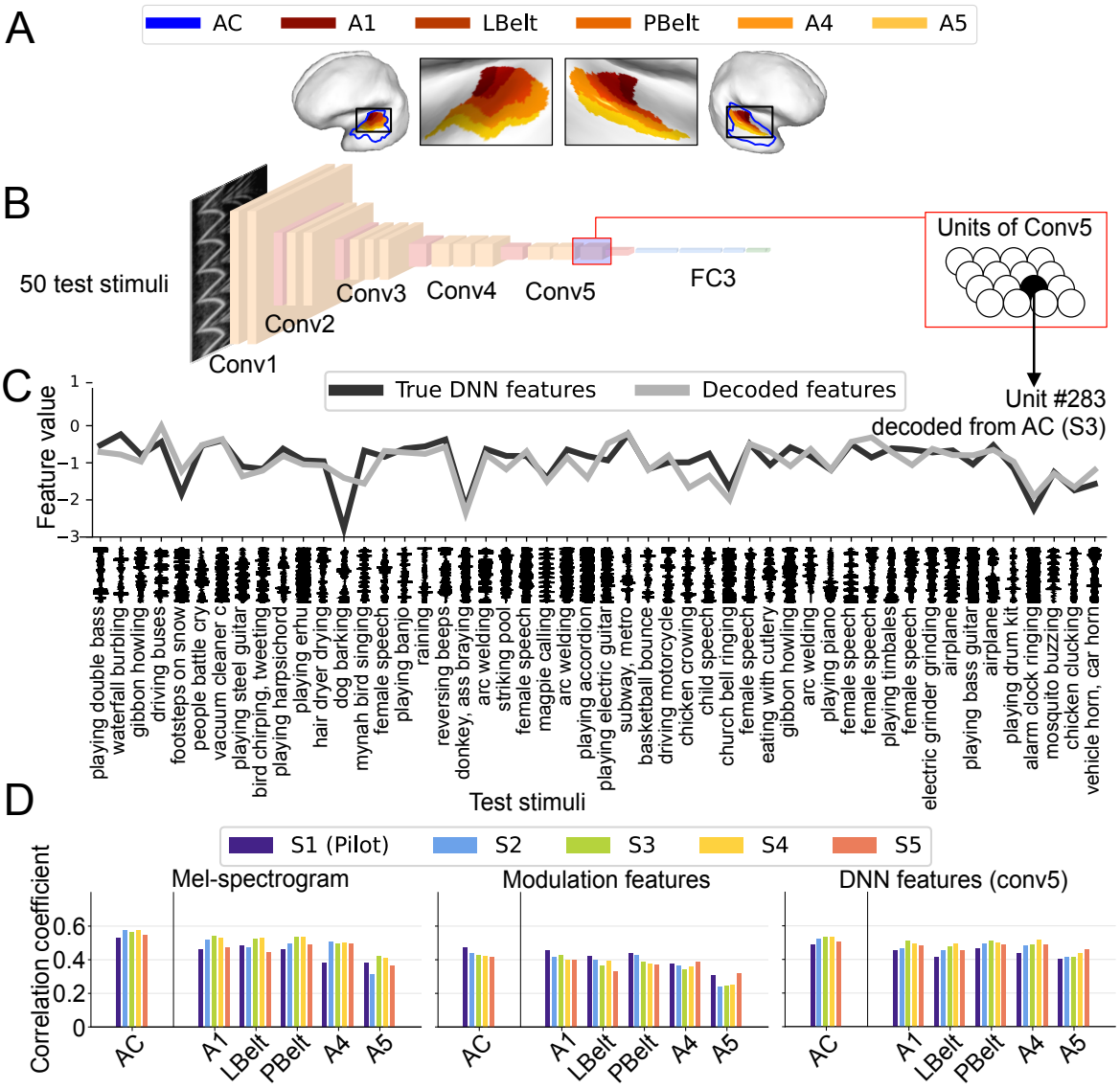

Fig. 4.1 Feature decoding analysis. (A) Region of Interest (ROI) selection. The auditory cortex (AC) is defined as a combination of the early auditory cortex and the auditory association cortex, following the classification system of the Human Connectome Project (HCP). I focused on A1, LBelt, and PBelt regions from the early auditory cortex, and A4 and A5 regions from the auditory association cortex, aligning with the ventral pathway. (B) Structure of the Sound Recognition Model. The general-purpose sound recognition model, VGGish-ish, is used in our analysis. The features derived from its highest convolutional layer (conv5) served as our Deep Neural Network (DNN) features. (C) Comparison between true and decoded DNN features. This illustration shows a comparison of the original DNN features and the features decoded from the AC for a set of 50 test sound stimuli. These features originated from the conv5 layer of the VGGish-ish model. (D) Evaluation of decoding performance for different auditory features. This bar chart displays the decoding performance of three auditory feature types: Mel-spectrogram, modulation features, and DNN features. Each bar corresponds to the average decoding accuracy of a subject, calculated across all feature units for each type.

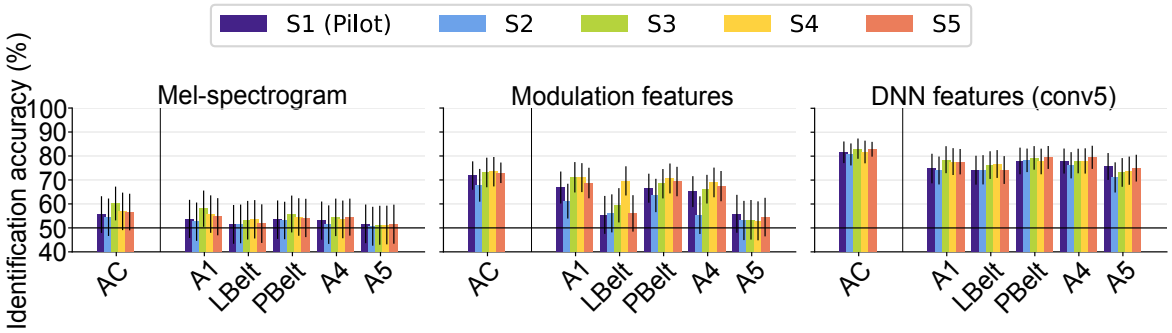

Fig. 4.2 Comparison of identification accuracy across three auditory feature types. Due to overlap among the samples, the identification accuracies from the 150 test samples were averaged and transformed into 50 separate data points. This was achieved by considering the results from the three fMRI samples corresponding to each stimulus. Each bar in the graph stands for the mean identification accuracy drawn from these 50 data points, and error bars denote the 95% confidence interval (CI). Different colors are used to distinguish individual subjects.

the early auditory cortical areas, such as A1, consistently demonstrated superior decoding accuracy across most DNN layers compared to other auditory regions. On the other hand, areas within the auditory association cortex, such as A4 and A5, showed slightly subdued performance for lower DNN layers compared to A1. However, their performance became on par with A1 for higher DNN layers. This pattern suggests that the different auditory cortical regions engage in a more distributed form of processing, rather than embodying a strict hierarchical structure that mirrors the architecture of the sound model.

In particular, I further examined the decoding performance of conv5 among five regions in the early auditory cortex (EAC: A1, LBelt, MBelt, PBelt, RI) and eight regions in the auditory association cortex (AAC: A4, A5, STSdp, STSda, STSvp, STSva, STGa, TA2) as defined in the HCP parcellation, considering both the left and right hemispheres (Figure 4.4). The results revealed that there wasn't a significant discrepancy between the hemispheres. The five regions of EAC exhibited a profile correlation of around 0.4, while in the AAC, the regions A4, A5, and TA2 demonstrated similar decoding performance around 0.4. However, in the STS and STG regions, performance was below 0.3, with some variation observed between participants.

Additionally, when I compared the decoding performance of DNN features trained on tasks other than sound recognition (Figure 4.5), the results showed lower decoding performance of the latent features from SpecVQGAN in not only the AC but also other

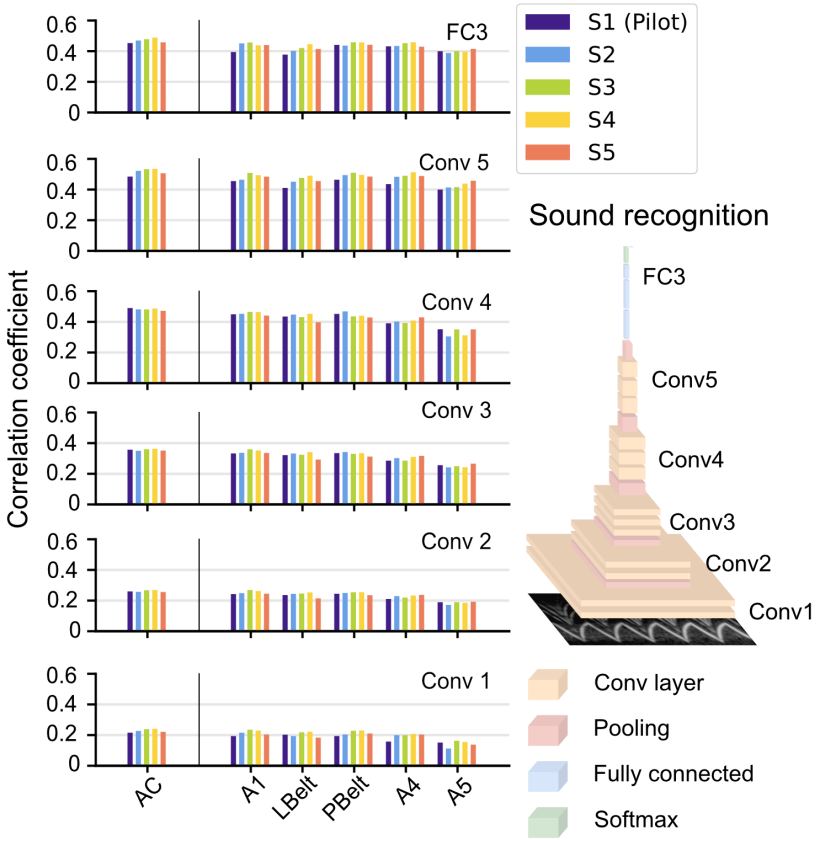

Fig. 4.3 Decoding performance of different layers in the VGGish-ish model. The bars represent the decoding accuracy of DNN features across six representative layers of the sound recognition model for each individual subject. Accuracy is calculated as the average across all units within each respective layer.

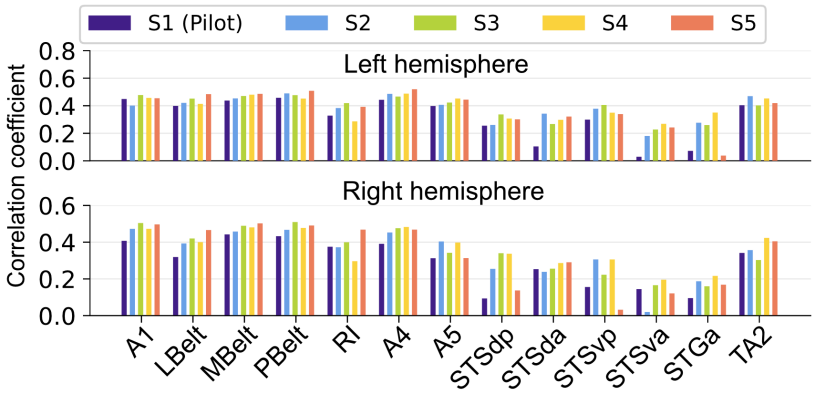

Fig. 4.4 Decoding accuracy of DNN features for individual ROI. This figure shows the decoding performance of DNN features derived from the conv5 layer of the VGGish-ish model across various regions of interest (ROIs) within the brain. The upper panel displays the decoding accuracy for individual ROIs in the left hemisphere, while the lower panel depicts the same for the right hemisphere. Each bar within the chart represents the average decoding accuracy for each subject, which was calculated across all units.

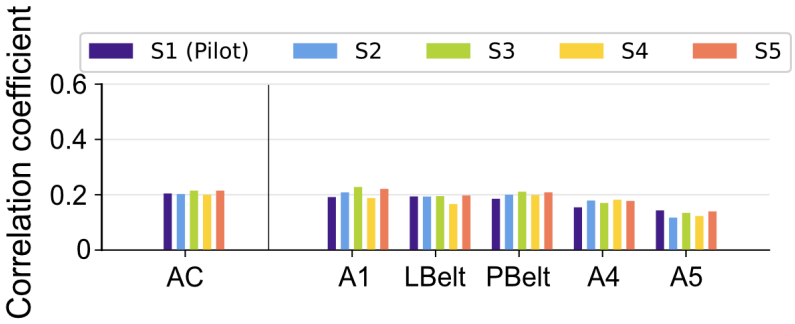

Fig. 4.5 Decoding accuracy of latent features from SpecVQGAN. This figure presents the decoding performance of latent features from SpecVQGAN from various brain regions. Each bar within the graph represents the average decoding accuracy for each individual subject, which was computed by averaging the results across all units.

regions, with decoding performance falling below 0.2. This performance was inferior to that of the DNN features calculated from the sound recognition model VGGish-ish.

4.4 Discussion

In this chapter, I performed a comprehensive investigation into the decoding of various auditory features from fMRI data, leveraging advancements in machine learning and neuroimaging. Our analysis employed several auditory features, including Mel-spectrograms,

modulation features, and Deep Neural Network (DNN) features. By utilizing a feature decoding method proposed by Horikawa and Kamitani (2017a), I were able to map brain activity to each of these auditory features.

Our feature decoding analysis highlighted the superior predictive performance of features originating from the hierarchical sound recognition model, as compared to Mel-spectrogram or modulation features. Consequently, the DNN features were identified as the most "brain-like," owing to their enhanced decoding capabilities. Significantly, features from sound recognition DNNs, which emulate the hierarchical processing inherent in the human auditory system, consistently outperformed other auditory features in terms of decoding performance. These findings align with prior encoding analyses illustrating systematic model-brain correspondence (Kell et al., 2018).

However, our decoding performance did not exhibit a clear hierarchical correlation between individual auditory ROIs and the layers within the DNN model, which stands in contrast to what prior encoding analyses with DNN have suggested (Kell et al., 2018; Li et al., 2022). Recent studies utilizing intracranial recordings propose a distributed functional organization within the human auditory cortex, suggesting the potential for parallel information processing across the auditory cortex (Hamilton et al., 2021; Nourski et al., 2014). These studies imply that auditory cortical ROIs participate in both distributed and hierarchical processing. In our decoding analysis, I utilized anatomically defined ROIs. Future studies, however, could benefit from employing voxels defined by tonotopic or encoding analysis, potentially offering deeper insights into the auditory hierarchy and its representations.

Especially when comparing the decoding performance of VGGish-ish and latent features from SpecVQGAN, I observed that even DNN models with the same hierarchical processing structure yielded varying results. Specifically, DNN features trained for sound recognition tasks outperformed the latent features from SpecVQGAN, which was trained to compute concise representations of Mel-spectrograms. This outcome is consistent with previous studies that observed not all DNNs exhibit homology with the human brain; their encoding performance can vary substantially depending on the DNN's structure and the tasks they're optimized for (Tuckute et al., 2023). Our results, which align with these findings from decoding analysis, further suggest that the similarity between DNN features and the brain can be task-dependent.

I noted a peculiar discrepancy in the case of Mel-spectrograms. Despite the consistently high correlation coefficients of individual pixels across various stimuli, the identification accuracy between the actual and predicted Mel-spectrograms was surprisingly low. This discrepancy could be explained by the tendency of decoded Mel-spectrograms to capture common variations across pixels, rather than accurately decoding each pixel's actual values. This observation is in line with prior studies using direct regression for spectrogram features from neuroimaging techniques, where decoded Mel-spectrograms were reported to be dominated by an indistinguishable broadband component, yielding a smoothed pattern across pixels (Défossez et al., 2022).

Chapter 5

Sound reconstruction from brain activity

5.1 Introduction

Sound reconstruction from brain activity is an emerging field that seeks to convert our brain's neural signals into audible sounds. This field essentially enables us to "hear" auditory experiences such as melodies, and unspoken words that reside solely in our brain signal, converting them into sounds others can perceive audibly.

Despite the enormous potential, this endeavor of reconstructing sound from brain activity presents significant challenges due to the complex temporal sequences inherent in sounds and the limited resolution of neuroimaging techniques. Traditional neuroimaging methods, such as electroencephalography (EEG) and magnetoencephalography (MEG), offer superior temporal resolution but can only record real-time electrical activity from scalp sensors or head-mounted devices. As a result, their application has largely been restricted to classifying predefined speech and reconstructing limited samples, like digits and select words (Akbari et al., 2019; Martin et al., 2018; Moses et al., 2019; Pei et al., 2011; Wang et al., 2018). Furthermore, functional Magnetic Resonance Imaging (fMRI), despite its inherent temporal resolution constraints, has mainly been utilized for classification approaches (Correia et al., 2015; Formisano et al., 2008).

Nonetheless, an encouraging solution to these challenges could lie in harnessing the hierarchical nature of auditory brain processing. Recent research has highlighted the parallels between the hierarchical structure of the human auditory system and deep neural network

(DNN) models (Kell et al., 2018). Additionally, advances in audio-generative models now allow compact representations to be converted back into high-resolution sounds (Iashin and Rahtu, 2021).

In this chapter, I introduce a new method for sound reconstruction that combines the decoding of auditory features from fMRI responses with an audio-generative model. Using fMRI responses to natural sounds, I discovered that the hierarchical sound features of a DNN model were decoded with greater accuracy than spectrotemporal features. Subsequently, I used an audio transformer to disentangle the compact temporal information in the decoded DNN features, enabling sound reconstruction. The proposed method has proven capable of reconstructing arbitrary sounds, capturing both the perceptual content and quality of these sounds. These findings suggest that our model offers a means of expressing and articulating auditory experiences derived from human brain activity. The contents of this chapter is based on the section Materials and methods: *Model components and reconstruction methods, and model components* and Results: *Sound reconstruction and model components* of (Park et al., 2023).

5.2 Methods

5.2.1 DNN model

For the sound reconstruction from fMRI response, I incorporated multiple DNN models that were initially established in prior studies by Iashin and Rahtu (Iashin and Rahtu, 2021). The pre-trained models and associated scripts are accessible at https://iashin.ai/SpecVQGAN. I specifically utilized the pre-trained models for VGGish-ish, Melception classifier, and the spectrogram vocoder due to their ability to function irrespective of sound length. In alignment with our fMRI experiments, I also trained models for the Spectrogram Vector Quantized Generative Adversarial Network (SpecVQGAN) and the audio transformer to generate 4-s sound segments.

VGGish-ish classifier

The VGGish-ish model, a convolutional neural network (CNN) composed of 13 convolution layers and three fully connected layers, was specifically trained for sound recognition tasks

using the VGGsound training dataset. This model was utilized to extract DNN features from the generated Mel-spectrograms. The unit responses from each layer within the Mel-spectrograms were calculated as DNN features, with dimensions represented as ($n_{spectral}$ $\times$ $n_{temporal}$ $\times$ $n_{channels}$). These extracted DNN features were then reshaped to the format ($n_{spectralchannel}$ $\times$ $n_{temporal}$) while retaining the temporal dimension, thus preparing them to serve as the conditioning input for the audio transformer model. I identified six layers that exhibited outstanding decoding performance within each convolutional and fully-connected layer block. These are: conv1_1 with dimensions of (5120 $\times$ 336), conv2_1 with dimensions of (5120 $\times$ 168), conv3_1 with dimensions of (5120 $\times$ 84), conv4_1 with dimensions of (5120 $\times$ 42), conv5_3 with dimensions of (2560 $\times$ 21), and fc3 with dimensions of (309). These representative layers were selected for further processing.

Audio transformer

I employed a transformer model to translate the compact temporal information present within the DNN features into codebook representations. Leveraging the success seen in autoregressive generative applications (Esser et al., 2021; Iashin and Rahtu, 2021; Vaswani et al., 2017), I trained an audio transformer to predict codebook representations for every temporal point across each spectral direction in an autoregressive manner. I utilized a GPT-2-medium, consisting of 24 layers, 1024 hidden units, and 16 attention heads. The input to this model comprised DNN features shaped as (nspectral $\times$ channels $\times$ ntemporal). The transformer processed the DNN features at each temporal point, translating them into a probability distribution for the subsequent codebook index via a 1024-way softmax classifier. The training objective of the transformer was to minimize the cross-entropy loss between the predicted and actual codebook representations. Ultimately, the audio-transformer translates the sequence of DNN features into a sequence of codebook representations, with dimensions of (nspectral $\times$ ntemporal = 5 $\times$ 21).

Spectrogram vocoder

I utilized MelGAN (Kumar et al., 2019), a fully convolutional feed-forward model that thakes a Mel-spectrograms as an input and produces a waveform as an output. The model employs four upsampling layer with a residual block to ultimately upscale the time-series information from the Mel-spectrogram by a factor of 256 to create a waveform. Differing from the

Griffin-Lim procedure (Griffin and Lim, 1984), MelGAN synthesizes audio waveforms in a non-autoregressive fashion. This technique enables audio reconstruction that is not only faster but also of higher fidelity.

Sound reconstruction from fMRI responses

Our sound reconstruction model, which synergizes the trained model components with the brain decoder, operates in a sequential manner. In chapter 4, we trained a brain decoder to predict the auditory features of the presented sound stimuli from fMRI response patterns (Figure 5.1A). These auditory features encompassed the pixels of a Mel-spectrogram, the modulation features, and the features extracted from a sound recognition DNN model. Reinforcing previous findings, we recognized the hierarchical features from the sound recognition DNN model as the most "brain-like", given their superior decoding performance from brain activity compared to other auditory features. Subsequently, we trained an audio-generative transformer, a sequence-to-sequence model, to predict the codebook representation (a concise depiction of a Mel-spectrogram) conditioned on the DNN features in an autoregressive fashion (Figure 5.1B). During the testing phase, we obtained decoded DNN features from fMRI responses using the feature decoders and then transformed them into codebook representations using the audio-generative transformer (Figure 5.1C). This process starts with the brain decoder, which decodes DNN features from fMRI responses in the test dataset. The decoded DNN features are subsequently transformed into codebook representations with the aid of the audio transformer. These codebook representations are then converted back into Mel-spectrograms using a codebook decoder. Finally, a spectrogram vocoder transforms these spectrograms into audio waveforms.

5.2.2 Evaluation of fidelity

To assess the accuracy and quality of our sound reconstructions, I implemented a pairwise identification analysis. This procedure involved extracting auditory features from both the original and the reconstructed sounds, then assessing the ability of the reconstructed sounds to correctly identify the original stimulus from a pair consisting of the true stimulus and each of the remaining test stimuli. I calculated the correlation coefficient between the auditory features of the reconstructed sounds and those of a pair of candidate stimuli: one of the candidates was the actual stimulus presented, while the other was one of the other test stimuli

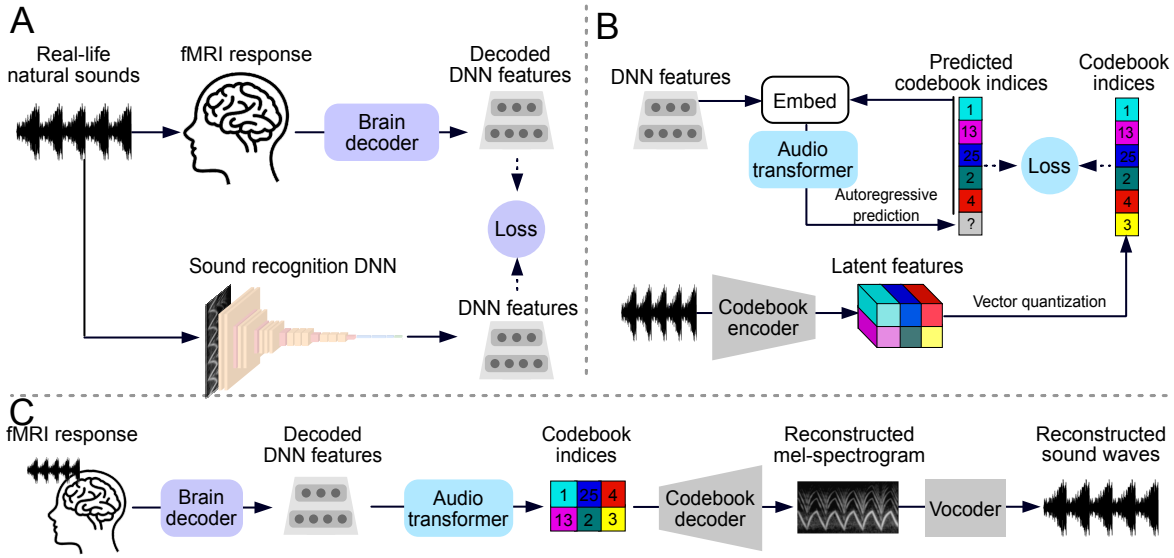

Fig. 5.1 Schematic overview of the sound reconstruction model from fMRI responses. (A) Training the brain decoder. Subjects are presented with real-world natural sounds during fMRI scans to record brain activity. Concurrently, identical sounds are processed through a sound recognition model to extract DNN features. This combined data is then utilized to train a brain decoder with the aim of predicting the corresponding auditory features from the fMRI signals. (B) Training the audio transformer. A codebook encoder is employed to generate codebook indices, which offer a succinct representation of the Mel-spectrogram. Following this, an audio transformer is trained to predict sequences of these codebook indices conditioned on the DNN features, employing an autoregressive approach. (C) Sound reconstruction from fMRI responses. The reconstruction process begins by computing decoded DNN features from the fMRI responses using the trained feature decoders. These decoded DNN features are subsequently transformed into codebook indices via the audio transformer. In the final stages, the codebook decoder and spectrogram vocoder convert these codebook indices into Mel-spectrograms, and eventually, into audible sound waves.

(out of 149 stimuli). I executed pairwise identification for all 149 pairs, defining identification accuracy as the proportion of instances where the presented stimulus had a higher correlation coefficient.

Pixels of Mel-spectrogram

The fidelity of the reconstructions was assessed using both Mel-spectrogram pixels and hierarchical representations. Initially, I evaluated fidelity using the pixels from the Mel-spectrogram, which are considered low-level or raw features.

Hierarchical representation

After evaluating the raw-level features, I used a pre-trained Melception classifier (Iashin and Rahtu, 2021) to analyze the fidelity of hierarchical representations. The Melception classifier, which is an audio classifier developed specifically for sound recognition tasks, was used to extract DNN features from both the reconstructed sounds and the original stimuli. These DNN features served as stand-ins for hierarchical sound representations. For this process, I selected six layers from the Melception classifier that demonstrated superior performance in sound classification tasks. These layers were conv1, conv5, mix5_d, mix6_d, mix7_c, and fc1.

5.2.3 Evaluation of quality

Despite basing proposed reconstructions on higher-level DNN features, it was vital to ascertain that reconstructed sounds encapsulated the perceptual and qualitative aspects of the original sounds. To verify this, I evaluated the reconstructed sounds using three acoustic metrics: Fundamental frequency (F0) - this measures perceived pitch and tonality, helping to differentiate between various sounds and voices. Spectral Centroid (SC) - this metric quantifies the 'brightness' of a sound, with a higher spectral centroid typically indicating a sound that is brighter or sharper. Harmonics-to-Noise Ratio (HNR) - this helps distinguish between tonal sounds and noise-like sounds (Alías et al., 2016).

Fundamental frequency

The fundamental frequency, often referred to as F0, is a crucial aspect of sound perception. Essentially, it denotes the lowest frequency of a periodic waveform and signifies the perceived pitch of the sound. For instance, in speech, the fundamental frequency can denote the intonation of a spoken sentence, its stress pattern, and even the mood of the speaker. For this study, the YIN algorithm (Mauch and Dixon, 2014) was employed to compute the F0 for both the original and reconstructed sounds using the Librosa toolbox (https://librosa.org). The F0 values were averaged across the time series, generating a representative F0 for each sound. The identification accuracy of reconstructed sounds was then evaluated through a comparison of these representative F0 values. However, there were circumstances where F0 could not be calculated from each segment or the entire stimulus of both the original and reconstructed sounds, especially in the case of non-harmonic sounds without a distinguishable pitch. In such scenarios, these stimuli were omitted from the analysis.

Spectral centroid

The spectral centroid (SC) is another essential feature of audio signals. It represents the center of mass of the spectrum and is used to quantify the brightness or sharpness of a sound. Typically, a higher spectral centroid value indicates a brighter or sharper sound. To evaluate how well our model preserved the spectral content of the original sounds, I computed the spectral centroid for both the original and reconstructed sounds using the Librosa toolbox (https://librosa.org). In this process, I chose the median of the SC across the time series, thereby generating a single representative SC value for each sound. The assessment of the reconstructed sounds' accuracy was then carried out by comparing these representative SC values.

Harmonics-to-Noise Ratio

The Harmonics-to-Noise Ratio (HNR) is an important metric that is used to distinguish between tonal or harmonic sounds and aperiodic or noise-like sounds. Sounds that have a high HNR are more tonal in nature, whereas those with a low HNR are more noise-like. In this study, I computed the HNR for both the original and reconstructed sounds using a Python package (https://github.com/brookemosby/Speech_Analysis). The HNR

was computed in a similar fashion to the F0; I averaged the HNR across the time series to generate a representative HNR for each sound. However, there were circumstances where it wasn't feasible to calculate the HNR for each segment or the entire stimulus, both for the original and reconstructed sounds. This was particularly the case when the sound contained a non-harmonic structure that did not have a discernible pitch. In such situations, these stimuli were not included in the analysis.

5.2.4 Comparison with other auditory features

Through our brain decoding analysis, I found that among various auditory features, the DNN features showed the highest decoding performance. Therefore, I utilized these for our reconstruction analysis. In order to provide a comparative perspective with previous research, I juxtaposed the sounds reconstructed by our proposed model with those derived from other reconstruction methodologies that use different auditory features.

Reconstruction from decoded Mel-spectrogram

For the Mel-spectrogram features, I trained the brain decoder for predicting the pixels of the Mel-spectrogram based on fMRI responses, with dimensions of ($n_{spectral} \times n_{temporal}$ = 80 $\times$ 336). In the test phase, decoded Mel-spectrograms were subsequently transformed into audible sound waves with the aid of a pre-trained spectrogram vocoder.

Reconstruction from spectro-temporal modulation features

For the spectro-temporal modulation features, I trained the brain decoder to predict each feature of the modulation representations, which have a dimension of ($n_{spectral} \times n_{temporal} \times n_{spectralmodulation} \times n_{temporalmodulation}$ = 60 $\times$ 10 $\times$ 6 $\times$ 10). To reconstruct sounds from the decoded modulation features, I followed the methodology outlined in a previous study by (Santoro et al., 2017). This reconstruction process involves two key steps: transforming from the modulation domain to the spectrogram, and subsequently from the spectrogram to the waveform. The model estimates the missing phase information part of the complex-valued modulation representation, and an iterative procedure is utilized to reconstruct spectrogram and waveform subsequently. The reconstruction process, from the modulation domain to the

spectrogram and then to the waveform, was implemented using the "NSL Tools" package in MATLAB, available at www.isr.umd.edu/Labs/NSL/Software.htm.

5.2.5 Comparison with other reconstruction methods

Given that our model generates various intermediate representations, it raised questions about which components significantly influence the reconstructed sounds to the reconstruction process. I utilized an ablation study to inspect how each component influences the reconstruction process. The method, brain-to-codebook reconstruction, avoids utilizing brain-like features. Instead, it directly predicts the codebook representations derived from brain responses. On the other hand, pixel optimization reconstruction bypasses the audio transformer, which is typically employed to unravel temporal information in decoded brain-like features. Instead, this method directly optimizes the Mel-spectrogram derived from decoded DNN features.

Brain-to-codebook reconstructon

Our reconstruction process began with training a decoder to directly predict latent features of SpecVQGAN encoder from fMRI responses, followed by computing decoded latent features for the test dataset. The next stage entailed transforming these decoded representations into quantized codebook representations. This was achieved by identifying the nearest representations in the pre-trained codebook dictionary. These quantized codebook representations were subsequently converted into Mel-spectrograms using the codebook decoder. In the final stage, these Mel-spectrograms were converted into audio waveforms utilizing the spectrogram model. This end-to-end process allowed us to capture the complex spectral-temporal patterns from fMRI responses and recreate them into high-quality audio signals.

Pixel optimization reconstruction

I employed an image feature-based optimization technique for our study (Shen et al., 2019b). This method optimized pixel values in 2D Mel-spectrogram images using a VGGish-ish mode. The reconstruction algorithm commenced with an initially noisy image, and iteratively optimized pixel values to align the DNN features extracted from the VGGish-ish model with those decoded from brain activity, across all DNN layers. Importantly, I encountered a challenge previously noted in the referenced paper: the absence of loss convergence when

attempting reconstruction using only features from a single layer, in this case, the conv5_3 layer of our proposed model. To address this, I optimized the loss across all features from the layers of the VGGish-ish model, compared to the decoded features, thereby circumventing the issue of non-convergence. All other parameters were kept at their default values, ensuring the consistency of our approach.

5.3 Results

In this section, I present the results of sound reconstruction from fMRI responses, along with an evaluation of the fidelity and quality of the reconstructed sounds. Our feature decoding analysis identified "brain-like" features extracted through a DNN model. This model mirrors the hierarchical structure of the auditory system, and these DNN features encapsulate compressed temporal information. To convert these features back into their original high-dimensional sound form, it's necessary to disentangle the compressed temporal information within the DNN features. For this task, I utilized a transformer model, known for its exceptional performance in sequential processing. I trained this audio transformer to convert the sequence of DNN features into a sequence of codebook representations in an autoregressive manner. Using the trained transformer, I transformed the decoded DNN features from fMRI responses into codebook representations. These codebook representations were then converted into Mel-spectrograms using a codebook decoder, and finally into audio waveforms with the aid of a spectrogram vocoder.

5.3.1 Reconstructed sounds

Following figures present examples of Mel-spectrograms reconstructed using proposed model. The figure demonstrates our capability to generate auditory experiences from fMRI responses. The reconstructed spectral and temporal patterns bear a striking resemblance to the original stimuli, suggesting that our model can effectively recreate auditory experiences from fMRI data. Remarkably, these reconstructed sounds exhibit consistent quality across different participants, underscoring the robust reproducibility of our approach.

Upon closer examination of the results, distinctive characteristics emerge across different categories. For instance, Figure 5.2 represents the animal category, revealing unique spectral patterns that make the reconstructed sounds easily recognizable compared to the original stimuli. The three fMRI samples computed from the same stimuli exhibit significant similarity.

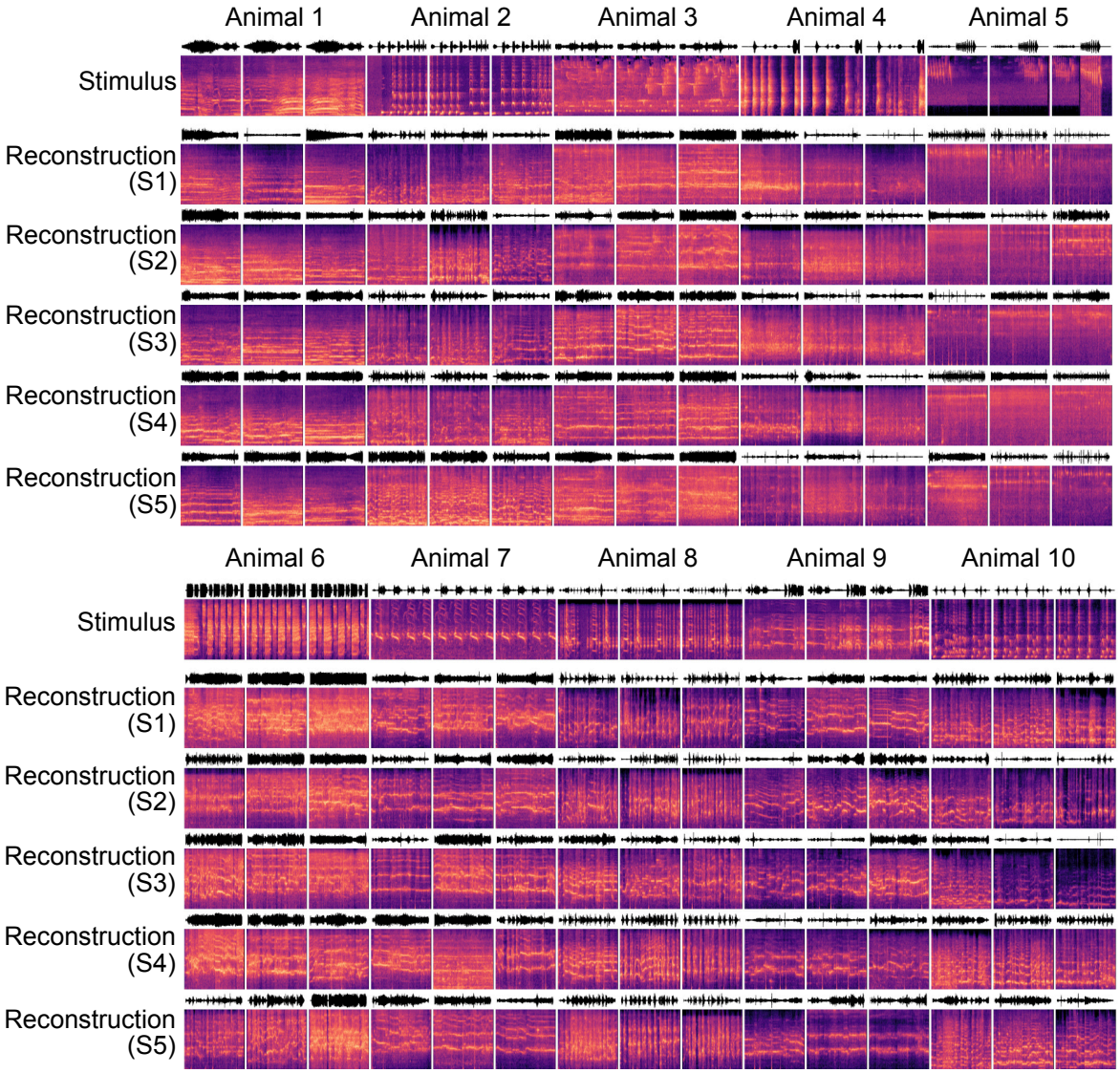

Fig. 5.2 Reconstructed Mel-spectrogram of 'Animal' category. The top row displays the original Mel-spectrogram of the presented sound. The following five rows show the Mel-spectrograms reconstructed from each subject using the AC and the conv5 layer. The three consecutive samples originate from a single stimulus.

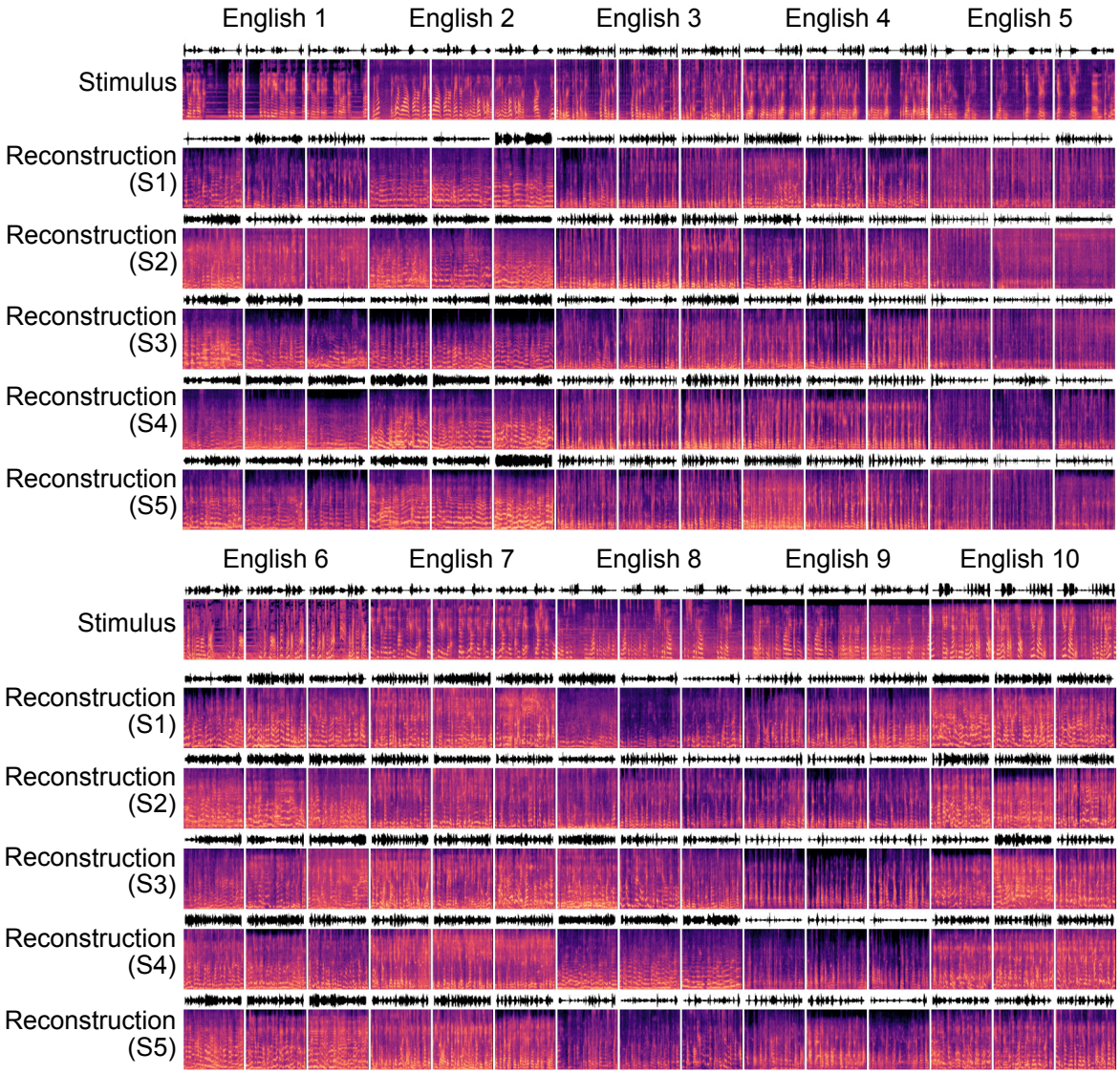

Fig. 5.3 Reconstructed Mel-spectrogram of 'Speech (English)' category. The top row displays the original Mel-spectrogram of the presented sound. The following five rows show the Mel-spectrograms reconstructed from each subject using the AC and the conv5 layer. The three consecutive samples originate from a single stimulus.

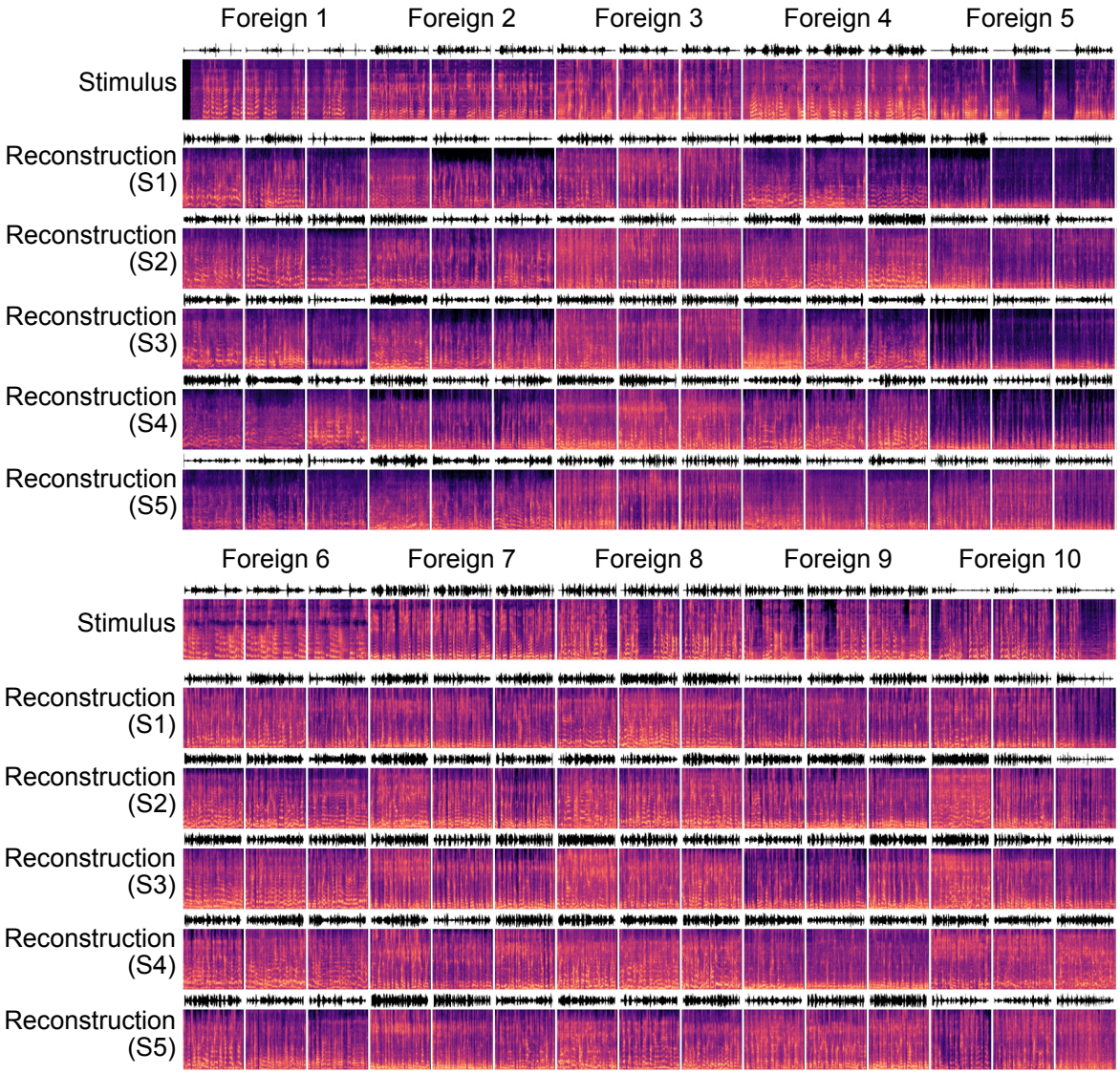

Fig. 5.4 Reconstructed Mel-spectrogram of 'Speech' category. The top row displays the original Mel-spectrogram of the presented sound. The following five rows show the Mel-spectrograms reconstructed from each subject using the AC and the conv5 layer. The three consecutive samples originate from a single stimulus.

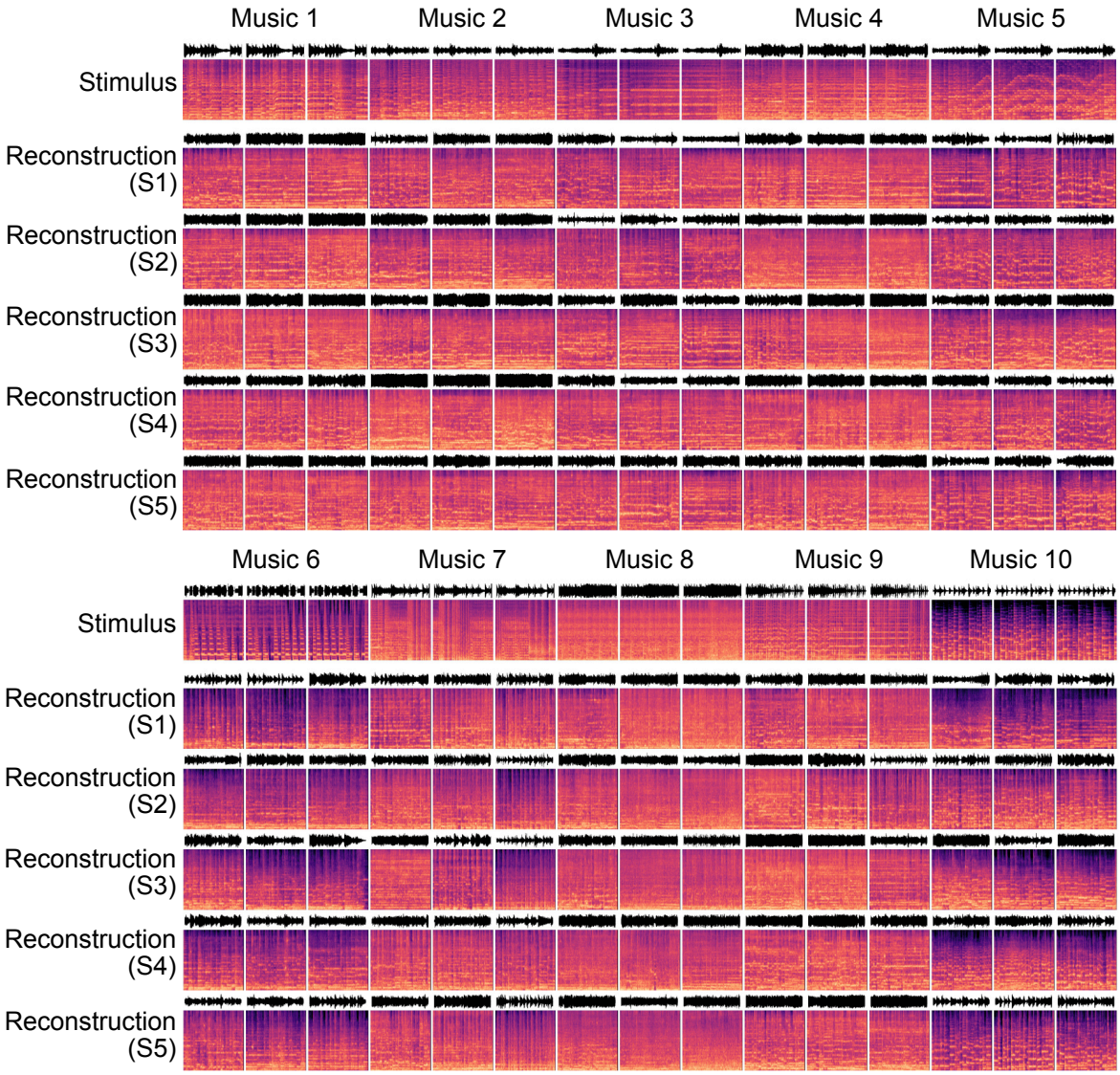

Fig. 5.5 Reconstructed Mel-spectrogram of 'Music' category. The top row displays the original Mel-spectrogram of the presented sound. The following five rows show the Mel-spectrograms reconstructed from each subject using the AC and the conv5 layer. The three consecutive samples originate from a single stimulus.

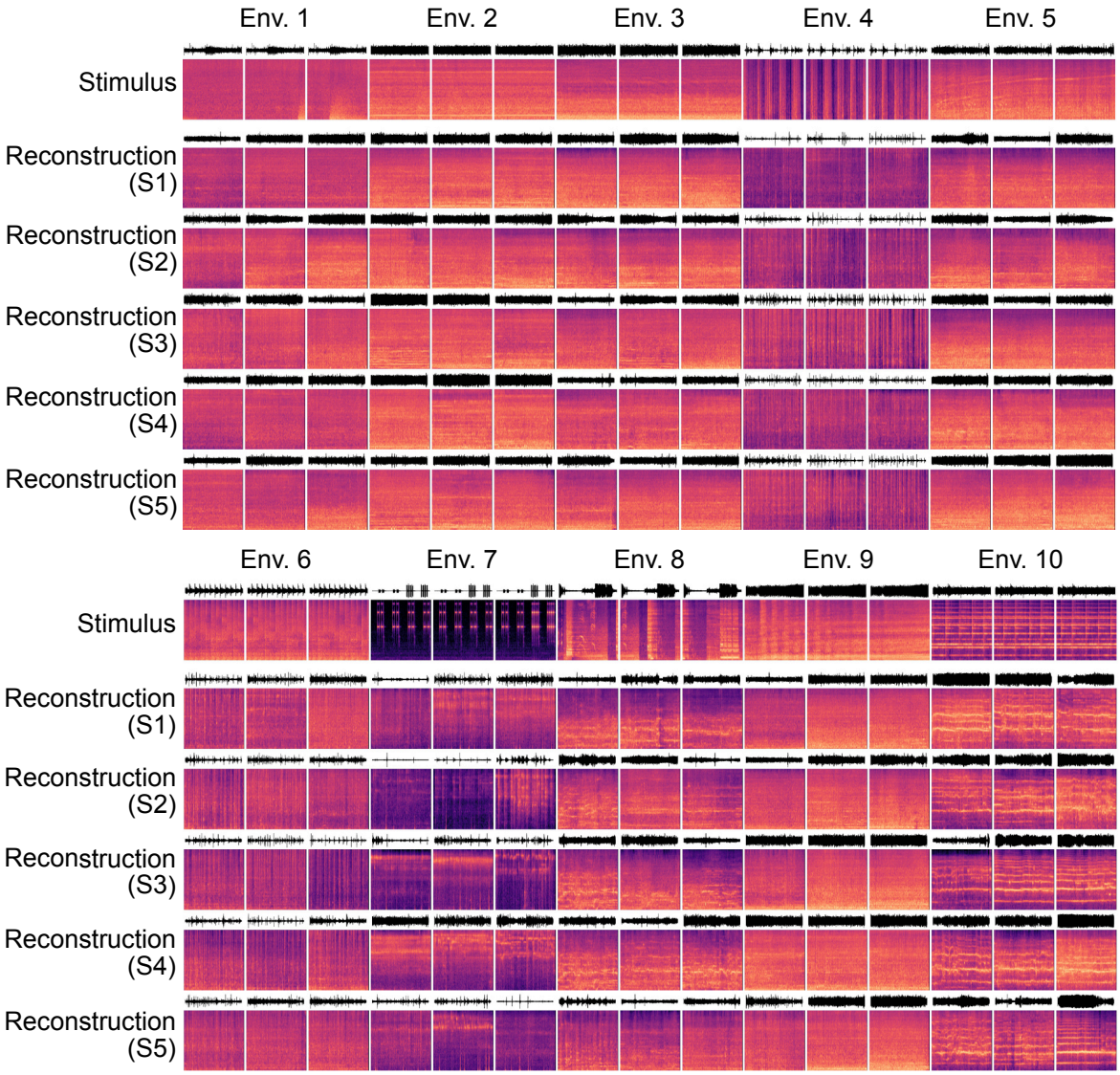

Fig. 5.6 Reconstructed Mel-spectrogram of 'Environmental' category. The top row displays the original Mel-spectrogram of the presented sound. The following five rows show the Mel-spectrograms reconstructed from each subject using the AC and the conv5 layer. The three consecutive samples originate from a single stimulus.

However, the model falls short of accurately reconstructing the details in cases with short temporal sequence variations, such as in animal sounds 2,4,7. The model also struggles with sounds that occur only in the first or second half of a 4-second stimulus, such as animal sound 5, failing to accurately reconstruct the sound onset and instead distributing it evenly across the 4-second duration. Figures 5.3 and 5.4, representing the speech category in English and non-English/non-native languages, respectively, exhibit clear harmonic patterns akin to human speech. These patterns distinguish them from other categories. Notably, examples where F0 and its harmonic structure appear across a wide range (e.g., female speaker in English2, 6, 7, 10) also show a harmonious structure in the reconstructed sounds at the high range. Figure 5.5, dedicated to the music category, illustrates complex patterns that span a wide frequency range. However, the model's reconstruction results are less sensitive to very short temporal sequence changes, such as in music6. Figure 5.6, allocated to the environmental category, depicts simpler but distinctive spectral patterns. Nevertheless, the reconstruction results for very short sequences, like Env. 7, generally show a smoothed pattern. Importantly, our model succeeds in reconstructing these complex spectro-temporal patterns while preserving the general content of each sound stimulus. This is a noteworthy improvement over previous fMRI-based reconstructions, which typically exhibited temporally smoothed patterns.

5.3.2 Evaluation of reconstructed sounds

To evaluate the quality and accuracy of our reconstructed sounds, I conducted a pairwise identification analysis. The process involved examining how accurately the reconstructed sounds could identify the original stimulus amongst pairs of stimuli. Auditory features were calculated from both the original stimuli and the reconstructed sounds, encompassing elements such as Mel-spectrogram pixels, hierarchical auditory representation, and acoustic features like Fundamental Frequency (F0), Spectral Centroid (SC), and Harmonic to Noise Ratio (HNR). An illustration of this evaluation is demonstrated in Figure 5.7A.

I calculated the correlation coefficient between the auditory features of the reconstructed sounds and those of the test stimuli. Identification accuracy was then determined by counting instances where the actual stimulus was correctly identified from the set of test stimuli. Initial evaluations, based on the pixel values of the Mel-spectrogram, yielded an average identification accuracy of about 70% across all subjects, suggesting that the reconstructions preserve a substantial amount of raw-feature information (Figure 5.7B).

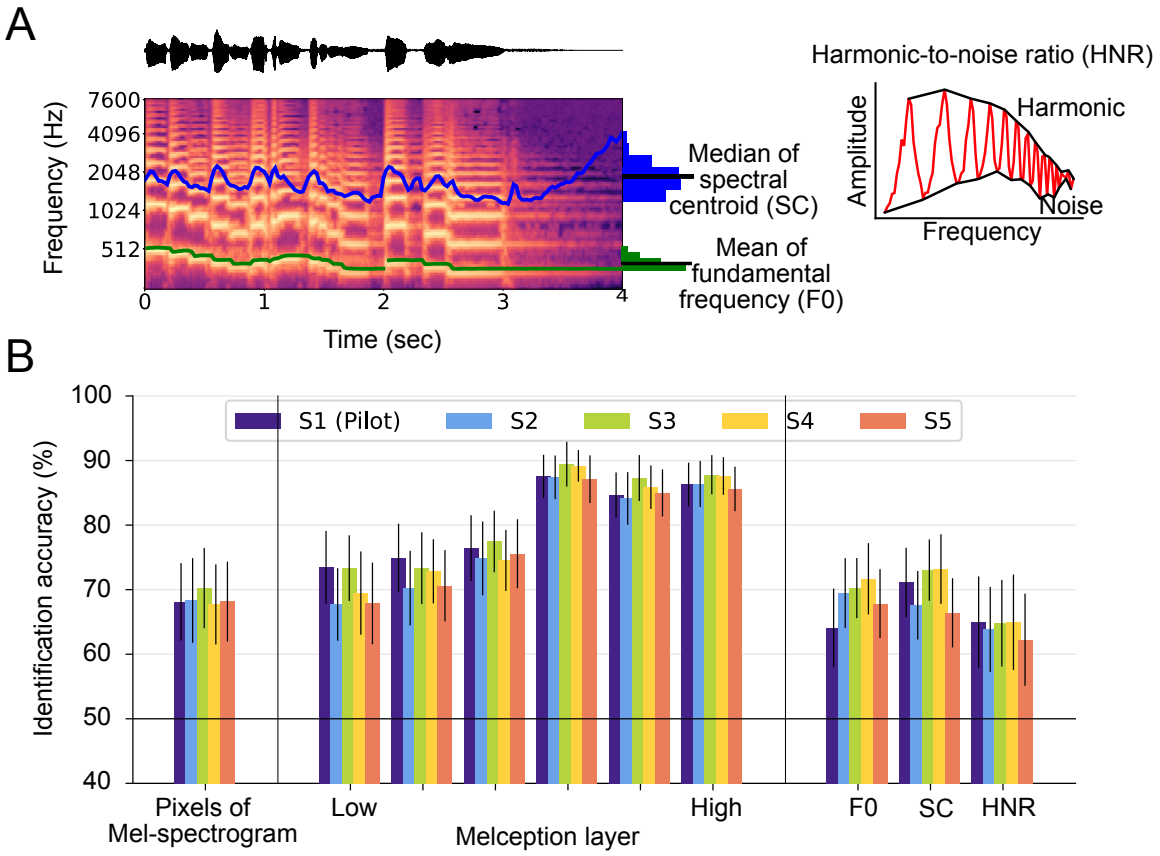

Fig. 5.7 Evaluation of reconstructed sound. (A) Acoustic properties. This panel showcases a sample Mel-spectrogram used in the evaluation of reconstructed sounds. I appraised three key acoustic properties: Fundamental Frequency (F0), Spectral Centroid (SC), and Harmonic to Noise Ratio (HNR). (B) Evaluating reconstructed sounds. The sound fidelity and quality of the reconstructions were assessed through an identification analysis involving the Mel-spectrogram pixels, hierarchical representation, and acoustic features. Each bar indicates the mean identification accuracy, while the error bar represents the 95% confidence interval, estimated using 50 data points. Different colors signify different subjects.

To assess the fidelity of the reconstructed sounds in terms of hierarchical auditory features, I employed the Melception classifier, a separate DNN model distinct from the one used in our decoding analysis. Our findings revealed that the identification accuracy of the reconstructed sounds surpassed the accuracy achieved with pixel-based analysis of the Mel-spectrogram, particularly with higher-level representations. In these instances, the average identification accuracy across all test stimuli exceeded 85% for each subject.

Although our reconstructions stem from high-level DNN features, it was paramount to ensure they also encapsulate the perceptual and qualitative facets of the original sounds. To validate this, I evaluated the reconstructed sounds based on three critical acoustic properties. Evaluations based on these acoustic features yielded mean identification accuracies comparable to those obtained using the pixels of the Mel-spectrogram. Specifically, I observed average identification accuracies of approximately 70% for both F0 and SC features, and about 65% for HNR across all subjects.

By-category identification analysis

I further analyzed the performance of sound reconstruction on a category-by-category basis, focusing on identification accuracy. Initially, I compared the identification accuracy of each category of reconstructed sound against all 150 test set candidates, as illustrated in Figure 5.8A. For the Animal category, there was no significant difference in identification accuracy between pixel and hierarchical representations when compared to the full test dataset. However, regarding acoustic quality, particularly F0, the performance was high, achieving 80% identification accuracy. HNR, on the other hand, resulted in a performance lower than 60%. The Environmental category results closely matched those of the full test dataset, showcasing an identification accuracy in the range of 70-80%. In contrast, the Speech category demonstrated a slightly lower performance in pixel and lower representations, under 70%, but still maintained high accuracy, around 85%, in higher representations. The music category exhibited a broad range of identification accuracies, from 70-90%, in pixel and hierarchical representations. However, performance declined somewhat in the intermediate layer and significantly dropped below 60% for F0.

Next, I evaluated the identification performance within the same category candidates, as shown in Figure 5.8B. In this scenario, both the animal and environmental categories demonstrated approximately 70% identification accuracy in pixel and hierarchical represen-

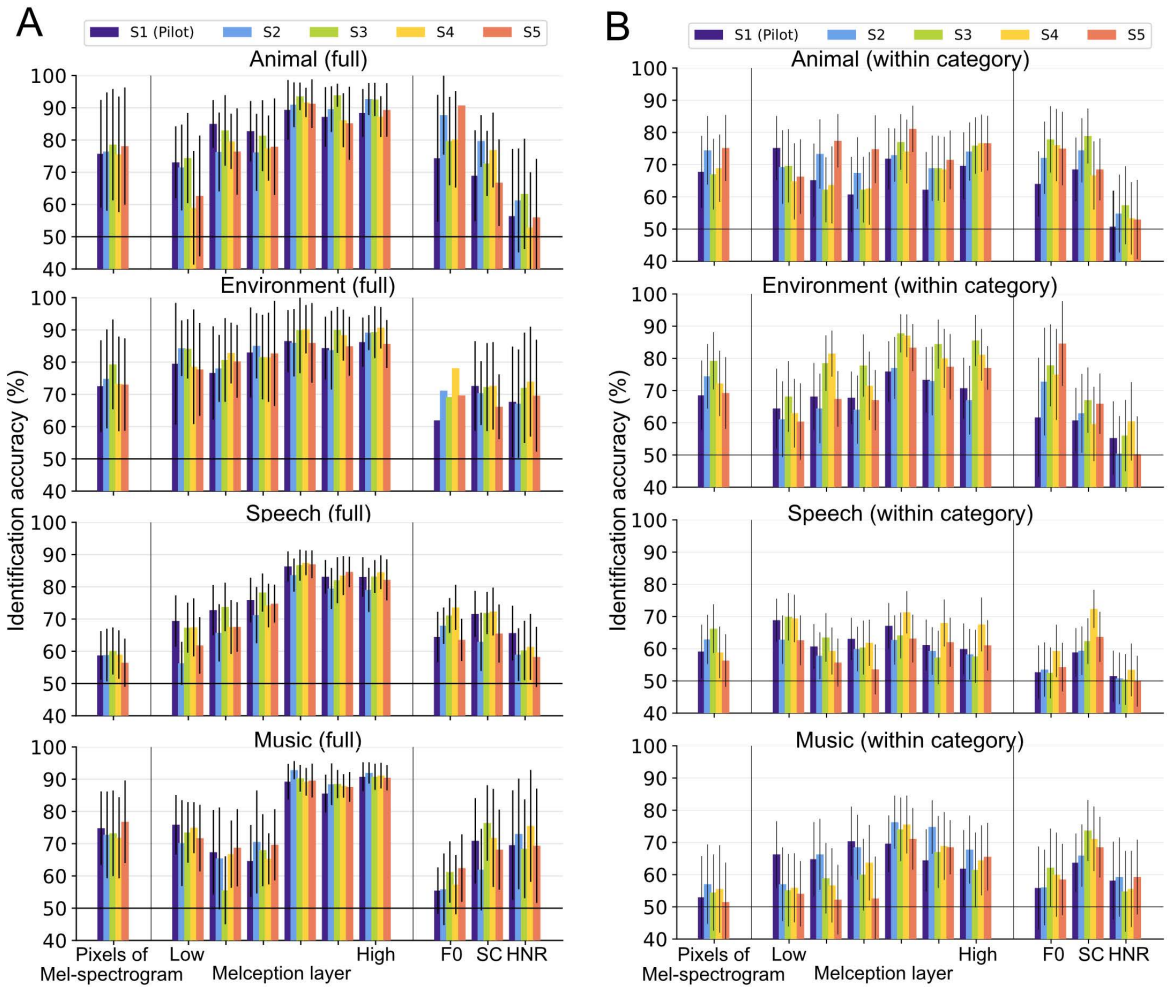

Fig. 5.8 Evaluation of reconstructed sound with by category analysis. (A) Identification accuracies with each test category set. Three different factors were evaluated: Mel-spectrogram pixels, hierarchical features of the Melception classifier, and acoustic features. Each bar shows the mean identification accuracy, averaged over 10 test stimuli for Environment, Animal, and Music categories, and 20 test stimuli for the Speech category. The error bars represent the 95% confidence interval. Each participant is symbolized by a unique color. (B) Identification accuracies with each test category set within category identification analysis. The results are displayed in a manner comparable to chart B, providing a comparative analysis of the same test data within each category.

tations, while performance was at chance levels for HNR. The speech and music categories showed performance levels around 60% for pixel and hierarchical representations, with both categories performing below 60% in terms of acoustic quality. These results indicate that while high-level content information was well reconstructed, improvements are needed for the reconstruction of detailed information.

5.3.3 Sound reconstruction from single trial fMRI sample

In our experiments, we primarily utilized multiple trial averages to enhance the signal-to-noise ratio (SNR) of the data, recognizing that this approach has inherent limitations for real-time or single-trial applications. The term 'single trial' refers to one instance or repetition of an experiment, for instance, presenting a specific auditory stimulus and recording the consequent brain activity. While most sound reconstruction studies have adopted the method of multiple trial averages, our research also delves into sound reconstruction from single trial fMRI samples.

Figure 5.9 presents examples of Mel-spectrograms reconstructed from a single trial fMRI sample. While inherently noisier compared to results derived from averaged fMRI samples, the reconstructed spectral and temporal patterns display similarities to the original stimuli. This suggests that our model effectively recreates auditory experiences even from single trial fMRI data.

Figure 5.9 also provides a quantitative evaluation of the sound reconstructed from a single trial fMRI sample. The results display an identification performance of around 60% when evaluated with Mel-spectrogram and low-level representations. However, as we ascend to higher hierarchical layers, there is a gradual increase in performance, showing an identification efficacy of 70-80%. For acoustic features, the model displays an identification performance of approximately 60%. These results, albeit showing a 10% performance drop compared to the sound reconstructed from eight-sample averaged fMRI samples, still demonstrate that our proposed model can achieve a reasonable reconstruction from single trial fMRI samples.

5.3.4 Sound reconstruction from actual features

In the next phase of the investigation, I assessed the boundaries of our model's capability for reconstructing sound with fidelity. To do this, I supplied the model with actual features at

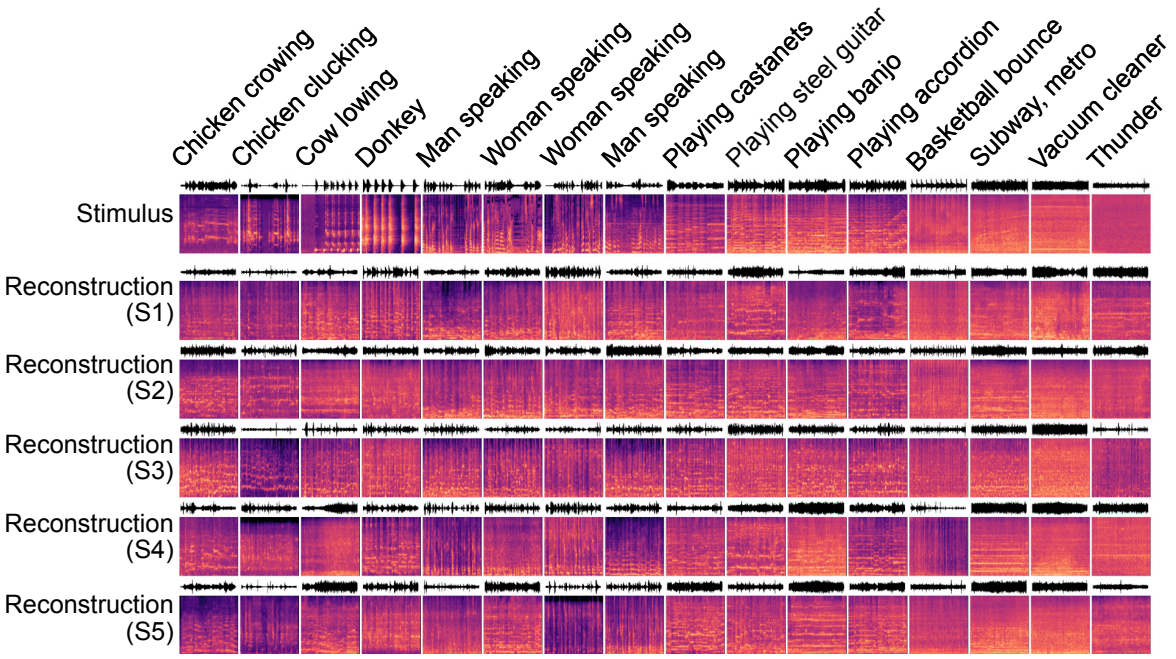

Fig. 5.9 Reconstructed sounds from single trial fMRI samples. The top row displays the original Mel-spectrogram of the presented sound. The following five rows show the Mel-spectrograms reconstructed from each subject using the AC and the conv5 layer.

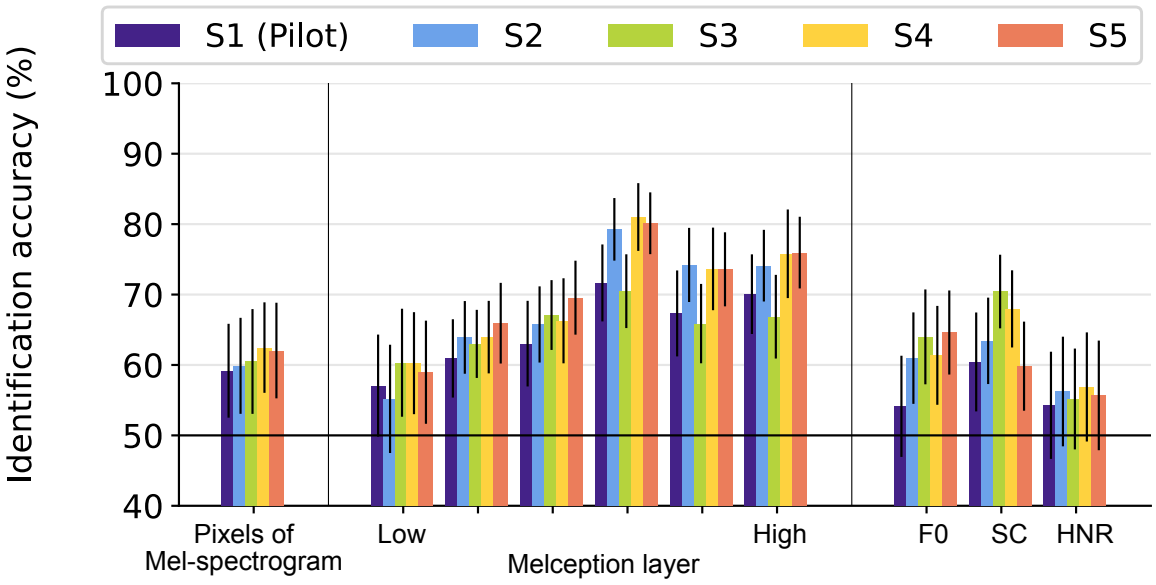

Fig. 5.10 Evaluation of reconstructed sound from single trial fMRI samples. The sound fidelity and quality of the reconstructions were assessed through an identification analysis involving the Mel-spectrogram pixels, hierarchical representation, and acoustic features. Each bar indicates the mean identification accuracy, while the error bar represents the 95% confidence interval, estimated using 50 data points. Different colors signify different subjects.

each stage of the process. When provided with actual codebook representations or actual DNN features, the performance of our model was outstanding, producing reconstructions of the original sound that were nearly perfect, as depicted in Figure 5.11A. The identification accuracies of these reconstructions were extremely high, reaching 99% and 95% respectively, as displayed in Figure 5.11B. These results underline that the fidelity of the reconstructed sounds is primarily determined by the decoding performance of the brain decoding analysis. In essence, the quality of our model's reconstructions depends on the decoding performance of the brain decoding analysis.

5.3.5 Auditory features

In order to delve deeper into the analysis, I performed additional reconstruction investigations using decoded Mel-spectrogram features and modulation features. In the case of Mel-spectrogram features, I extracted pixel values directly from the fMRI responses and converted them into an audible sound wave using a spectrogram vocoder. In contrast, for the modulation features, I implemented a two-step iterative reconstruction approach. This method initially transitions from modulation features to a spectrogram and then from the spectrogram to an audible sound wave. This technique is based on methodologies used in previous studies, particularly the work by Santoro et al. (2017). The resultant sounds from the reconstructed Mel-spectrogram and modulation features exhibited temporally smoothed patterns of the original spectrogram, as illustrated in Figure 5.12A. Upon a more quantitative evaluation (as depicted in Figure 5.12B), it became apparent that while the sounds reconstructed from Mel-spectrogram and modulation features were capable of identifying the actual stimuli above chance level, their effectiveness was particularly pronounced in the mid-level hierarchical layers. However, the model proposed in this study, which uses DNN features, surpassed the other auditory features in terms of performance across all evaluated metrics, clearly indicating the superiority of our methodology.

5.3.6 Model components

In order to better understand the impact of different components within our proposed model, I conducted a comparative analysis of reconstructions using two alternative methods Figure 5.13. The first, referred to as brain-to-codebook reconstruction, used a linear regression model to predict the codebook representation based on fMRI responses. The second method, known as pixel optimization reconstruction, iteratively optimized the pixel values of the

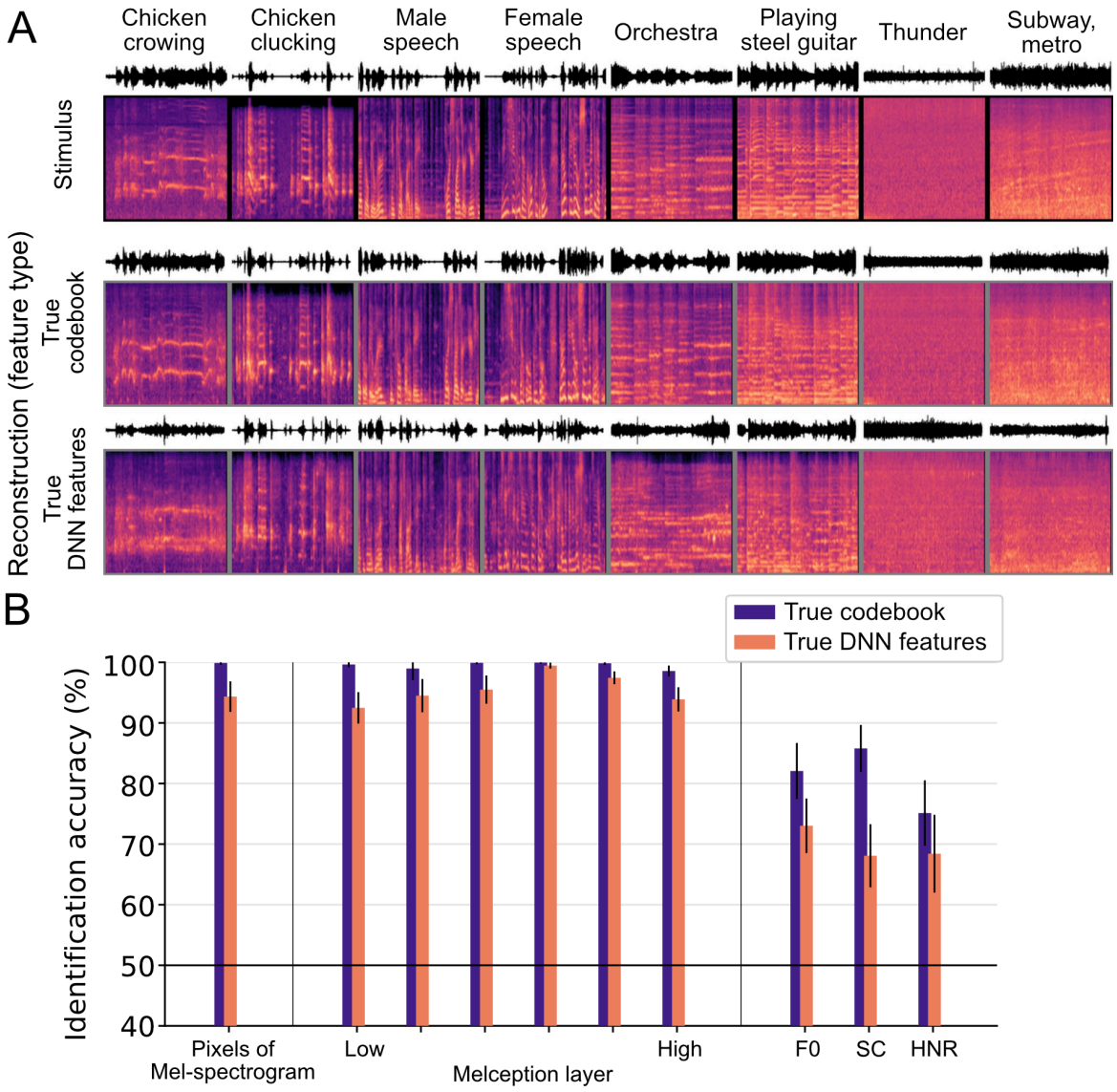

Fig. 5.11 Reconstructed sounds using true features. (A) Reconstructed Mel-spectrogram using true codebook and DNN features. The first row displays the Mel-spectrogram of the original sound. The second row showcases Mel-spectrograms reconstructed from authentic codebook representations. In contrast, the third row exhibits Mel-spectrograms reconstructed using authentic DNN features. (B) Evaluation of reconstructed sounds using true features. Each category of features used in the reconstruction is represented by a unique color.

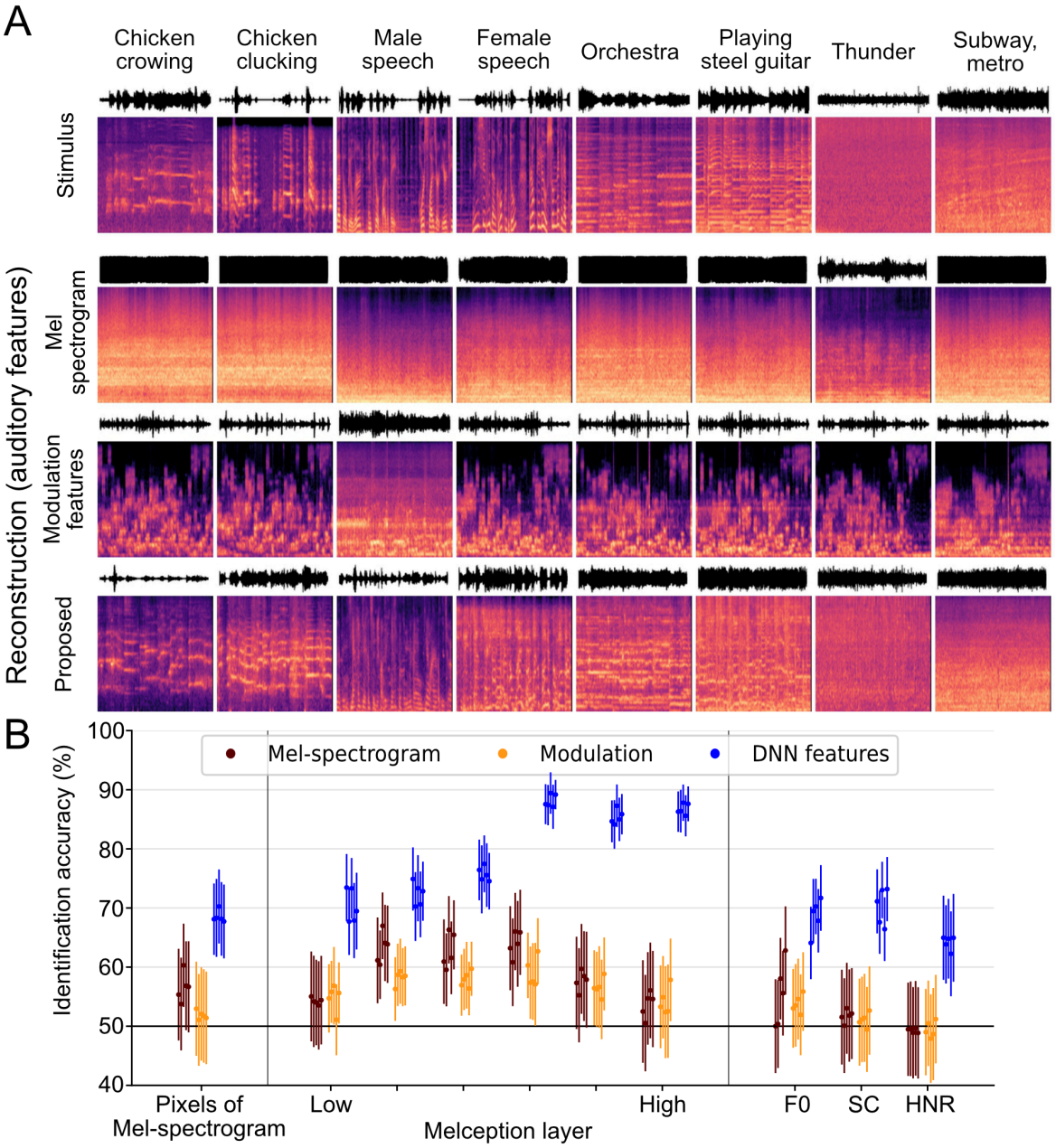

Fig. 5.12 Comparison of sound reconstruction using different auditory features. (A) Reconstructed Mel-spectrograms: The first row illustrates the Mel-spectrogram of the original sound, while rows two to four showcase the reconstructed Mel-spectrograms. These reconstructions use different auditory features and were obtained from the auditory cortex (AC) data of subject S3. (B) Evaluation of Sound Reconstruction Using Different Auditory Features: Each unique color represents a different auditory feature used in the reconstruction. Each dot indicates the mean identification accuracy for each subject, with the error bar denoting the 95% CI estimated using 50 data points.

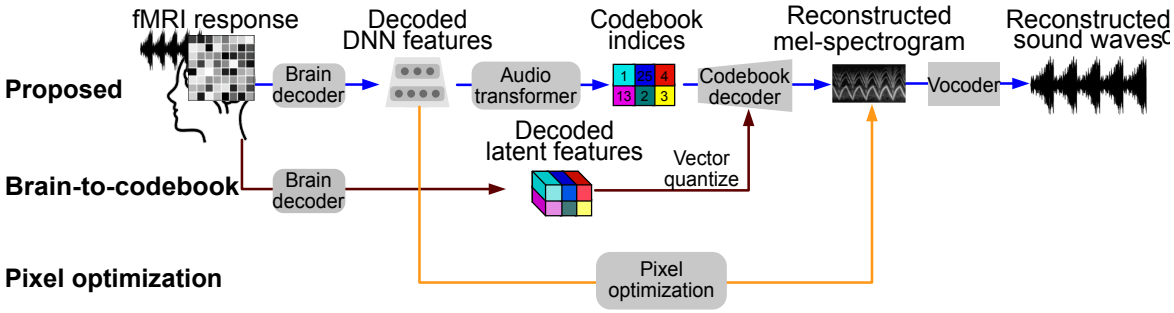

Fig. 5.13 Schematic of sound reconstruction with ablated model components. The diagram presents two different approaches to sound reconstruction. The first, termed 'brain-to-codebook reconstruction', is shown by the brown line. This method predicts codebook representations directly from fMRI responses. The second approach, known as 'pixel optimization reconstruction', is indicated by the orange line. It involves the iterative optimization of pixel values in Mel-spectrograms to align with the decoded DNN features, which are inferred from fMRI responses.

Mel-spectrogram by aligning them with the decoded DNN features derived from brain responses.

Our exploration into the relative effectiveness of different components within our model involved analyzing and comparing sound reconstructions via two alternative methods: the brain-to-codebook reconstruction and the pixel optimization reconstruction. The brain-to-codebook reconstruction involves predicting codebook representations from fMRI responses using a trained brain decoder. These predicted representations are then converted into a Mel-spectrogram via a codebook decoder and transformed into sound waves with a spectrogram vocoder. The Mel-spectrogram resulting from this process closely mirrored a temporally smoothed version of the original spectrogram, a phenomenon previously observed with direct regression on modulation or physical features (Figure 5.14A). The pixel optimization reconstruction, meanwhile, uses all decoded DNN features drawn from the VGGish-ish model to optimize the spectrogram pixels for sound reconstruction. This method successfully reproduces distinct spectral patterns but struggles to capture the detailed temporal patterns inherent in the original spectrogram.

In our quantitative evaluation (Figure 5.14B), the brain-to-codebook reconstruction yielded identification accuracies of approximately 70%, based on the pixel values of the Mel-spectrogram, which are comparable to the results from our proposed model. However, identification performance fell as I ascended the hierarchical layers of sound representation, dipping below 60% for all subjects at the highest layer of the Melception model. Evalu-

ations using acoustic features revealed similar patterns. Spectral Centroid (SC) displayed identification accuracy akin to that of Mel-spectrogram pixels, but all subjects recorded identification accuracies below 60% for both Fundamental Frequency (F0) and Harmonic to Noise Ratio (HNR) metrics. The pixel optimization reconstruction demonstrated identification performance, based on the pixels of the Mel-spectrogram, comparable to other reconstruction methods. However, when evaluated based on hierarchical representations, this method surpassed the brain-to-codebook reconstruction but lagged behind our proposed method, especially in higher layers. Additionally, evaluations based on acoustic features saw identification accuracy for F0 and HNR metrics barely exceed chance levels, with the exception of SC.

These findings indicate that while all three reconstruction methods were able to approximate spectral patterns to some degree, only our proposed model's reconstructed sound retained perceptual qualities closely resembling the actual stimuli. In conclusion, our proposed model outperformed other reconstruction methods in producing sound reconstructions that were both more perceptible and perceptually similar to the original sounds. These findings highlight the value of integrating brain-like features and separating temporal information from DNN features in our sound reconstruction approach.

5.4 Discussion

In this chapter, I introduced a novel approach to unrestricted neural sound reconstruction from fMRI activity, leveraging DNN features and an audio-generative model. This process effectively transformed decoded DNN features into high-quality audio signals, thanks to an audio-generative transformer. This device predicted a simplified version of a Mel-spectrogram (a codebook representation) based on the decoded DNN features. Our model successfully managed to reconstruct complex spectral-temporal patterns that approximated the content and quality of the original sound stimulus, as confirmed by both qualitative and quantitative evaluations. However, I acknowledged room for improvement, especially concerning the fine details in speech or music sequences. Crucially, our sound reconstruction demonstrated resilience even when specific categories were not included in the training phase.

While our model adeptly reconstructed sounds using true features (Figure 5.11), reconstruction from decoded features fell short on detail for complex sequences such as speech and music. Bearing these limitations in mind, future research could investigate more advanced

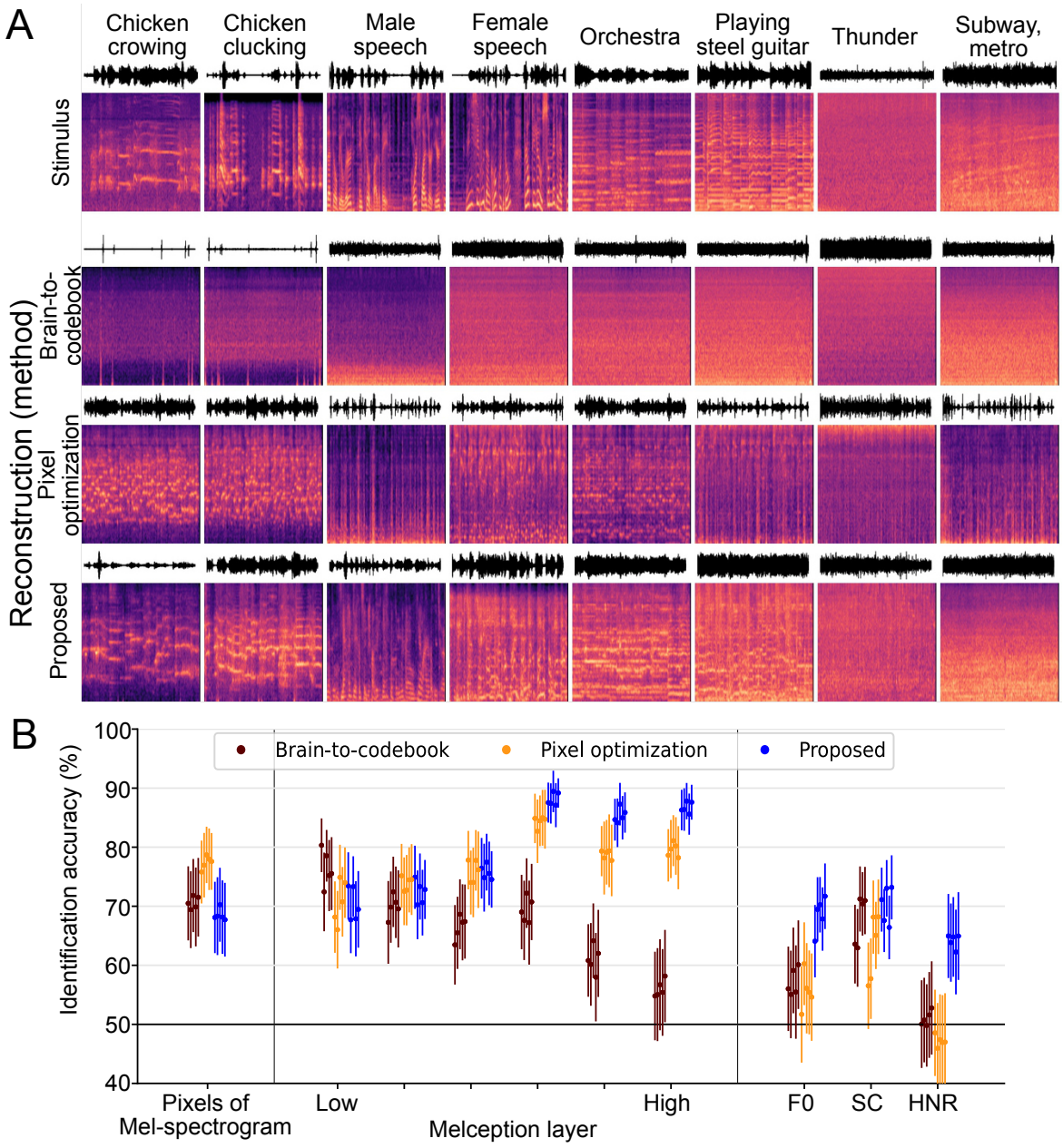

Fig. 5.14 Effect of model components. (A) This panel displays the reconstructed Mel-spectrograms. The first row shows the original Mel-spectrogram of the sound presented. The subsequent rows, from two to four, present Mel-spectrograms reconstructed through different methods, using the AC from subject S3. (B) Evaluation of the sounds reconstructed from different methods. Each method is represented by a distinct color. Each method is represented by a distinct color. Every dot illustrates the mean identification accuracy computed for each subject, with the associated error bar denoting the 95% confidence interval, estimated using 50 data points.

decoding methods incorporating sequential processing. Techniques like Long Short-Term Memory networks (LSTM), Recurrent Neural Networks (RNN), and other recursive models have been widely used in auditory decoding tasks across different neuroimaging modalities (Daly, 2023; Szabó and Barthó, 2022; Yoo et al., 2021). Moreover, efforts have been made to utilize the inherent temporal information within fMRI (Loula et al., 2018; Wang et al., 2019). Given our success with the transformer model in untangling compressed features, applying a transformer to decode temporal information in fMRI signals could be a promising approach to enhance the quality and detail of reconstructed sounds.

Upon comparing the sound reconstructed using different auditory features, it became clear that DNN features outshined both the Mel-spectrogram and modulation features in terms of decoding and reconstruction. Interestingly, an attempt to replicate reconstruction using modulation features, as illustrated in a previous study (Santoro et al., 2017), failed to produce satisfactory results. This is noteworthy, despite the previous study's objective to reconstruct short stimuli durations of 1 second using high spatial resolution 7T fMRI. The standout decoding performance of DNN features, calculated through hierarchical processing, aligns with findings from previous encoding analyses (Kell et al., 2018; Tuckute et al., 2023). This not only highlights the merit of using brain-like features in sound reconstruction but also suggests that DNN features, characterized by their brain-like properties and hierarchical processing capabilities, are a promising prospect for improving the quality and accuracy of sound reconstruction from fMRI data.

When comparing sound reconstructions achieved by removing different model components, it was clear that both DNN features—with their brain-like hierarchical properties—and the process of translating these features back into sound through codebook representation substantially enhanced reconstruction performance, both quantitatively and qualitatively. Yet, this raises the question of what precisely the role of the codebook representation is in the reconstructed results.

In the forthcoming chapter, I will dive into a comprehensive analysis that explores the model's ability to reconstruct sounds from categories not included in the training set. This exercise will serve as a rigorous test of the model's generalization capability. A key feature of this analysis will be an in-depth interpretation of the codebook representation, a crucial component of our sound reconstruction model. I will demonstrate how our model pieces together these elemental features to synthesize sound. Through this examination, I aim to

garner deeper insights into the mechanisms underpinning our model's performance, especially its capacity for generating high-fidelity sound reconstructions.

Chapter 6

Generalization beyond trained categories

6.1 Introduction

Most previous studies on sound reconstruction often avoided the task of reconstructing arbitrary sounds from brain responses due to the vast diversity of sounds and their complex sequences, not to mention the limited resolution of brain imaging techniques. As a result, these studies were mainly focused on classifying specific types of speech (Chakrabarti et al., 2015; Martin et al., 2018; Moses et al., 2019; Pei et al., 2011) or reconstructing a limited set of examples such as digits (Akbari et al., 2019) and words (Wang et al., 2018). However, as I showcased in the previous chapter, it is indeed possible to reconstruct a wider variety of sounds by using the spatial patterns of fMRI responses, thereby capturing both the content and quality of sounds.

In this chapter, I will dig deeper into the abilities of our model to reconstruct sounds that fall outside the categories used in the training data. I conducted an experiment where I trained the decoder by excluding one category at a time from the training data. I then evaluated how well the model could reconstruct sounds from the omitted category using only the test data. The fact that our model could generalize to categories not present in the training data indicates that it isn't simply matching brain data to the training examples.

Further, I performed an in-depth analysis of the "codebook" representations used by our model, demonstrating that the reconstructed sounds are synthesized from a combination of basic sound features, defined by a sequence of codebook entries. This finding provides

valuable insights into how our model works and its potential to recreate a wide variety of sounds. The contents of this chapter is based on the Results: *Generalization beyond trained categories* of (Park et al., 2023).

6.2 Methods

To begin with, I labeled all the stimuli in the training dataset, categorizing them into one of four categories: animal, environments, human speech, or music. There were cases where each training stimulus could belong to multiple categories or not fit into any of these four. Subsequently, for the brain decoder training, I excluded the category in question and calculated the decoded DNN features and reconstructed sound using the test set of the excluded category. In this scenario, when the animal category was ablated, the training for the decoder utilized 12,036 fMRI samples computed from 1,003 stimuli. For the environments category, 9,552 fMRI samples were used, calculated from 796 stimuli. For human speech, 8,844 samples were derived from 737 stimuli, and for music, 9,516 fMRI samples were drawn from 793 stimuli. Following the computation of the reconstructed sound, I conducted an identification analysis using test stimuli from all categories. Then, I compared the results of the ablated training category set with those from the complete training category set. Finally, I conducted a within-category analysis, considering only the evaluation candidates from the same category. The results of this analysis were also compared with those obtained using all the categories in the training set.

6.3 Results

6.3.1 Sound reconstruction with ablated category training sets

First, an analysis was conducted where the decoder was trained by excluding one category at a time from the training data. I then evaluated the reconstruction results for the excluded category from the test dataset. As depicted in Figure 6.1A, the results indicate that our model can still reconstruct spectral and temporal patterns similar to the actual stimuli, even without training on a specific category. Particularly for animal and environmental sound categories, the model effectively maintained the reconstruction performance even when these categories were not part of the training set. Although the reconstructed speech sounds were somewhat noisy, the model managed to extract and emphasize distinct characteristics of human voices. However, a significant drop in performance was observed when the music category was

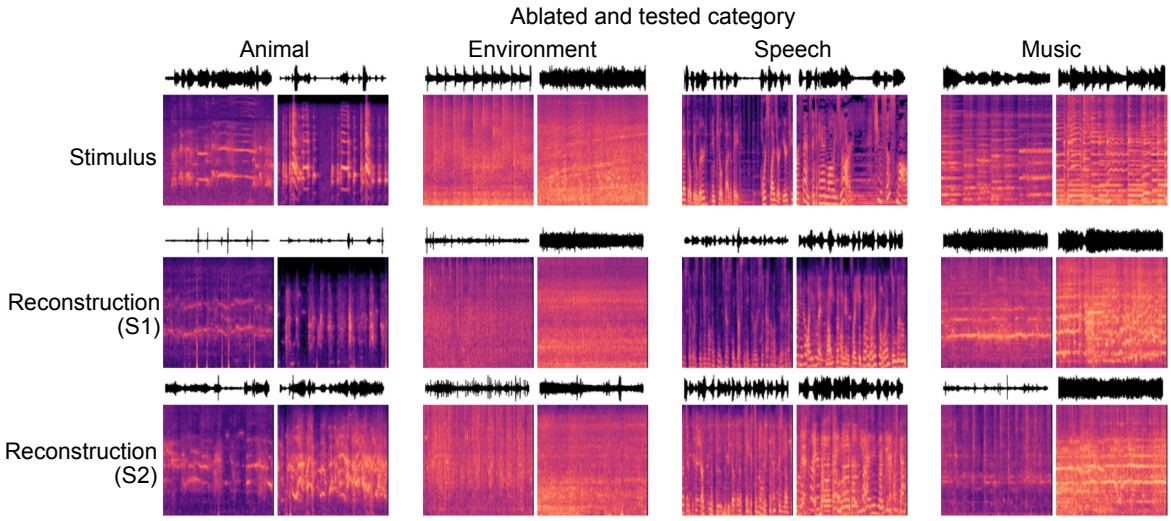

Fig. 6.1 Reconstructed sounds with ablated training category sets. The image presents examples of reconstructed Mel-spectrograms for four categories (Environment, Animal, Speech, and Music) obtained from decoders trained on a dataset where data from the corresponding category was excluded. The top row displays the Mel-spectrograms of the original presented sounds. The second and third rows show the reconstructed Mel-spectrograms from two different subjects, using the AC and conv5 layer respectively.

excluded from the training set, resulting in the reconstructed sounds resembling rhythmic environmental noise rather than actual music.

For quantitative evaluation, I compared the identification accuracies derived from the ablated training category set with those from the complete training category set. Despite a slight dip in performance, the overall identification accuracies from the ablated training set remained comparable to those from the full training set for each category. In a category-specific analysis, the animal and environmental sound categories demonstrated identification accuracies above 70% for most metrics in the ablated training set. The music and speech categories showed performances around or below 60% when using pixels of Mel-spectrograms, F0, and HNR, but achieved approximately 70% accuracy at higher hierarchical representations. These findings mirror the trends observed in the full training set.

6.3.2 Interpretation of codebook representations

In previous sections, I demonstrated that proposed model can reconstruct arbitrary sounds, not used in the training dataset, from fMRI patterns. These reconstructions are translated from DNN features to a sequence or combination of codebook representations through an

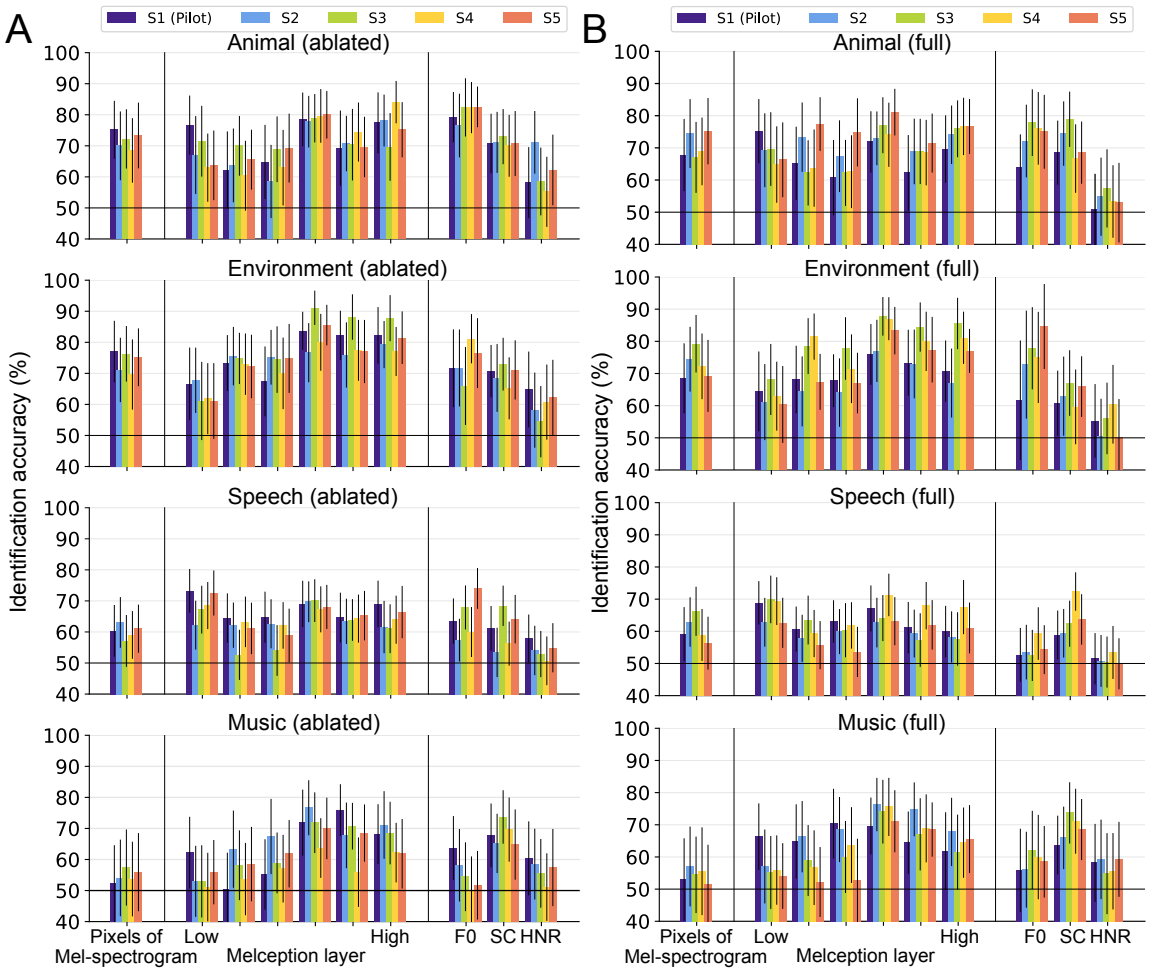

Fig. 6.2 Evaluation of reconstructed sounds with ablated training category sets. (A) Identification accuracies with the ablated training set. This depicts the mean identification accuracy for each category using Mel-spectrogram pixels, hierarchical features of the Melception classifier, and acoustic features. Each bar represents the mean identification accuracy, averaged across 10 test stimuli for the Environment, Animal, and Music categories, and 20 test stimuli for Speech. The error bars denote the 95% confidence interval. Different colors represent different subjects. (B) Identification accuracies with the full training set. This shows the results of the same test data for each category, presented for comparison in the same format as in (A).

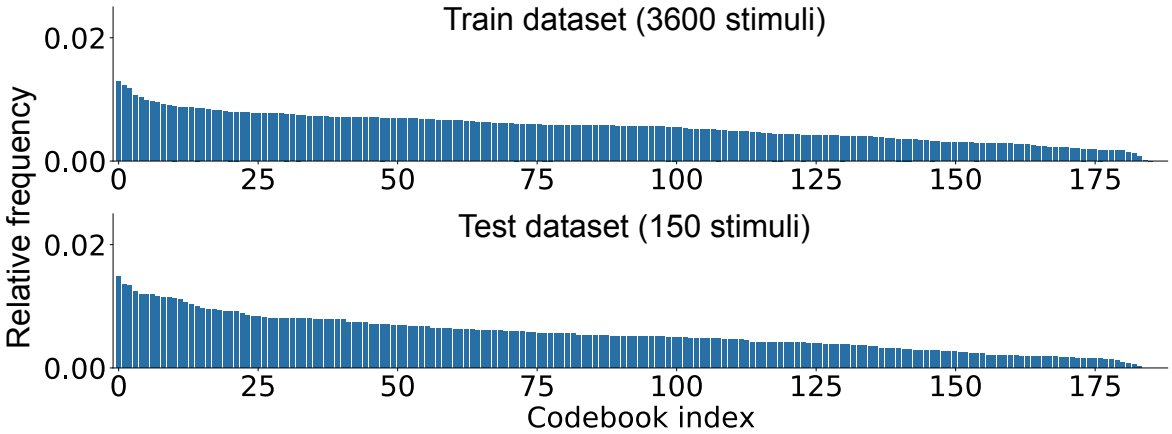

Fig. 6.3 Histogram of the codebook indices used for representing the training dataset and test set. The upper panel displays the histogram of codebook indices when mapping the 4800 stimuli used for training the brain decoder. The lower panel exhibits the histogram of codebook indices when mapping the 150 stimuli used for the test set, using the same codebook.

audio transformer, and then converted back into a Mel-spectrogram. This process revealed that each code serves as elements of auditory features. In this chapter, I conducted an interpretation analysis to understand how each code is used in actual training and testing, and what patterns they represent in patches cropped from the Mel-spectrogram.

I examined the distribution of the codebook indices used in sound generation. Figure 6.3 displays the histogram of codebook indices for the true stimuli of the training and test datasets used in the experiment. This illustration demonstrates that the codebook indices are broadly distributed. It also confirms that the test dataset has a distribution fairly similar to the training dataset. Figure 6.4 presents the histogram of codebook indices from the sound generated from the brain response. The top panel is the histogram of all test samples and five subjects, showing that like the histograms of the train/test dataset, all codebook indices are generally distributed.

On analyzing by category, it was evident that the codebook indices were widely distributed in the animal, speech, and music categories as well. Although a few codebook indices were frequently used in environmental sounds due to their monotonous nature, overall other codebook indices were also fairly distributed. These results show that the codebook indices used in our sound generation are not specifically used in any category, but generate sound through a combination of various codes. Each code represents these elements of acoustic features.

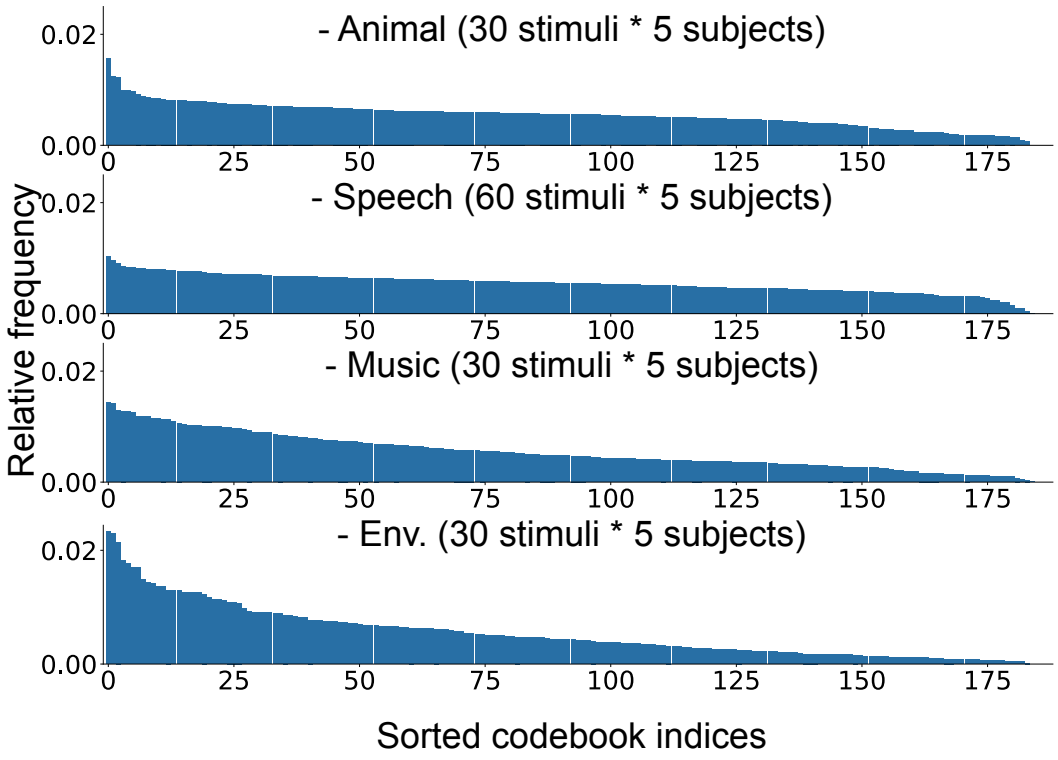

Fig. 6.4 Histogram of the codebook indices used for representing the reconstructed sounds. Each panel displays the histograms of codebook indices used to represent the sounds reconstructed for each category, derived from five test subjects.

6.4 Discussion

In this chapter, I addressed the reconstruction of arbitrary sounds from brain activity using a model trained on a multitude of sound categories. The ability of the model to reconstruct sounds even when one category was systematically excluded from the training set is indicative of a generalizable approach that does not merely memorize the training examples. This capacity for sound reconstruction seems to hinge on the codebook representations learned by the model, which seemingly act as elemental features for the reconstruction of a diverse array of sounds.

The different performance levels of the model in reconstructing sounds from various categories, despite excluding those categories from the training set, provide an intriguing perspective on the nature of the learned codebook representations. It appears that some categories, such as animals and environmental sounds, are more robustly represented, or more easily inferred from the remaining categories during training. This could be due to these categories having more shared features or patterns with the other categories included in the training set. Conversely, the significant performance drop in music reconstruction when this category was excluded from the training set suggests that music possesses unique characteristics that are not as easily extrapolated from the other sound categories.

Chapter 7

Hierarchical auditory areas and features

7.1 Introduction

In Chapter 5, I demonstrated how our proposed approach effectively reconstructs sounds, capturing both their perceptual content and quality. A fundamental tenet of our hypothesis was that auditory features are processed hierarchically, mirroring the functional organization of the human brain. The human auditory system and DNN models exhibit analogous hierarchical structures, as evidenced by research conducted by Kell et al. (2018). They implemented an exhaustive brain encoding analysis, predicting human auditory responses from DNN model responses, which revealed a striking hierarchical parallel between the DNN model and fMRI data. They designed a DNN architecture for sound recognition, crafted to reflect the human auditory system's hierarchical processing. Their DNN structure distinguished between common layers, responsible for low-level processing akin to early auditory stages, and branching layers, handling task-specific processing like speech recognition or music genre classification. By leveraging this trained DNN model and fMRI data for natural sounds, their encoding analysis signified a hierarchical correspondence between brain and DNN models. Early auditory cortex responses were better predicted from the DNN features derived from common layers, whereas responses from nonprimary regions were better aligned with DNN features from the branched layer. Despite these findings, a subsequent study noted that not all DNNs emulate the human brain's organization. Encoding performance could differ substantially depending on the DNNs' structure and optimization tasks (Tuckute et al., 2023). Moreover, in chapter 4, I were unable to discern a clear hierarchical correspondence

between DNN layers and auditory ROIs based on feature decoding performance. In this chapter, I delve into the impact of individual ROIs and hierarchical auditory features on the sound reconstruction process. I present the results of a meticulous investigation into how these individual ROIs and DNN layers, under various conditions, influence the generation of sounds from fMRI responses. The sounds produced under these diverse conditions were assessed using an array of metrics. This comprehensive examination seeks to shine a light on the influence of individual ROIs and hierarchical auditory features on the reconstruction process, thereby fostering a more profound understanding of our reconstruction model's outcomes. The contents of this chapter is based on the Results: *Auditory ROIs and DNN layers* of (Park et al., 2023).

7.2 Methods

7.2.1 Sound reconstruction from individual ROIs

To examine the variation in sound reconstruction across individual auditory regions, I conditioned the brain decoder's training solely on individual ROIs to predict DNN feature units. For this purpose, I employed regions following the ventral pathway from the A1, LBelt, PBelt, A4, and A5, using each hemisphere independently and in a combined fashion. Prior to the training of each decoder, I computed the correlation between voxel responses and DNN features, selecting the top 200 voxels to train the L2-regularized linear regression model. In the test phase, the decoded features calculated were employed as inputs to the audio transformer, codebook decoder, and spectrogram vocoder, trained as detailed in Chapter 5. This process allowed us to generate the reconstructed sound from each individual ROI.

7.2.2 Sound reconstruction from DNN layers

To investigate the variance in sound reconstruction stemming from different DNN layers, I conditioned the brain decoder's training on the AC to predict DNN feature units from selected layers within the VGGish-ish model. Drawing upon the feature decoding results from our pilot study with subject S1, I chose six representative layers (conv1_1, conv2_1, conv3_1, conv4_1, conv5_3, fc1) that demonstrated the highest decoding performance from the convolutional blocks of VGGish-ish. Additionally, I trained individual audio transformers to translate DNN features into codebook representations, with the translation adapted according to each layer. For the convolutional layers, the temporal dimension of the

given input was resampled to 21 points to align with the output temporal dimension across all DNN features. Employing the codebook decoder and spectrogram vocoder, as detailed in Chapter 5, I reconstructed the sound from each layer.

7.3 Results

7.3.1 Auditory ROIs

Figure 7.2A summarizes the reconstructed Mel-spectrogram from individual ROIs (refer to Supplementary Figure 7.4 for hemisphere separation). Our investigation found that the reconstructed sounds bore a resemblance to the original sounds, irrespective of the auditory ROI from which they were derived. The first set of stimuli, representing the category of animal sounds, show unique spectral patterns, making the reconstructed sounds easily distinguishable when compared to the original stimuli. Following this, the speech category stimuli reveal clear harmonic patterns, distinctive of human speech, setting them apart from other sound categories. The next set of stimuli, designated to the music category, depict complex patterns spanning a wide frequency range. Finally, the environmental sound stimuli present simpler but characteristic spectral patterns. Notably, our model manages to reconstruct these intricate spectro-temporal patterns, preserving the general content of each sound stimulus. This fidelity is a marked improvement over prior fMRI-based reconstructions, which often displayed temporally smoothed patterns. Moreover, the reconstructed sounds generated from each ROI exhibit a high degree of consistency.

Figure 7.2B offers a quantitative evaluation of the reconstruction performance across different ROIs. Our analysis confirmed that both the Mel-spectrogram and the low-level features within the hierarchical representation were reconstructed more accurately when the decoder was trained on the early auditory cortex, specifically region A1. On the other hand, the decoder performed better on the reconstruction of intermediate or high-level features within the hierarchical representation when it was trained on the auditory association cortex, such as region A4. Furthermore, when comparing the performance of the decoder on the early auditory cortex and the auditory association cortex, the decoder demonstrated higher accuracy in identifying acoustic features when trained on the early auditory cortex. These findings suggest that our proposed reconstruction model successfully leverages the distributed neural responses across the auditory region to generate accurate reconstructions.

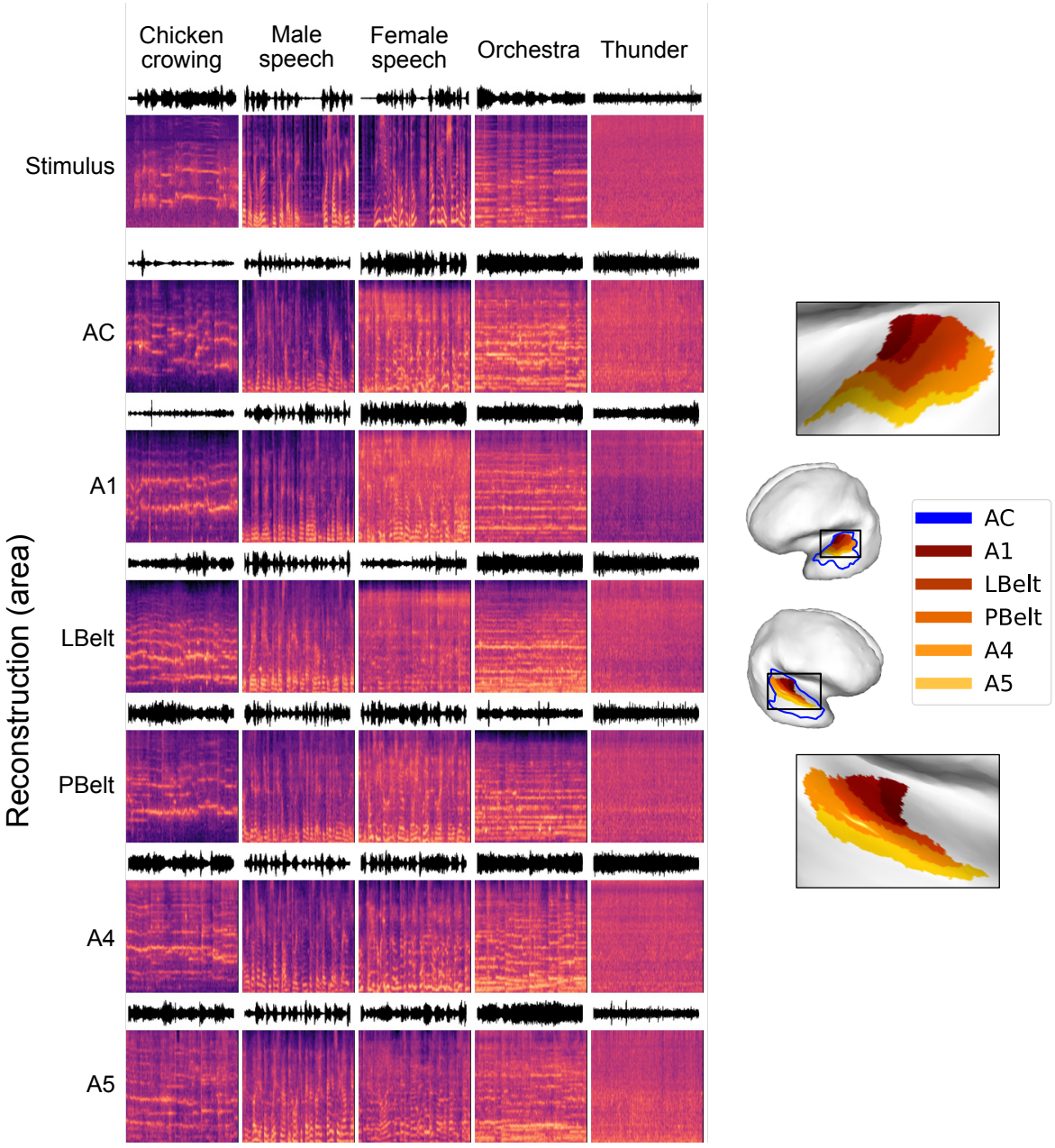

Fig. 7.1 Sound reconstruction from individual ROIs. The first row illustrates the original Mel-spectrogram of the presented sound. Rows two to seven sequentially display the Mel-spectrograms that have been reconstructed from each separate ROI using the conv5 layer. These reconstructions are based on the data from subject S3.

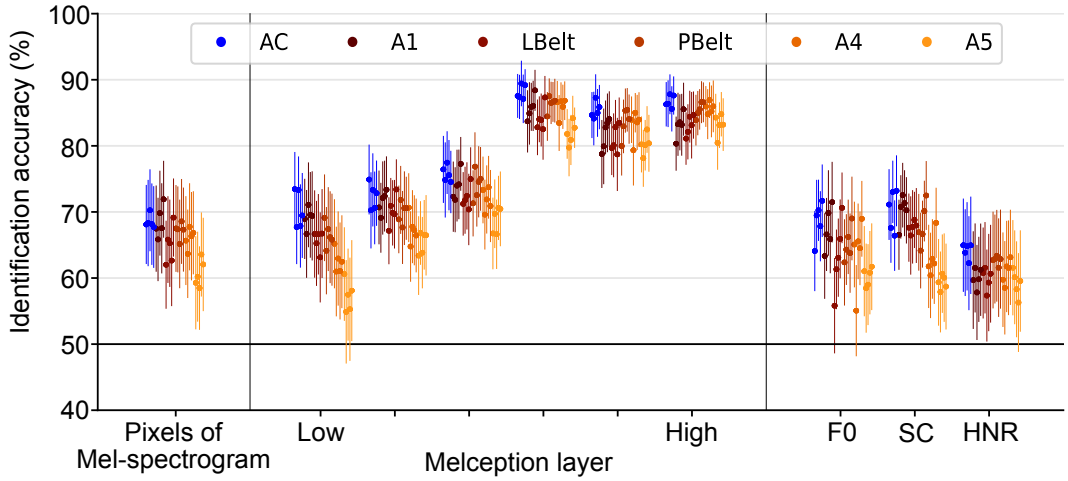

Fig. 7.2 Evaluation of reconstructed sounds from individual ROIs. Each ROI is represented by a unique color for easy distinction. Each dot on the graph denotes the average identification accuracy obtained from individual subjects. The associated error bars indicate the 95% confidence interval, which has been estimated using a sample of 50 data points.

7.3.2 DNN layers

Figure 7.6A presents an overview of our investigation into the influence of hierarchical representations within the DNN model on sound reconstruction. Our study showed that sounds reconstructed from lower layers of the DNN, such as Conv1 and Conv2, resulted in reconstructions of lower perceptual quality, tainted by significant noise. Conversely, the intermediate layers, namely Conv3 and Conv4, yielded reconstructions with spectral patterns resembling those of the original stimuli. Remarkably, sounds reconstructed from the higher layers, especially Conv5, accurately reflected the spectral patterns of the actual stimuli. Interestingly, even without a temporal dimension, the FC3 layer managed to replicate distinct spectral patterns of the original stimuli within the reconstructed Mel-spectrogram. I found that sounds reconstructed from the higher layers preserved perceptual content more effectively than those reconstructed from the lower layers.

Figure 7.6B provides a quantitative evaluation of the influence of different DNN layers on the sound reconstruction. Our analysis revealed that sound reconstructions from higher layers demonstrated superior identification accuracy across all evaluation metrics compared to those from lower layers. The top convolution layer (Conv5) notably achieved the highest performance. On the other hand, the FC3 layer's sound reconstructions displayed comparable performance to Conv5 in replicating high-level hierarchical representations. However, it

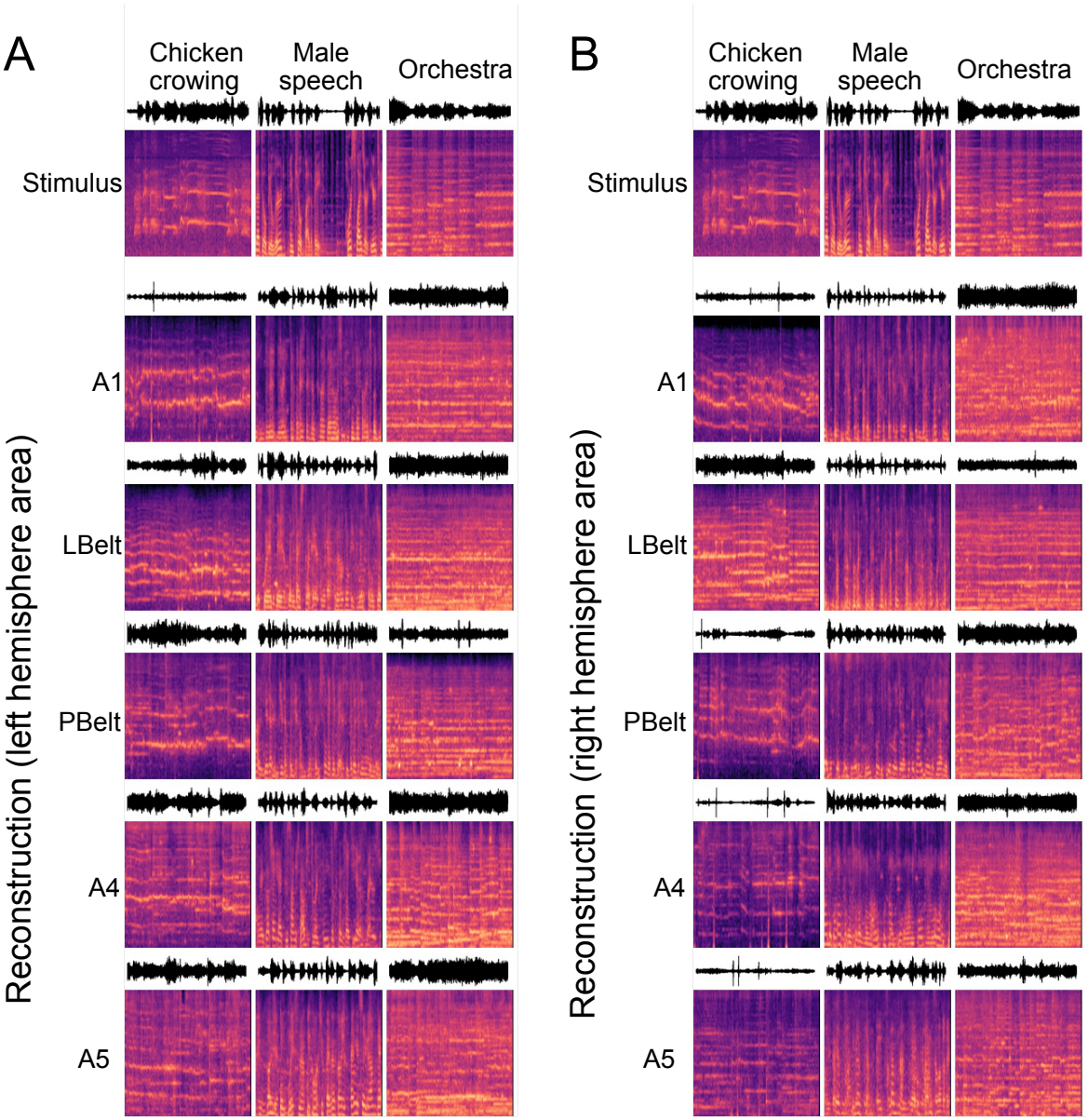

Fig. 7.3 Hemispheric comparison of reconstructed sounds using individual ROIs. The left panel illustrates the spectrograms reconstructed from the left hemisphere, whereas the right panel represents those from the right hemisphere. The top row shows the Mel-spectrogram of the original sound presented to the subject. Subsequent rows, from two to six, display the Mel-spectrograms reconstructed from each individual ROI, leveraging the conv5 layer from subject S3.

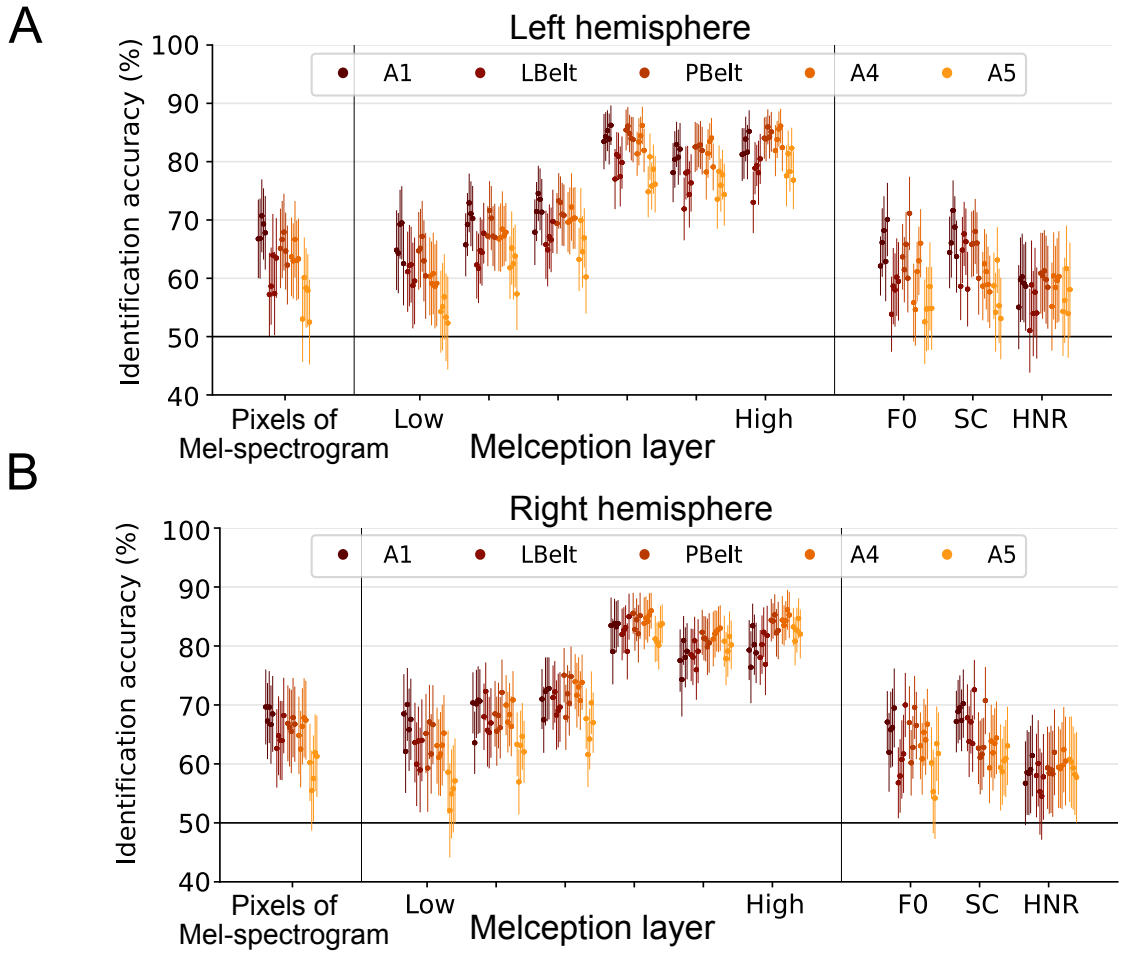

Fig. 7.4 Comparative evaluation of reconstructed sounds using individual ROIs from separate hemispheres. Panel A illustrates the evaluation of reconstructed sounds from individual ROIs in the left hemisphere. Each ROI is represented by a unique color, and each dot signifies the mean identification accuracy across 150 test stimuli for each subject. The error bar indicates the 95% confidence interval. Panel B showcases the evaluation of reconstructed sounds from individual ROIs in the right hemisphere, following the same methodology as in Panel A.

failed to match Conv5's performance in terms of Mel-spectrogram and lower-level representations. These findings suggest that incorporating DNN features with a temporal dimension significantly enhances the accuracy of lower-level features in the sound reconstruction process.

7.4 Discussion

This chapter delved into the roles of hierarchical auditory areas and DNN features in our sound reconstruction process. In contrast to the indistinguishable feature decoding performance across individual ROIs discussed in Chapter 4, our sound reconstruction exhibited clear differences amongst individual ROIs. More specifically, the core region outperformed peripheral areas in terms of low-level representations and acoustic features' identification performance. Yet, this performance dwindled as I moved towards the peripheral regions. Conversely, the high-level representations demonstrated a slight improvement in the PBelt and A4 compared to A1. These findings align with prior encoding studies, such as the one by Kell et al. (2018), which posits that neurons in the primary auditory cortex's vicinity are more focused on local integration, reflecting a preference for generic acoustic representations. In contrast, neurons near nonprimary regions participate in longer-time integration and are more inclined towards category selectivity. These results highlight the potential of our proposed model to reconstruct the information encoded in individual ROIs.

In examining the reconstruction results from the different DNN layers, I found that the higher layers generally exhibited superior reconstruction performance. This trend was broadly consistent with decoding performance. However, despite showing similar decoding performance, the Conv5 and FC3 layers produced slightly different reconstructed sounds. Both layers achieved comparable reconstruction performance in the intermediate and high layers of the hierarchical representation, but when it came to lower layers or pixels of Mel-spectrogram and acoustic features, Conv5 outperformed FC3. These observations suggest that incorporating DNN features with a temporal dimension significantly enhances the reconstruction accuracy of lower-level features in the sound reconstruction process.

In summary, this chapter underscores the pivotal role that hierarchical auditory areas and DNN features play in the sound reconstruction process. I elucidate the importance of the parallel hierarchical structures that are present both in the human auditory system and DNNs. These structures can be harnessed to improve sound reconstruction efficiency, demonstrating the potential of our proposed model to reconstruct the information encoded in individual ROI.

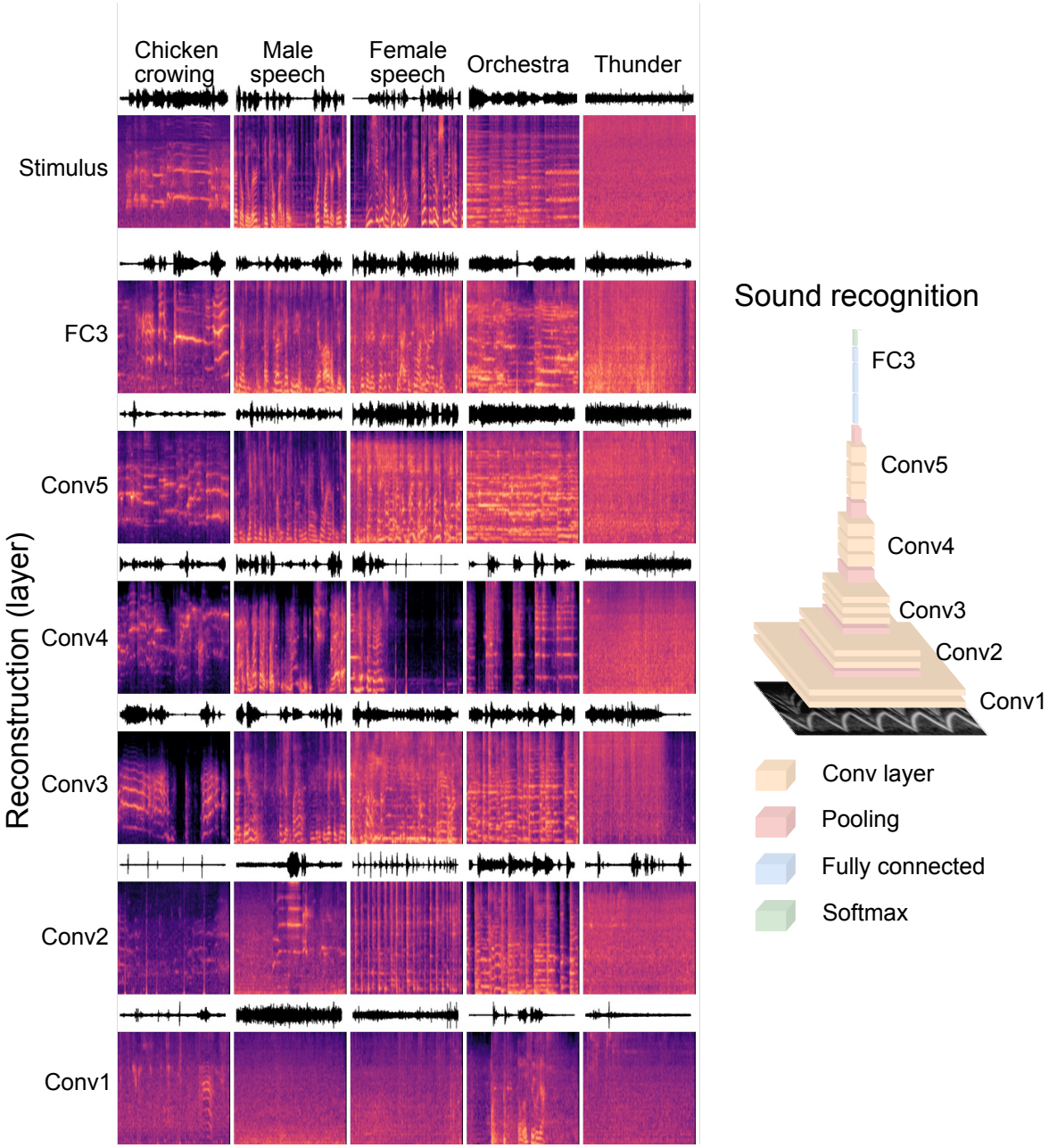

Fig. 7.5 Effect of DNN layers on sound reconstruction. The figure showcases reconstructed Mel-spectrograms resulting from different layers within the VGGish model. The topmost row displays the original Mel-spectrogram of the presented sound. Subsequent rows, two through seven, feature Mel-spectrograms reconstructed from various layers of the VGGish model, employing the AC data from subject S3.

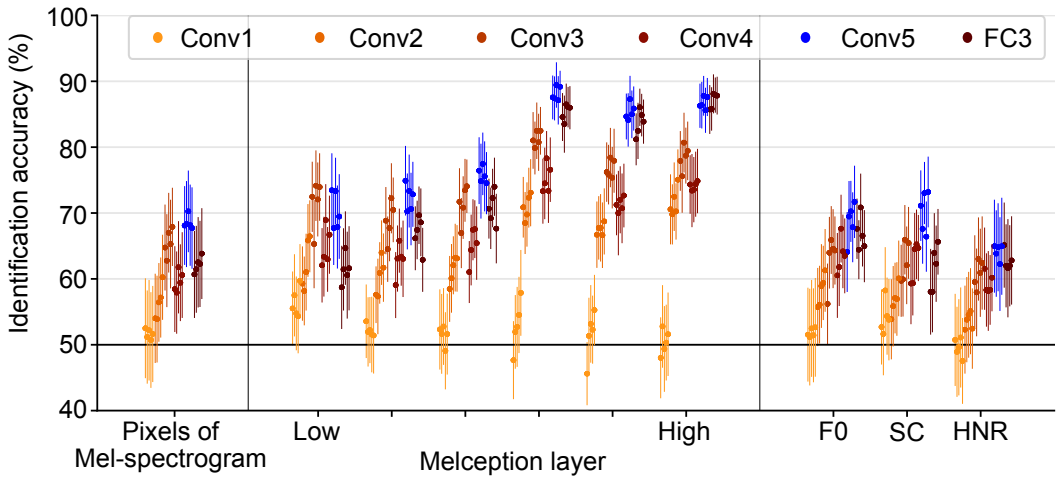

Fig. 7.6 Evaluation of reconstructed sound using different DNN layers. Each layer of the DNN is denoted by a unique color. Each dot in the plot symbolizes the mean identification accuracy derived from each subject, while the error bars denote the 95% confidence interval, estimated using 50 data points.

This finding contributes to our understanding of auditory processing and could pave the way for more advanced reconstruction models. Further investigations into these interrelations may uncover additional insights that can help to enhance this understanding and to refine our model further. Building on these findings, the chapter 8 will explore the versatility of our proposed model. Specifically, I will investigate whether our model can reflect subjective perceptual experiences in the reconstructed sounds.

Chapter 8

Auditory attention

8.1 Introduction

In a world filled with complex and overlapping sounds, our brains have the impressive capability of distinguishing individual sounds from a composite auditory input and selectively attend to specific stimuli. This phenomenon is often referred to as the "cocktail party problem"—a term that encapsulates our ability to concentrate on a single conversation amidst a cacophony of ambient noise. To address this problem, researchers have suggested two main solutions (McDermott, 2009). The first is a bottom-up approach that involves segregating auditory information from the mixed sounds. There is compelling psychological evidence supporting this mechanism. For instance, studies such as Bregman's Auditory Scene Analysis (Bregman, 1990), propose that our auditory system can segregate individual auditory 'objects' from the superimposed sound mixture based on physical characteristics such as pitch, location, and timbre. On the other hand, top-down modulation is a strategy where attention is directed towards familiar or important information while ignoring the surrounding noise. This mechanism has also been widely explored and demonstrated in research. For example, studies have found that meaningful content, such as one's own name, can capture attention even when unattended, suggesting the role of higher cognitive processes in auditory attention (Cherry, 1953). However, while both processes, bottom-up segregation and top-down modulation, are well-documented in isolation, the interaction between these two processes remains relatively unexplored. This raises questions about how our brains

integrate bottom-up and top-down information to efficiently process complex auditory scenes in real-world environments.

Recent advancements in neuroimaging techniques promise significant potential for tackling these challenges. For instance, Ding and Simon (2012b) utilized magnetoencephalography (MEG) and applied a multivariate analysis technique to decipher the neural encoding of attended and unattended speech. Their findings showed that cortical representations of the attended speech were notably enhanced compared to the unattended speech. In parallel, Degerman et al. (2006) used functional magnetic resonance imaging (fMRI) to delve into the neural mechanisms of selective auditory attention. In their study, participants listened to two concurrent auditory narratives and were instructed to focus on one. Analysis of cortical responses during these tasks revealed that selective attention modulated responses predominantly in the secondary auditory cortex, rather than the primary auditory cortex.

Simultaneously, the advent of machine learning has offered a novel perspective to this domain. Through these models, it's now possible to decode the attended speaker from a listener's brain activity, a stride that unlocks deeper understanding of the computational mechanisms of the cocktail party problem. O'Sullivan et al. (2015) conducted a study to determine if attentional selection in a cocktail party environment can be decoded from single-trial electroencephalography (EEG). They developed a novel decoding approach based on neural data and showed that attentional selection can indeed be decoded from single-trial EEG data. They discovered that there was a remarkable degree of trial-by-trial correspondence between the listener's attentional state and the pattern of neural responses. Their work provides evidence that EEG could be used as a non-invasive readout of a listener's focus of attention. In a follow-up study, O'Sullivan et al. (2017) examined how the brain could decode attentional selection in multi-speaker environments even without access to clean sources. They used a closed-loop experimental setup that tracked the EEG responses to two competing speech signals and demonstrated that EEG responses track the attended speaker and not the ignored one. This study established that our brains could segregate and select attention in complex acoustic environments even without clear, separate sources. Bednar and Lalor (2020) conducted a study to determine the brain's ability to locate different sound sources in a dynamic auditory scene. They utilized EEG data from participants who were tasked with tracking multiple moving sound sources. The researchers decoded the location of attended and unattended sound sources in real-time from the EEG data. Their study suggested that our brain contains detailed spatial information about multiple sound sources and can flexibly shift attention between these sources in complex auditory scenes. These studies highlight how

integrative use of advanced technologies could potentially revolutionize our understanding of selective auditory attention,

Despite these advancements, previous research primarily focused on classifying attended from unattended sounds using decoded auditory features, due to limitations in neuroimaging temporal resolution. Few studies have explored the direct reconstruction of sounds from neural activity under auditory attention task. In this chapter, I utilize a novel framework for sound reconstruction based on fMRI responses under selective auditory attention task. The objective is to verify whether these reconstructed sounds mirror subjective perceptual experiences. The study of auditory attention is pivotal to understanding how the brain navigates our intricate acoustic environment. The contents of this chapter is based on the Materials and methods: *Experimental design* and Results: *Attention* of (Park et al., 2023).

8.2 Methods

8.2.1 Subjects

Five non-native English-speaking subjects with normal hearing abilities, including one female subject, participated in our study. The average age of the subjects was 27.6 years. One subject (S1) was used for the exploratory analysis to establish the reconstruction model, while the other four subjects were used to validate the results independently. All subjects provided informed consent prior to the scanning sessions. The study protocol received approval from the Ethics Committee of the Advanced Telecommunications Research Institute International (approval no: 106) and was conducted following the principles of the Declaration of Helsinki.

8.2.2 Sound stimuli

During our auditory selective attention experiment, I simultaneously presented pairs of sounds from distinct categories. In the pilot study, I used 10-second stimuli from various categories: animal sounds, environmental sounds, music, and male and female speech. I randomly selected two representative stimuli from each category, totaling 26 pairs. However, during our analysis, I decided to exclude any pairs that included environmental sounds due to the inherent difficulty in focusing on these sounds when they overlapped with others.

In the subsequent experiments conducted with four other subjects, I stuck to the four remaining categories: male speech, female speech, animal sounds, and music. From each category, I selected two exemplars that showed the highest reconstruction performance during the pilot study with subject S1. I then created all possible pairs of these sounds, excluding pairs from the same category, to form a set of 24 stimuli.

During each attention trial, subjects were instructed to focus on one of the two simultaneously presented sounds, yielding 48 unique attention trials for the dataset. To ensure consistency, the energy levels of these superimposed sounds were normalized to match each other.

For the attention experiment, our aim was to conduct a binomial test to ascertain whether the proportion of correct identification surpassed the chance level of 50%. The sample size, decided prior to the experiment, was set at N = 48, larger than the required sample size to detect an effect size of g = 0.2 (correct rate = 0.7) at a significance level of 0.05 (N = 37). Despite using data samples from a 4-second time window for decoder training and reconstruction, statistical evaluations were conducted on data points corresponding to 8-second stimulus blocks. This decision was to accommodate the lack of independence among the three samples derived from an 8-second stimulus block. In the analysis of the attention test samples, the binary outputs from the three samples of the same stimulus and attention condition were pooled by majority voting to define a single binary data point. This resulted in 48 data points in each condition per subject.

8.2.3 Experimental design

In the selective auditory attention experiment, subjects were asked to focus on one out of two overlapping sound stimuli played simultaneously, under diotic listening conditions. Each participant took part in a single session which comprised eight functional runs, and each run included 48 attention trials.

In each trial, a pair of sound categories played simultaneously for a duration of 8 seconds. I presented visual cues that represented both the target (attended) and the non-target (unattended) sound categories. The word that corresponded to the target sound category was distinctly marked with a dash ("-"), signaling it as the sound to be focused on.

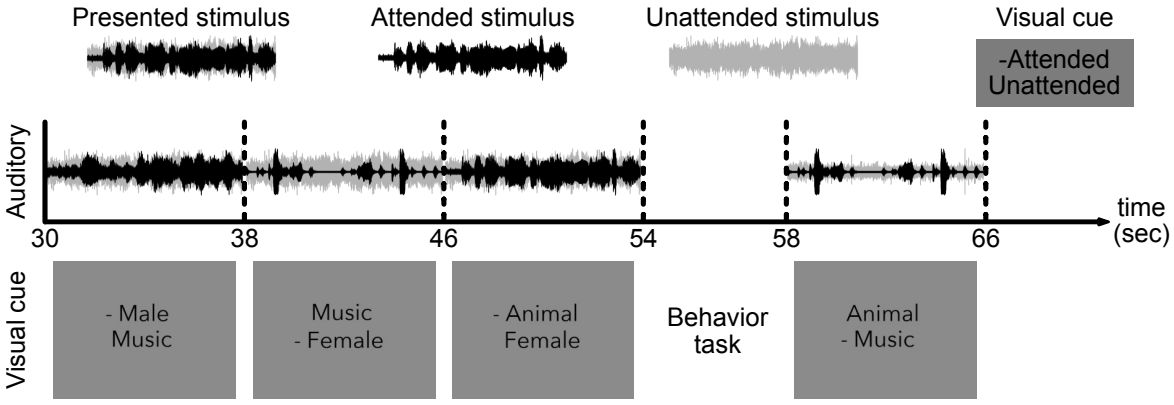

Fig. 8.1 Schematic of the experimental design for the selective auditory attention experiment. The diagram presents an 8-second period during which two different sound categories play concurrently. To assist the participant, visual cues corresponding to both the target (attended) and the non-target (unattended) sound categories are displayed. A dash ("-") distinctly marks the cue associated with the target sound category, indicating the sound the participant should concentrate on.

To ensure the participant's active involvement, I interspersed a behavioral task randomly four times in each run. In this task, subjects were asked to identify the sound category they were instructed to focus on in the immediate preceding trial.

8.3 Results

8.3.1 Feature decoding under attention task

To ascertain whether the reconstructed sounds encapsulate subjective listening experiences, fMRI responses were collected under a selective auditory attention task, colloquially referred to as "cocktail party conditions." This scenario mimics our auditory system's complex ability to separate a single sound from an array of sounds within a noisy environment. The goal was to determine if the decoded DNN features and the resulting reconstructed sounds could reflect this selective attention process.

To that end, decoded DNN features were derived from the fMRI responses collected during a selective auditory attention task. This process used a brain decoder that was trained under natural sound listening conditions. These decoded features were then utilized to create sound reconstructions under auditory attention tasks.

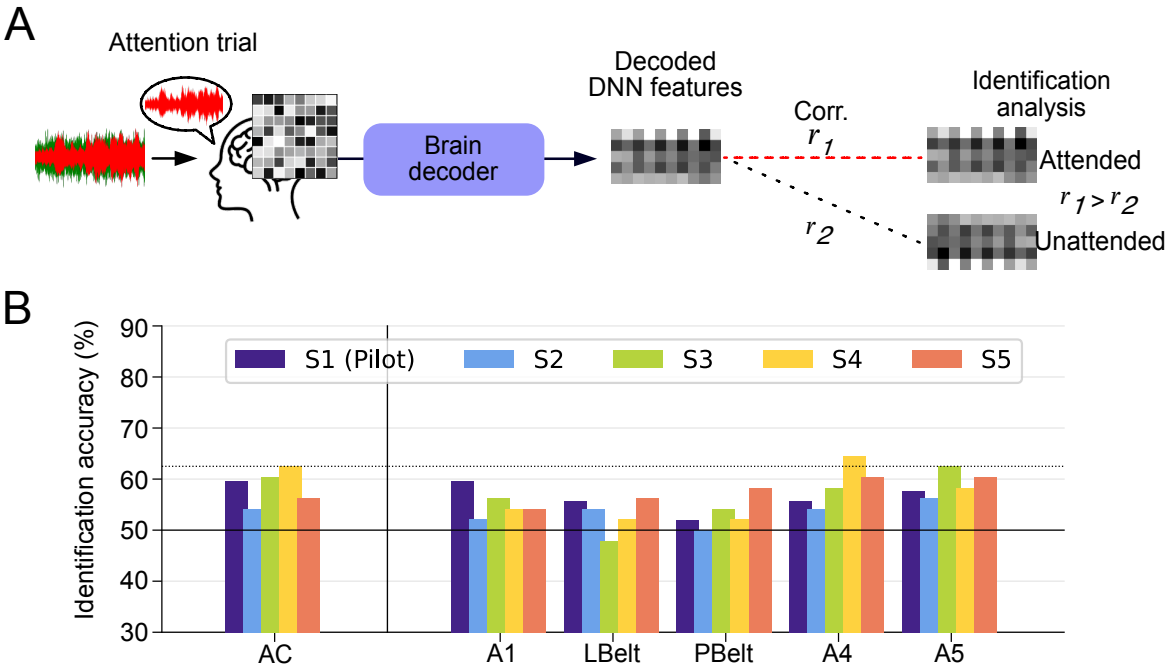

Fig. 8.2 Identification analysis of attended stimuli based on decoded DNN features. (A) Schematic of feature decoding and evaluation of selective auditory attention. Decoded features from the brain response during the selective auditory attention task was computed by decoder trained on passive listening conditions. Subsequently, decoded features was assessed by identification performance by comparing correlation coefficients with DNN features of attended and unattended stimuli. (B) Identification accuracy of attended stimuli from the unattended. Each bar represents the mean identification accuracy with an error bar indicating the 95% confidence interval estimated with 48 data points. Each subject is represented by a different color.

The decoding performance under the selective auditory attention task was evaluated via an identification analysis. This analysis differentiated the attended stimuli from the unattended stimuli based on their correlation with decoded DNN features (Figure 8.2A). As shown in Figure 8.2B, decoded features from the AC successfully identified the attended stimuli from the unattended stimuli with approximately 60% identification accuracy. When examining the results for individual ROIs, most subjects showed a mean identification accuracy of less than 60% in the A1 region. However, three subjects exhibited a mean identification accuracy of more than 60% in the A5 region. These results highlight individual variability, but the overall trend indicates significant sound identification capability from decoded DNN features, particularly in the higher auditory regions.

8.3.2 Sound reconstruction under attention

To generate the reconstructed sounds under auditory attention tasks, the decoded features were employed. Figure 8.3 encapsulates the resultant sounds when the subjects attended to one of the two superimposed sounds, presented as identical stimuli. Generally, the reconstructed sounds and Mel-spectrograms show a stronger resemblance to the attended stimulus than to the unattended one, even though they largely mirror the actual combined stimuli. Notably, when attention is directed towards speech, the resulting reconstructed sounds manifest a harmonic structure akin to the spectral pattern of speech.

An identification analysis based on the audio features extracted from the reconstructed sounds was conducted to evaluate the capability of these sounds to distinguish the attended stimulus from the unattended one (Figure 8.4). This process involved comparing the correlation of attended and unattended stimuli with the auditory features extracted from the reconstructed sounds. A sound was correctly identified if it exhibited a higher correlation with the attended stimulus than the unattended one. From three samples of the same stimulus and attention task, binary results were collected and analyzed by majority voting to compute a single binary data point, culminating in 48 binary data points. The mean identification accuracy was then computed from these 48 data points for each condition and subject, and this was compared with the level of significance derived from a binomial test.

Performance was evaluated using three types of extracted features, mirroring the analysis process used for single sound samples. As displayed in Figure 8.4, when employing Mel-spectrogram and low-layer features, the reconstructed sounds of most subjects showcased an identification accuracy below 60%. Nevertheless, a performance improvement was noted with the intermediate and high-layer features in the hierarchical representation, raising identification accuracy to around 60% for the majority of subjects. In the intermediate and higher layers, a few subjects even displayed performance surpassing the significance threshold. However, when evaluating the reconstructed sounds based on acoustic features, the identification accuracy for most subjects fell below 60%, thus not reaching the level of statistical significance.

8.3.3 Sound reconstruction from individual ROIs

Further in our analysis, a reconstruction examination was carried out using individual auditory regions of interest (ROIs). Figure 8.5 captures the summarized reconstructed

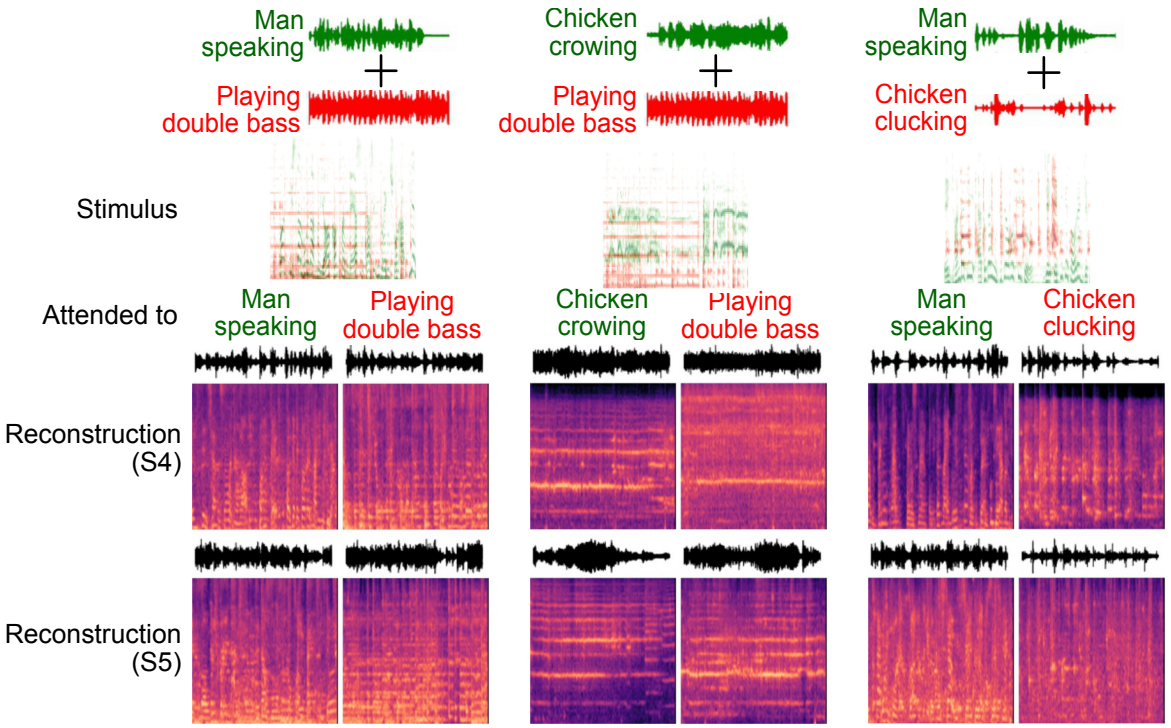

Fig. 8.3 Reconstructed Mel-spectrograms under selective auditory attention tasks. The upper panel showcases the original Mel-spectrogram of the sound that was presented to the subjects during the task, wherein they were tasked with focusing on one specific sound within an array of superimposed auditory stimuli. The lower panel, conversely, demonstrates the reconstructed Mel-spectrograms from two distinct subjects (S4 and S5), generated using the AC and onv5 responses.

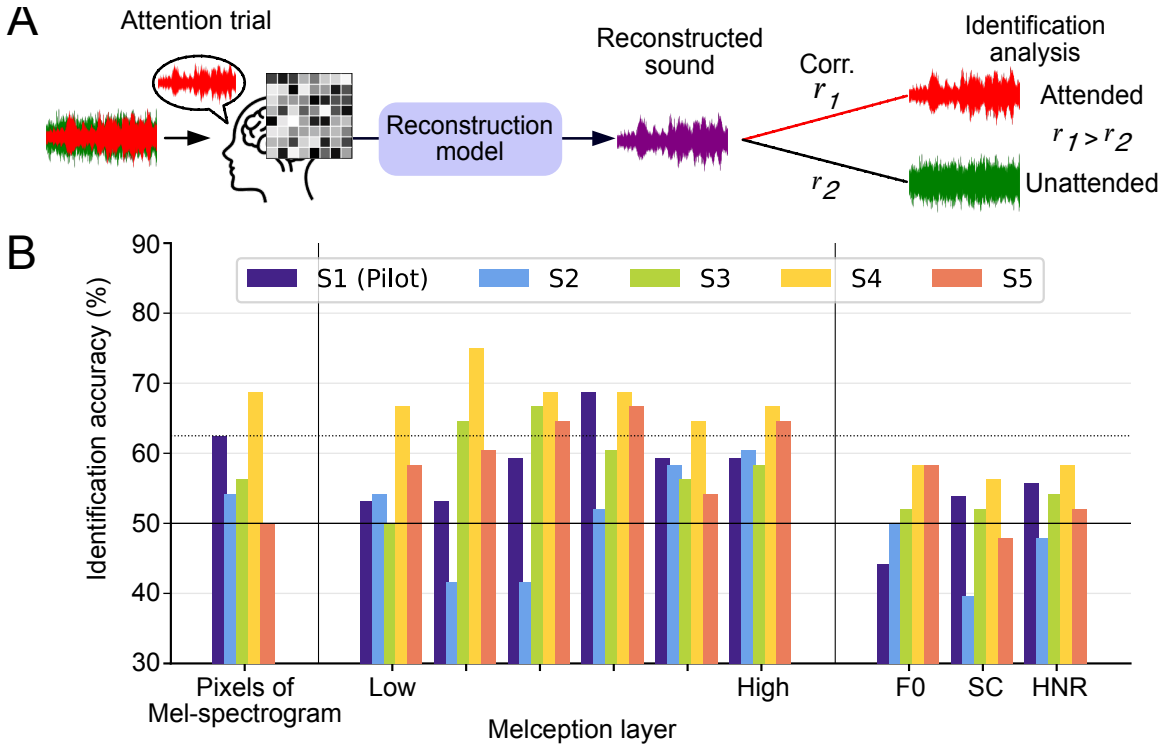

Fig. 8.4 Evaluation of sound reconstruction under selective attention tasks. (A) Schematic representation of the process for reconstruction and evaluation during selective auditory attention. The process involved comparing the fidelity and quality of the attended and unattended stimulus reconstructions, to evaluate how accurately attention was represented. (B) Results of the evaluation of sound reconstruction under selective attention tasks. Identification results from three samples of the same stimulus and attention task were collated through majority voting, generating 48 data points for each condition and subject. Each bar corresponds to the mean identification accuracy from these 48 data points. The dashed line illustrates the level of accuracy deemed significant ($p < 0.05$) in the binomial test ($N = 48$). Please note, stimuli that were not applicable in the calculation of F0 and HNR were excluded from the statistics. In such instances, a higher level of significant accuracy is necessitated, which is not depicted here.

sounds from each ROI during the selective attention task. While the derived sounds from different auditory ROIs displayed general similarity, the spectral patterns echoing the attended stimuli, especially the harmonic spectral pattern of speech, were more pronounced in the reconstructed sounds from boundary regions, notably A4 and A5, compared to those from the core regions.

Figure 8.6 provides a quantitative assessment of the reconstructed sounds during the selective attention task across various ROIs. Evaluations based on Mel-spectrograms and low hierarchical representations revealed no noticeable variation in identification accuracy among auditory ROIs. However, a consistent incremental trend in identification performance from A1 to A5 was observed when evaluations were conducted using intermediate and high hierarchical representations in the reconstructed sounds from three subjects. In contrast, evaluations founded on acoustic features did not reveal significant differences between ROIs.

These observations hint that subjects might have given more weight to the categorical features of the attended stimuli. Moreover, the higher auditory regions seemed to be more intimately tied to attentional modulation.

8.4 Discussion

This chapter explores the ability of reconstructed sounds to encapsulate subjective listening experiences by collecting fMRI responses under cocktail party conditions, which showcase our auditory system's remarkable capacity to isolate specific sounds amidst numerous sound sources. The aim is to investigate if decoded DNN features and reconstructed sounds can accurately reflect the process of selective auditory attention. Our results suggest that sounds reconstructed from the auditory cortex mirror the attended sound more than the unattended one, but overall, they tend to resemble the perceived superimposed sounds. This implies that while bottom-up processing is predominant in the auditory cortex, the ability to distinguish attended sounds from unattended ones improves at higher levels of the auditory hierarchy, suggests that subjects were more attentive to the category-specific elements of the stimuli they were asked to focus on.

Decoded DNN features are extracted from fMRI responses during a selective auditory attention task, using a brain decoder trained under natural sound listening conditions. These decoded features are then utilized to generate sound reconstructions during auditory attention tasks. The results reveal that reconstructions derived from auditory cortical activity during

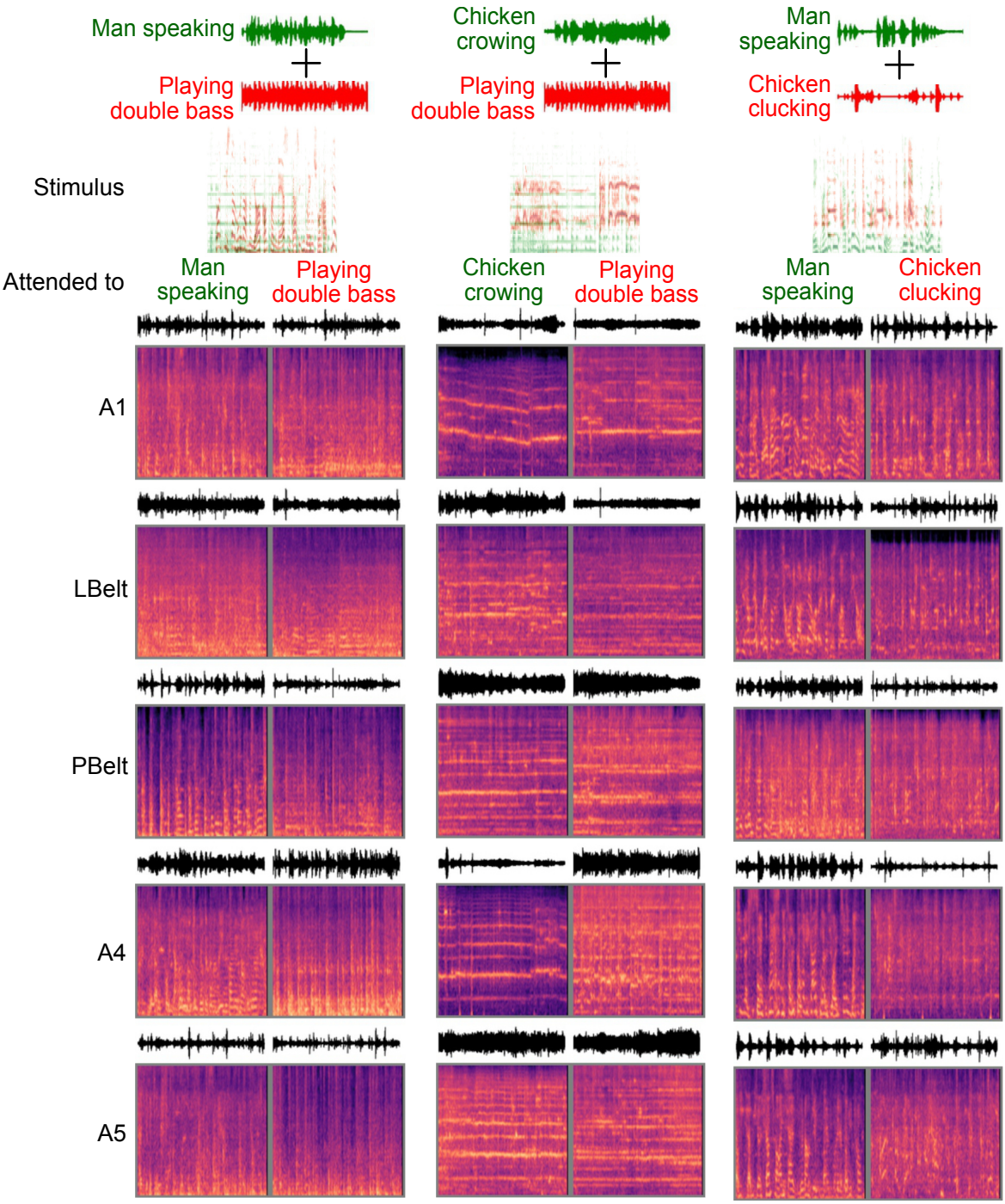

Fig. 8.5 Reconstructed sound from Individual ROIs under selective auditory attention tasks. The top panel displays the Mel-spectrogram of the sound stimulus used during the task, where subjects were instructed to concentrate on a specific sound within a superimposed soundscape. The bottom panel showcases the Mel-spectrograms reconstructed from each individual ROI from subject S4, using the conv5 layer.

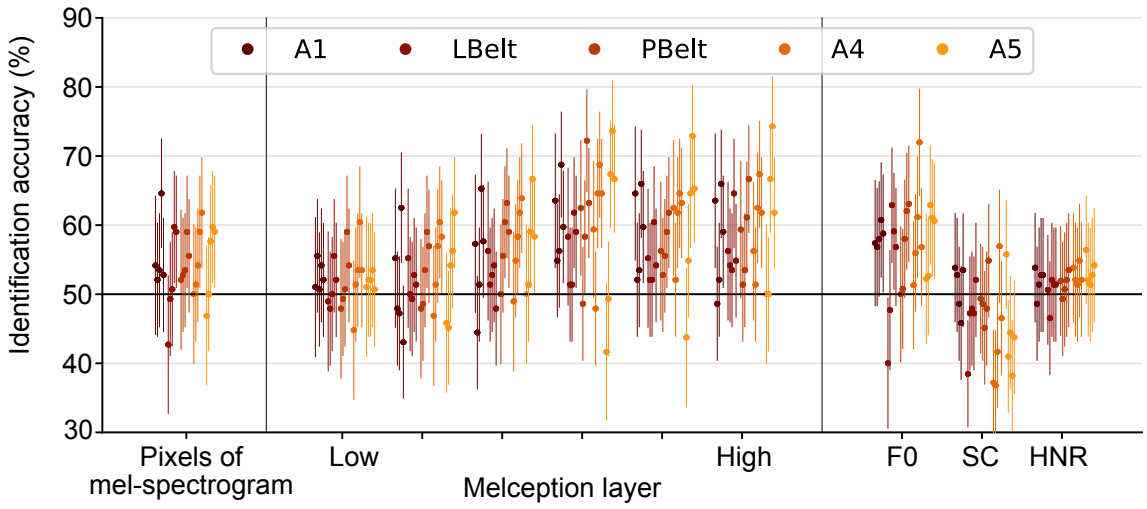

Fig. 8.6 Evaluation of the reconstructed sounds from individual ROIs. Each ROI is represented by a unique color. Each dot corresponds to the mean identification accuracy derived from each subject, calculated from 48 data points. The dashed line indicates the accuracy level deemed significant ($p < 0.05$) by the binomial test.

top-down, selective attention are closer to the attended sound than the unattended one. Moreover, when differentiating attended from unattended stimuli based on the reconstructed sound, the identification accuracy was higher in the boundary area compared to the core area. This difference was significant for three out of four participants, demonstrating the robustness of our proposed framework for reconstructing subjective auditory content.

To validate whether our reconstruction accurately reflects subjective listening experiences, I expanded our model to simulate cocktail party conditions. The results indicated a tendency for the reconstructed sounds to resemble the attended sound more than the unattended one, in line with previous studies suggesting that decoded auditory features, such as envelope, spectrogram, and trajectory, derived from brain responses during attention tasks, better reflect the attended sound (Bednar and Lalor, 2020; Ding and Simon, 2012a; O'Sullivan et al., 2015; O'Sullivan et al., 2017). Despite the limited temporal resolution inherent to fMRI, our study illustrates the possibility of reconstructing sound during attention-demanding tasks. Furthermore, our reconstructed sounds were more similar to the attended sound than the unattended one, providing evidence that our reconstruction method echoed the perceptual content of subjective listening experiences under complex soundscapes.

However, the results also highlight the role of individual differences in personal experience and task ability in decoding performance and reconstruction quality of selective

attention. For instance, one participant displayed chance-level performance across all auditory ROIs, while another participant with musical experience showed relatively higher identification performance. This implies that factors such as individual expertise may influence the decoding accuracy of auditory attention. These findings underscore the necessity to consider individual differences in future research ventures exploring the psychological and neural correlates of attentional modulation. This approach will deepen our understanding of the neural mechanisms underlying selective auditory attention.

Chapter 9

General discussion

9.1 Summary

In this study, I introduce a novel approach for unconstrained neural sound reconstruction from fMRI activity, utilizing "brain-like" auditory features and an audio-generative model. In chapter 4, I conducted a decoding analysis using various auditory features from anatomically defined regions of interest (ROIs) in the auditory cortex. This analysis led us to identify "brain-like" features derived from the sound recognition DNN model that mirror the hierarchical structure of the auditory system.

In chapter 5, I employed an audio-generative transformer to convert these decoded DNN features into high-fidelity audio signals. This approach enabled us to predict a concise representation of a Mel-spectrogram, a codebook representation, based on the decoded DNN features. Our model was capable of reconstructing complex spectral-temporal patterns, broadly maintaining content and quality similar to the actual sound stimulus. This finding was validated through both qualitative and quantitative evaluations. However, I acknowledge room for improvement, specifically in detailing the content within speech or music sequences.

In chapter 6, I found that our sound reconstruction remained robust even when certain categories were not present during the training phase. This finding suggests that our proposed model does not merely match brain data to training examples. Instead, it appears to synthesize sounds using a combination of elemental features.

Chapter 7 underscores the critical role that hierarchical auditory areas and DNN features play in the sound reconstruction process. I elucidate the importance of parallel hierarchical structures present in both the human auditory system and DNNs. These structures can be harnessed to improve the efficiency of sound reconstruction, demonstrating the potential of our proposed model to reconstruct the information encoded in individual ROIs.

In chapter 8, I extended our study to "cocktail party conditions", illustrating the potential of our model to reconstruct the subjective content of top-down auditory attention. The reconstructed sounds were more likely to reflect the attended stimulus rather than the unattended one. These findings provide evidence that our reconstruction method mirrored the perceptual contents of subjective listening experiences under complex soundscapes.

9.2 Hierarchical nature of brain auditory processing

The reconstruction model proposed in this research makes use of only the spatial patterns of voxels, without time information, to reconstruct sound. This approach builds upon the findings of several previous studies. For instance, research has indicated that the brain is capable of processing temporal information in a compact form and that fMRI is able to capture this compact information. In addition, the detailed patterns of these voxels could potentially facilitate the decoding or reconstruction of more nuanced temporal information. For instance, the study by Hasson et al. (2008) demonstrated that the hierarchical processing in visual system to compactly process temporal information from early visual area to higher area when subject see the silent movies. Moreover, the research conducted by Nishimoto et al. (2011) showed that the temporal information of natural scenes a person was observing could be decoded from voxel patterns in fMRI data. Furthermore, Santoro et al. (2017) developed a computational model that decoded the physical features of natural sounds from high spatial resolution 7T fMRI responses. These findings suggest that it might be possible to reconstruct unconstrained sounds without needing an exact temporal alignment between neural recordings and auditory stimuli.

In our analysis, we identified "brain-like" features from sound recognition deep neural networks (DNNs) that exhibit hierarchical processing analogous to the human auditory system. These features consistently outperformed other auditory features in decoding performance, aligning with previous encoding analyses of systematic model-brain correspondence (Tuckute et al., 2023). Interestingly, we did not observe a clear hierarchical correspondence between individual auditory regions of interest (ROIs) and the layers within the DNN model

in our decoding performance. This contradicts what previous encoding analyses with DNNs have suggested (Kell et al., 2018). Recent studies using intracranial recordings have pointed towards a distributed functional organization within the human auditory cortex, suggesting a potential parallel information processing across the auditory cortex (Hamilton et al., 2021; Nourski et al., 2014). This aligns with our results, which indicate that auditory cortical ROIs participate in both distributed and hierarchical processing. While our decoding analysis utilized anatomically defined ROIs, future studies could benefit from employing voxels specified by tonotopic or encoding analyses, providing further insight into the auditory hierarchy and its representations.

Despite no discernable difference in feature decoding performance across individual ROIs, we observed variations among individual ROIs concerning the quality of the reconstructed sound. We found that the core region showed superior identification performance in low-level representations and acoustic features, while this performance declined when moving towards the peripheral areas. In contrast, high-level representations performed slightly better in the PBelt and A4 compared to A1. This pattern is consistent with previous encoding studies (Norman-Haignere et al., 2022), which propose that neurons near the primary auditory cortex engage in local integration and show selectivity towards generic acoustic representations. Simultaneously, neurons near non-primary regions partake in longer-time integration and exhibit a preference for category selectivity. These findings suggest the potential of our proposed model to reconstruct the information encoded in individual ROIs.

However, in our pursuit to improve the signal-to-noise ratio (SNR) of fMRI samples, we computed each fMRI sample by averaging three consecutive functional volumes from the onset of the stimuli and taking the trial average of the same test sound stimuli. These methods enhanced both the decoding performance of features and the quality of sound reconstruction. Interestingly, it was observed that sounds generated from non-averaged fMRI samples still retained perceptual content and quality. This observation suggests that the application of advanced decoding methods could provide possible avenues for improving single-trial reconstructions. These advancements could have far-reaching implications, particularly for real-time sound reconstruction and applications in the domain of brain-machine interfaces (BMIs).

9.3 Externalization of auditory perceptual experiences

In chapter 8, we delved further into the capability of reconstructed sounds to encapsulate subjective listening experiences. Our model was expanded to cover "cocktail party" conditions to examine whether our reconstructions mirrored subjective listening experiences. We discovered that the reconstructed sounds tended to reflect the attended sound more than the unattended one. This outcome aligns with previous studies that proposed decoded auditory features, such as envelope (Ding and Simon, 2012a; O'Sullivan et al., 2015), spectrogram (O'Sullivan et al., 2017), and trajectory (Bednar and Lalor, 2020), derived from brain responses during attention tasks, are more akin to the attended sound rather than the unattended one. Despite the limited temporal resolution inherent to fMRI, our study demonstrated the feasibility of sound reconstruction during attention-demanding tasks. Furthermore, our reconstructed sounds were more reflective of the attended sound than the unattended one. These findings provide evidence that our reconstruction method accurately mirrored the perceptual contents of subjective listening experiences under complex auditory environments.

Notably, in our auditory attention analysis, we observed individual differences in reconstruction performance. Subject S4, for instance, substantially outperformed other subjects in feature decoding and sound reconstruction. This subject, who has an extensive musical background, reported ease in focusing on a specific sound in a multi-speaker environment. This finding suggests that personal experiences and task strategies could potentially influence the decoding performance of selective attention. Hence, it is beneficial to consider individual differences and task strategies in future research that seeks to explore the psychological and neural correlates of attention modulation.

9.4 Bridging the gap between sound and neuroimaging using DNN

A notable finding in our research is the effectiveness of codebook representation and audio transformer in bridging the gap between high-dimensional spectrograms and the low temporal resolution of neuroimaging data. The SpecVQGAN effectively performed dimensionality reduction and served as a perceptually rich prior, facilitating sound reconstruction (Dhariwal et al., 2020; Iashin and Rahtu, 2021; Liu et al., 2021; Zhao et al., 2020). Visualization of patch patterns associated with each code revealed that the SpecVQGAN captures a broad

range of spectral and temporal characteristics of sounds. This finding supports the model's versatility in reconstructing a variety of sounds from brain responses.

However, as the performance of reconstruction using only the codebook representations was suboptimal, it is clear that identifying brain-like features is critical. The decoding of these features from the audio transformer significantly enhanced the temporal resolution of the reconstructed sound by adapting to the time-varying spectral characteristics captured by the codebook. Especially, the audio transformer, with its inherent self-attention mechanism, allows it to effectively model the dependencies between different parts of the sound signal, regardless of their position in time. This characteristic endows the transformer with the capability to understand the intricate, time-varying spectral patterns in the sound data, which is instrumental in successfully decoding the temporal dynamics from fMRI data. This suggests that the audio transformer not only aids in the sound reconstruction process but also facilitates the temporal decoding of brain activity, providing a closer approximation to real-time auditory experience. The key here is the transformer's ability to adapt and reconstruct the temporal dynamics encoded in the brain's response, which is an aspect not captured by the codebook alone. This underscores the complementary roles played by the codebook and the audio transformer in achieving high-quality sound reconstruction.

Looking forward, the effectiveness and generalizability of our model open up potential avenues for future research. One such direction could be the exploration of our model's ability to reconstruct pure tones. Furthermore, the unique insight provided by the codebook could be exploited to understand better and model the neural processes underlying auditory perception.

9.5 Future applications

This thesis has demonstrated our model's capability in reconstructing arbitrary sounds, effectively capturing both their perceptual content and quality from human brain activity. As I move forward, I extend our investigations into two key potential areas: the domain of musical experiences and the intriguing interplay between vision and audition.

Firstly, I will explore the application of these methods in the reconstruction of musical experiences. By leveraging diverse auditory stimuli, specifically focusing on drum beats and rhythm parts, I will engage a broad range of musical experiences across various genres

and instruments. This will serve to further test our model's capabilities and expand our understanding of how complex musical sounds are processed and represented in the brain.

Secondly, I will delve into the fascinating crossmodal interactions between vision and audition, particularly the phenomenon of sound-induced visual reconstruction. By employing state-of-the-art deep neural network models, I aim to investigate how auditory stimuli can influence the reconstruction of visual experiences, thus shedding light on the depth and extent of intermodal perceptual interactions within the human brain.

Through these explorations, I hope to not only showcase the potential versatility and applicability of our proposed sound reconstruction model but also to further our understanding of the brain's processing and representation of complex sensory information. This will pave the way for future developments in neuroscience, psychology, artificial intelligence, and multimedia arts, demonstrating how our model can serve as a powerful tool in these fields. I invite readers to join us as I venture into these exciting new territories in the following sections.

9.5.1 Music loop

One promising application of our sound reconstruction model lies in the arena of music synthesis from brain activity. This exciting intersection of neuroscience and music technology presents a unique opportunity to decode and recreate the intricate neural processes that underpin our musical experiences.

In the past, numerous studies have made significant strides in exploring the neural basis of music perception and production. For example, Alluri et al. (Alluri et al., 2012) conducted an investigation into the neural underpinnings of timbral, tonal, and rhythmic features of a naturalistic musical stimulus. Their findings suggest a broad brain network involved in the processing of individual acoustical features during the listening experience, encompassing cognitive, motor and limbic brain circuitry. This work provides a comprehensive insight into the neural representation of different musical features, emphasizing the importance of considering music as a multi-faceted, holistic experience. Furthermore, (Sturm et al., 2015) explored the neural correlates for the differential perception of chord progressions and note onsets in music, and the effects of these on perceived musical tension. multivariate regression-based method successfully extracted onset-related brain responses from ongoing EEG data, highlighting the dynamic relationship between the stimulus's sharpness, spectral

centroid, rhythmic complexity, and the listeners' EEG response. These findings not only underline the utility of individual EEG responses in assessing musical feature perception but also contribute to our understanding of the factors that influence musical tension. However, the task of reconstructing musical experiences from human brain activity itself has historically been viewed as a challenging endeavor due to the intricate nature of these experiences.

By applying our sound reconstruction model to music, I aim to delve deeper into these complex auditory experiences. From perceived melodies to rhythm patterns, I aim to decode these experiences from fMRI responses and transform them into high-quality musical sounds. This approach promises to extend our understanding of the neural underpinnings of music and open up exciting prospects in a variety of fields, including music technology, neuroscience, neurorehabilitation, and assistive technologies.

Methods

Subject For this preliminary analysis, I engaged the participation of a subject (S1) who had previously been involved in the exploratory stage of establishing our sound reconstruction model. This participant agreed to further contribute to this new phase of research. Before the scanning sessions began, the participant was fully informed about the purpose and procedures of the study, and they provided their informed consent. Our study protocol was approved by the Ethics Committee of the Advanced Telecommunications Research Institute International (approval no: 106), ensuring our adherence to the ethical guidelines and principles established in the Declaration of Helsinki.

Stimuli I selected rhythmically music pieces from the Apple Loop dataset (all royalty-free musical loops copyright © 2011, Apple Inc). These stimuli underwent meticulous assessment by human listeners to ensure sound quality. For the training set, I chose 200 unique rhythmic patterns, and for the test set, I selected 50. All audio clips incorporated into the fMRI dataset were resampled to a frequency of 22050 Hz, center-cropped to a duration of 8 seconds, and normalized to ensure equivalent energy levels.

fMRI experiments In the fMRI experiments, subjects passively listened to various audio clips of natural sounds. I recorded whole-brain fMRI responses while subjects listened to 200 stimuli for the training dataset and 50 for the test dataset. Each subject underwent a training

session followed by a separate test session. Each session included eight functional runs, none exceeding 90 minutes. Each run started with a 30-second rest period, followed by 55 blocks of 8-second stimulus presentations (consisting of 50 unique sound stimuli and five randomly interspersed behavioral task blocks), and ended with a 10-second rest period. This resulted in a total run duration of 8 minutes. To maintain subject concentration, I incorporated a one-back repetition detection task in which subjects were required to press a button if the subsequent stimulus was identical to the previous one. These repetition blocks (five per run) were excluded from the analysis. Experiments for all training and test sets were repeated four times.

Similar to the sound reconstruction experiments, I adjusted the preprocessed functional data by shifting them forward by 2 seconds to account for the hemodynamic delay. To augment the number of available data samples, I slid a 4-second time window across the original 8-second stimulus at 2-second intervals. For each 4-second sound stimulus, an fMRI sample was created by averaging the three consecutive functional volumes after the stimulus onset. This procedure resulted in three data samples from each original 8-second trial, yielding a total of 800 training samples. For the test datasets, I improved the Signal-to-Noise Ratio (SNR) by averaging the fMRI responses to identical sound stimuli across multiple repetitions. This method resulted in a total of 150 test samples (50 stimuli $\times$ 3 samples per stimulus = 150 samples).

DNN models While our natural sound reconstruction model could generate sound that contains some perceptual contents belonging to the music category, it was missing details crucial for the reconstruction of music loops. To resolve this, I retrained our Deep Neural Networks (DNN) model using datasets from different rhythmic music pieces.

In this study, I employed the Expanded Groove Midi Dataset (EGMD) (Callender et al., 2020). The EGMD is an augmented version of the Groove Midi Dataset, containing approximately 45000 drum sound stimuli. I used this dataset to train our SpecVQGAN for the codebook and the Audio Transformer. For the Audio Transformer, I utilized the conv5_1 layer of a VGGish-ish model.

Following this, I used the fMRI dataset to train a decoder that predicts conv5_1 from the VGGish-ish, calculated from the fMRI data. For the reconstruction of sounds, the process starts with the brain decoder, which decodes DNN features from the fMRI responses in the

test dataset. The decoded DNN features are then transformed into codebook representations with the aid of the Audio Transformer. These codebook representations are subsequently converted back into Mel-spectrograms using a codebook decoder. Lastly, a spectrogram vocoder converts these Mel-spectrograms into audio waveforms.

Results and discussion

Our research involved conducting a sound reconstruction analysis of a music loop (Figure 9.1). The results indicate that the reconstructed spectral and temporal patterns exhibit a resemblance to the original stimuli, particularly the spectral patterns. However, the model was unable to capture precise beats, rhythms, and subtle variations in timbre. Quantitative evaluations demonstrated fidelity performances below 60% and quality evaluations, particularly when using acoustic features, were around chance level, at approximately 50%.

Despite these modest performance results, this application could still serve as a useful tool for unveiling how rhythmic patterns and temporal information are encoded within the brain. Even though the outcomes were not optimal in both quantitative and qualitative terms, this endeavor paves the way for further exploration into the neural mechanisms underlying our interactions with music. This application not only contributes to our understanding of the neural encoding of music but could also serve as the foundation for innovative applications in music technology and neurorehabilitation. The potential for these advances extends beyond the scope of scientific understanding, paving the way for novel musical compositions drawn directly from our neural responses to music.

9.5.2 Reconstruction of crossmodal interaction

Crossmodal interaction, a key concept in neuroscience and cognitive science, presents the idea that our perception isn't simply an amalgamation of independent sensory responses. Instead, it is an integrated process where information from diverse sensory modalities can reciprocally influence each other. This dynamic interaction between our sensory systems disrupts the traditional concept of independent sensory pathways, offering a more nuanced understanding of human perception.

This integrative sensory processing has been investigated in several studies. For instance, Meyer et al. (Meyer et al., 2010) employed multivariate pattern analysis of fMRI data to

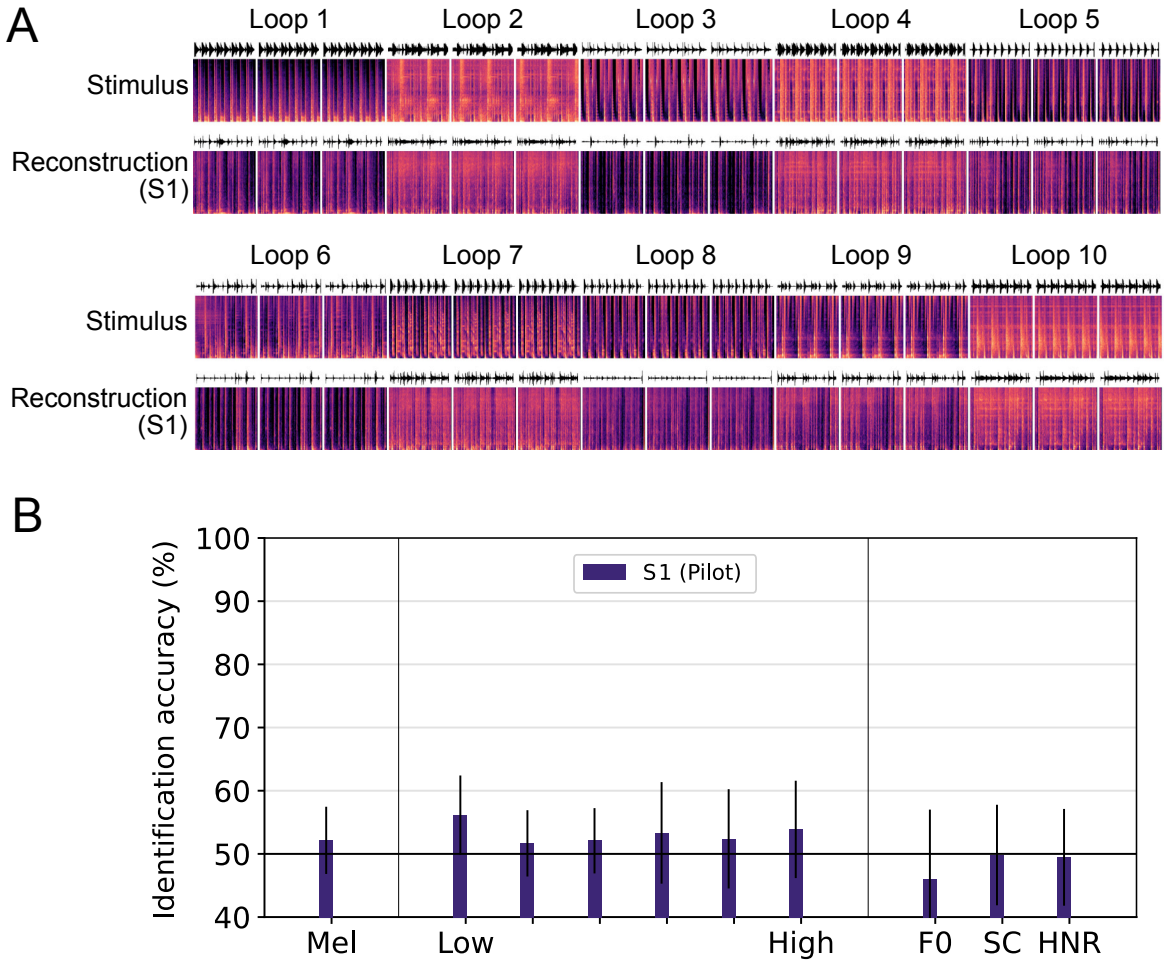

Fig. 9.1 Sound reconstruction of music loops. (A) Reconstructed Mel-spectrogram of music loops. The top row showcases the original Mel-spectrogram of the presented sound. The subsequent five rows display the Mel-spectrograms reconstructed from the AC and the conv5_1 layer. The three consecutive samples seen in the diagram originate from a single stimulus. (B) Evaluating reconstructed sounds. The fidelity and quality of the reconstructions were gauged through an identification analysis which incorporated Mel-spectrogram pixels, hierarchical representation, and acoustic features. Each bar in the diagram signifies the mean identification accuracy, while the error bar denotes the 95% confidence interval, which was estimated using 50 data points.

demonstrate that even in the absence of auditory stimulation, early auditory cortices' activity in humans was linked to the subjective experience of sound. Subjects were exposed to silent visual stimuli that suggested sound, and the resulting activity in the auditory cortex varied, reflecting sounds related to distinct categories such as different animals, musical instruments, and objects. This study underlined that early sensory cortex activity is more representative of perceptual experience than just sensory stimulation.

In a similar vein, Vetter et al. (Vetter et al., 2014) explored the impact of auditory perception and imagery on brain activity patterns in the early visual cortex without any feedforward visual stimulation. Their work showed that category-specific information from both complex natural sounds and imagery could be inferred from early visual cortex activity in blindfolded participants. Importantly, this coding of non-retinal information was common across actual auditory perception and imagery and could potentially be mediated by higher-level multisensory areas. The coding process was found to be resilient to minor shifts in attention and working memory, but it was susceptible to cognitively demanding visuospatial processing. Significantly, the information fed back to the early visual cortex was category-specific and could be generalized to sound exemplars of the same category. This suggested the presence of abstract information feedback rather than precise pictorial feedback. The study provides compelling evidence that the early visual cortex receives non-retinal input from other brain areas generated by auditory perception and/or imagery, bearing common abstract information.

These research findings highlight the intricate relationship between our sensory systems and the role of crossmodal interactions in shaping our perceptions. In the next section, I endeavor to extend these insights, applying the latest Deep Neural Network techniques to experiment with the generation of images from brain responses during auditory perception. The exploration of such crossmodal interactions holds substantial implications for understanding sensory processing disorders, advancing the development of sensory prosthetics, and enhancing human-computer interfaces.

Methods

Stimuli I utilized the same sound dataset for this study as used in the previous brain encoding analysis (Norman-Haignere et al., 2015). This dataset comprises 165 stimuli that are frequently encountered in everyday life, each with a stimulus duration of 2 seconds.

fMRI experiments In our fMRI experiments, I engaged a single participant (S1) who passively listened to various audio clips of natural sounds. I recorded whole-brain fMRI responses while the subject listened to a set of 165 unique sound clips. I utilized a sparse sampling technique in our fMRI experiments. This method ensured that no scans were conducted during the audio playback, creating an environment for the participant to listen to the sound stimuli without any scanner noise interference. The Repetition Time (TR) was set at 4.4 seconds, during which 2 seconds were allocated for scanning and 2.2 seconds for the stimulus playback while the scanner was inactive.

The experiment was conducted over eight sessions involving a single participant. Each session was composed of eight functional runs, with none exceeding a duration of 90 minutes. Every functional run began with a pre-rest period of 30.8 seconds, followed by a stimulus period that was repeated five times. Each repetition of the stimulus period consisted of a 2-second stimulus, a 0.2-second silence gap, and 2 seconds of scanner time, making up a total of 4.4 seconds per repetition. After the conclusion of the stimulus repetitions, the run ended with a post-rest period of 13.2 seconds. The experiments for all the training and test sets were performed four times. In the preprocessing stage of the functional data, I made adjustments by shifting the preprocessed functional data forward by 4.4 seconds (1 functional volume), to account for the hemodynamic delay.

DNN models The DNN model used in this study is rooted in the framework of the Contrastive Language–Image Pre-training (CLIP) model (Radford et al., 2021). The original CLIP model is designed to capture the relationships between image and text data by training on a vast number of text-image pairs. The model uses noise-contrastive estimation where text-image pairs from the same sample are used as positive examples, and all other pairs in the same batch are treated as negatives. This creates a shared embedding space where both image and text representations are projected, enabling the model to map the continuous conceptual space of images to the discrete symbolic space of text. This has resulted in a successful application in tasks such as zero-shot classification and cross-modal retrieval, and has also been extended to cross-modal generation.

The unique feature of Wav2CLIP is that it extends the CLIP's capabilities to include the audio modality (Wu et al., 2021). The central principle here is the creation of shared representations across different modalities - audio, image, and text - by mapping them to a common latent space. This multi-modal capability is realized by training the model on

a variety of tasks across different modalities using a vast dataset derived from the internet, with a contrastive loss function being used to distill image embeddings from CLIP into audio encoders. The result is a powerful model that is adept at handling tasks involving the interaction and integration of multiple sensory modalities, making it an ideal tool for exploring cross-modal interactions.

The architecture of Wav2CLIP includes additional multi-layer perceptron (MLP) projection layers, adding flexibility to the distillation process. Notably, the weights of the original CLIP model are kept frozen throughout distillation. The image encoder is then traused as a frozen feature extractor, allowing audio encoder to be trained to have The shared embedding representations. These in this model are not only versatile and rich but also resilient against minor shifts in attention and working memory, making it well-suited for demanding cognitive tasks such as those investigated in our study.

Upon training the audio encoder, I computed the shared embedding representations using the sound stimuli from our fMRI experiments. I then used the fMRI dataset to train a decoder that could predict the shared representations from the Wav2CLIP audio encoder based on the fMRI data. Following this, I used a pre-trained Vector-Quantized Variational AutoEncoder Generative Adversarial Network (VQGAN) model, a method capable of generating images from the CLIP embedding space (Esser et al., 2021). By using this model, I were able to generate images based on the decoded features predicted from the fMRI data. In essence, our approach transformed neural activity, recorded during the presentation of auditory stimuli, into shared representations, and finally into visual images. This process illustrates the power of these multimodal models and showcases the potential of using deep neural networks in understanding cross-modal cognitive processing.

Results and discussion

In this study, I used fMRI data acquired while the subject was exposed to natural sounds to train a brain decoder aimed at predicting multimodal embedding representations. The decoding performance from both the auditory cortex (AC) and the visual cortex (VC) is displayed in Figure 9.3. The results showed a modest correlation of 0.2 from the AC, whereas the VC presented a nearly negligible correlation. This indicates a rather restrained response from the sensory cortex when predicting multimodal shared representations, aligning with

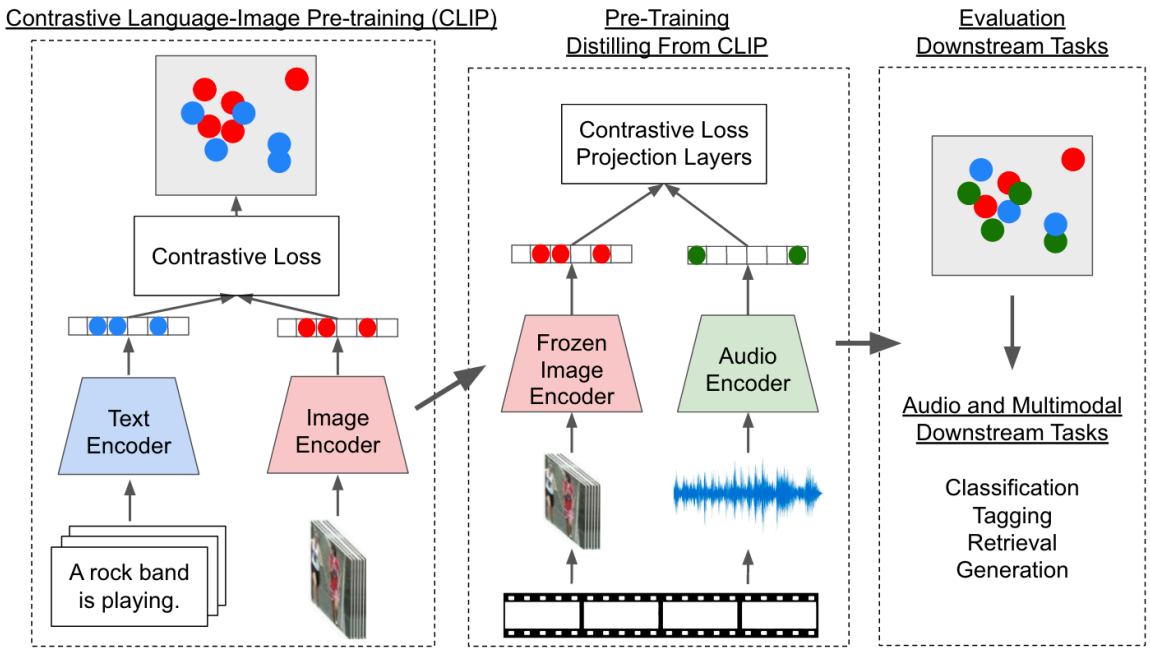

Fig. 9.2 Training strategy for multimodal shared embedding representation using Contrastive Language-Image Pre-training (CLIP). This figure illustrates the process of training the cross-modal embedding representation, which makes use of image, text, and sound. The approach entails freezing the previously trained CLIP image encoder, obtained from text and image datasets. This frozen image encoder then functions as a feature extractor, enabling the training of the audio encoder to generate shared embedding representations. This strategy fosters an interconnected network of cross-modal representations, enhancing the overall depth and versatility of the model. The figure was adapted from Wu et al. (2021), with copyright held by IEEE in 2021.

earlier findings that suggest limited crossmodal interaction from sensory regions in category classifications (Vetter et al., 2014).

Following this, I proceeded to synthesize images from the decoded features. As shown in Figure 9.2, the examples of images synthesized from crossmodal brain responses seem to possess meaningful semantic properties. This implies that the Wav2CLIP audio embeddings project into a significant shared space where individual components from the mixture can be distinguished. Through this crossmodal application exploration, I can visually interpret the brain's responses to various auditory stimuli.

Our perceptual experiences involve interactions across multiple modalities. The methods proposed in this research are particularly relevant in this context. As shown in previous research (Iashin and Rahtu, 2021), using machine learning to train a transformer can create a new spectrogram from a pre-trained spectrogram codebook given a set of video features. This could allow the generation of sounds from images that are perceptually related, expanding the possibilities beyond sound-induced visual reconstruction to visual sound-induced reconstruction.

In the framework I propose for sound reconstruction, training an audio transformer using this multimodal shared representation could indeed make such reconstructions possible. This process could translate our complex perceptual experiences into an audible format that can be shared, heard, and interpreted by others. This could pave the way for a more comprehensive understanding of our perceptual experiences and contribute to the field of neurorehabilitation, music technology, and beyond.

9.6 Concluding remarks

In this thesis, I explored the potential for reconstructing sounds from human brain activity. The challenge lay in the intricate nature of temporal sequences in sounds and the constraints posed by neuroimaging modalities. This research contributes to overcoming these hurdles by harnessing the hierarchical structure of brain auditory processing, finding parallels with DNN models, and leveraging recent advancements in audio-generative models. The proposed method combines brain decoding of auditory features with an audio-generative model, offering a promising approach to sound reconstruction.

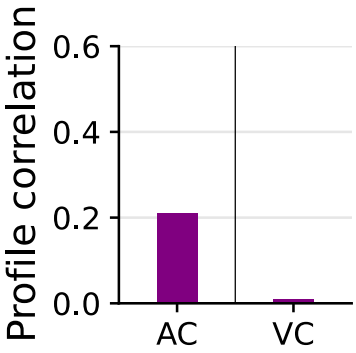

Fig. 9.3 Decoding performance of multimodal embedding representations derived from natural sounds. This bar chart depicts the decoding performance drawn from the auditory cortex (AC) and visual cortex (VC). Each bar represents the mean decoding performance for subject S1, calculated across all feature units of the embedding representations. This visual representation showcases the efficacy of our model in deciphering the complex interplay of features within natural sound stimuli across different brain regions.

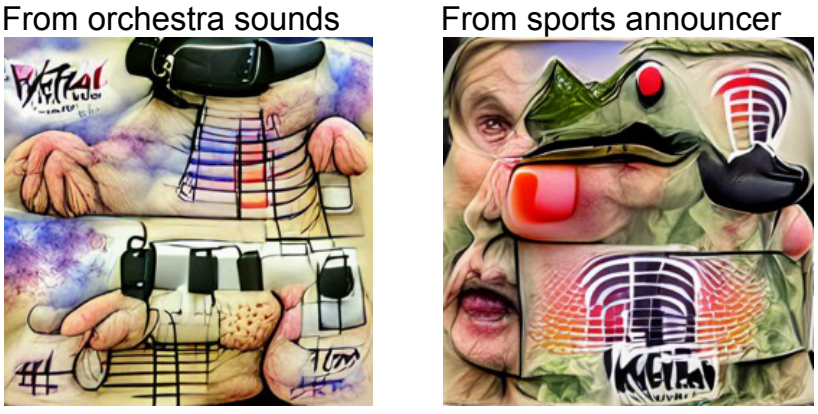

Fig. 9.4 Examples of synthesized images derived from decoded embedding representations in the auditory cortex (AC) when a subject hears natural sounds. The left figure presents samples generated when the subject is exposed to orchestra sounds, while the right figure showcases results when the subject hears sports announcers.

My research first showcased that our model could reconstruct complex spectral-temporal patterns closely mirroring the original sound stimulus's content and quality. It was resilient in synthesizing sounds, even when specific categories were not part of the training phase, suggesting that the model is not merely fitting the brain data to training examples. However, there is room for enhancing the reproduction of finer details, particularly in speech or music sequences.

Next, I analyzed the variations in sound reconstruction across individual ROIs. The core region was superior in identifying low-level representations and acoustic features, although its performance decreased towards peripheral regions. High-level representations, on the other hand, showed some improvement in certain peripheral regions. This highlighted the importance of hierarchical auditory areas and DNN features in sound reconstruction, suggesting that parallel structures in the human auditory system and DNNs can enhance sound reconstruction efficiency.

I also explored how reconstructed sounds encapsulate subjective listening experiences under cocktail party conditions. The sounds reconstructed from the auditory cortex tended to reflect the attended sound more than the unattended one, suggesting a more nuanced process of selective auditory attention that emphasizes category-specific elements of focused stimuli.

The ability to externalize subjective auditory experiences has far-reaching potential. It bridges the internal cognitive processes with their external manifestations, creating opportunities for advancements in diverse areas like communication through imagined sounds, diagnostics for auditory hallucinations in mental health, artistic creation, and basic neuroscience research. This research contributes to a deeper understanding of the neural basis of auditory perception, providing a valuable tool to study the human mind and fostering collaborative efforts in the field of neuroscience.

References

Akbari, H., Khalighinejad, B., Herrero, J. L., Mehta, A. D., and Mesgarani, N. (2019). Towards reconstructing intelligible speech from the human auditory cortex. *Scientific Reports*, 9(1).

Alluri, V., Toiviainen, P., Jääskeläinen, I. P., Glerean, E., Sams, M., and Brattico, E. (2012). Large-scale brain networks emerge from dynamic processing of musical timbre, key and rhythm. *NeuroImage*, 59(4):3677–3689.

Alías, F., Socoró, J. C., and Sevillano, X. (2016). A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds. *Applied Sciences*, 6(5):143.

Andersson, P., Ragni, F., and Lingnau, A. (2019). Visual imagery during real-time fMRI neurofeedback from occipital and superior parietal cortex. *NeuroImage*, 200:332–343.

Avants, B. B., Epstein, C. L., Grossman, M., and Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, 12(1):26–41.

Barbour, D. L. and Wang, X. (2003). Contrast tuning in auditory cortex. *Science, New Series*, 299(5609).

Bednar, A. and Lalor, E. C. (2020). Where is the cocktail party? decoding locations of attended and unattended moving sound sources using eeg. *NeuroImage*, 205:116283.

Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. The MIT Press.

Callender, L., Hawthorne, C., and Engel, J. (2020). Improving perceptual quality of drum transcription with the expanded groove MIDI dataset. arXiv:2004.00188 [cs].

Chakrabarti, S., Sandberg, H. M., Brumberg, J. S., and Krusienski, D. J. (2015). Progress in speech decoding from the electrocorticogram. *Biomedical Engineering Letters*, 5(1):10–21.

Chen, H., Xie, W., Vedaldi, A., and Zisserman, A. (2020). Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725, Barcelona, Spain. IEEE.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, 25(5):975–979.

Chi, T., Ru, P., and Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, 118(2).

Chittka, L. and Brockmann, A. (2005). Perception space—the final frontier. *PLOS Biology*, 3(4).

Correia, J. M., Jansma, B. M. B., and Bonte, M. (2015). Decoding articulatory features from fMRI responses in dorsal speech regions. *Journal of Neuroscience*, 35(45):15015–15025.

Cox, R. W. (1996). Afni: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, 29(3):162–173.

Daly, I. (2023). Neural decoding of music from the EEG. *Scientific Reports*, 13(1).

Degerman, A., Rinne, T., Salmi, J., Salonen, O., and Alho, K. (2006). Selective attention to sound location or pitch studied with fMRI. *Brain Research*, 1077(1).

Deng, L., Hinton, G., and Kingsbury, B. (2013). New types of deep neural network learning for speech recognition and related applications: an overview. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8599–8603, Vancouver, BC, Canada. IEEE.

Dhariwal, P., Jun, H., Payne, C., Kim, J. W., Radford, A., and Sutskever, I. (2020). Jukebox: A Generative Model for Music. arXiv:2005.00341 [cs, eess, stat].

Ding, N. and Simon, J. Z. (2012a). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences*, 109(29):11854–11859.

Ding, N. and Simon, J. Z. (2012b). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of Neurophysiology*, 107(1):78–89.

Donahue, C., McAuley, J., and Puckette, M. (2019). Adverarial audio synthesis. *arXiv:1802.04208 [cs]*.

Défossez, A., Caucheteux, C., Rapin, J., Kabeli, O., and King, J.-R. (2022). Decoding speech from non-invasive brain recordings.

Esser, P., Rombach, R., and Ommer, B. (2021). Taming transformers for high-resolution image synthesis. *arXiv:2012.09841 [cs]*.

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., and Gorgolewski, K. J. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, 16(1):111–116.

Fischl, B. (2012). FreeSurfer. *NeuroImage*, 62(2).

Formisano, E., De Martino, F., Bonte, M., and Goebel, R. (2008). "who" is saying "what"? brain-based decoding of human voice and speech. *Science*, 322(5903):970–973.

Formisano, E., Kim, D.-S., Di Salle, F., van de Moortele, P.-F., Ugurbil, K., and Goebel, R. (2003). Mirror-symmetric tonotopic maps in human primary auditory cortex. *Neuron*, 40(4).

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C. F., Jenkinson, M., Smith, S. M., and Van Essen, D. C. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. The MIT Press, Cambridge, Massachusetts.

Graves, A., Mohamed, A.-r., and Hinton, G. (2013). Speech recognition with deep recurrent neural networks. *arXiv:1303.5778 [cs]*.

Griffin, D. and Lim, J. (1984). Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):236–243.

Hamilton, L. S., Oganian, Y., Hall, J., and Chang, E. F. (2021). Parallel and distributed encoding of speech across human auditory cortex. *Cell*, 184(18):4626–4639.e13.

Han, Y., Park, J., and Lee, K. (2017). Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification. In *Workshop on Detection and Classification of Acoustic Scenes and Events*.

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., and Ng, A. Y. (2014). Deep Speech: Scaling up end-to-end speech recognition. *arXiv:1412.5567 [cs]*.

Hassabis, D., Spreng, R. N., Rusu, A. A., Robbins, C. A., Mar, R. A., and Schacter, D. L. (2014). Imagine all the people: How the brain creates and uses personality models to predict behavior. *Cerebral Cortex*, 24(8).

Hasson, U., Yang, E., Vallines, I., Heeger, D. J., and Rubin, N. (2008). A hierarchy of temporal receptive windows in human cortex. *Journal of Neuroscience*, 28(10):2539–2550.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., and Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430.

Heelan, C., Lee, J., O'Shea, R., Lynch, L., Brandman, D. M., Truccolo, W., and Nurmikko, A. V. (2019). Decoding speech from spike-based neural population recordings in secondary auditory cortex of non-human primates. *Communications Biology*, 2(1).

Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R. J., and Wilson, K. (2017). Cnn architectures for large-scale audio classification. arXiv.1609.09430[cs.SD].

Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507.

Horikawa, T. and Kamitani, Y. (2017a). Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, 8:15037.

Horikawa, T. and Kamitani, Y. (2017b). Hierarchical neural representation of dreamed objects revealed by brain decoding with deep neural network features. *Frontiers in Computational Neuroscience*, 11:4.

Horikawa, T. and Kamitani, Y. (2022). Attention modulates neural representation to render reconstructions according to subjective appearance. *Communications Biology*, 5(1):34.

Horikawa, T., Tamaki, M., Miyawaki, Y., and Kamitani, Y. (2013). Neural decoding of visual imagery during sleep. *Science*, 340(6132):639–642.

Humphries, C., Liebenthal, E., and Binder, J. R. (2010). Tonotopic organization of human auditory cortex. *NeuroImage*, 50(3):1202–1211.

Iashin, V. and Rahtu, E. (2021). Taming visually guided sound generation. *arXiv:2110.08791 [cs, eess]*.

Ince, R. A. A., Kay, J. W., and Schyns, P. G. (2022). Within-participant statistics for cognitive science. *Trends in Cognitive Sciences*, 26(8):626–630.

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., and Smith, S. M. (2012). Fsl. *NeuroImage*, 62(2):782–790.

Kaas, J. H. and Hackett, T. A. (2000). Subdivisions of auditory cortex and processing streams in primates. *Proceedings of the National Academy of Sciences*, 97(22):11793–11799.

Kamitani, Y. and Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, 8(5):679–85.

Kell, A. J., Yamins, D. L., Shook, E. N., Norman-Haignere, S. V., and McDermott, J. H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644.e16.

Kumar, K., Kumar, R., de Boissiere, T., Gestin, L., Teoh, W. Z., Sotelo, J., de Brebisson, A., Bengio, Y., and Courville, A. (2019). MelGAN: Generative adversarial networks for conditional waveform synthesis. *arXiv:1910.06711 [cs, eess]*.

Lewis, J. W. and Van Essen, D. C. (2000). Corticocortical connections of visual, sensorimotor, and multimodal processing areas in the parietal lobe of the macaque monkey. *Journal of Comparative Neurology*, 428(1):112–137.

Li, Y., Anumanchipalli, G. K., Mohamed, A., Lu, J., Wu, J., and Chang, E. F. (2022). Dissecting neural computations of the human auditory pathway using deep neural networks for speech. preprint, Neuroscience.

Liu, X., Iqbal, T., Zhao, J., Huang, Q., Plumbley, M. D., and Wang, W. (2021). Conditional sound generation using neural discrete time-frequency representation learning. arXiv:2107.09998 [cs, eess].

Loula, J., Varoquaux, G., and Thirion, B. (2018). Decoding fMRI activity in the time domain improves classification performance. *NeuroImage*, 180:203–210.

Martin, S., Iturrate, I., Millán, J. d. R., Knight, R. T., and Pasley, B. N. (2018). Decoding inner speech using electrocorticography: Progress and challenges toward a speech prosthesis. *Frontiers in Neuroscience*, 12.

Mauch, M. and Dixon, S. (2014). Pyin: A fundamental frequency estimator using probabilistic threshold distributions. pages 659–663, Florence, Italy. IEEE.

McDermott, J. H. (2009). The cocktail party problem. *Current biology : CB*, 19(22):R1024–1027.

McDermott, J. H. (2018). Audition. In Wixted, J. T., editor, *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Meyer, K., Kaplan, J. T., Essex, R., Webber, C., Damasio, H., and Damasio, A. (2010). Predicting visual stimuli on the basis of activity in auditory cortices. *Nature Neuroscience*, 13(6).

Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M.-a., Morito, Y., Tanabe, H. C., Sadato, N., and Kamitani, Y. (2008). Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, 60(5):915–929.

Moerel, M., De Martino, F., and Formisano, E. (2014). An anatomical and functional topography of human auditory cortical areas. *Frontiers in Neuroscience*, 8.

Moses, D. A., Leonard, M. K., Makin, J. G., and Chang, E. F. (2019). Real-time decoding of question-and-answer speech dialogue using human cortical activity. *Nature Communications*, 10(1):3096.

Moshitch, D., Las, L., Ulanovsky, N., Bar-Yosef, O., and Nelken, I. (2006). Responses of neurons in primary auditory cortex (A1) to pure tones in the halothane-anesthetized cat. *Journal of Neurophysiology*, 95(6).

Munoz-Lopez, M. M., Mohedano-Moriano, A., and Insausti, R. (2010). Anatomical pathways for auditory memory in primates. *Frontiers in Neuroanatomy*, 4.

Nishimoto, S., Vu, A., Naselaris, T., Benjamini, Y., Yu, B., and Gallant, J. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 21(19):1641–1646.

Norman-Haignere, S., Kanwisher, N., and McDermott, J. (2015). Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. *Neuron*, 88(6).

Norman-Haignere, S. V., Long, L. K., Devinsky, O., Doyle, W., Irobunda, I., Merricks, E. M., Feldstein, N. A., McKhann, G. M., Schevon, C. A., Flinker, A., and Mesgarani, N. (2022). Multiscale temporal integration organizes hierarchical computation in human auditory cortex. *Nature Human Behaviour*, 6(3):455–469.

Nourski, K. V., Steinschneider, M., McMurray, B., Kovach, C. K., Oya, H., Kawasaki, H., and Howard, M. A. (2014). Functional organization of human auditory cortex: Investigation of response latencies through direct recordings. *NeuroImage*, 101:598–609.

Ogawa, S., Lee, T. M., Kay, A. R., and Tank, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences*, 87(24):9868–9872.

Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). WaveNet: A generative model for raw audio. *arXiv:1609.03499 [cs]*.

Oord, A. v. d., Vinyals, O., and Kavukcuoglu, K. (2018). Neural discrete representation learning. *arXiv:1711.00937 [cs]*.

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., Slaney, M., Shamma, S. A., and Lalor, E. C. (2015). Attentional selection in a cocktail party environment can be decoded from single-trial eeg. *Cerebral Cortex*, 25(7):1697–1706.

O'Sullivan, J., Chen, Z., Herrero, J., McKhann, G. M., Sheth, S. A., Mehta, A. D., and Mesgarani, N. (2017). Neural decoding of attentional selection in multi-speaker environments without access to clean sources. *Journal of Neural Engineering*, 14(5):056001.

Park, J.-Y., Tsukamoto, M., Tanaka, M., and Kamitani, Y. (2023). Sound reconstruction from human brain activity via a generative model with brain-like auditory features. arXiv:2306.11629 [cs, eess].

Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Crone, N. E., Knight, R. T., and Chang, E. F. (2012). Reconstructing speech from human auditory cortex. *PLoS Biology*, 10(1):e1001251.

Pei, X., Barbour, D. L., Leuthardt, E. C., and Schalk, G. (2011). Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans. *Journal of Neural Engineering*, 8(4):046028.

Petkov, C. I., Kayser, C., Augath, M., and Logothetis, N. K. (2006). Functional imaging reveals numerous fields in the monkey auditory cortex. *PLoS Biology*, 4(7).

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. arXiv:2103.00020 [cs].

Radford, A., Metz, L., and Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv:1511.06434 [cs]*.

Rauschecker, J. P. and Scott, S. K. (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nature Neuroscience*, 12(6).

Recanzone, G. H., Guard, D. C., Phan, M. L., and Su, T.-I. K. (2000). Correlation between the activity of single auditory cortical neurons and sound-localization behavior in the macaque monkey. *Journal of Neurophysiology*, 83(5).

Salamon, J. and Bello, J. P. (2017). Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 24(3).

Sankaran, N., Thompson, W. F., Carlile, S., and Carlson, T. A. (2018). Decoding the dynamic representation of musical pitch from human brain activity. *Scientific Reports*, 8(1).

Santoro, R., Moerel, M., De Martino, F., Valente, G., Ugurbil, K., Yacoub, E., and Formisano, E. (2017). Reconstructing the spectrotemporal modulations of real-life sounds from fMRI response patterns. *Proceedings of the National Academy of Sciences*, 114(18):4799–4804.

Seeliger, K., Güçlü, U., Ambrogioni, L., Güçlütürk, Y., and van Gerven, M. A. J. (2018). Generative adversarial networks for reconstructing natural images from brain activity. *NeuroImage*, 181:775–785.

Shen, G., Dwivedi, K., Majima, K., Horikawa, T., and Kamitani, Y. (2019a). End-to-end deep image reconstruction from human brain activity. *Frontiers in Computational Neuroscience*, 13:21.

Shen, G., Horikawa, T., Majima, K., and Kamitani, Y. (2019b). Deep image reconstruction from human brain activity. *PLOS Computational Biology*, 15(1):1006633.

Stowell, D., Giannoulis, D., Benetos, E., Lagrange, M., and Plumbley, M. D. (2015). Detection and classification of acoustic scenes and events. *IEEE Transactions on Multimedia*, 17(10):1733–1746.

Sturm, I., Dähne, S., Blankertz, B., and Curio, G. (2015). Multi-variate eeg analysis as a novel tool to examine brain responses to naturalistic music stimuli. *PLOS ONE*, 10(10):1–30.

Sweet, R. A., Dorph-Petersen, K.-A., and Lewis, D. A. (2005). Mapping auditory core, lateral belt, and parabelt cortices in the human superior temporal gyrus. *The Journal of Comparative Neurology*, 491(3):270–289.

Szabó, P. and Barthó, P. (2022). Decoding neurobiological spike trains using recurrent neural networks: a case study with electrophysiological auditory cortex recordings. *Neural Computing and Applications*, 34(4):3213–3221.

Tian, B. (2001). Functional specialization in rhesus monkey auditory cortex. *Science*, 292(5515).

Tuckute, G., Feather, J., Boebinger, D., and McDermott, J. H. (2023). Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions. bioRxiv.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *arXiv:1706.03762 [cs]*.

Vetter, P., Smith, F., and Muckli, L. (2014). Decoding sound and imagery content in early visual cortex. *Current Biology*, 24(11):1256–1262.

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 1096–1103, New York, NY, USA. Association for Computing Machinery.

Walker, J., Razavi, A., and Oord, v. d. A. (2021). Predicting video with vqvae. *arXiv:2103.01950 [cs]*.

Wang, L., Li, K., Chen, X., and Hu, X. P. (2019). Application of convolutional recurrent neural network for individual recognition based on resting state fMRI data. *Frontiers in Neuroscience*, 13:434.

Wang, R., Wang, Y., and Flinker, A. (2018). Reconstructing speech stimuli from human auditory cortex activity using a WaveNet-like network. *arXiv:1811.02694[cs.SD]*.

Wu, H.-H., Seetharaman, P., Kumar, K., and Bello, J. P. (2021). Wav2CLIP: Learning robust audio representations from CLIP. *arXiv:2110.11499 [cs, eess]*.

Yoo, S.-H., Santosa, H., Kim, C.-S., and Hong, K.-S. (2021). Decoding multiple sound-categories in the auditory cortex by neural networks: An fnirs study. *Frontiers in Human Neuroscience*, 15:636191.

Zhang, J., Zhang, G., Li, X., Wang, P., Wang, B., and Liu, B. (2018). Decoding sound categories based on whole-brain functional connectivity patterns. *Brain Imaging and Behavior*.

Zhao, Y., Li, H., Lai, C.-I., Williams, J., Cooper, E., and Yamagishi, J. (2020). Improved prosody from learned f0 codebook representations for vq-vae speech waveform reconstruction. arXiv:2005.07884 [eess.AS].

Appendix A

Publications

A.1 Manuscript

- Park, J., Tsukamoto, M., Tanaka M., and Kamitani Y. (2023). Sound reconstruction from human brain activity via a generative model with brain-like auditory features. arXiv. http://arxiv.org/abs/2306.11629

A.2 Poster presentation

- Park, J., Horikawa, T., Majima K., and Kamitani Y. (2018). Brain decoding of auditory-induced cortical activity with deep neural network features. Annual meeting on the Korean Society for Cognitive Science - **Best poster presentation award**