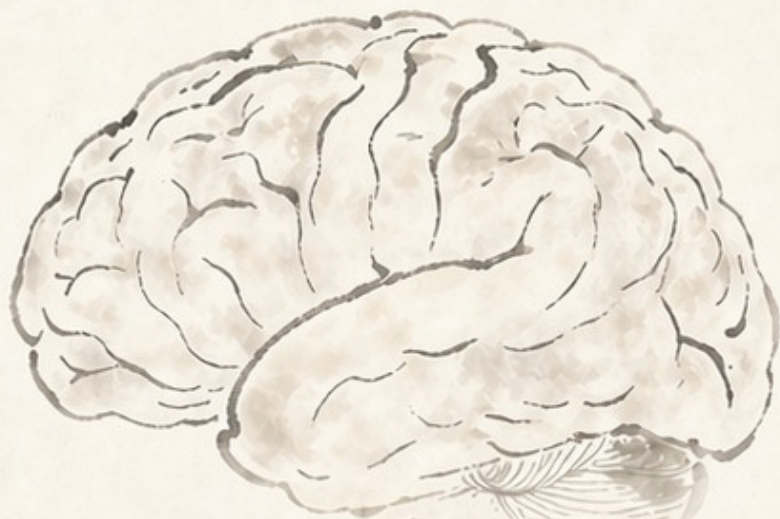




Identifying **Sound** Features from Brain Activity

Lidongsheng Xing

邢立冬生



PROPOSITIONS
Belonging to the thesis

Identifying Sound Features from Brain Activity

1. Neural responses to sound can be systematically modulated using auditory signals reconstructed from brain activity. (this thesis)
2. Deep neural network-based sound synthesis provides a powerful framework for dissociating acoustic features from perceptual categories such as speech and music. (this thesis)
3. Category-selective neural responses in the auditory system are not solely determined by natural acoustic statistics, but can also be driven by artificially synthesized sounds. (this thesis)
4. Hemispheric asymmetry in auditory processing is better explained by cross-band temporal dependencies than by classical spectro-temporal modulation models. (this thesis)
5. Combining computational modeling with neuroimaging techniques enables a more comprehensive understanding of complex sensory representations than either approach alone.
6. In neuroscience, hypothesis-free approaches can reveal critical features that remain overlooked within theory-driven frameworks.
7. Machine learning methods in neuroscience should prioritize interpretability alongside predictive performance.
8. Advances in artificial intelligence are increasingly reshaping how we study, model, and understand the human brain.

Doctoral thesis

Identifying Sound Features from Brain Activity

Lidongsheng Xing

2026

©Lidongsheng Xing, Maastricht, 2026.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without prior written permission of the author.

Layout: Lidongsheng Xing

Cover design: Lidongsheng Xing

Production: ProefschriftMaken, www.proefschriftmaken.nl

ISBN: 978-94-6534-392-1

Identifying Sound Features from Brain Activity

DISSERTATION

to obtain the degree of Doctor at Maastricht University,
on the authority of the Rector Magnificus, Prof. dr. Jan Smits
in accordance with the decision of the Board of Deans,
to be defended in public
on Wednesday, 6th of May 2026, at 10:00 hours

by
Lidongsheng Xing

Supervisor(s)

Dr. Lars Riecke Maastricht University

Prof. Dr. Elia Formisano Maastricht University

Assessment Committee

Chair:

Prof. Dr. Sonja Kotz Maastricht University

Members:

Prof. Dr. Yi Du Chinese Academy of Sciences

Dr. Bruno L. Giordano Centre National de la Recherche Scientifique and Université Aix-Marseille

Dr. Christian Herff Maastricht University

Contents

CHAPTER 1	1
GENERAL INTRODUCTION	1
1 THE HUMAN AUDITORY SYSTEM: FROM SOUND WAVES TO SEMANTICS	3
2 CONVENTIONAL METHODOLOGY FOR LINKING SOUND FEATURES TO NEURAL RESPONSES	4
3 THE FEATURE IDENTIFICATION APPROACH: FROM BRAIN ACTIVITY TO SOUND FEATURES	5
4 ASPECTS OF AUDITORY PROCESSING ADDRESSED IN THIS THESIS	7
5 THESIS OUTLINE	12
CHAPTER 2	15
NEURAL AND AUDITORY FIDELITY OF BASIC TEMPORAL STIMULUS PATTERNS RECONSTRUCTED FROM INDIVIDUAL EEG RECORDINGS	15
1 INTRODUCTION	18
2 METHODS	19
3 RESULTS	26
4 DISCUSSION	33
CHAPTER 3	37
DEEP SOUND SYNTHESIS MATCHED TO BRAIN ACTIVITY RECAPITULATES PREFERENTIAL RESPONSES TO SPEECH AND MUSIC	37
1 INTRODUCTION	40
2 METHODS	42
3 RESULTS	48
4 DISCUSSION	60
5. SUPPLEMENTARY MATERIALS	64
CHAPTER 4	71
DEEP SOUND SYNTHESIS REVEALS SOUND FEATURES UNDERLYING FUNCTIONAL ASYMMETRY OF THE AUDITORY CORTICES	71
1 INTRODUCTION	74
2 METHODS	75
3 RESULTS	82
4 DISCUSSION	85
5 SUPPLEMENTARY MATERIALS	90
CHAPTER 5	99
GENERAL DISCUSSION	99
1 SUMMARY AND IMPLICATIONS OF THE FINDINGS	101
2 THE IMPLICATIONS OF THE FEATURE IDENTIFICATION APPROACH	104
3 POTENTIAL APPLICATIONS OF THE FEATURE IDENTIFICATION APPROACH	107
4 LIMITATIONS AND FUTURE DIRECTIONS	108

5 CONCLUSION	110
REFERENCES	111
APPENDIX	125
 IMPACT PARAGRAPH.....	127
 PUBLICATIONS.....	134

Abbreviations

AST – Asymmetric Sampling in Time
DNN – Deep Neural Network
EC – Envelope Correlation
EEG – Electroencephalography
ERP – Event-Related Potential
fMRI – Functional Magnetic Resonance Imaging
fROI – Functional Region of Interest
HG – Heschl's Gyrus
MEG – Magnetoencephalography
MMN – Mismatch Negativity
MTF – Modulation Transfer Function
PET – Positron Emission Tomography
PP – Planum Polare
PT – Planum Temporale
RSA – Representational Similarity Analysis
SI – System Identification
STG – Superior Temporal Gyrus
STM – Spectrotemporal Modulation
STS – Superior Temporal Sulcus
STRF – Spectrotemporal Receptive Field
SVM – Support Vector Machine
TES – Transcranial Electrical Stimulation
TMS – Transcranial Magnetic Stimulation
TRF – Temporal Response Function
VOT – Voice Onset Time

Chapter 1

General Introduction

1 The Human Auditory System: From Sound Waves to Semantics

Sound perception and interpretation are fundamental to human cognition, supporting spoken communication, environmental awareness, and emotional engagement. These functions involve the gradual transformation of acoustic waveforms into semantically meaningful representations. The neural pathway that implements this transformation can be broadly divided into three hierarchical stages: The early stage, involving the peripheral auditory nervous system (e.g., cochlea) and the subcortex (e.g., brainstem and thalamus), which extracts basic acoustic features¹ such as frequency, intensity, and temporal patterns (Winer and Schreiner, 2011; Pickles, 2013); The intermediate stage, mediated by the auditory cortex (e.g., primary and secondary auditory areas), involves the transformation of acoustic information into increasingly abstract neural representations that support the perception of complex sound attributes such as pitch, timbre, and spatial location (Bizley and Cohen, 2013); The late stage, involving higher-order semantic networks (e.g., prefrontal and associative cortices), which integrates auditory information with memory and meaning to enable object recognition and comprehension (Hickok and Poeppel, 2007).

The auditory cortex is the focus of this thesis. The primary auditory cortex, primarily located on Heschl's Gyrus (HG), is the first cortical destination of sound-evoked neural signals and exhibits tonotopic organization (Moerel et al., 2014). More specifically, it is thought to analyze the spectrotemporal modulation content of the auditory cochleagram, which it receives from previous stages, through a bank of specialized filters (Chi et al., 2005). These filters exhibit varying selectivity across both spectral domain (narrow to broad bandwidths) and temporal domain (slow to fast modulation rates). The resulting spectrotemporal receptive fields (STRFs) are systematically organized along the tonotopic axis, creating a comprehensive representation of sounds in terms of their spectro-temporal content (Chi et al., 1999). Beyond HG, auditory information is processed in a hierarchy of higher-order auditory regions, including the planum temporale (PT) and planum polare (PP), and at a still higher level, the superior temporal gyrus (STG), and superior temporal sulcus (STS). These nonprimary auditory areas receive input from primary auditory cortex and progressively transform low-level acoustic features into higher-level representations. Through this integrative processing, the auditory system extracts perceptual dimensions such as pitch, rhythm, and timbre, and supports the analysis of speech-specific features such as phonemes, ultimately enabling speech comprehension (Griffiths and Warren, 2002).

Functional and anatomical brain studies have supported a framework in which the higher-order auditory regions are organized into parallel hierarchical pathways: a dorsal stream for spatial analysis ("where") and a ventral stream for non-spatial feature extraction ("what") (Rauschecker and Tian, 2000; Rauschecker and Scott, 2009; Bizley and Cohen, 2013). While this framework outlines the functional-neuroanatomical pathways underlying sound processing, the mechanisms by which the brain transforms acoustic signals into meaningful semantic representations remain surprisingly unclear (Lewis et al., 2011; Elliott et al., 2013; Giordano et al., 2013; Santoro et al., 2014, 2017; Kell et al., 2018; Norman-Haignere and McDermott, 2018).

¹ In the context of this thesis, the term 'feature' is a measurable, computationally defined property that captures relevant temporal and spectral characteristics of acoustic input and serves sound recognition.

This transformation is thought to rely on the encoding of multiple sound-related features across distributed neural populations. A reason for the knowledge gap is certain limitations of the research methodology that is conventionally applied.

2 Conventional Methodology for Linking Sound Features to Neural Responses

Investigations in auditory neuroscience have primarily sought to understand how simple acoustic properties of sound—such as frequency, intensity, timing, and spatial location—are represented in neural activity across the auditory pathway. These studies relied on experimental manipulation of specific basic sound features and neural measurements to reveal how the acoustic information is encoded across subcortical and cortical levels (Theunissen and Elie, 2014; King et al., 2018). Typically, sound manipulation was guided by narrowly defined hypotheses, often isolating one acoustic parameter at a time (e.g., pure tone frequency, amplitude modulation rate, or binaural disparity) to examine its effects on neural activity.

In this thesis, we attempt to validate a different research approach that extracts sound features from the temporal dynamics and spatial patterns of brain responses. Before introducing the approach, the conventional methodologies that have been employed are briefly reviewed in the following sections.

2.1 Temporal Dynamics

Noninvasive recording techniques such as electroencephalography (EEG) and magnetoencephalography (MEG) enable researchers to examine the millisecond-level dynamics of auditory perception. Many EEG/MEG studies of human auditory processing have focused on brief, isolated stimuli, largely because assessment of event-related potentials (ERPs) requires neural responses to be time-locked to discrete sensory events (e.g., Handy, 2004; Luck, 2014). Using this approach, the impulse response function of the auditory system can be effectively approximated by averaging transient neural responses across hundreds of identical, time-locked trials.

More recent research has adopted extended and naturalistic stimuli to investigate how neural oscillations align with ongoing auditory rhythms. One common approach is to assess the coherence between cortical responses and the temporal dynamics of the stimulus using various metrics, such as cerebro-acoustic coherence (Peelle et al., 2013) and inter-trial correlation of neural responses in the delta or theta frequency bands (Ding et al., 2014). In addition, advanced computational approaches have been used that model the relationship between neural activity and input sounds, allowing researchers either to predict neural responses from acoustic features (forward) or to reconstruct stimulus temporal properties from neural signals (backward). The backward direction to reconstruct the temporal features is the focus of the first empirical chapter of this thesis (further details of this direction are introduced in Section 3).

2.2 Spatial Response Patterns

Complementary advances in functional neuroimaging have provided a spatially resolved perspective on auditory encoding in the human auditory system. Early Positron Emission Tomography (PET) and Functional Magnetic Resonance Imaging (fMRI) studies aimed to reveal the tonotopic organization of the auditory cortex, wherein neuronal populations could be arranged according to their preferred frequency responses (Talavage et al., 2000; Formisano et al., 2003) or pre-defined sound categories (e.g., Belin and Zatorre, 2003; Mesgarani et al., 2014; Overath et al., 2015). A key limitation of these studies is their focus on a restricted set of pre-defined acoustic properties or sound categories.

Modern computational approaches have extended these investigations by enabling the prediction of voxel-wise neural responses (Kell et al., 2018) or spatial response patterns (Giordano et al., 2023), as well as the reconstruction of sound features from distributed spatial patterns of activity (Santoro et al., 2017). These approaches allow the use of naturalistic, everyday sounds as stimuli without the need for strictly pre-defined hypotheses. In contrast, systematic model comparison enables efficient and principled characterization of the functional organization of the auditory cortex. In the second and third empirical chapter of this thesis, we focus on the forward modeling direction and take a further step toward identifying sound features from the predicted neural responses (further details are introduced in Section 3).

3 The Feature Identification Approach: From Brain Activity to Sound Features

Recent advances in computational modeling have transformed the field by moving from conventional descriptive analyses toward predictive frameworks. By combining computational modeling—especially using multivariate regression and deep neural networks (DNNs)—with machine learning, the complexity of the human auditory pathway can be handled more intuitively and higher-order nonlinear interactions between sound attributes and neural activity can be revealed. More specifically, forward encoding models (also known as stimulus–response or feature–response models) aim to predict neural activity patterns from known acoustic features of a stimulus, thereby characterizing how complex sound attributes are represented in the brain (Theunissen and Elie, 2014; King et al., 2018). This approach has been instrumental in synthesizing novel stimuli that drive specific neural patterns according to the predictions of a forward encoding model (Bashivan et al., 2019; Ponce et al., 2019; Ratan Murty et al., 2021; Tuckute et al., 2024). Conversely, backward decoding models (or response–stimulus models) reconstruct or infer features of the auditory stimulus from measured neural signals. These decoding approaches (also known as feature reconstruction techniques) have revealed that distinct acoustic features—such as amplitude envelope or spectrotemporal modulation—can be discriminated from cortical activity patterns (Bialek et al., 1991; Kay et al., 2008; Miyawaki et al., 2008; Mesgarani et al., 2009; Naselaris et al., 2009; Santoro et al., 2017). The two aforementioned frameworks are explained in more detail in the next sections.

3.1 Forward Synthesis: Engineering Sounds from Model Predictions

Forward synthesis refers to the use of computational encoding models to design stimuli that elicit targeted neural responses, based on model predictions. This approach involves two key steps: fitting the forward encoding models and optimizing the stimulus. An important first prerequisite for this approach to be successful is that the fitted encoding model must not only provide strong explanatory power for observed neural responses but also demonstrate robust generalizability by accurately predicting responses to novel inputs. Moreover, the optimization step requires sufficient computational power, as it typically involves a large number of iterations, during which pixel-level values or latent codes in a deep neural network of an initial stimulus are adjusted to drive the predicted neural response toward a target pattern—often the maximization of activity in a specific cortical region. Pioneering work in the visual domain has shown that the approach allows for generating novel artificial images to selectively activate specific neuronal sites in visual cortical area V4 (Bashivan et al., 2019). The approach has also been applied to explore functional regions of interest in the visual cortex (Ponce et al., 2019; Ratan Murty et al., 2021) and semantic cortical networks (Tuckute et al., 2024).

The approach has enabled abovementioned studies to explore causal relations between the stimulus features in the computational forward encoding model and their evoked neural responses, as well as to characterize the stimulus features to which brain regions are tuned. Furthermore, beyond basic neuroscience, forward synthesis holds promise for clinical applications, particularly in treating brain dysfunctions. For example, by generating stimuli that selectively modulate pathological neural activity (e.g., hyperactive circuits in tinnitus or maladaptive plasticity in hearing loss), this approach could pave the way for targeted, non-invasive neuromodulation therapies (Peter and Kleinjung, 2019).

Applications of the forward synthesis approach in the auditory literature are currently still missing, although they could help addressing relevant questions. For example, can we synthesize sounds that optimally drive functional regions in the auditory cortex? Or can we even evoke categorical percepts by selectively modulating distinct auditory cortical regions? Answering these questions would deepen our understanding of the computational mechanisms in the auditory cortex, which could further inform aforementioned clinical applications.

3.2 Backward Reconstruction: Decoding the Neural Signature of Sound

Backward stimulus reconstruction, also termed model-based neural decoding, seeks to reverse-engineer stimulus features from recorded brain activities. Unlike more common classification-based decoding methods—which are limited to distinguishing among a small set of pre-defined stimulus categories—this approach incorporates a representational model based on elementary stimulus features. This allows for the reconstruction of individual stimuli from neural response patterns.

This model-based decoding approach has previously been used to reconstruct the spectrotemporal characteristics of brief segments of natural sounds, such as real-life sounds (Santoro et al., 2017) and speech (Pasley et al., 2012), by analyzing spatial activity patterns in the auditory cortex. These studies demonstrated that neural response patterns contain

sufficient information to recover critical acoustic features (e.g., spectrogram and spectrotemporal modulation properties) enabling the reconstruction of perceptually identifiable auditory stimuli.

Additionally, the temporal dynamics of longer sound sequences, such as continuous speech or music, have been successfully reconstructed from EEG and MEG recordings of brain activity, which provide millisecond-level temporal precision (Ding and Simon, 2012a, 2012b; Wang et al., 2019). These studies revealed that low-frequency components of the speech signal, e.g., amplitude envelope and phonemic features, can be decoded from neural signals tracking these dynamic features, highlighting the brain's ability to synchronize with the temporal trajectory of the unfolding auditory input.

Advances in decoding cortical speech representations are paving the way for transformative applications in brain–computer interfaces (Herff et al., 2015, 2020), particularly in reconstructing inner speech imagery—the silent, internal articulation of words without overt movement or sound (Luo et al., 2022; Berezutskaya et al., 2023; Su and Tian, 2025). Combined with real-time feedback such applications could revolutionize communication for individuals with speech impairments (e.g., locked-in syndrome) and enhance human-machine interaction (Brumberg et al., 2018; Angrick et al., 2024).

Reconstruction quality is typically assessed using objective metrics, such as the correlation between the reconstructed and the original stimulus, as well as subjective perceptual evaluations. Perceptual ratings have been successfully applied to reconstructions of speech and other everyday sounds, demonstrating that decoded signals can be perceptually identifiable and behaviorally meaningful (Pasley et al., 2012; Santoro et al., 2017). However, perceptual evaluation of reconstructed temporal patterns themselves remains largely unexplored. In particular, it is still unclear to what extent reconstructed temporal features—such as energy fluctuation patterns—are perceptually salient or interpretable to human listeners. Addressing this gap is essential for validating model-based neural decoding approaches beyond purely statistical measures and for establishing their relevance to auditory perception of temporal features such as rhythm. The first empirical chapter of the present thesis aims to bridge this gap by systematically evaluating both objective and perceptual measures of temporal pattern reconstruction from neural signals.

4 Aspects of Auditory Processing Addressed in This Thesis

As aforementioned, while auditory cortical representations of sound are still not fully understood given the shortcomings of the conventional research methods, the more recent modeling approach could possibly advance auditory neuroscience to a better understanding of more complex feature processing. This dissertation explores the application of the feature identification approach to the temporal dynamics and functional-anatomical spatial patterns of auditory cortical processing. First, it examines the perceptual validity of temporal patterns reconstructed from neural responses to sounds (Chapter 2). Second, it explores distinct speech- and music-related neural features within specialized auditory cortical regions, with the aim of revealing how functional selectivity emerges across the auditory cortex (Chapter 3).

Third, it investigates the acoustic underpinnings of hemispheric asymmetries in auditory processing, addressing long-standing questions about lateralization in higher-order sound analysis (Chapter 4). This section provides a more detailed overview of these three empirical studies.

4.1 Identifying Temporal Stimulus Patterns from Neural Temporal Dynamics

The accurate perception of naturalistic sounds, especially continuous speech and music, depends critically on the brain's ability to encode temporal dynamics (Riecke et al., 2018). These dynamics encompass millisecond-scale acoustic fluctuations (e.g., formant transitions) to slower rhythmic patterns (e.g., syllabic or prosodic rhythms) (Ding et al., 2017). Central to this temporal processing is the phenomenon of neural tracking, wherein endogenous cortical oscillations synchronize their phase and frequency to external rhythms across modalities (Bartlett et al., 2011; Ito et al., 2011).

A direct approach to studying neural tracking to continuous stimuli is through system identification (SI; Marmarelis, 2004), which mathematically models how specific stimulus properties are mapped onto neural responses. Among the various SI frameworks (reviewed in Wu et al., 2006), the most straightforward are linear time-invariant systems in which the system is characterized by its impulse response. Based on this framework, Crosse et al. (2016) introduced a temporal response function (TRF) toolbox, which allows linking recordings of neural signals to continuous stimuli. Although the brain is neither fully linear nor time-invariant, these assumptions are often reasonable in some contexts (Boynton et al., 1996) and allow the system to be characterized by its impulse response. Using reverse correlation (Ringach and Shapley, 2004), TRFs can be estimated either to predict neural responses from continuous input or to reconstruct stimulus features from brain activity (Crosse et al., 2016).

In prior work, the fidelity of the reconstructed temporal auditory features has often been evaluated by computing the correlation between the reconstructed features and the original stimulus. While informative, this metric primarily captures the degree of neural tracking of the temporal feature and does not guarantee that reconstructions preserve perceptually meaningful information. Thus, it remains unclear whether reconstructions of such patterns elicit equivalent perceptual or neural responses as the original inputs.

To address this, Chapter 2 of this thesis examines whether it is feasible to reconstruct salient, perceptible amplitude modulation patterns from non-invasive recordings of neural activity. Beyond this methodological validation, this work also explores whether cortical tracking-based stimulus reconstructions might be utilized to enhance sensory processing.

4.2 Identifying Sound Features Underlying Sound Categorization in Auditory Cortical Regions

The human auditory system exhibits remarkable specialization in processing speech and music—two biologically and culturally fundamental forms of communication that differ profoundly in their acoustic structure and perceptual demands. It has been long debated how

the auditory cortex represents sound categories. The localist view postulates a one-to-one correspondence between highly selective neuronal populations and semantic categories (Norman-Haignere et al., 2015, 2022), while the compositional coding view argues that semantic information is represented as spatially distributed activity patterns (Staeren et al., 2009).

Before examining the neural mechanisms underlying sound category processing, it is essential to characterize the distinct acoustic signatures of the individual categories, as these may shape how the brain encodes and categorizes the sounds. Speech and music can be distinguished in terms of both, their low-level acoustics (e.g., spectrotemporal resolution) and their higher-order domain-specific attributes (e.g., acoustic-phonetic structure and musical tones) (Zatorre et al., 2002).

4.2.1 Low-Level Acoustics

At the foundational level, speech and music diverge in their spectrotemporal structure (Zatorre et al., 2002; Ding et al., 2017). Speech arises from the highly constrained biomechanics of the human vocal tract, resulting in predictable spectrotemporal patterns optimized for linguistic communication. The articulatory dynamics of speech production generate a complex acoustic signal characterized by rapid transitions between periodic (voiced) and aperiodic (unvoiced) components. The temporal patterning of consonants—particularly stop consonants like /p/, /t/, and /k/—has emerged as a critical contributor to speech perception. These phonemes require listeners to detect acoustic events separated by mere tens of milliseconds, such as the voice onset time (VOT) distinguishing /ba/ (~0 ms VOT) from /pa/ (~60 ms VOT). Therefore, speech relies heavily on rapid temporal modulations (20–50 Hz) for phoneme discrimination and slower spectral cues (formants) for vowel identity (Shannon et al., 1995).

In contrast, music exploits a vast array of sound sources, from vibrating strings to resonating air columns, creating rich acoustic diversity governed by principles of harmony, rhythm, and timbre rather than linguistic constraints. Thus, music prioritizes precise spectral resolution (harmonic relationships) and broader temporal integration (rhythmic patterns spanning seconds). Melodic pitch changes typically occur in small, precise intervals ranging from 1/12 to 1/6 of an octave (Vos and Troost, 1989), whereas speech intonation often spans more than half an octave (Patel et al., 1998). Meanwhile, musical notes generally maintain substantially longer durations than consonant phonemes, with melodic recognition becoming significantly impaired when note durations fall below approximately 160 ms (Warren et al., 1991).

4.2.2 Higher-Order Attributes

Beyond acoustics, domain-specific features drive categorical representations. Speech processing engages hierarchical networks for extracting phonological, syntactic, and semantic content (Hickok and Poeppel, 2007; Keshishian et al., 2023). Music recruits right-lateralized circuits for pitch contour analysis, melodic expectation, and emotional valence (Koelsch, 2011; Sankaran et al., 2024). Critically, these higher-order features are not merely perceptual but

shaped by learned schemata: for instance, phonetic boundaries in speech or tonal hierarchies in music are culture-dependent (McClelland and Elman, 1986; Patel, 2008). Such top-down influences suggest that auditory category representations arise from dynamic interactions between innate biophysical constraints and experience-dependent plasticity.

4.2.3 Intermediate Features

While spectrotemporal modulation provides a powerful framework for understanding early auditory processing, recent studies reveal its limitations in explaining neural responses in nonprimary auditory cortex (Norman-Haignere et al., 2015; Norman-Haignere and McDermott, 2018). On the other hand, a localist view struggles to specify how many—or which—semantic categories are encoded and requires ad hoc mechanisms for exemplar coding (Giordano et al., 2023). This unresolved gap underscores the need to investigate intermediate features that bridge raw sensory input and higher-order meaning.

Recent work by Giordano et al. (2023) has addressed this gap by leveraging DNNs trained on sound-to-event categorization. Their findings revealed that intermediate DNN layers, which neither purely encode acoustics nor semantics, best predict neural responses in the STG. These layers were found to capture abstract internal representations that were found to be behaviorally salient but distinct from categorical labels like "speech" or "music". For instance, STG responses to sounds like a meowing cat or a buzzing bee are better explained by intermediate DNN features than by either acoustic modulations or semantic embeddings, suggesting a hierarchical transformation where spectrotemporal patterns are progressively integrated into perceptually meaningful units in the STG.

Critically, these intermediate features may incorporate source-related attributes—such as the material of a sound-generating object (Gaver, 1993) or the biomechanics of vocal production (Mesgarani et al., 2014; Hamilton et al., 2021). This duality explains why DNNs outperform traditional models: they implicitly learn statistical regularities (e.g., harmonicity in music, abrupt onsets in speech) that are crucial for recognition but not explicitly categorical (not specific to speech or music). However, the opaque nature of the DNN prevents these results from revealing the computational fingerprints of the abstract features. Future research should portray these features to concretely resolve how the brain transforms waveforms into percepts. In Chapter 3, we addressed this problem utilizing the forward synthesis approach to sonify these features.

4.3 Sound Features Accounting for Hemispheric Asymmetry

The hemispheric lateralization of language functions stands as one of the most robust and clinically significant phenomena in human neuroscience. Since Broca's (1861) and Wernicke's (1874) pioneering observations of left-hemisphere lesions causing specific language impairments, the scientific community has sought to understand why and how the brain organizes these critical communicative functions asymmetrically. The hemispheric differences have been explained through multiple theoretical frameworks, each offering explanations for

how temporal and spectral information is differentially processed in the two hemispheres. The most prominent models include the temporal versus spectral processing hypothesis (Zatorre et al., 2002), the asymmetric sampling in time theory (AST) (Poeppel, 2003), and the spectrotemporal modulation framework (Flinker et al., 2019). While these models successfully explain many aspects of basic auditory processing, they face limitations when accounting for higher-order auditory functions.

4.3.1 Temporal versus Spectral Processing

The temporal versus spectral processing hypothesis (Zatorre et al., 2002), which emphasizes the trade-off between temporal and spectral resolution, proposes that the left hemisphere is specialized for rapid temporal changes in sound (Penhune et al., 1999; Ehrlé et al., 2001), while the right hemisphere excels in spectral resolution (Samson and Zatorre, 1994; Liégeois-Chauvel et al., 1998). Patients with left-hemisphere lesions often struggle with discriminating rapid auditory sequences, such as consonant-vowel transitions, whereas right-hemisphere damage impairs pitch and timbre perception (Robin et al., 1990). PET studies have consistently shown that cortical responses to temporal features are weighted towards the left hemisphere, while responses to spectral features are weighted towards the right one (Zatorre, 2001).

However, this model has been criticized for oversimplifying hemispheric differences. Natural speech and music contain intertwined temporal and spectral cues, making it difficult to attribute processing solely to one hemisphere. Additionally, the model does not fully explain why certain speech elements (e.g., vowels with steady spectral profiles) still engage the left hemisphere, suggesting that linguistic relevance may modulate these asymmetries (McGettigan and Scott, 2012).

4.3.2 Asymmetric Sampling in Time

Poeppel (2003) proposed the AST hypothesis arguing that hemispheric differences arise from distinct integration windows. According to AST, the left auditory cortex operates within shorter temporal windows (~20–80 ms), making it ideal for tracking phonemes and syllabic rates, while the right hemisphere uses longer windows (~150–300 ms) to capture prosodic and melodic contours.

This framework is supported by neural oscillation studies showing that left-lateralized gamma-band activity (~30–50 Hz) aligns with rapid speech cues, whereas right-hemisphere theta oscillations (~4–8 Hz) track slower rhythmic patterns (Giraud and Poeppel, 2012). The framework has also received substantial empirical support from functional neuroimaging studies, with numerous fMRI experiments demonstrating hemispherically asymmetric activation patterns corresponding to different temporal processing scales (Boemio et al., 2005; Jamison et al., 2006; Obleser et al., 2008). Notably, such asymmetries appear to be fundamental properties of the auditory system, observable both during resting states in adults (Giraud et al., 2007; Morillon et al., 2010) and in early developmental stages during infant auditory processing (Telkemeyer et al., 2009).

4.3.3 Spectrotemporal Modulation as A Unifying Framework

Flinker et al. have proposed a framework for auditory cortical asymmetry that partially unifies the above theories (Flinker et al., 2019). In their framework, auditory signals are decomposed into temporal modulations (amplitude fluctuations over time) and spectral modulations (frequency-band oscillations) (Chi et al., 2005). The left hemisphere preferentially processes temporal modulations in the 20–50 Hz range, which are critical for speech intelligibility, while the right hemisphere shows bias for spectral modulations (0.5–1.5 cycles/octave) underlying pitch and timbre perception.

The spectrotemporal modulation framework provides a quantifiable, biologically grounded computational model for auditory processing. The model has been applied to characterize cortical responses to various sounds (Shamma, 2001; Chi et al., 2005; Santoro et al., 2017) and to account for speech intelligibility (Chi et al., 1999; Elliott and Theunissen, 2009), thereby linking neural and perceptual representations of speech. It unifies aforementioned theories by showing that hemispheric asymmetries emerge from differential integration of modulation rates, offering a scalable model for both basic and clinical research.

Recent work shows that the neural computations in nonprimary auditory cortex may be characterized better by a DNN than modulation-based models (Norman-Haignere and McDermott, 2018). This suggests that the DNN might outperform modulation-based models also in accounting for brain asymmetry, a question that is explored in the Chapter 4.

5 Thesis Outline

This dissertation employs feature identification approaches to investigate cortical representations of sound features across two dimensions of neural activity: temporal dynamics and spatial organization.

In Chapter 2, we addressed the question of whether different modulation patterns with the same temporal rate can be reconstructed from neural responses. We also addressed whether the reconstructions could facilitate modulation of sensory processing. To this end, we employed a linear temporal response function to backward-reconstruct the 1-Hz amplitude modulations imposed on 4-Hz tone sequences. Two modulation patterns—repeatedly ascending and descending—were designed to assess whether the reconstructions hold perceptual and neural significance for listeners.

In Chapter 3, we investigated the sound features that primarily drive speech- and music selectivity in auditory cortical regions. We first mapped the layer-wise representation of natural sounds embedded in a state-of-the-art DNN onto neural responses, then synthesized novel sounds that maximized the neural responses predicted by the forward mapping. Subsequently, we conducted perceptual and neural validations of the synthesized sounds, along with

representational dissimilarity analysis, testing for a causal relation between the DNN-based synthetic sound features and sound categorization in the human auditory system.

In Chapter 4, we explored the sound features that account for the brain asymmetry of auditory processing. Building on the approach from Chapter 3, we synthesized novel sounds that maximized the difference between the predicted responses to these sounds in the left vs. right hemisphere (i.e., auditory cortical asymmetry). In a subsequent acoustic stimulus analysis, we transformed the synthesized sounds into various acoustical representations to identify the sound features that are most explanatory for auditory brain asymmetry.

Chapter 2

Neural and Auditory Fidelity of Basic Temporal Stimulus Patterns Reconstructed from Individual EEG Recordings

Based on

Xing, L., Riecke, L. (submitted). Neural and auditory validity of basic temporal stimulus patterns reconstructed from individual EEG recordings with a linear model

Abstract

This study investigated the neural and perceptual validity of temporal sound patterns reconstructed from individual scalp EEG recordings. Twenty healthy participants were presented with ascending, descending, and random loudness sequences (4-Hz tone repetitions carrying a 1-Hz intensity modulation) while EEG was recorded. A linear temporal response function (TRF) model was used to reconstruct stimuli from EEG responses, which were subsequently tested in behavioral and neural validation experiments with the same participants. Results demonstrate successful reconstruction of temporal stimulus patterns from individual EEG recordings, preserving 18.5% (ascending) and 31.4% (descending) of original loudness variations. Neural validation showed that reconstructed stimuli evoked activity patterns resembling those evoked by the original sequences (ascending: 26%, descending: 27%), whereas “inverted” stimuli (reconstructed from EEG responses with artificially inverted polarity) significantly reduced this resemblance (ascending: 14%, descending: 26%). Behaviorally, participants identified temporal pattern-matched stimuli significantly above chance with accuracies reaching on average up to 65.3%. These results demonstrate that auditory stimuli reconstructed from individual scalp EEG recordings have sufficient fidelity to induce neural and perceptual temporal patterns replicating those of their original counterparts. This insight suggests that individualized, EEG-based auditory stimulus reconstruction could serve as a non-invasive technique for targeted neuromodulation applications.

Keywords

Neural entrainment; Envelope reconstruction; Perception validation

1 Introduction

Neural activity in the auditory cortex evoked by auditory stimulation follows the slow temporal variations of the stimulation, in particular its broadband energy fluctuations (i.e., envelope). This phenomenon, usually termed as neural tracking (Lakatos et al., 2019), is thought to emerge from the sequential phase-resetting of neural oscillations by the rhythmic sensory input (Lakatos et al., 2005, 2007; Guo et al., 2017). Cortical tracking of dynamic sensory stimuli is often investigated with a linear model, known as temporal response function (TRF), that bridges the neural dynamics and the time-varying features of the evoking stimulation (Crosse et al., 2016). When such a model can be accurately fitted to stimulus-evoked brain response data, it can be further utilized in two ways: as a forward model to predict neural temporal responses to novel stimuli not included in the fitting procedure, or as a backward model to reconstruct dynamic stimulus features from novel neural activity patterns not included in the fitting procedure.

The reconstruction accuracy achieved by backward models is usually taken as a quantitative measure of the fidelity with which the neural response pattern represents the dynamics of a given stimulus feature. This approach has proven particularly valuable in studying cognitive modulation of neural tracking, including investigations of the influence of attentional selection (Lakatos et al., 2008; Ding and Simon, 2012a, 2012b; Zion Golumbic et al., 2013), context (Mesgarani et al., 2009), and prior knowledge (Wang et al., 2019). A common aspect of previous studies is that they validated the auditory stimulus reconstructions in the acoustic domain, by computing the correlation between the reconstructions and the original stimulus. Thus, these studies assessed only whether the reconstructed stimulus preserved *the physical temporal properties* of the original stimulus. While this approach allows inference about the neural representation of the dynamic stimulus feature, it leaves unclear whether the reconstructed stimulus also preserves the *neural and perceptual temporal properties* of the original stimulus, i.e., whether the reconstructed stimulus can reproduce the original brain-activity pattern and subjective sound experience.

We considered this an interesting question for the following reasons: if stimulus reconstructions can reproduce the neural and perceptual pattern of the original stimulus with sufficient accuracy as hypothesized, then the stimulus-reconstruction approach might be utilized to induce also other specific 'target' patterns. This presumed 'brain control' would be desirable in various neuromodulation contexts. In particular, it might serve the optimization of stimulation parameters in both, fundamental research aiming to establish a functional role for a specific brain-activity pattern, and clinical research aiming to restore or suppress that specific brain-activity pattern for treatment purposes (Mesgarani et al., 2009; Peter and Kleinjung, 2019).

The goal of the present study was to assess the accuracy with which stimulus reconstructions obtained with widely used methods (non-invasive EEG recordings and a multivariate linear regression model based on the TRF) can replicate the neural and perceptual temporal pattern of the original stimulus. We applied the following two-step approach individually to each participant:

In the first step (training phase), we induced a cortical activity pattern and aimed to identify stimuli that can potentially control (support or suppress) this pattern. To this end, we presented

a loudness sequence—a regularly repeating tone with loudness variations that were either patterned (ascending or descending sequence) or random (shuffled sequence, used as a baseline)—and recorded the induced cortical activity pattern with EEG. Using the TRF, we determined the mapping between stimulus loudness and the induced EEG activity, and reconstructed loudness sequences from the activity patterns. From these ‘brain-based’ reconstructions, we generated individual stimuli that we expected to enable control over the activity pattern induced by the original loudness pattern. Specifically, we expected that the reconstructed loudness sequence would replicate—and thereby potentially support—the activity pattern induced by the original sequence. We further explored whether artificial inversion of the polarity of EEG recordings from which the sequence was reconstructed (resulting in an anti-correlated sequence) can induce the opposite—a suppressive effect—on the originally induced activity pattern.

In the second step (validation phase), we verified these expectations. That is, we investigated whether the individually reconstructed stimuli can indeed support or suppress the activity pattern induced by the original loudness pattern. To this end, we mixed the reconstructed loudness sequence with the original loudness pattern (as presented in the training phase) and measured cortical and behavioral responses to this mixture and to the original loudness pattern. We expected that responses to the putatively supportive mixture would resemble responses to the original loudness pattern significantly better than the baseline (i.e., responses to a mixture including a reconstruction of the shuffled sequence). For the putatively suppressive mixture, we expected the opposite, i.e., a significantly reduced resemblance to original loudness pattern-evoked responses compared to baseline.

2 Methods

2.1 Participants

Twenty healthy university students (19 females, 1 male) participated in the study. Their age ranged from 17 to 28 years. All participants reported having normal hearing and no neurological or psychiatric disorders. All participants gave their written informed consent and received study credits or monetary vouchers for their participation. The experimental procedure was approved by the local ethics committee of the Faculty of Psychology and Neuroscience, Maastricht University (ERCPN # _233_09_02_2021).

2.2 Stimulation

The loudness sequence consisted of a 250-Hz tone repeating at a rate of 4 Hz (tone duration: 200 ms including 5-ms on/off ramps, followed by a 50-ms inter-tone gap). The intensity of the tone was varied uniformly within each sequence across four equidistant levels: 47, 50, 53, and 56 dB SPL. As shown in Figure 1, three different types of sequence were generated. In the original *shuffled* sequence (referred to as ORI_{SHU}), these intensities were arranged randomly across the entire sequence. In the original *ascending* and *descending* sequences (referred to as ORI_{ASC} and ORI_{DES}), the four intensities were monotonically increased or decreased across

every set of four consecutive tones, resulting in an ascending or decreasing loudness pattern that cycled at 1 Hz across the entire sequence. The auditory stimulation was presented diotically using an audio mixer (DigiLive 16, Studiomastr), and insert earphones (ER-3A, EAR-TONE).

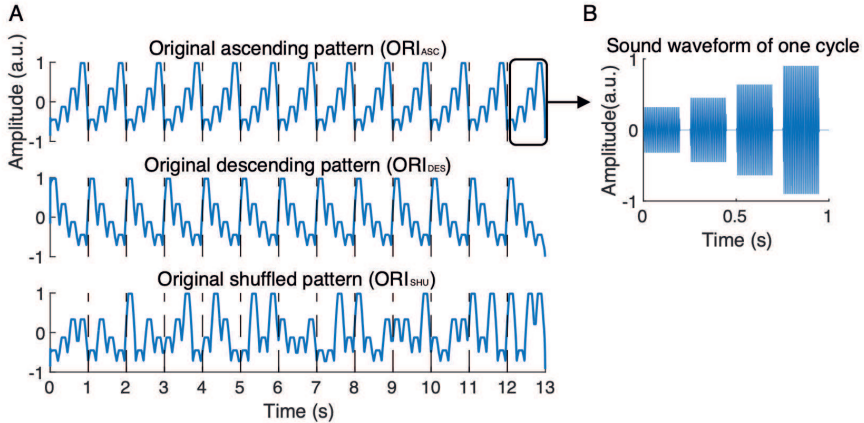


Figure 1. General stimulus design. Panel A shows from top to bottom the amplitude envelope of the ascending loudness sequence (ORI_{ASC}), descending sequence (ORI_{DES}), and random sequence (ORI_{SHU}) over an interval of 13 s. Vertical dashed lines delineate cycles of intensity patterns. Panel B shows an enlarged version of the sound waveform within a single cycle of the ascending pattern.

2.3 EEG Recording

EEG was recorded from the scalp with 31 Ag/AgCl electrodes mounted in an elastic cap and referenced to the left mastoid, using a BrainAmp amplifier (Brain Products, Munich, Germany). The electrodes were placed according to the International 10-20 system: Fz, FCz, Cz, CPz, Pz, POz, FP1/FP2, F3/F4, F7/F8, FC1/FC2, FC5/FC6, C3/C4, T7/T8, CP1/CP2, CP5/CP6, P3/P4, P7/P8, O1/O2, and A2 (right mastoid). The vertical EOG was recorded with an additional electrode placed below the left eye. Electrodes were connected to the skin using conductive gel, and the impedance of each electrode was reduced to below 5 k Ω . The recorded EEG signals were low-pass filtered (cutoff frequency: 250 Hz, analog filter) and digitized using a sampling rate of 1000 Hz.

2.4 Statistical Analysis

Participants' individual measures were submitted to second-level group analyses using statistical tests for repeated measures. The assumption of normality was verified with Lilliefors tests; no distribution was found to deviate significantly from normality. Parametric statistical tests (paired t-tests) and a significance criterion $\alpha = 0.05$ were used. Type-I error probabilities

inflated by multiple comparisons were corrected by controlling the false-discovery rate (Benjamini and Hochberg, 1995).

2.5 Procedure

The two phases of the experiment (training and validation) were conducted within a single session on the same day in a sound-attenuating and electrically shielded chamber. In between these two phases, EEG data from the training phase were analyzed and stimuli for the validation phase were reconstructed from these data, while participants took a break with the EEG cap staying on their head.

2.5.1 Specific Methods in the Training Phase

2.5.1.1 Procedure

After participants gave informed consent, they were prepared for the EEG recording and seated in the recording chamber. They were asked to sit still and focus attention on a muted natural documentary video that was presented continuously on a PC screen. While watching the video, they were presented with six blocks of loudness sequences while EEG was recorded, interleaved with short breaks. As illustrated in Figure 2, each block contained 48 sequences (16 ORI_{ASC}, 16 ORI_{DES}, and 16 ORI_{SHU}) in random order, where each sequence consisted of 52 tone repetitions and lasted in total 13 s. Each block was preceded by a 2-s excerpt from the shuffled sequence (not shown in Figure 2); this dummy interval served to reduce stimulation-onset responses at the onset of the first sequence within the block and was excluded from the EEG-data analysis. Dummy interval and sequences within blocks were presented continuously, resulting in an overall isochronous stimulation.

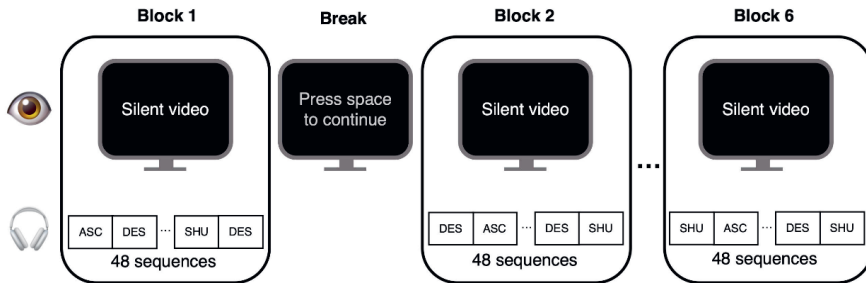


Figure 2. Design of the training phase. Each box represents a single block of stimulation. Each block involved the continuous and random presentation of 16 ascending (ORI_{ASC}), 16 descending (ORI_{DES}), and 16 random (ORI_{SHU}) loudness sequences each lasting 13 s; see Figure 1A. During the auditory stimulation, participants watched a silent video while their EEG was recorded. Blocks were interleaved with stimulation-free breaks. In total, 288 sequences (96 ORI_{ASC}, 96 ORI_{DES}, and 96 ORI_{SHU}) were presented in the training phase.

2.5.1.2 EEG Data Analysis

EEG data were processed in MATLAB (2019b, The MathWorks, Inc., Natick, MA) using Fieldtrip (Donders Institute for Brain, Cognition and Behavior, Radboud University Nijmegen, The Netherlands: <http://www.ru.nl/neuroimaging/fieldtrip/>) and custom scripts. The continuous EEG data from each block were band-pass filtered (Butterworth filter, cutoff frequencies: 0.5 and 10 Hz, filter order: 4) and segmented into 13-s epochs resembling the loudness sequences (from -50ms until 12950ms relative to the onset of the first tone of the sequence). The average amplitude in the interval from -50 to 0ms was subtracted from each epoch. The epoched data were resampled at 200 Hz and re-referenced to the average of the two mastoids. To reduce artifacts, independent component analysis (ICA) was applied and artifactual components resembling blinks and eye or head movements were identified by visual inspection and rejected (on average 6.5% of all components). The artifact-reduced EEG channel data were low-pass filtered (Butterworth filter, cutoff frequency: 8 Hz, filter order: 4) and normalized using z-scores. The data were split into two sets that were used respectively for model fitting and stimulus reconstruction (see next section). Each of these datasets contained 144 epochs (equally distributed across ORI_{ASC} , ORI_{DES} , and ORI_{SHU}) that were separately averaged to obtain a single average waveform for each sequence type (ORI_{ASC} , ORI_{DES} , and ORI_{SHU}).

2.5.1.3 Stimulus Reconstruction

Auditory stimuli were reconstructed from the EEG recordings using a multivariate linear model based on the temporal response function, as implemented in the mTRF toolbox (Crosse et al., 2016) in MATLAB.

In the first step, the mapping of the amplitude envelope of the presented loudness sequences onto the recorded spatiotemporal EEG-response patterns was determined. For this, the temporal response function was fitted for each EEG channel at 60 time points between 0 ms and 300 ms (representing the timing of EEG activity relative to the loudness sequence). This was done using regularized linear regression and a six-fold cross-validation procedure. The model-fitting data (consisting of three 13-s waveforms, see previous section) were concatenated and split into six segments each spanning an interval of 6.5s (i.e., 26 tone presentations). The first segment was regarded as test data while the other five segments were regarded as training data and fed into the cross-validation procedure. On each iteration of the procedure, a different set of four training segments was used to fit the model for a range of values of the regularization parameter ($\lambda = 10^{-6}, 10^{-4}, \dots, 10^6$). The left-out training segment was used to validate the fitted model for each λ -value, which was done by reconstructing amplitude envelopes from the EEG data and assessing the reconstruction accuracy. The latter was quantified as the linear correlation (Pearson's R) of the reconstructed amplitude envelope (referred to as REC_{ASC} , REC_{DES} , or REC_{SHU}) and the amplitude envelope of the respective original loudness sequence (ORI_{ASC} , ORI_{DES} , or ORI_{SHU}). At the end of the cross-validation procedure (after each training segment had served for model validation), reconstruction accuracies were averaged across iterations and the λ -value that yielded the maximum reconstruction accuracy was retained to train the model with all training data. Then the model was used to reconstruct the test data and implement the second step below.

In the second step, the mapping identified in the first step was applied to the EEG recordings of the stimulus-reconstruction data (see previous section) to reconstruct stimulus-amplitude envelopes. This yielded a 13-s long amplitude envelope for each sequence type (REC_{ASC} , REC_{DES} , and REC_{SHU}) that was rescaled to fit the amplitude range from 0 to 1. To assess the validity of these reconstructed sequences in the conventional, acoustic domain, the reconstruction accuracy was computed using Pearson's R as described above.

2.5.2 Specific Methods in the Validation Phase

2.5.2.1 Rationale and Stimuli

The auditory stimulation in the validation phase was composed of two simultaneous components: the original loudness pattern (ORI_{ASC} or ORI_{DES}), which served to induce the same cortical activity pattern as in the training phase, and an individually reconstructed loudness sequence (REC_{ASC} , REC_{DES} , or REC_{SHU} , obtained from the training phase), which we expected to selectively modulate that activity pattern. We specifically expected that adding putatively pattern-inducing stimulation (e.g., REC_{ASC}) to its original stimulation (ORI_{ASC}) would result in an activity pattern that largely preserves the original pattern induced by ORI_{ASC} , whereas adding random stimulation (REC_{SHU}) would fail to preserve that pattern. We refer to this as the 'support' hypothesis.

We further explored a 'reduction' hypothesis, according to which a temporal brain-activity pattern is largely reduced by anti-phasic stimulation. We generated anti-correlated stimulation by reconstructing loudness sequences from EEG recordings, of which we artificially inverted the polarity (by inverting the sign of each amplitude value). We refer to these stimuli as $-REC_{ASC}$, $-REC_{DES}$, and $-REC_{SHU}$. Under the reduction hypothesis, adding the putatively pattern-reducing stimulation (e.g., $-REC_{ASC}$) to its original stimulation (ORI_{ASC}) would result in an activity pattern that largely reduces the original pattern induced by ORI_{ASC} , whereas adding random stimulation ($-REC_{SHU}$) should do so to a smaller degree.

The stimuli were generated as shown in Figure 3. To enable testing the support hypothesis, REC_{ASC} was added to ORI_{ASC} , resulting in a stimulus we refer to as 'support mixture' ($SUPMIX_{ASC}$). Analogously, to enable testing the reduction hypothesis, $-REC_{ASC}$ was added to ORI_{ASC} , resulting in a stimulus we refer to as 'reduction mixture' ($REDMIX_{ASC}$). The reconstructed random sequence was used to establish a baseline; for this, REC_{SHU} was added to ORI_{ASC} , resulting in a 'control mixture' ($CTRMIX_{ASC}$). All mixtures were normalized for peak amplitude by dividing each mixture by its absolute maximum. The mixtures were then applied to 12-s long 4-Hz tone sequences to obtain the stimuli for the validation part. Each mixture was split into four segments of 3-s duration to obtain four unique instances of each mixture. All of these steps were also applied to the descending sequence type to obtain corresponding descending stimuli ($SUPMIX_{DES}$, $REDMIX_{DES}$, and $CTRMIX_{DES}$).

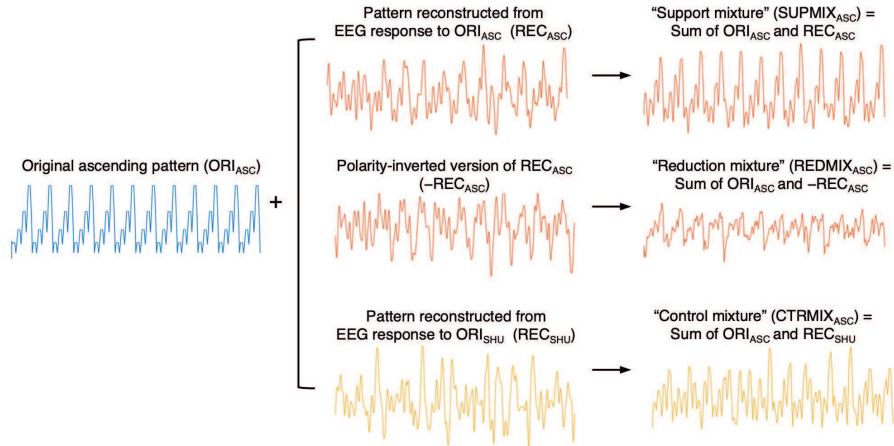


Figure 3. Stimulus design in validation phase exemplified for the ascending sequence type. The original loudness pattern (e.g., ORI_{ASC} , left column, blue) was mixed with an individual brain-based reconstruction of a loudness sequence (middle column) to verify whether the latter sequence can selectively modulate the cortical activity pattern induced by the original loudness pattern. The mixture was designed to either support the original pattern (stimulus $SUPMIX_{ASC}$, for testing a 'support' hypothesis, top row, red) or reduce it (polarity-inverted stimulus $REDMIX_{ASC}$, for testing a 'reduction' hypothesis, middle row, red). A control mixture based on a reconstruction of a shuffled loudness sequence (stimulus $CTRMIX_{ASC}$, bottom row, orange) was used as baseline. Stimuli of the descending sequence type were obtained analogously.

2.5.2.2 Procedure and Task

In the validation phase, participants performed a three-interval two-alternative forced-choice match-to-sample task while EEG was recorded. The trial design is sketched in Figure 4. Each trial started with the presentation of an original loudness sequence (e.g., ORI_{ASC} , test stimulus, duration: 3 s), which was visually indicated to the participant as "Interval 0". This was followed by the consecutive presentation of the corresponding support or reduction mixture (e.g., $SUPMIX_{ASC}$, comparison stimulus 1) and the corresponding control mixture (e.g., $CTRMIX_{ASC}$, comparison stimulus 2). The two comparison stimuli (duration: 3 s each) were presented in random order and visually indicated as "Interval 1" and "Interval 2". To reduce temporal smearing of ERPs or temporal predictions from one interval into the next, the three intervals were separated by gaps, the duration of which was randomly chosen from 800, 900, 1100 and 1200 ms. Participants were instructed to focus on the auditory stimulation and compare the sounds in intervals 1 and 2 with the sound in interval 0. They were further instructed to report which comparison interval (1 or 2) sounded more similar to the test interval 0, by pressing a key (labelled 1 or 2) on a keyboard during a visually indicated response interval, which was presented immediately after interval 2. The next trial started 1 s after the button press. Participants performed 64 trials of the task, with a short break after 32 trials. Sequence type

(ascending or descending) and modulation type (support or reduction) were presented in random order and equally distributed across trials. Each trial constituted a unique pairing of the comparison stimuli, i.e., each of the four instances of a main mixture was combined only once with each of the four instances of the corresponding control mixture; this was done for each sequence type (ASC and DES) and each modulation type (support or reduction).

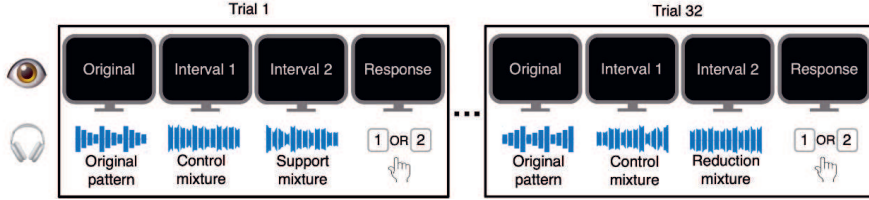


Figure 4. Design of the validation phase. Each rectangle represents a single trial of the match-to-sample task. Each trial consisted of four visually cued intervals. During the first interval, the test stimulus was presented, which was an original loudness pattern (ORI_{ASC} or ORI_{DES} ; see exemplary trials 32 and 1, respectively). During the second and third interval, the individual comparison stimuli were randomly presented. Each comparison stimulus was a unique mixture of the corresponding original loudness pattern and a reconstructed loudness sequence. One of the comparison stimuli was expected to selectively support the cortical activity pattern induced by original loudness pattern (support mixture $SUPMIX_{ASC}$ or $SUPMIX_{DES}$; see exemplary trial 1) or reduce it (reduction mixture $REDMIX_{ASC}$ or $REDMIX_{DES}$; see exemplary trial 32). The other comparison stimulus (control mixture $CTRMIX_{ASC}$ or $CTRMIX_{DES}$, reconstructed from EEG responses to a shuffled loudness pattern) served as baseline. During the last interval, participants reported which of the two comparison stimuli matched better with the test stimulus. In total, 64 unique trials were presented in the validation phase while EEG was recorded.

2.5.2.3 EEG Data Analysis

EEG recordings from the validation phase were preprocessed as in the training phase, except for the following differences: artifact reduction was done by re-applying the ICA weights that were obtained in the training phase and removing the artifactual components as before (see above). Epochs started 100 ms before the onset of a given stimulus (original pattern, main mixture, or control mixture) and lasted 3100 ms. The pre-stimulus interval (first 100 ms) was treated as baseline. Channels that tracked the loudness pattern (i.e., showed a prominent 1-Hz peak in the EEG power spectrum elicited by ORI_{ASC} and ORI_{DES}) in the training phase were selected and averaged. ERPs were computed for each stimulus type by averaging the artefact-reduced epochs across trials; this was done separately for each of the four conditions (two sequence types [ASC, DEC] $\times$ two modulation types [support, reduction]).

To assess the validity of the reconstructed loudness sequences in the neural domain, we quantified the degree to which the cortical activity pattern induced by the reconstructed sequence resembled the activity pattern induced by the original loudness pattern. We

computed the linear correlation (Pearson's R) between the ERP induced by the comparison stimulus (e.g., SUPMIX_{ASC} or CTRMIX_{ASC}) and the ERP induced by the corresponding test stimulus (e.g., ORI_{ASC}) for each condition. To test the support hypothesis, we statistically compared the correlation coefficient obtained with stimulus SUPMIX_{ASC} vs. the correlation coefficient obtained with stimulus CTRMIX_{ASC}. Correspondingly, we tested the reduction hypothesis by comparing correlation coefficients obtained with REDMIX_{ASC} vs. CTRMIX_{ASC}. The same analyses were applied to the descending sequence type to explore the generalizability of the results across loudness patterns.

2.5.2.4 Behavioral Data Analysis

To assess the validity of the reconstructed loudness sequences in the perceptual domain, we quantified the degree to which the perceived loudness pattern of the reconstructed sequence resembled the originally perceived loudness pattern. Under the support hypothesis, loudness sequences reconstructed from cortical responses to loudness patterns should match the original pattern better than sequences reconstructed from responses to no-patterns (i.e., shuffled loudness sequences). This expectation is met when participants perceive the stimulus SUPPMIX (rather than the stimulus CTRMIX) to sound alike the original stimulus; thus, we labelled behavioral responses representing the stimulus SUPPMIX as 'correct'. CTRMIX responses on support trials were labeled as 'incorrect'.

Under the reduction hypothesis, the opposite is expected, i.e., anti-correlated loudness sequences (reconstructed from polarity-inverted cortical responses to loudness patterns) should match the original pattern *less* than sequences reconstructed from responses to no-patterns. This expectation is met when participants perceive the stimulus CTRMIX (rather than the stimulus REDMIX) to sound alike the original stimulus; thus, we labelled behavioral responses representing the stimulus CTRMIX on reduction trials as 'correct'. REDMIX responses were labeled as 'incorrect'.

Accuracy was computed as the percentage of trials on which participants gave a 'correct' response. To test the aforementioned hypotheses, the observed accuracies were statistically compared to the 50% chance level (the average accuracy that participants would achieve when only guessing). As before, the same analyses were applied to the descending sequence type to explore the generalizability of the results.

3 Results

3.1 Acoustic Stimulus Validation

Figure 5 shows results from the stimulus reconstruction. In Figure 5A, it can be seen that the loudness sequence that was reconstructed from EEG responses to the ascending sequence in an individual participant (REC_{ASC}) resembled well the ascending pattern of the original sequence (ORI_{ASC}). From the sequence reconstructed from responses to the random sequence (REC_{SHU}), this resemblance was less visible. Figure 5B shows the average reconstruction accuracy of the ascending sequence, which was found to be 0.44 ± 0.03 (Pearson's R , mean

$\pm$ SEM across participants). As suggested above, the reconstruction of the random sequence resembled the original ascending sequence significantly less well ($t_{(19)} = 6.59$, $p_{FDR} = 1.31 \times 10^{-6}$), although a positive average correlation ($R = 0.13 \pm 0.02$) could still be observed. The latter was attributable to the common tone-repetition rate, which was inevitably enhanced by the way in which the reconstructed loudness sequence were generated (see Methods). To obtain better estimates of the accuracy of the reconstruction of the actual loudness pattern vs. the reconstructed tone repetitions, we bandpass-filtered the EEG data into two separate 0.1-Hz wide frequency bands centered respectively on 1 Hz (resembling the loudness pattern) and 4 Hz (resembling the tone repetitions; Butterworth filter, filter order: 4) and repeated the stimulus-reconstruction analysis. The results in Figure 5C show that tone repetitions could be reconstructed with very high accuracy of 0.92 ± 0.01 , while the superimposed loudness pattern could still be reconstructed with an accuracy of 0.43 ± 0.07 . The latter value indicates that the reconstructed stimulus preserved on average 18.5% of the original loudness variations. This indicates that the brain-based loudness sequence resembled the ascending loudness pattern well beyond the tone-repetition rate on which it was superimposed.

As shown in Figures 5D–F, a consistent pattern of results could be observed for the descending loudness sequence: the reconstruction of the descending sequence (REC_{DES}) resembled the original descending sequence (ORI_{DES}) significantly better than the reconstruction of the random sequence (REC_{SHU} , $t_{(19)} = 6.77$, $p_{FDR} = 1.31 \times 10^{-6}$). The reconstruction accuracy of the 1-Hz loudness pattern reached on average a value of 0.56 ± 0.07 , indicating that the reconstructed stimulus preserved on average 31.4% of the original loudness variations. The 4-Hz pattern could be reconstructed with an average accuracy of 0.93 ± 0.01 .

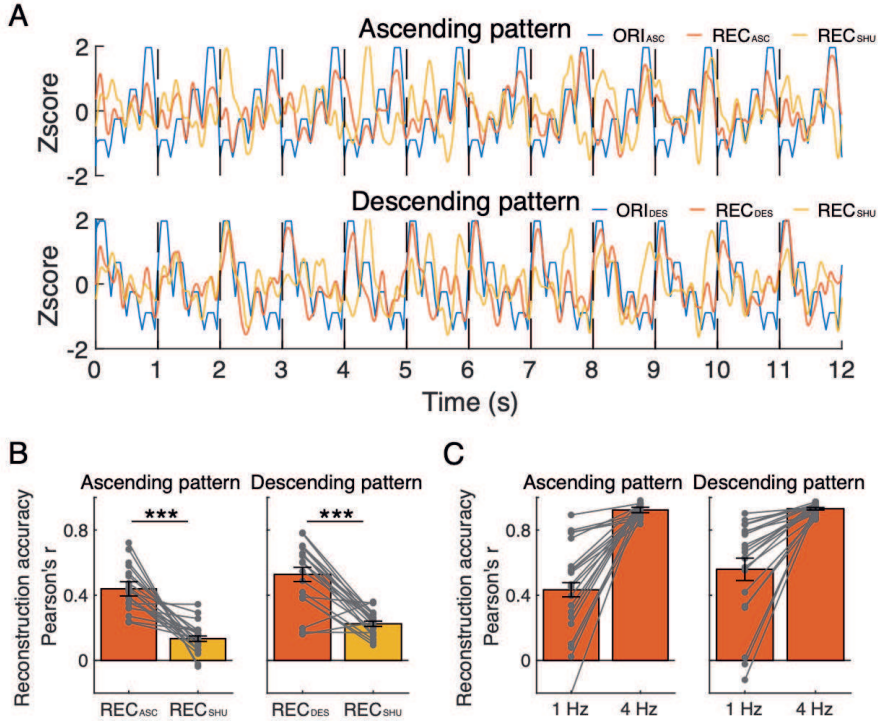


Figure 5. Stimulus-reconstruction results. **A.** The waveform shows the original ascending loudness pattern (blue), the brain-based reconstruction of this pattern (red), and the reconstruction of a random sequence (orange) for a single participant. Vertical dashed lines delineate cycles of loudness patterns. **B.** The bar plot shows participants' average reconstruction accuracy (0.44 ± 0.03 across participants) for the ascending pattern. The red bar represents the correlation between the original ascending pattern (ORI_{ASC}) and the reconstruction of this pattern (REC_{ASC}). The orange bar represents the correlation between the original ascending pattern (ORI_{ASC}) and the reconstruction of the random sequence (REC_{SHU}). Gray lines represent data from individual participants. *** represents $P < 0.001$. **C.** The bar plot shows the same data as the red bar in panel B, but after filtering the loudness sequences and recorded brain activity into a narrow frequency band centered on 1 Hz (left bar, resembling the repetition rate of the loudness pattern) or 4 Hz (right bar, resembling the repetition rate of the individual tones).

3.2 Neural Stimulus Validation

One goal of the present study was to explore the neural effects of stimuli reconstructed from individual EEG recordings. We specifically aimed to assess the degree to which these stimuli can be utilized to support or hamper the brain-activity patterns from which the stimuli were originally reconstructed. Based on our observation of good reconstruction accuracies (see

previous section), we expected that the reconstructed loudness sequences would evoke spatiotemporal EEG patterns that resemble well the patterns evoked by the original sequences.

Results on the support hypothesis are shown in Figures 6A–C. Figure 6A shows ERPs evoked respectively by the original ascending loudness sequence (ORI_{ASC}), the mixture of ORI_{ASC} and reconstructed ascending sequence ($SUPMIX_{ASC}$), and the mixture of ORI_{ASC} and reconstructed random sequence ($CTRMIX_{ASC}$) for a single participant. Contrary to our expectation, the waveforms induced by ORI_{ASC} and $SUPMIX_{ASC}$ did not strongly resemble each other based on visual inspection. However, a correlation analysis confirmed that the waveforms were positively correlated as expected: as shown in Figure 6B, the correlation between the ERP evoked by the original ascending sequence (ORI_{ASC}) and the ERP evoked by the reconstructed ascending sequence ($SUPMIX_{ASC}$) was on average 0.51 ± 0.01 (Pearson's R , mean $\pm$ SEM across participants). This correlation was significantly higher ($t_{(19)} = 12.087$, $p_{FDR} = 1.529 \times 10^{-10}$) than that between the ERP evoked by the original ascending sequence (ORI_{ASC}) and the ERP evoked by the reconstructed random sequence ($CTRMIX_{ASC}$), which was on average only 0.37 ± 0.02 . This indicates that the reconstructed ascending stimulus evoked a brain-activity pattern that resembled the originally induced activity pattern above chance. The result did not generalize to the descending pattern: the ERP evoked by the reconstructed descending sequence did not resemble the originally induced ERP better than the ERP evoked by the reconstructed shuffled sequence ($t_{(19)} = 4.259$, $p_{FDR} = 1.0$). In sum, these neural results provide partial evidence for the support hypothesis.

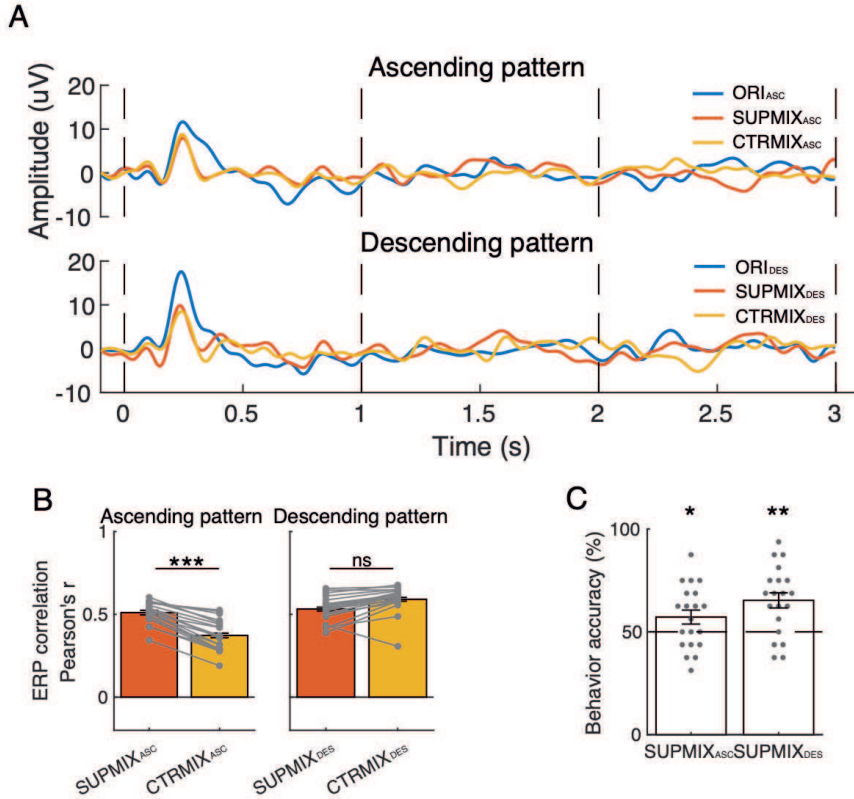


Figure 6. Neural and behavioral results on the support hypothesis. **A.** The waveforms show ERPs evoked by the original ascending loudness pattern (ORI_{ASC}, blue), ORI_{ASC} and the brain-based reconstruction of this pattern (SUPMIX_{ASC}, red), and ORI_{ASC} and the reconstruction of a random sequence (CTRMIX_{ASC}, orange) for a single participant, averaged across EEG channels whose activity in the training phase strongly tracked the loudness pattern. Vertical dashed lines delineate cycles of loudness patterns. **B.** The bar plot shows participants' average correlation (0.51 ± 0.01 across participants) between the ERP evoked by the original ascending pattern and the ERP evoked by the reconstructed stimuli. The red bar represents the correlation for the stimuli hypothesized to support the originally evoked ERP (SUPMIX_{ASC}). The orange bar represents this correlation for the control stimuli (CTRMIX_{ASC}). Gray lines represent data from individual participants. **C.** The bar plot shows participants' average behavioral accuracy (mean $\pm$ SEM across participants) for the ascending pattern (left bar) and descending pattern (right bar). The horizontal dashed line represents the chance level. Dots represent data from individual participants. Asterisks represent significant differences from chance. N.S.: non-significant; *, **, ***: $P < 0.05, 0.01, 0.001$.

Neural results on the reduction hypothesis are shown in the same format in Figures 7A–C. As shown in Figure 7B, the correlation between the ERP evoked by the original ascending

sequence (ORI_{ASC}) and the ERP evoked by the inverted reconstructed ascending sequence (REDMIX_{ASC}) was on average 0.38 ± 0.03 (Pearson's R , mean $\pm$ SEM across participants). In line with our hypothesis, this correlation was significantly *lower* ($t_{(19)} = 20.160$, $p_{\text{FDR}} = 1.679 \times 10^{-13}$) than that between the ERP evoked by the original ascending sequence (ORI_{ASC}) and the ERP evoked by the reconstructed random sequence (CTRMIX_{ASC}), which was on average 0.70 ± 0.03 . This indicates that the polarity-inverted, reconstructed ascending stimulus evoked a brain-activity pattern that resembled the originally induced activity pattern rather poorly. This result was found to generalize to the descending pattern: the ERP evoked by the polarity-inverted reconstructed descending sequence resembled the originally induced ERP significantly less than the ERP evoked by the reconstructed random sequence ($t_{(19)} = 16.135$, $p_{\text{FDR}} = 3.783 \times 10^{-12}$). In sum, these results provide neural support for the reduction hypothesis.

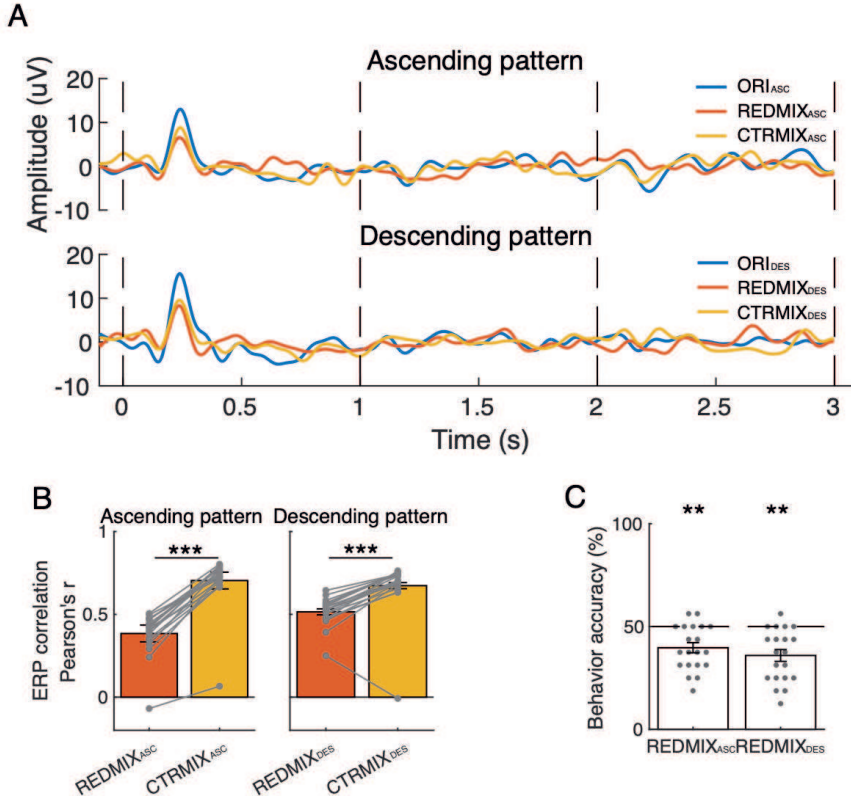


Figure 7. Neural and behavioral results on the reduction hypothesis. Panels A, B, and C are shown in the same format as in Figure 6.

3.3 Behavioral Stimulus Validation

Another goal of the present study was to assess the perceptual quality of the stimuli reconstructed from the individual EEG recordings. Based on our observation of good reconstruction accuracies, we expected that the reconstructed loudness sequences would elicit auditory percepts that preserve the salient loudness pattern of the original sequences. Results from the match-to-sample task addressing this “support hypothesis” are shown in Figure 6C. On average, participants perceived $57.19 \pm 3.40\%$ (mean $\pm$ SEM across participants) of the putatively pattern-supporting stimuli to match best the original ascending loudness pattern. This value was found to be significantly higher than the chance level ($t_{(19)} = 2.11$, $p_{FDR} = 0.024$). The result generalized to the descending pattern (accuracy: $65.31 \pm 3.70\% > 50\%$, $t_{(19)} = 4.14$, $p_{FDR} = 3.693 \times 10^{-4}$), therewith providing converging behavioral evidence for the support hypothesis.

Behavioral results on the reduction hypothesis are shown in Figure 7C. On average, participants perceived $60.31 \pm 2.45\%$ (mean $\pm$ SEM across participants) of the random stimuli to match best the original ascending loudness pattern, which was significantly above chance

($t_{(19)} = 4.203$, $p_{FDR} = 3.693 \times 10^{-4}$). This result also generalized to the descending pattern (accuracy: $64.06 \pm 2.86\% > 50\%$, $t_{(19)} = 4.911$, $p_{FDR} = 1.938 \times 10^{-4}$), therewith providing converging behavioral evidence for the reduction hypothesis.

4 Discussion

The goal of the present study was to explore the neural and perceptual fidelity of auditory stimuli reconstructed from individual EEG recordings. We first induced a cortical activity pattern from which we reconstructed stimulation designed to control that specific pattern. We then aimed to verify whether the reconstructed stimulation could indeed modulate both the aforementioned activity pattern and the accompanying auditory percept. Using basic temporal patterns of a repeating tone and a linear model, we confirmed that stimuli reconstructed from individual EEG recordings preserve the temporal patterns of the original stimuli with sufficient accuracy. More importantly, using additional EEG and behavioral measurements, we confirmed that the reconstructed stimuli replicate neural and perceptual responses to the original stimuli with above-chance accuracy. Overall, these observations provide insights into the validity of individual EEG-based reconstructions of basic auditory temporal patterns at multiple (acoustic, neural, and behavioral) levels. This may form a basis for further exploration of the proposed approach in the context of stimulation optimization for targeted brain modulation.

4.1 Acoustic Validity

We observed that a linear model can reconstruct basic, quasi-periodic loudness patterns of a repeating tone from individual EEG recordings of 62.4 min with satisfactory accuracy: the brain-based reconstructions preserved on average 18.5% and 31.4% of the original across-tone loudness variations (for ascending and descending pattern, respectively). Thus, although a major portion of these original variations got buried probably in EEG-measurement noise and model inaccuracies, a significant portion could still be successfully recovered. This recovered portion constituted a solid basis for the neural and behavioral validation.

The more general auditory variations induced by the presence vs absence of auditory input (i.e., the regular 4-Hz occurrence of the tone) was preserved to on average 84.6% and 82.5% (for ascending and descending pattern, respectively). The observed reconstruction accuracies are considerably higher than those usually reported in EEG or MEG studies using speech stimuli, which are more in the range from 1% to 9% (Ding and Simon, 2012a, 2012b; Zion Golumbic et al., 2013). A likely reason is the fact that the loudness pattern in our stimuli repeated at a single fixed rate of 4 Hz, whereas amplitude-modulation rates of auditory speech are quasi-periodic and span a broader frequency range (including delta and theta bands) that may be more difficult to reconstruct from brain activity (Ding and Simon, 2014; Lakatos et al., 2019; Duecker et al., 2024).

4.2 Neural Validity

We specifically aimed to assess the degree to which the reconstructed stimuli can be used to support or hamper the brain-activity patterns from which these stimuli were originally reconstructed. Based on our observation of good reconstruction accuracies, we expected that the reconstructed loudness sequences would evoke spatiotemporal EEG patterns that resemble well the patterns evoked by the original sequences.

Results show that the reconstructed ascending stimulus evoked a brain-activity pattern that resembled the originally induced activity pattern significantly above chance. The result did not generalize to the descending pattern, perhaps because of temporal asymmetries in auditory coding and perception that may result from stronger sensitivity to up-ramping sounds (Neuhoff, 1999; Deneux et al., 2016; Irsik et al., 2021). Similarly, the inverted, reconstructed ascending stimulus evoked a brain-activity pattern that resembled the originally induced activity pattern significantly *below* chance. This result generalized to the descending pattern.

In sum, our neural results provide partial evidence for the support hypothesis and converging evidence for the reduction hypothesis.

4.3 Perceptual Validity

Participants perceived the putatively ascending pattern-supporting stimuli to match best the original ascending loudness pattern significantly above chance. They also perceived the *inverted* ascending stimuli to match best the original ascending loudness pattern significantly *below* chance. Both results generalized to the descending pattern. Thus, these results provide converging behavioral evidence for both the support and reduction hypothesis.

4.4 Novel Aspects and Limitations

A novel aspect of our study is the individualized stimulus validation at neural and behavioral levels. This allowed exploring the potential usefulness of the EEG-based stimuli for brain control, which might be potentially used for clinical treatment (e.g., stimulation for intractable tinnitus) (Chang and Zeng, 2012). A strength of the proposed approach is that the validation is based on statistical comparison with a dedicated no-pattern condition, whereas most previous studies relied on surrogate (permutation-based) null distributions.

Our interpretation of the results should be treated with caution for two reasons. First, our stimuli for neural and behavioral validation comprised two components, of which one served to induce the brain activity pattern, and the other served to modulate (i.e., support or reduce) that pattern. It should be noted that these two components were mixed digitally (i.e., in the ‘pre-acoustic’ domain), implying that they could interact before entering participants’ ears. It means the mixtures were processed as a single sound object, instead of their interaction in the neural or perceptual domain and consequently hampers firm conclusions about neural or perceptual modulation. To address this limitation and extend the physiological interpretation of the observed suppression effects, a follow-up study employing dichotic presentation would be

highly informative. In such a design, the support and reduction sounds would be delivered separately to the left and right ears via headphones, rather than being combined into a single channel. Second, another limitation is that our sample included almost exclusively females, leaving unclear whether our findings generalize to the non-female population.

4.5 Conclusion

Beyond the aforementioned limitations, our study confirms that auditory stimuli reconstructed from individual, non-invasive, 62-min long EEG-recordings of neural activity can replicate the original neural and perceptual patterns significantly above chance. This provides useful insights into the fidelity of EEG-based stimulus reconstructions and can inspire their further development in future studies that aim to modulate specific brain-activity patterns with optimized stimulation. Such studies should take into account the limitations of the present work, e.g., by presenting the two stimulation components—one for pattern induction and the other for pattern modulation—through independent sensory channels, which would allow for a more direct assessment of the neural and perceptual modulation potential of EEG-based auditory stimulation.

Chapter 3

Deep Sound Synthesis Matched to Brain Activity Recapitulates Preferential Responses to Speech and Music

Based on

Xing, L., Formisano, E., & Riecke, L. (2026). Deep Sound Synthesis Matched to Brain Activity Recapitulates Preferential Responses to Speech and Music. *The Journal of neuroscience*, 46(9), e1651252025. <https://doi.org/10.1523/JNEUROSCI.1651-25.2025>

Abstract

The human auditory system extracts meaning from sounds in the environment by transforming acoustic input signals into semantic categories, such as speech and music. Although distinct acoustic features give rise to these categorical percepts and to preferential responses in spatially segregated regions in the auditory cortex, the nature of the internal representations underlying this transformation remains poorly understood. Here, we combined neuroimaging, a deep neural network (DNN), brain-based sound synthesis, and psychophysical testing in human participants of either sex to investigate the internal sound features encoded in speech- and music-selective regions of the auditory cortex and their functional role in sound categorization. We found that sounds synthesized from cortical activity patterns—though acoustically dissimilar to natural speech and music sounds—nonetheless elicited similar categorical cortical and behavioral responses. These results suggest that the auditory cortex relies on internal, abstracted representations of category structure that are not reducible to the natural acoustic properties of speech and music. Our findings provide new insights into intermediate sound features, as captured by DNNs that may support categorization in the human auditory system.

Keywords

Auditory cortex; Speech perception; Music perception; Deep neural network

Significance Statement

Speech and music are two uniquely human sound categories. While their distinctive acoustic features and semantic attributes have been studied extensively, how the human auditory system differentially transforms the acoustics into meaning remains largely unknown. In this study, we used a deep neural network (DNN) to sonify sounds from activity patterns in the auditory cortex. We found that these synthetic sounds, though acoustically dissimilar to natural speech and music, elicit similar categorical cortical and behavioral responses. Our findings indicate that categorization in the human auditory system relies on internal representations of category as captured by DNNs that are irreducible to natural acoustic speech and music features.

1 Introduction

Speech and music are among the most fundamental communicative signals for humans. They emerge from different natural and human-made sound sources (e.g., human voice or musical instruments) and consequently differ in their characteristic acoustic properties, such as spectrotemporal modulations (Ding et al., 2017, Elliott and Theunissen, 2009). Speech and music stimuli elicit preferential responses in distinct, partially overlapping regions of the non-primary auditory cortex (Binder et al., 1994; Leaver and Rauschecker, 2010; Rogalsky et al., 2011; Peretz et al., 2015) with some hemispheric lateralization (Zatorre et al., 2002; Poeppel, 2003) (speech: left; music: right), suggesting they are processed by partially specialized mechanisms. Previous studies have attributed these category-preferring responses to distinct spectrotemporal modulation rates (Albouy et al., 2020) and higher-level attributes (Norman-Haignere et al., 2015), suggesting that speech- and music-preferential cortical processing emerges from the acoustic features and/or categorical percepts that speech and music induce.

Despite several studies have addressed this topic, it is unclear how category-preferring cortical regions internally represent the preferred sound. By “internal sound representation” we refer to (the level of abstraction of) category-related information that is processed within the region, which could range from basic fine-grained spectrotemporal information to abstract semantic labels.

A conventional way of addressing these questions involves an exhaustive testing of a large set of pre-defined, experimentally controlled sound features hypothesized to contribute to categorization. A more efficient way may be to explore possibly novel, categorization-relevant sound features within the vast feature space of an artificial deep neural network (DNN) trained to perform sound classification. The latter DNN transforms an acoustic input signal into a sound-category label across several hierarchical processing stages (i.e., layers), where consecutive layers represent the original acoustic input as a combination of increasingly abstract features. The representation of natural sounds in such a DNN has been shown to correlate with the responses that these sounds evoke in the human auditory cortex (Kell et al., 2018; Giordano et al., 2023), implying that DNNs may be utilized to predict auditory cortical response to natural sounds. Indeed, its ability to represent sounds at various abstraction levels and to predict cortical responses to these sounds make the DNN potentially suitable for identifying the internal sound representation of a given cortical region from the DNN layers. In this approach, the DNN's predictive power is exploited to determine stimuli that effectively elicit a response in a given brain region (according to the DNN's prediction), and the features of those stimuli can be inferred to contain the information that is primarily processed within that region. Application of this approach in monkey electrophysiology and human neuroimaging studies has yielded novel insights into stimulus features underlying specific brain functions. For instance, human functional magnetic resonance-imaging (fMRI) studies have determined visual objects (Ratan Murty et al., 2021) and sentences (Tuckute et al., 2024) that a DNN predicted to maximally activate category-selective regions in the visual cortex and cortical language network, respectively. Moreover, monkey electrophysiology found novel, DNN-generated visual stimulus features that activate visual cortical neurons (Ponce et al., 2019) to higher levels than natural stimulus features do (Bashivan et al., 2019), and demonstrated the tuning of cortical “face cells” to non-facial stimulus features (Vinken et al., 2023).

In the present study, we applied the DNN-based feature-exploration approach to inform fMRI and behavioral experiments aimed at testing the contribution of DNN-based sound features to sound categorization in the human auditory system (Figure 1). We first established a linear mapping from representations of natural categorical sounds (as modeled by individual DNN layers) onto fMRI responses to the same sounds (recorded from speech- and music-preferring auditory cortical regions, ROIs). Utilizing this mapping in a sound synthesis, we obtained novel auditory stimuli predicted to effectively activate the speech- or music-ROI. In subsequent behavioral, fMRI, and in-silico experiments, we tested the ability of these novel sounds to elicit speech- and music categorization in the human auditory system, and attempted to characterize their categorical features.

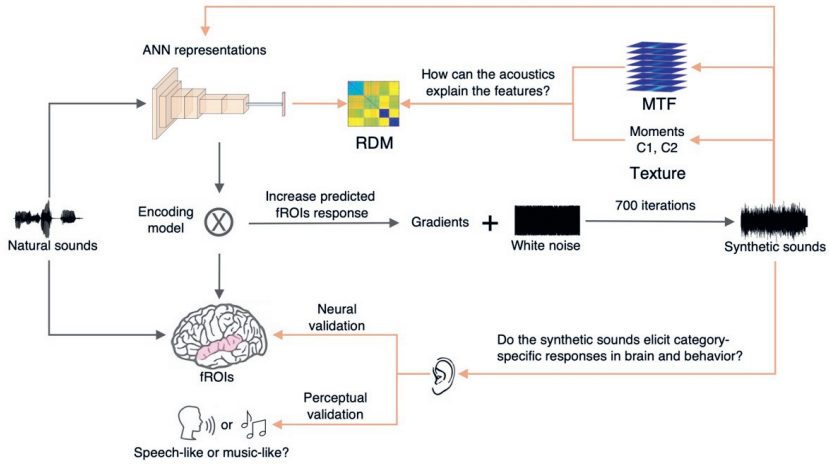


Figure 1. Overview of the approach. Black arrows indicate steps involved in the DNN-based sound-feature exploration. Orange arrows indicate steps involved in the characterization of the sound features and the testing of their putative functional role in sound categorization. In the first part, natural categorical sounds were presented to both DNN and participants, and the DNN and fMRI responses to these sounds were recorded. A linear encoding model was fitted that mapped the DNN responses onto the fMRI responses within speech and music fROIs in the auditory cortex. Utilizing this mapping and a sound-synthesis approach, features embedded in the DNN that maximized the predicted response of the speech fROI or music fROI were sonified, resulting in novel synthetic sounds. In the subsequent experimental part, the individually synthesized sounds were presented to both the respective participants and the DNN, and the behavioral responses in a sound-categorization task, fMRI responses, and DNN responses were recorded. To test the functional relevance of the synthesized sound features for sound categorization, participants' behavioral and fROI responses to sounds synthesized to activate the speech fROI were compared vs those synthesized to activate the music fROI. To further characterize the categorical nature of the synthesized sound features, DNN responses to the synthetic sounds were compared to DNN responses to natural speech or music (using an analysis of representational dissimilarity matrices, RDM) and their similarity to acoustic features known to distinguish sound categories (MTF and acoustic texture) was assessed.

2 Methods

2.1 Participants

Eight healthy participants (median age = 29, two males, six females) participated in the fMRI experiment. They also participated in the subsequent behavioral experiment, together with 10 additional healthy participants (median age = 24, four males, six females). All participants reported normal hearing and no history of a hearing disorder or neurological disease and gave informed consent before undergoing the experimental procedures. The experimental procedure was approved by the local ethics committee of the Faculty of Psychology and Neuroscience, Maastricht University (ERCPN #_233_09_02_2021).

2.2 Extraction of Functional Region of Interest (fROI) Responses to Natural Sounds

2.2.1 Auditory Stimuli

In the first two fMRI sessions, the auditory stimulation consisted of 288 recordings of real-life sounds (Santoro et al., 2017), with 48 sounds from each of the following six semantic categories: speech, (nonspeech) voice, animal cries, musical instruments, nature scenes, and tools. The speech sounds included an equal number of male and female speakers, and the vocal sounds included an equal number of male, female, and children voices. All sounds had a 1-s duration including 10-ms ramps at their onset and offset. They were sampled at 16 kHz, matched for RMS, filtered to correct for non-flatness of the frequency response of the MR-compatible in-ear system (Sensimetrics S14), and presented at a comfortable sound level.

2.2.2 fMRI Procedure

To establish six unique fMRI runs, the 288 natural sounds were randomly split into six sound sets with equal, uniform category distributions. Within each set, the sounds were randomly ordered, with the constraint that a given category was never repeated immediately. Sounds were presented in blocks of five, so that each block included five presentations of the same sound. To encourage participants to focus on the sounds, they were instructed to press a button when they detected the target. The latter was defined as a 7-dB level decrease applied to a single, randomly chosen sound repetition in every block. Loudness-change detection performance ranged from 72.5% to 96.1% (mean $\pm$ SEM: $84.9 \pm 0.03\%$) across participants given a chance level of 25%, suggesting that all participants paid attention to the auditory stimulation. After every fourth block (or “trial”) a baseline block was presented that was identical to the aforementioned stimulation blocks, except that no sound or task was presented. For participant P1, the stimulation protocol included a larger number of blocks with shorter durations (blocks of four sounds alternated with baseline blocks of half the duration of sound blocks). The aforementioned procedure was conducted twice (i.e., in two fMRI session) with a gap of 1-7 days in-between. The fMRI data collected in these two sessions served the definition of individual fROIs and encoding models (explained in detail below).

MRI data were acquired with a 3T scanner (Siemens) at Scannexus (<https://scannexus.nl/>).

Anatomical T1-weighted images were acquired using a modified magnetization-prepared rapid-acquisition gradient-echo sequence (number of slices = 192, TR = 2250 ms, TE = 2.21 ms, flip angle = 9°, voxel size = 1 mm isotropic, GRAPPA acceleration factor = 3). Blood oxygenation level-dependent signal changes were recorded using a modified gradient-echo echo-planar imaging (EPI) sequence isotropic (number of slices = 33 without gap, TR = 2400 ms, TA = 1000 ms, TE = 30 ms, flip angle = 82°, voxel size = 2 mm isotropic, GRAPPA acceleration factor = 3). Functional image acquisition and sound presentation alternated to ensure clear audibility of the sounds in the absence of scanner noise. Functional slices were centered on the Sylvian fissures to cover the temporal lobes. All functional and anatomical images were acquired in anterior-posterior phase encoding direction. To enable subsequent correction for EPI distortions, an additional pair of EPI volumes with reversed phase-encoding direction was acquired before the first functional run (Andersson et al., 2003).

2.2.3 fMRI Data Preprocessing

Imaging data were preprocessed with BrianVoyager 22.4 (Brain Innovation). Functional data were processed in four steps including slice scan-time correction (using cubic spline interpolation), 3D motion correction (using trilinear/sinc interpolation), temporal high-pass filtering (cutoff frequency: 7 cycles per run), and EPI distortion correction (using the TOPUP method). The anatomical and functional data from all fMRI sessions were spatially coregistered and normalized to Talairach space.

The fMRI signal of each voxel was z-scored along the time course within each run and delayed by one TR relative to the stimulation protocol to compensate for the hemodynamic delay. Functional volumes acquired in response to the four sound repetitions within a given block were averaged to obtain a single, time-averaged response measure for that block (the volume acquired immediately after the first sound was rejected to reduce carryover effects from the preceding block). The same approach was applied to baseline blocks to obtain a time-averaged baseline response.

2.2.4 fROI Selection

The fROIs were selected in three steps (Norman-Haignere et al., 2015): first, voxels showing significant sound-evoked responses were identified for each session by comparing sound blocks vs. baseline blocks ($p < 0.001$, one-tailed t-test). Second, from these sound-responsive voxels, the ones showing a response profile that was most reliable across sessions were selected. Reliability r was assessed as follows:

$$r = 1 - \frac{\| \mathbf{v}_1 - \text{proj}_{\mathbf{v}_2} \mathbf{v}_1 \|_2}{\| \mathbf{v}_1 \|_2},$$

$$\text{proj}_{\mathbf{v}_2} \mathbf{v}_1 = \left(\frac{\mathbf{v}_2^\top \mathbf{v}_1}{\|\mathbf{v}_2\|_2} \right) \mathbf{v}_2,$$

where vector $\mathbf{v}_1$ is the response of a given voxel to the 288 sounds in the first session, and $\mathbf{v}_2$ is the same voxel's response to the same sounds in the second session. Only voxels fulfilling a reliability criterion of $r > 0.3$ were retained (except for P1, we retained top 10% voxels). Finally, to identify a speech fROI, the retained voxels were submitted to statistical analysis contrasting speech blocks vs. no-speech sound blocks (one-tailed t-test) and the 5% of voxels showing the largest t-value were selected (all $p < 0.001$). Correspondingly, a music fROI was identified using the contrast music blocks vs. no-music sound blocks. Voxels within a given fROI were pooled to obtain a single, spatially averaged response measure.

2.3 Extraction of DNN Responses to Sounds

We focused on the VGGish (Hershey et al., 2017), a DNN that has been pre-trained to classify 10-s YouTube audio tracks with a set of 527 labels from the Audioset ontology (Gemmeke et al., 2017). The VGGish has a relatively simple architecture and has been proven successful in predicting human auditory cortex response patterns to natural sounds (Giordano et al., 2023). The VGGish takes audio segments of 0.975-s duration (converted to a stabilized log-mel spectrogram) as input and feeds it through an architecture consisting of an input layer, four convolutional blocks (convolution, ReLU, MaxPooling), and three fully connected blocks (fully connected, ReLU) that progressively reduce dimensionality, resulting in a 128-dimensional semantic embedding (output layer). We fed the 288 natural sounds into the VGGish and extracted representations of these sounds from its convolutional layers and fully connected layers as follows: after generating a cochleagram of each sound, we passed it through the VGGish, and extracted the response of each model unit, from each layer. The convolutional layers represented the sound along three dimensions: time, frequency, and kernel. To facilitate prediction of the time-averaged fROI responses, we averaged the time dimension yielding a time-averaged vector whose length (4096 points) equaled the product of the number of frequency points and number of kernels. A vector of equal length was directly extracted from the fully connected layers (i.e., without time dimension). Thus, this procedure resulted in eight sets of fixed-size (4096 points) sound representations (conv1, conv2, conv3_1, conv3_2, conv4_1, conv4_2, fc1_1, fc1_2). Considering the computational load and the time limit of the fMRI recording, we limited the validation experiments on four layers covering the early to late processing stages of the DNN architecture: conv1, conv3_1, conv4_1 and fc1_1.

2.4 Encoding Models: Mapping DNN Responses onto fROI Responses

We next trained an encoding model to map the DNN representations of the sounds onto the measured fROI responses to these sounds, separately for each fROI and layer. We modeled each fROI's time-averaged response to a given sound as a linear combination of the 4096 points defining that sound's time-averaged representation in a given layer. The model was fitted using regularized linear regression and a six-fold cross-validation procedure. The six unique sound sets from the six fMRI runs were split into five training sets (used to fit the model) and

one test set (used to evaluate the accuracy of the prediction of the fitted model). Each set contained responses to 48 unique sounds and was z-scored independently. On each iteration of the procedure, a different sound set was defined as test set. Given that the number of regressors (4096) exceeded the number of training sounds (240), L2-regularized regression was used. Introducing the L2-penalty on the weights results in a closed-form solution to the regression problem, which is similar with the least-squares regression normal equation:

$$\mathbf{w} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y},$$

where $\mathbf{w}$ was a 4096-point column vector containing the regression weights, $\mathbf{y}$ was a 240-point column vector representing the response variable (i.e., the fROI's response to each sound), $\mathbf{X}$ was a 240×4096 matrix representing the predictor variable (i.e., the layer-specific representation of each sound), and $\mathbf{I}$ was the 4096×4096 identity matrix. The ridge parameter was optimized within the range $\lambda = 10^2, 10^1, \dots, 10^{-6}, 10^{-7}$. The value yielding the most accurate predictions (maximum mean predictive variance) was retained. The predictive variance R^2 was estimated as $1 - \text{SSE}_{\text{test}} / \text{SST}_{\text{test}}$, where SSE_{test} represents the sum of the squared error of the trained model's prediction of the test-set data and SST_{test} represents the total sum of squares of the test-set data. To obtain an empirical chance level, we computed 1000 permutations across the test-set sounds for each split. The same permutations were applied to the corresponding sounds in the training set. The aforementioned procedure was conducted separately for each fROI × layer combination (2×4).

2.5 Sound Synthesis

We utilized the VGGish and the fitted encoding model to synthesize sounds via a gradient-ascent optimization approach. The procedure started with random stimuli (30 white noise samples) (Bashivan et al., 2019), the RMS level of which was matched to the natural sounds. The noise was fed into the VGGish, its representation was extracted from each layer, and the fROI response to this noise was predicted using the fitted encoding model. The value of each time point of the noise waveform was adjusted in the direction that increased the predicted fROI response, using a cost function based on the negation of the predicted fROI response and the total variation between directly neighboring time points (the latter served to reduce acoustic transients). The learning rate was set to a value of 0.001. The gradient of the cost function was calculated across 700 iterations using the function built in TensorFlow 2.13 (www.tensorflow.org), after which the optimized sounds showing the minimum costs were chosen (Figure S1). In each iteration, the novel sound from the previous iteration was temporally shifted in random direction (forward or backward) with a maximum of 1 ms to generate the adversarial component; the latter served to promote the synthesis of sounds that elicit predicted fROI response levels unaffected by such small shifts. The above procedure was conducted separately for each fROI × layer combination (2×4), each time starting with the same set of 30 (frozen) white noise samples. Overall, this resulted in a set of 240 (2×4×30) unique synthetic sounds.

2.6 Representational Analysis of Synthetized Sounds

Representational dissimilarity analysis (RSA) (Kriegeskorte et al., 2008) was applied to representations of the sounds in the DNN, the MTF model, and the sound texture model. VGGish representations of the synthetic sounds were extracted as described above (see section *Extraction of DNN responses to sounds*). The MTF was computed using NSLtools (Chi et al., 2005) as follows: first, the sound cochleagram was computed, which consisted of the time-varying output (temporal resolution = 8 ms) of 128 cochlear filters with log2 spaced frequencies (179–7246 Hz). Then, the magnitude of the spectrotemporal modulations (scale from 0.25 to 8 cycles/octave; unsigned rate from 2 to 32 Hz) in each channel of the cochleagram representation was assessed and averaged across time points. The sound texture was analyzed using the Sound Texture Synthesis Toolbox v.1.7 (McDermott and Simoncelli, 2011). The texture model considered here includes four separate components measuring different summary statistics of the time-varying amplitude envelopes of the cochleagram (32 frequency bands evenly spaced on an ERB-rate scale from 20 Hz to 10 kHz), or of the modulation analysis specific to each of the frequency bands (six modulation filters with center frequencies log-spaced between 30 and 100 Hz): (1) the marginal statistics of the band-specific envelopes (mean, EM; variance, EV; skewness, ES; kurtosis, EK); (2) the pairwise correlations C between band-specific amplitude envelopes (EC); (3) the pairwise correlation between the modulation analysis of each frequency band at a constant modulation filter frequency (C1 correlations); and (4) the correlation between the modulation analysis of the same frequency band at adjacent modulation filter frequencies (C2 correlations). For each type of representation, between-stimulus cosine distances were computed using a customized function in MATLAB, and all parameters were given equal weight.

We used ridge linear regression to predict the group-averaged RDM of the DNN from the group-averaged RDMs of the acoustic models (cross-validated across both stimuli and participants). For the stimuli, we focused on natural speech, natural music, speech activator, and music activator. More specifically, we focused on the RDM regions that revealed the categorical nature of the synthetic sounds, i.e., where the speech activator was close to natural speech and distant to natural music; and the reverse for the music activator. All the sounds were split into a training set and a test set each including half of the sounds of each category, and the sample was split into a training group and test group consisting respectively of six and two participants. All possible combinations of sound splits and sample splits were tested, resulting in 56 folds. The predictive variance R^2 was estimated as described above (see section on encoding models). Each fold included the estimation of a null distribution (10,000 row $\times$ column permutations of the RDM) as well as a cross-validated measure of noise ceiling that captured the maximum predictable variance in the group-averaged DNN distances. The latter was estimated as $1 - \text{SSD}_{\text{test-train}} / \text{SST}_{\text{test}}$, where $\text{SSD}_{\text{test-train}}$ represents the sum of the squared differences between the group-averaged DNN distances in the test vs. training set.

Statistical significance was assessed with a one-sided Mantel's test comparing the observed R^2 values with the permutation-based null distributions followed by a maximum-statistics correction for multiple comparisons (Maris and Oostenveld, 2007).

2.7 Neural Validation of Synthesized Sounds

The synthesized sounds were validated at the neural level in a third fMRI session that matched the two preceding fMRI sessions in all regards (participant, stimulation protocol, fMRI protocol, task, and fMRI data preprocessing), except for the following: the stimulation included the 240 synthesized sounds (2 categories $\times$ 4 layers $\times$ 30 samples) and only 48 natural sounds (24 speech, 24 music). The synthetic sounds were preprocessed identically to the natural sounds (see above). The 24 natural speech and music sounds were the ones observed to elicit the strongest speech-fROI and music-fROI response in the two preceding fMRI sessions. The analysis focused on fMRI responses extracted from the fROIs defined in the preceding fMRI sessions.

2.8 Behavioral Validation of Synthesized Sounds

The behavioral experiment used the same study design as the fMRI experiment, additionally including a synthetic-sound condition that served to control for potential bias in participants' category ratings of synthetic sounds. For that condition, synthetic sounds were generated as before, but with the difference that a random-weights VGGish with Kaiming He weights initialization was used and only layer c4 was considered. Thus, the design contained twelve conditions: 2 fROIs $\times$ 4 layers, plus 2 natural sounds, and 2 control sounds from the random VGGish. Participants who had undergone the fMRI part received the same sounds as in the fMRI experiment (see section above), except that 30 (instead of 24) unique natural sounds were presented to match the number of corresponding unique synthetic sounds. For each of the other, naïve participants, the sounds were the same as for a pseudorandomly chosen fMRI participant. The stimuli were presented via a sound card and headphones (Sennheiser, HD650) at a sound level of 65 dB SPL.

On each trial of the speech/music categorization and rating task, participants were presented with a sound and instructed to categorize the stimulus as either 'speech-like' or 'music-like' and rate their confidence in this categorization on an 11-point Likert scale (0 = not confident at all; 10 = totally confident). Responses were entered via a GUI that was continuously displayed throughout the experiment. The position (left or right) of the 'speech-like' and 'music-like' response buttons on the interface was counterbalanced across participants. After participants entered their rating, the next trial started. Trials were presented in blocks of 72, each including six unique sounds from each condition and lasting approximately seven minutes. In total, five blocks were presented, resulting in an overall number of 360 unique trials (30 per condition).

2.9 Statistical Analysis

All statistical analyses were performed using MATLAB 2023 (MathWorks, Natick, MA, USA) and SPSS 23 (SPSS, Chicago, ILL). Sound-responsive and category-preferring voxels were identified within each participant using two-sample paired t-tests (see fROI selection for details). The significance of the fit of the encoding models was assessed within each participant using one-sided Mantel tests that compared the observed R^2 values vs. permutation-based R^2 values (averaged across the 6 folds of the cross-validation procedure). Probability values were

corrected for multiple comparisons across DNN layers using a maximum-statistics correction (Maris and Oostenveld, 2007).

In the behavioral validation part, the statistical analysis was conducted at the group level ($n = 18$). The significance of categorization performance was assessed with one-sample t-tests comparing the observed accuracy vs. the theoretical chance level (50%). The significance of perceptual bias was assessed with one-sample t-tests comparing the observed bias vs. zero. The significance of differences in confidence level, which was measured on an ordinal scale, was assessed with two-sample Wilcoxon signed-rank tests. Probability values were corrected for multiple comparisons across DNN layers using Bonferroni's method.

In the neural validation part, the significance of category-preferential responses (natural sounds: $n = 24$; synthesized sounds: $n = 30$) was assessed within each participant using a 2 (sound category: speech vs. music) $\times$ 2 (fROI: speech vs. music) random-effect repeated-measures ANOVA for each DNN layer. Probability values were corrected for multiple comparisons across DNN layers using Bonferroni's method. At the group level ($n = 8$), non-parametric statistical tests were used given the small sample size. The significance of category preference was assessed within each DNN layer with one-sample Wilcoxon signed-rank tests comparing the observed preference index vs. zero. Overall differences across layers were tested with a Kruskal–Wallis H test, and post-hoc pairwise multiple comparisons. Probability values were corrected for multiple comparisons across DNN layers or layer pairs using Bonferroni's method. A linear mixed-effects model (LMM) was used to examine the relationship between fROI's category preference and behavioral categorization performance. The model included category preference as a fixed effect, and subject as a random intercept to account for within-subject dependence arising from repeated observations of the same individuals across layer conditions. Significance of the fixed effect was assessed with a t-test.

In the representational similarity analysis, the significance of differences in representational distance was assessed at the group level using two-sample paired t-tests. Probability values were corrected for multiple comparisons across categories using Bonferroni's method. The significance of the similarity between RDMs was assessed in the same manner as the fit of the encoding models (see above).

3 Results

3.1 Prediction of Auditory Cortical Responses from DNN Responses to Natural Sounds

Figure 2A shows the cortical location of the speech- and music-fROIs that were identified, which occupied multiple portions of the STG in both hemispheres. Figure 2B shows the cross-validated prediction accuracy of the encoding model used to establish the DNN-to-fROI mapping. Each layer could predict each fROI's overall response to the natural sounds with an accuracy significantly above chance level (estimated from 1000 permutations, number of cross-validation folds = 6, Mantel test: $p = 0.001$, corrected across DNN layers). For each fROI, the best prediction was achieved with layer 4 (median proportion explained variance in speech fROI: $R^2 = 0.70$, music fROI: $R^2 = 0.50$) and the poorest with layer 1 (speech fROI: $R^2 = 0.54$, music fROI: $R^2 = 0.36$).

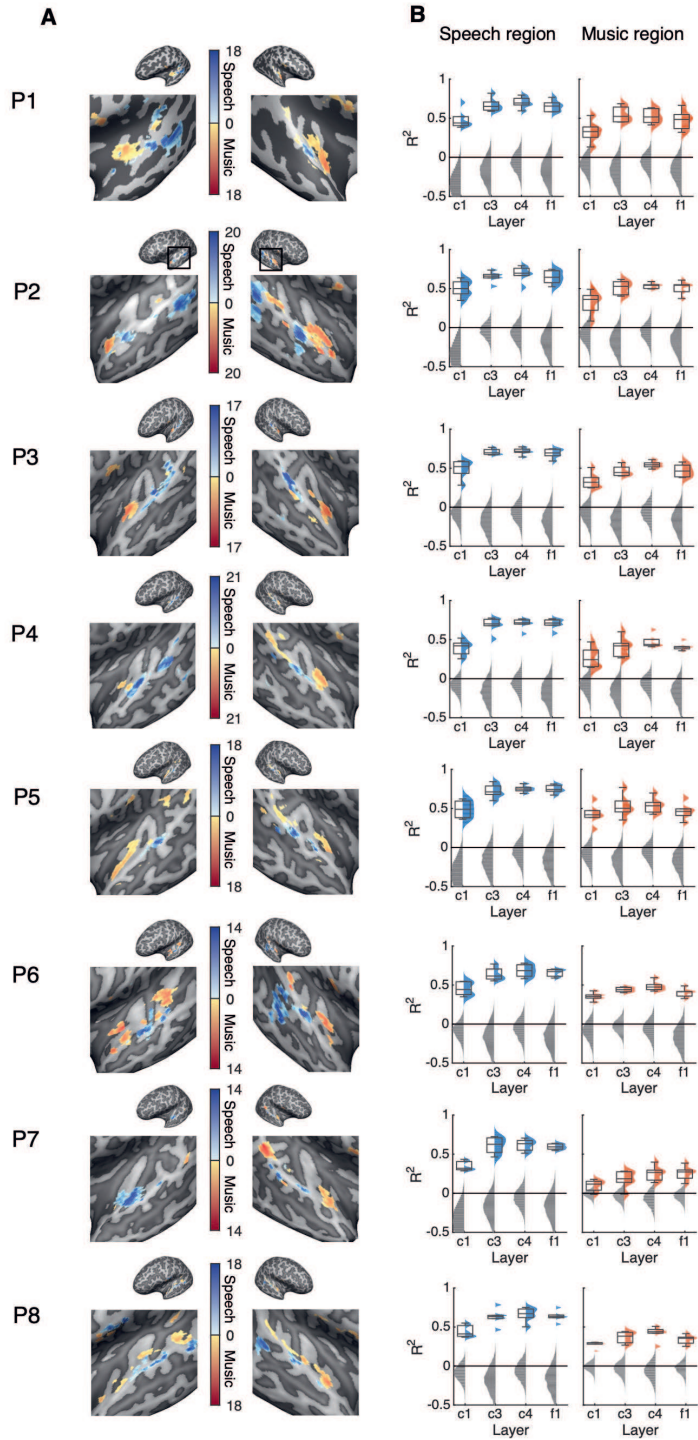


Figure 2. fROI localization and encoding results. **A.** Cortical map of the speech-preferring voxels (blue) and music-preferring voxels (red) that constituted the speech fROI and music fROI, respectively. The color scale represents t-values (denoted sound > all other categories). **B.** Prediction accuracy of the encoding model. The model was used to predict fROI responses to the natural sounds based on DNN layer representations of these sounds. Cross-validated prediction accuracy was found to be significantly above chance level (permutation-based null distribution, gray) for each layer and fROI. Summary statistics across trials: center, median; lower/upper box limits, first/third quartile; bottom/top whisker, data within 1.5 interquartile ranges from first and third quartiles, respectively.

3.2 DNN-based sound synthesis

We refer to sounds synthesized to activate the speech fROI and music ROI respectively as “speech activators” and “music activators”. Figure 3 shows the spectrogram of an exemplary synthesized sound for each participant, fROI, and DNN layer. Supplementary audio files are provided in Multimedia files. Supplementary audios 1–8 are demonstration sounds synthesized from participant 2. Auditory inspection of these sounds revealed that they elicit highly unnatural, unintelligible percepts, but with audible differences across fROIs (speech activator vs. music activator) and layers. Consistently, visual inspection of the spectrograms indicated changes in spectral and temporal structure across fROIs and layers. For example, speech activators synthesized from layers 3 and 4 exhibited relatively smooth spectrotemporal transitions, whereas music activators synthesized from those layers showed more “tone”-like features with higher spectral modulations.

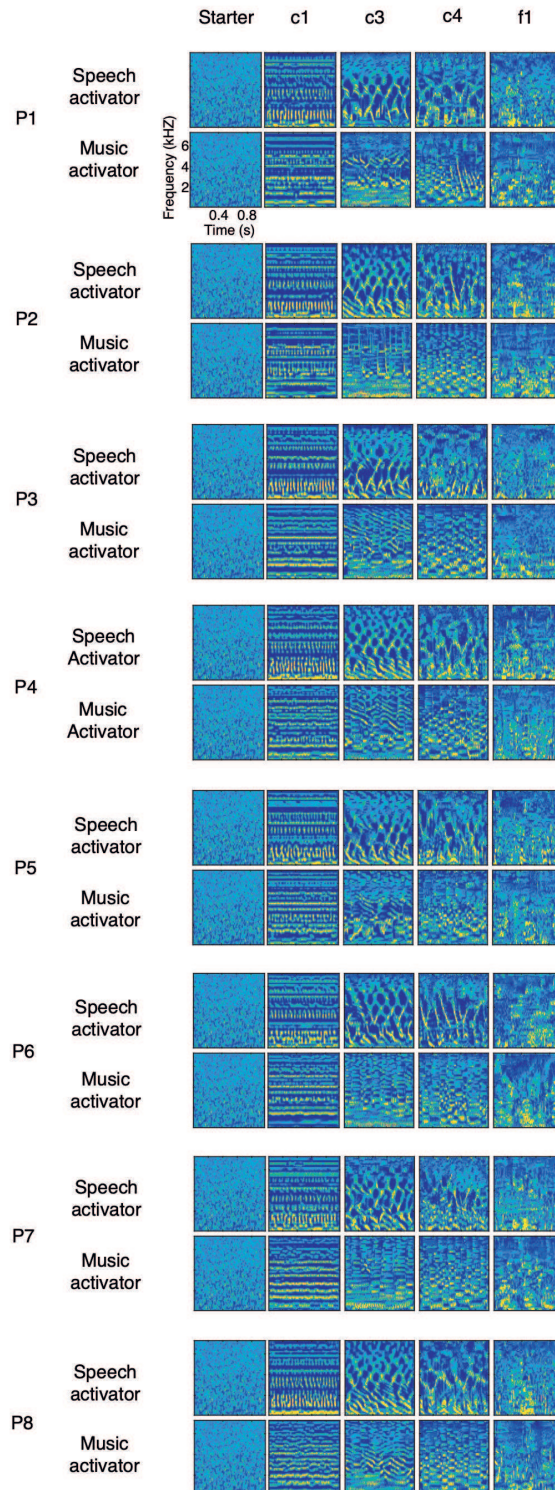


Figure 3. Examples of the DNN-based synthetic sounds. The leftmost column shows the spectrogram of a frozen sample of white noise that served as starting sound waveform for the iterative synthesis procedure. The same sample was used to synthesize one sound that activates the speech fROI (“speech activator”, upper row) and another sound that activates the music fROI (“music activator”, lower row). The other columns represent the DNN layer whose sound-evoked response was used in the synthesis procedure to predict and maximize the fROI response. Data are shown for each participant. For audio files, see Supplementary audio.

3.3 Effects of Synthesized Sound Features on Sound-Category Perception

To test whether the synthesized sounds can elicit sound-category perception, we presented them to the respective fMRI participants (each of whom received their own, individually-tailored set of sounds), as well as a group of naïve participants, in a speech/music categorization task with subsequent confidence ratings, and we assessed the effects on behavioral responses.

We presented 240 individually-tailored synthetic sounds (60 synthesized from each layer; for the naïve participants, these sounds matched those from a randomly chosen fMRI participant), 30 natural speech sounds, and 30 natural music sounds. We also included 60 control sounds that were synthesized from layer c4 as before but using a “naïve” (i.e., untrained) DNN. On each trial of the task, participants listened to a sound and subsequently categorized it as “speech-like” or “music-like” and rated their confidence in their categorization. We deemed “speech-like” responses to natural speech or speech activator as “correct”, and the same for “music-like” responses to natural music or music activator. Categorization performance was defined as the percentage of all trials on which participants gave a correct response (chance level: 50%; Figure 4A).

We observed no performance difference between fMRI participants vs. naïve participants; therefore, we pooled their data in subsequent analyses. Categorization performance for natural sounds was 100% accurate, consistently across all participants, indicating that they executed the task correctly. In line with our hypothesis, we found that participants categorized synthetic sounds from layer c4 correctly (median performance: 71%) and significantly above chance (one-tailed t-tests: $t(17) = 5.99$, $p = 3.6 \times 10^{-5}$, Bonferroni-corrected across DNN layers). Sounds synthesized from the other layers and control sounds failed to show this effect ($t(17) < 1.94$, $p > 0.172$).

To control for potential category bias, we calculated the performance difference between speech activator vs. music activator. Positive values of this measure indicate a bias to speech; negative values indicate a bias to music (Figure 4B). We found a significant bias toward music for synthetic sounds from layers c1 (two-tailed t-test: $t(17) = -85.87$, $p = 3.6 \times 10^{-23}$, Bonferroni-corrected across DNN layers) and c3 ($t(17) = -4.35$, $p = 0.002$), and a significant speech bias for synthetic sounds from layer f1 ($t(17) = 3.73$, $p = 0.008$). Thus, sounds synthesized from the early layers elicited significantly more music- than speech-like percepts, and vice versa for sounds from the final layer, which might explain our observation of an overall inaccurate categorization of the sounds from these layers. We observed no bias for synthetic sounds from layer c4 ($t(17) = -0.02$, $p = 1.000$), control sounds ($t(17) = -2.59$, $p = 0.096$), or natural sounds

(all bias = 0), indicating that the accurate categorization of sounds from layer c4 observed above could not be attributed to bias.

We further explored participants' subjective confidence in their categorization (Figure 4C). We observed the highest confidence ratings for the natural sounds, and the lowest confidence ratings for the control sounds; the latter may be attributed to the fact that these sounds did not incorporate the sound-classification ability of the pre-trained DNN. Confidence ratings for the other synthetic sounds fell in-between, with highest rating for sounds synthesized from layers c4 and c1. In statistical support of these observations, we found significantly higher confidence for synthetic sounds from the pre-trained DNN than for the control sounds, for all layers, except the final one (one-tailed Wilcoxon signed-rank test: c1: $W = 165$, $p = 0.001$; c3: $W = 152$, $p = 0.008$; c4: $W = 163$, $p = 0.002$; f1: $W = 133$, $p = 0.077$, Bonferroni-corrected across DNN layers).

In sum, these behavioral results validate that the sounds synthesized from layer 4 reliably evoked categorical percepts qualitatively similar with their natural counterparts. This supports our hypothesis that these synthetic sounds contain features that contribute to sound categorization in the human auditory system.

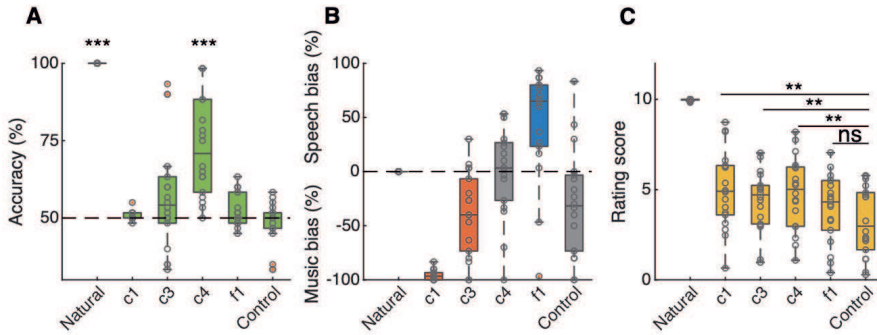


Figure 4. Perceptual validation of the synthesized sounds. **A.** Categorization performance was defined as the percentage of trials on which participants categorized the sound correctly (“speech-like” response to the natural speech or speech activator, or “music-like” response to the natural music or music activator). Categorization of natural sounds and sounds synthesized from layer c4 was found to be significantly better than chance level (50%), but not for sounds synthesized from the other layers or a naïve DNN. **B.** Category bias was computed as the performance difference speech minus music; positive values indicate bias to speech, negative values indicate bias to music. The sounds synthesized from layers c1 and c3 were found to induce significant bias to music (red), whereas sounds synthesized from layer f1 induced significant bias to speech (blue). Sounds synthesized from layer c4 or an untrained DNN induced no such bias (gray). **C.** Participants’ subjective confidence in their categorization. Participants showed the strongest confidence in their categorization of the natural sounds, and the weakest confidence for the control sounds (synthesized from a naïve DNN). Summary statistics across participants: center, median; lower/upper box limits, first/third quartile; bottom/top whisker, data within 1.5 interquartile ranges from first and third quartiles, respectively. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; n.s., non-significant.

3.4 Effects of Synthesized Sound Features on Sound Categorization in Auditory cortex

We conducted another fMRI experiment to assess whether the speech activator and music activator were sufficiently distinct to elicit sound categorization in the auditory cortex. To this end, we tested whether the differences between the DNN-based speech and music activators induce significant fROI response differences and whether these differences resemble natural speech/music-preferring responses.

As before, each participant received their own, individually-tailored set of sounds (60 synthesized from each layer c1, c3, c4, and f1; 24 natural speech, and 24 natural music), except for participant P7 who accidentally was presented with sounds tailored for P6. Figure 5 renders the results for each participant. We first confirmed that the fROIs showed significant responses to both, natural sounds as expected based on the initial fMRI part of the study (speech fROI: all $t(112) > 8.52$, $p < 10^{-13}$; music fROI: all $t(112) > 12.60$, $p < 10^{-22}$), and to the synthetic sounds as predicted by the encoding model (speech fROI: all $t(304) > 4.21$, $p < 10^{-4}$; music fROI: all $t(304) > 10.07$, $p < 10^{-20}$), compared to baseline. We observed that the synthesized sounds indeed significantly activated the fROIs as predicted by the encoding model, although their activation level was lower than that induced by the fROI's most-preferred natural sounds.

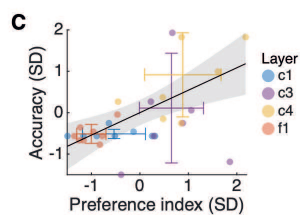
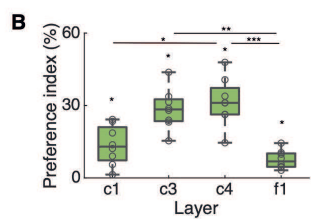
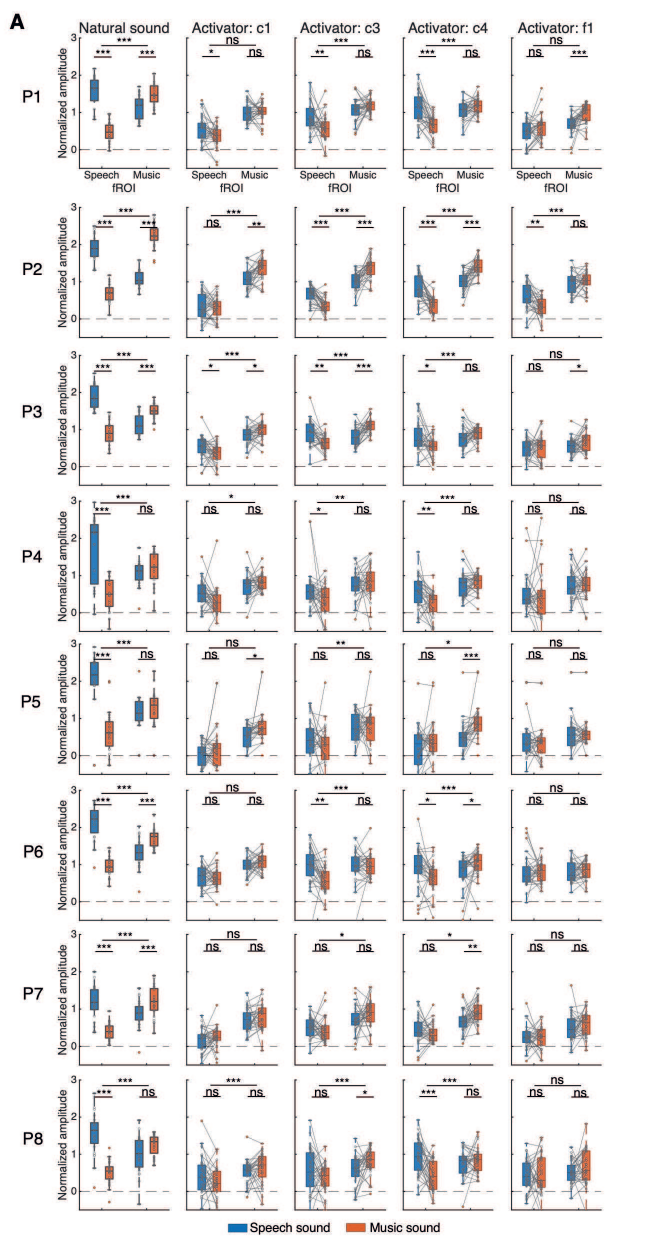


Figure 5. Neural validation of the synthesized sounds. **A.** BOLD responses to natural and synthetic sounds (individual results). The rows show individual results for each participant. In each boxplot, the upper significance marker indicates a significant interaction (sound category $\times$ fROI), whereas the lower markers show the simple effect of sound category per fROI, i.e., response difference between speech and music sounds. The solid lines highlight response differences within pairs of sounds synthesized from the same initial noise sample. The interaction between sound categories and the fROIs was significant for natural sounds as expected. Importantly, for synthetic sounds, the same interaction was found, consistently across all layers from which these sounds were synthesized. **B.** Group results ($N = 8$). Category preference (group level results, $N = 8$). The preference index (PI) quantifies the pairwise response difference between speech activator and music activator (speech fROI: speech activator minus music activator; music fROI: music activator minus speech activator) averaged across pairs and the two fROIs, expressed as a percentage of the PI observed with natural sounds. PI was significantly above 0% for all layers, and significantly higher for sounds synthesized from the middle layers (c4) vs. sounds synthesized from the early layer (c1) and later layer (f1). **C.** Brain-behavior correspondence. The scatterplot shows the bivariate distribution of PI and behavioral accuracy (expressed as z-scores; SD, standard deviation) pooled across DNN layers (color-coded). The black line represents the fixed effect of the mixed-effect linear model, showing a positive association between PI and behavioral accuracy. Crosses represent summary statistics (mean and SD across participants) of PI and behavioral accuracy for each given layer. It can be seen that sounds synthesized from the middle layers (c3 and c4) induced the strongest categorization in both brain and behavior. Descriptive statistics across trials (panel A top) or across participants (panel B): center, median; lower/upper box limits, first/third quartile; bottom/ top whisker, data within 1.5 interquartile ranges from first and third quartiles, respectively. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; n.s., non-significant.

We further verified that the fROIs responded preferentially to the natural speech vs. music as expected per fROI definition from the initial fMRI part of the study. A 2×2 (sound category $\times$ fROI) ANOVA confirmed a significant interaction for natural sounds (all participants: $F(1,23) > 50.18$, $p < 10^{-6}$); the speech fROI responded preferentially to natural speech vs. music (all participants: mean diff > 0.77 ; simple effect: $F(1,23) > 25.80$, $p < 10^{-4}$), and the music fROI responded preferentially to natural music vs. speech (all participants: mean diff > 0.13 ; simple effect of five participants: $F(1,23) > 18.30$, $p < 0.002$; see Figure 5A).

After this initial data quality check, we tested our main hypothesis using the same analyses as with the natural sounds above, i.e., we assessed whether the fROIs exhibited the category-preference as observed above also for the synthetic sounds. Similar to the natural sounds, we found a significant interaction effect (sound category $\times$ fROI) for sounds synthesized from the middle layers c3 and c4. Remarkably, we found this double dissociation reliably in every participant, including the participant who received another participant's synthetic sounds (c3: all $F(1,30) > 7.623$, $p < 0.039$; c4: all $F(1,30) > 7.308$, $p < 0.044$; Bonferroni-corrected across DNN layers; Table S1). We observed this effect also with sounds synthesized from the earlier or later layers, depending on the participant (Figure 5A). Simple-effect analysis confirmed for every participant that the speech fROI preferred the speech activator vs. music activator, and

the music fROI preferred the music activator vs. speech activator, from the middle layers c3 and c4, although this difference was not statistically significant in every participant (Table S2).

To allow generalizing our individual finding of category preference to the population level, we tested the category preference with group-level statistics. We quantified category preference with a preference index (PI), defined for the speech fROI as the pairwise difference “response to speech activator minus response to music activator” (both synthetic sounds generated from the same starting waveform), and vice versa for the music fROI. We pooled PI across all pairs (i.e., starting waveforms) and the two fROIs, and expressed it as a percentage of the observed natural category preference of the fROIs (i.e., a PI value defined analogously based on responses to the natural sounds). As expected, based on our individual double-dissociation results, we found that PI was significantly larger than 0% for every layer (Figure 5B; one-tailed Wilcoxon signed-rank test: $W = 36$, $p = 0.016$, Bonferroni-corrected across DNN layers). We further observed a main effect of layer on PI (Kruskal-Wallis test: $\chi^2(3) = 21.25$, $p < 10^{-4}$); a post-hoc analysis revealed significantly higher PI values in the middle layers than in early layer c1 or late layer f1 (Post-hoc comparisons: $c4 > c1$: mean diff = 13.50, $p = 0.024$; $c3 > f1$: mean diff = 16.25, $p = 0.003$, $c4 > f1$: mean diff = 18.13, $p = 0.001$, Bonferroni-corrected across layer pairs).

To test whether our observations of perceptual and cortical categorization of the activator sounds are coupled, we fitted participants' behavioral categorization performance with a LMM comprising participants' PI as predictor (see Methods). We found a significant positive correlation between PI and behavioral categorization performance (Figure 5C; $R^2 = 0.34$, $t(30) = 3.76$, $p = 0.001$), indicating that the two measures are coupled.

In sum, these fMRI results show that the synthetic sounds significantly activated the fROIs although the activation level fell below that induced by the fROI's most-preferred natural sounds. More importantly, the results reveal that the synthetic sounds reliably evoked preferential responses in speech and music fROIs qualitatively similar to their natural counterparts. Further in line with our behavioral results above, they show that this category preference is induced most strongly by sounds synthesized from the middle DNN layers. This lends further support to our hypothesis that the synthetic sounds contain features that contribute to sound categorization in the human auditory system.

3.5 Representational Analysis of Synthesized Sound Features

Figure 6A visualizes the representational dissimilarity matrix (RDM) of the layer c4 representations of all synthetic and natural sounds, after its projection into a two-dimensional space using multidimensional scaling (MDS). The closer two sounds (dots) are in the MDS plot, the more similar are their representations in the representational space. We observed that synthetic sounds were distant from natural sounds, probably reflecting their unnatural origin. Despite this large representational distance, the synthetic speech activator was closer to the natural speech, and the synthetic music activator was closer to natural music, in line with our behavioral and fMRI results. This observation was confirmed by a statistical analysis comparing the distance from speech activator (or music activator) to natural speech (or music) vs. its

distance to the other five categories of natural sounds (Figure 6B; one-tailed t-tests: speech: $t(478) > 21.06$, $p < 1.7 \times 10^{-91}$; music: $t(478) > 49.47$, $p < 3.0 \times 10^{-112}$; Bonferroni-corrected across categories). These results indicate that the DNN-based sound synthesis promoted acoustic features that differ from those of natural sounds but distinguish natural sound categories.

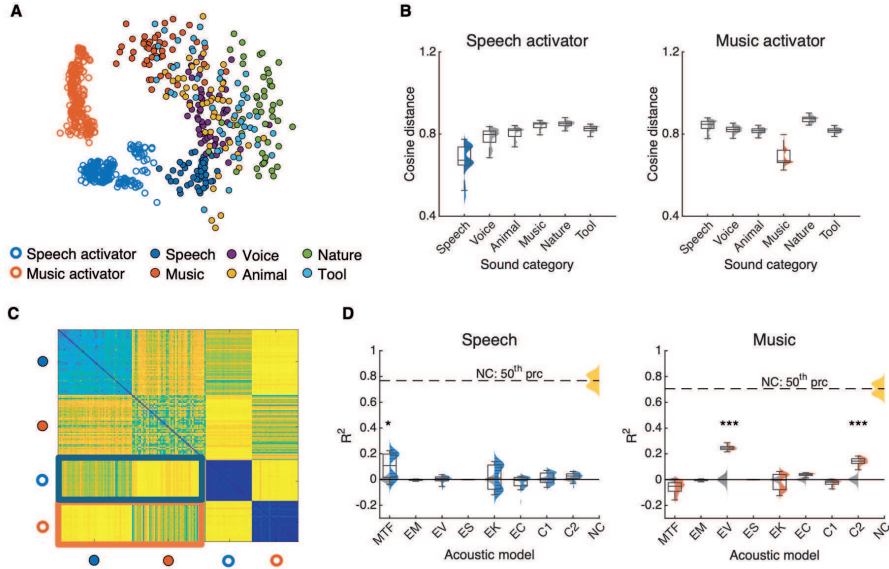


Figure 6. Results from analysis of sound representations in DNN layer 4. **A.** Two-dimensional (MDS-based) visualization of the representational similarity of the synthetic and natural sounds in DNN layer c4. Each dot represents the representation of a given sound in DNN layer c4, and the distance between dots represents the representational dissimilarity of the sounds. It can be seen that synthetic sounds (open circles with colored outline) are distant from natural sounds (filled circles with black outline) and the synthetic speech activator and music activator are visually closer to natural speech and music, respectively. **B.** Representational distance of speech activator (left) and music activator (right) from the six categories of natural sounds in layer c4. In each plot, the colored distribution is significantly lower than the gray distribution. These results confirm our hypothesis that the synthesized sounds possess categorical features. Summary statistics across participants: center, median; lower/upper box limits, first/third quartile; bottom/ top whisker, data within 1.5 interquartile ranges from first and third quartiles, respectively. **C.** RDM regions of interest. The different circles at the margins represent different sounds as labeled in panel A. Representational similarity (median across participants) is color-coded, with blue color indicating highest similarity and yellow color indicating lowest similarity. The blue outline highlights the region of the RDM showing that the speech-activator representation is more similar to the natural speech (vs. music) representation in layer c4. The red outline highlights the corresponding region for the music-activator representation. **D.** Linear regression results quantifying the proportion of the layer c4-based representational dissimilarity pattern that could be explained by MTF- or

*texture-based representational dissimilarity patterns, separately for each synthetic sound category / RDM region (speech activator: left plot, blue outline in panel C; music activator: right plot, orange outline in panel C; MTF, modulation transfer function; EM, envelope mean; EV, envelope variance; ES, skewness; EK, kurtosis; EC, envelope correlation; C1 and C2, modulation correlations, see method). Gray distributions are null distributions obtained from 10000 permutations; Orange distributions represent noise ceiling (NC, dashed line, median of NC) across participants. It can be seen that dissimilarity patterns of MTF- and texture-based sound representations could explain a significant, relatively small proportion of the dissimilarity patterns of layer c4-based speech- and music-activator representations, respectively. Summary statistics across participants: center, median; lower/upper box limits, first/third quartile; bottom/top whisker, data within 1.5 interquartile ranges from first and third quartiles, respectively. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; n.s., non-significant.*

We next assessed the uniqueness of the synthetic features by comparing their categorical nature with that of acoustic features known to distinguish sound categories, i.e., those promoted by modulation-transfer function (MTF) (Chi et al., 2005) and acoustic texture models (McDermott and Simoncelli, 2011). The MTF represents the spectrotemporal modulation of sounds, while the acoustic texture represents the marginal moments of the amplitude envelope (e.g., EM, EV, ES, EK) and pairwise correlations across frequency channels (EC, C1 and C2, see methods). We reasoned that observation of a second-order isomorphism, i.e., a resemblance between the DNN-based RDM and an RDM based on MTF or texture model, would indicate a functional link between the sound features promoted by the models (Kriegeskorte et al., 2008). We focused on the two RDM regions reflecting the categorical nature of the synthetic sounds in the DNN, see Figure 6C: the “speech-activator region” shows that the speech-activator representation is more similar to the natural speech representation vs. natural music representation, and the “music-activator region” shows the corresponding pattern for the music-activator representation. Using ridge linear regression, we quantified the resemblance between the DNN-based RDM and the MTF- or texture-based RDM, separately for each RDM region and each acoustic model (see Methods). Statistical comparison with chance level (estimated from 10000 permutations) showed that the dissimilarity pattern of MTF-based representations, but not that of texture-based representations, could significantly explain the dissimilarity pattern of the DNN-based representations in the speech-activator region (Mantel test: $p = 0.011$, corrected across models), and vice versa for the music-activator region (EV: $p = 10^{-4}$, C2: $p = 10^{-4}$, corrected across models). The proportion explained variance was 0.11 and 0.25 in speech-activator region and music-activator region, given a theoretical maximum of 0.78 and 0.70, respectively. These results indicate that the categorical features of the synthesized sounds resemble those promoted by MTF and acoustic texture models. More specifically, the features of the speech activators resemble more those of MTF, and the features of the music activators resemble more those of acoustic texture, suggesting that MTF and acoustic texture contribute respectively to speech and music categorization. While we found the resemblance between the synthesized features and MTF and acoustic texture to be significant, it was overall weak given that MTF and texture failed to account for a major portion of the theoretically explainable variance.

Applying identical analyses to the sounds synthesized from the other DNN layers largely replicated our observations for sounds from layer c4 (Figures S2 and S3): activator sounds were distant from natural sounds, speech activators were closer to natural speech, and music activators were closer to natural music (compared with other natural sound categories) in all DNN layers. Moreover, only small proportions of the dissimilarity patterns of the DNN-based synthetic sound representations could be explained by the dissimilarity patterns of MTF- or texture-based representations (Table S3). This indicates that the categorical characteristics reported above were not limited to layer c4, but emerged in all DNN layers tested.

4 Discussion

The aim of this study was to explore the nature of intermediate internal sound representations underlying categorization in the human auditory system. Using a DNN-based feature-exploration approach, we first synthesized novel auditory stimuli to activate cortical regions that preferentially encode natural speech- or music-related information. We then experimentally confirmed the categorical nature of the features of the synthesized sounds and their causal relevance for speech- and music categorization in the auditory cortex and perception. We found that these categorization-relevant, DNN-based features overlap only partially with those proposed by previous models based on MTFs or sound textures. Audio renderings of these features are provided in the Multimedia files.

4.1 Encoding of Natural Speech- and Music-Related Information in Auditory Cortex

The cortical regions that we observed to preferentially process natural speech or music were widespread across the STG in both hemispheres. Although their exact spatial pattern varied visibly across participants (see Figure 2), there appeared a tendency for the speech fROI to occupy more posterior portions of STG, and for the music fROI more anterior regions, in line with previous studies (Leaver and Rauschecker, 2010; Norman-Haignere et al., 2015). Our definition of speech and music fROIs resulted in small volumes of brain tissue (on average 0.4 cm³) that encoded speech- or music-related acoustic and/or categorical information reliably across fMRI sessions. This approach could not specify these fROIs' representational level and functional role in categorization, therefore we conducted the synthesized-sound experiments.

4.2 The DNN as a Valid Model of Sound-Evoked Processing in Speech and Music fROIs

We first established a DNN-to-fROI mapping, which confirmed that fMRI responses to natural sounds in the speech- and music ROIs can be predicted from DNN representations of these sounds with significant accuracies of up to 0.7, underscoring that the DNN provides an accurate model for the processing of natural categorical sounds in human auditory cortex (Giordano et al., 2023). Complementary to the previous findings based on spatially distributed activation patterns, our results show that the DNN can be utilized to predict also the average activation level within those patterns, which we found to be more suitable for the sound synthesis and yield higher prediction accuracies in pilot data. Importantly, beyond assessing the DNN's

predictive ability, our study goes a step further by utilizing this ability for experimentally testing a specific function of the auditory cortex—the preferential encoding of speech- and music-related information—thereby extending previous studies that aimed to explain category-preferring responses with analytical models (Norman-Haignere et al., 2015; Albouy et al., 2020).

4.3 Activator Sounds Elicit Speech/Music Categorization in the Human Auditory System

Instead of testing a set of pre-defined features hypothesized to be relevant for sound categorization, our study exploited a DNN-to-fROI mapping to explore previously unknown features within the feature space of the DNN. As this feature exploration was steered by the (predicted) fROI activation level, the resulting synthetic sounds can be inferred to contain information that is primarily processed within those regions. While the sound synthesis extracted (yet ill-defined) DNN-based sound features that may be primarily encoded in the fROIs, it left unclear whether these features contribute critically to sound categorization in the human auditory system.

Resolving this, our experimental results show that human participants could categorize the sounds synthesized from layer 4 “correctly”, i.e., in line with the preferred category of the cortical region that these sounds were synthesized to activate. Therefore, these sounds contained features relevant for perceptual sound categorization. No such effect could be observed with control sounds synthesized from a naïve DNN, suggesting that the categorization-relevant features emerged from the sound-classifying ability of the trained DNN.

At the cortical level, we found speech activators to activate the speech fROI significantly more than music activators, and vice versa for the music fROI. Thus, these synthetic sounds evoked preferential sound processing qualitatively similar to that of natural speech and music, reminiscent of previously reported speech- and music-selectivity of the auditory cortex. This outcome is non-trivial as the observed selectivity was not an objective of the sound synthesis. Instead, the latter aimed to merely activate a given fROI regardless of other regions, which theoretically could have resulted in sounds activating the entire cortex non-selectively.

In sum, our experimental results demonstrate that the distinct features of the novel activator sounds are sufficient to elicit naturalistic sound categorization at perceptual and cortical levels. Therefore, these sounds can be considered as synthetic speech and synthetic music, and their features contribute causally to sound categorization in the human auditory system.

4.4 Characteristics of the Activator Sounds

Unlike fMRI-DNN studies that synthesized natural-looking images or words from meaningful higher-order features, our study synthesized sounds from more basic, low-level features (instantaneous sound levels, comparable to individual image pixels (Bashivan et al., 2019)). Perhaps as a consequence, the resulting sounds have a highly artificial quality, as reflected by their large distance from natural sounds in the DNN. Despite this unnatural appearance, these sounds exhibit features distinguishing speech from music, as shown by their higher similarity

to natural sounds of the same (vs. different) category in the DNN. Therefore, the synthesized sounds contain naturalistic categorical features that reliably distinguish speech from music. Meanwhile, we observed that the speech and music activators from all participants formed two distinct clusters (Figure 6A), indicating that the synthesized sounds were largely consistent across participants. This suggests that the identified features capture universal category-relevant properties.

Unfortunately, the DNN's opaqueness prevents a conventional analytical description of the categorical features of the novel sounds. To address this current limitation, we sonified these features, allowing to appreciate the categorical nature of the novel sounds. To approximate an analytical description, we related these sounds to established categorical features of natural sounds, which revealed that they share some features with corresponding natural sounds of the same category despite being composed of mostly other, unnatural features. Moreover, their features are functionally linked with the categorical features promoted by MTF and texture models (Feather and McDermott, 2018), reflecting their shared ability to distinguish sound categories. However, we found the strength of the latter link to be poor compared to noise ceiling. This indicates that while MTF and texture models can significantly explain acoustic differences between speech and music, these differences are explained more accurately by features in the DNN. In sum, our representation-analysis results indicate that the DNN represents sound categories based on features that largely differ from those employed by the other models. Our human findings further suggest that the DNN provides a biologically plausible model of categorization in the human auditory system.

4.5 DNN Layer Contributions to Categorization

Our results from representational similarity analysis indicate that the synthetic sounds contain categorical information, regardless of the layers that were utilized for their synthesis. Thus, all layers appear to embed categorical sound features, probably reflecting their ability to jointly classify sounds. In contrast, not every layer's embedded features appear to be sufficient for sound categorization: our behavioral and fMRI results consistently show the strongest categorization effect for sounds synthesized from the middle layers, especially layer 4. This suggests that features embedded in the DNN up to layer 4 contribute to sound categorization in a cumulative way, whereas subsequent layers may add relatively little to it (Feather et al., 2023). Therefore, the sound representation in layer 4 appears to be the most suitable candidate for the internal sound representation in the category-preferring auditory cortex.

4.6 Limitations

Per our natural stimulus set, our interpretation is limited to short-term (< 1 s) sound features, leaving possible contributions from longer-term features (e.g., syntax) unclear (Li et al., 2023). Moreover, the activator sounds induced lower activation levels than the most-activating natural sounds. While this did not prevent theoretical insights into sound categorization, it suggests

that the DNN failed to encode some fROI-driving features specific to natural sounds. The fROIs might still be driven to supernatural activation levels with electrophysiologic recordings and a DNN that allow for a stimulus synthesis targeting a richer set of neural processing features (e.g., neural excitation/inhibition, cell types, or cortical layers, e.g., Bashivan et al., 2019; Ponce et al., 2019). Moreover, electrophysiologically-defined target neural responses might encode category-related features that differ from (and are possibly more specific to natural sounds than) those defined with fMRI. Future studies applying these methods may help further characterizing the internal representation of category-preferring auditory cortical regions and allow assessing the generalizability of our fMRI-based findings.

4.7 Conclusion

The main insight from our study is that a DNN trained for sound classification embeds sound features that are causally relevant for speech and music categorization in the human auditory system. While the opaqueness of the DNN hampers an analytical characterization, our findings show that these DNN-based features differ from previously reported MTF- and texture-like acoustic features for sound categorization, and may provide a more accurate description of the internal sound representation of speech- and music-preferring regions in the human auditory cortex. This representation might contain category-related information at a level that is intermediate between those of the low-level acoustic (domain-general) and semantic (domain-specific) representations reported previously. Beyond these theoretical insights into sound categorization in DNN and humans, our study highlights the utility of combined DNN-fMRI recordings for tailoring sensory stimulation to selectively activate cortical regions, which may be useful for non-invasive investigations of a variety of human brain functions.

Acknowledgments

We thank J. Hewett, M. Leonard, and B. Sorger for assistance with data collection. This work was supported by Maastricht University, China Scholarship Council (CSC202109110036 to L.X.), Dutch Research Organization (NWO Rubicon grant 04520232310005 to Min Wu and NWO Open Competition grant 406.20.GO.030 to E.F.). The authors have no conflicts of interest.

Author Contributions

L.X. and L.R. conceptualized the study. L.X., E.F. and L.R. developed the methodology. L.X. wrote the software. L.X., E.F. and L.R. validated the results. L.X. conducted the formal analysis and investigation. E.F. and L.R. provided resources. L.X. curated the data. L.X. and L.R. wrote the original draft of the manuscript. L.X., E.F. and L.R. wrote, reviewed and edited the manuscript. E.F. and L.R. supervised and administered the project and acquired funding.

5. Supplementary Materials

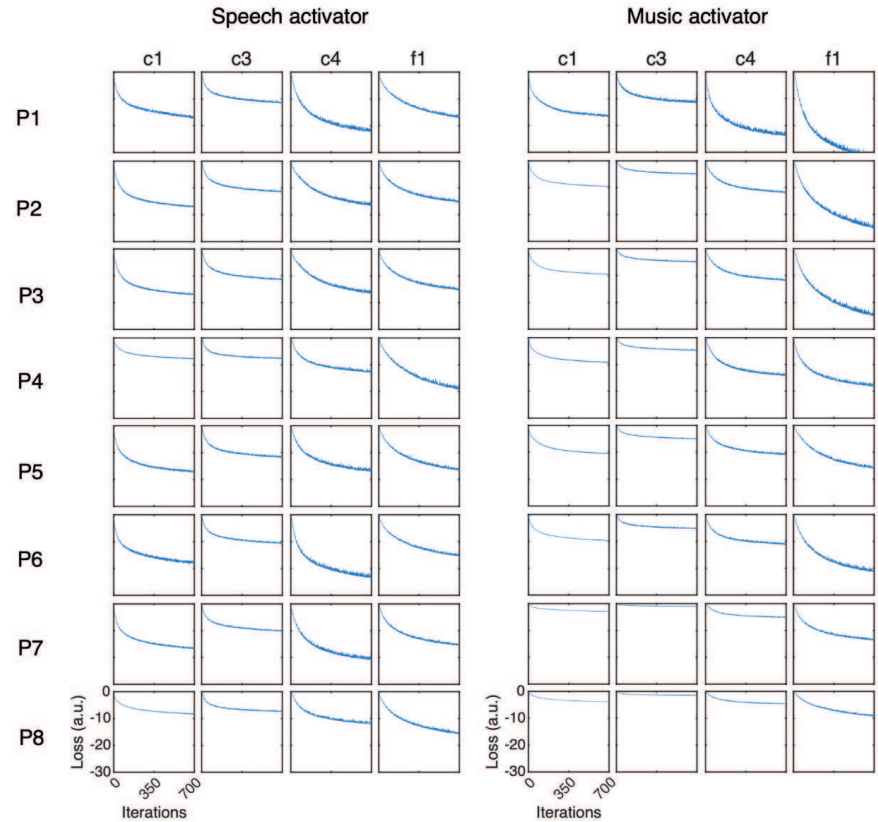


Figure S1. Loss function curves of the iterative sound synthesis procedure. The plots show the model's loss value as a function of the iteration number for each participant (rows: P1–P8), exemplary activator sound (columns: speech, music), and DNN layer (sub-columns: c1, c3, c4, f1). It can be seen that loss exhibited an overall smooth, monotonic decrease. For most participants, sounds, and layers, this decrease largely reduced toward later iterations, indicating that the optimization procedure converged on a stable solution.

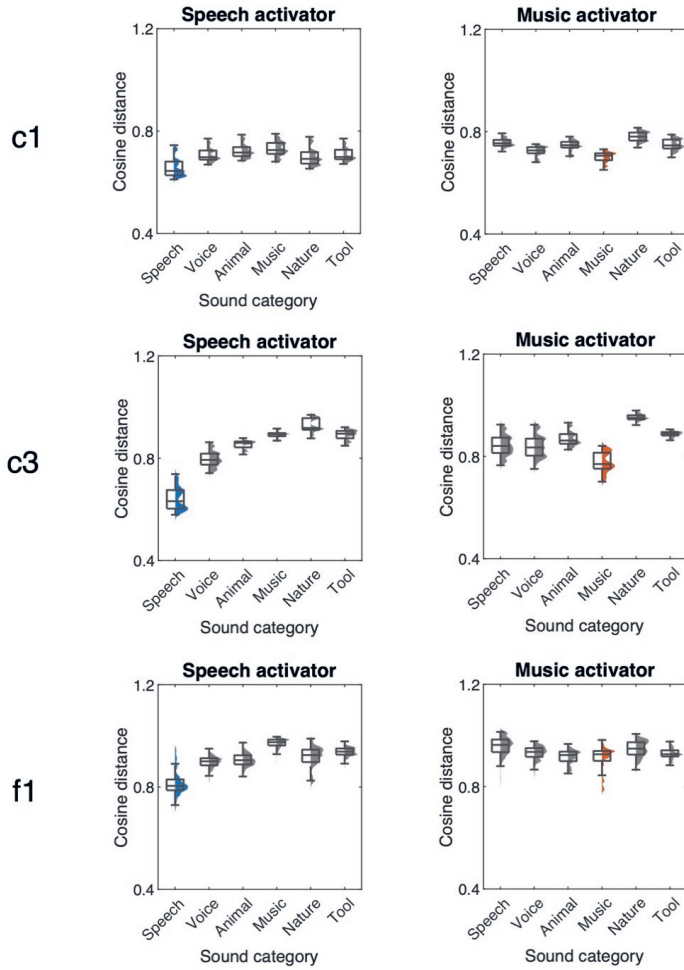


Figure S2. Representational distance of the speech activator (left column) and the music activator (right column) from the six categories of natural sounds for DNN layers c1, c3, and f1 (rows). For layer c4 and details, see the corresponding main Figure 6B. The synthetic sounds were significantly more similar to their natural counterparts than to natural sounds from other categories (colored distribution < gray distribution; $p < 10^{-9}$) regardless of the layer from which the sounds were synthesized, except for music synthesized from layer f1.

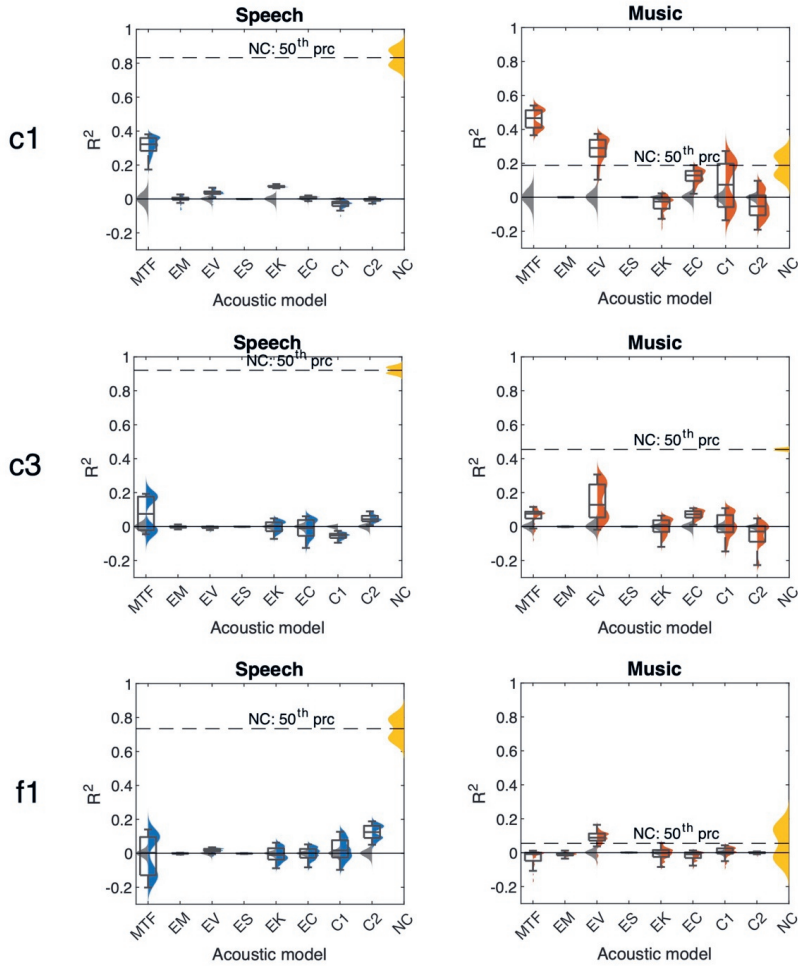


Figure S3. Linear regression results quantifying the proportion of the DNN layer-based representational dissimilarity pattern that could be explained by MTF- or texture-based representational dissimilarity patterns, separately for each synthetic sound category / RDM region (columns). Results are shown for DNN layers c1, c3, and f1 (rows). For layer c4 results and further details, see the corresponding main Figure 6D. For inferential statistics, see Table S1.

	c1		c3		c4		f1	
	speech	music	speech	music	speech fROI	music fROI	speech fROI	music fROI
MTF	0.33 (0.001)	0.46 (0.001)	0.08 (0.003)	0.075 (0.003)	0.10 (0.001)	-0.06 (1.000)	-0.02 (1.000)	-0.02 (1.000)
EM	0.00 (0.930)	-0.00 (0.979)	-0.04 (1.000)	-0.00 (0.993)	-0.00 (0.999)	-0.00 (0.999)	-0.00 (0.996)	0.02 (1.000)
EV	0.04 (0.1525)	0.29 (0.001)	-0.03 (1.000)	0.15 (0.001)	0.00 (0.801)	0.25 (0.001)	0.02 (0.274)	0.09 (0.001)
ES	-0.00 (0.992)	0.00 (0.947)	-0.01 (0.994)	-0.00 (0.978)	-0.00 (0.983)	-0.00 (0.984)	-0.00 (0.999)	0.00 (0.957)
EK	0.07 (0.030)	-0.04 (1.000)	0.02 (0.913)	0.00 (0.937)	0.02 (0.290)	-0.02 (1.000)	-0.01 (1.000)	-0.01 (1.000)
EC	0.01 (0.618)	0.13 (0.005)	-0.01 (1.000)	0.07 (0.004)	-0.02 (1.000)	0.04 (0.092)	-0.00 (0.974)	-0.01 (1.000)
C1	-0.03 (1.000)	0.07 (0.096)	-0.06 (1.000)	0.02 (0.349)	0.01 (0.455)	-0.02 (1.000)	0.02 (0.268)	0.01 (0.316)
C2	-0.01 (1.000)	-0.05 (1.000)	0.05 (0.017)	-0.04 (1.000)	0.03 (0.158)	0.14 (0.001)	0.12 (0.001)	0.00 (0.978)

Table S1. Statistical results from ridge linear regression analysis of RDM regions. Proportion explained variance (median R^2 value across cross-validation folds) and associated p -value (in parentheses, permutation-based) are shown for each acoustic model (rows) and synthetic sound/RDM region as represented in layer c1, c3, c4, and f1 (shown respectively in columns 1 and 2, columns 3 and 4, columns 5 and 6, and columns 7 and 8). P -values were adjusted using a maximum-statistics correction for multiple comparisons across acoustic models.

	natural	conv1	conv3_1	conv4_1	fc1_1
P1	6E-16	0.056	0.001	1E-6	0.177
P2	4E-17	4E-4	4E-10	1E-8	4E-4
P3	9E-15	2E-5	6E-9	1E-4	0.247
P4	1E-7	0.033	0.005	2E-4	0.577
P5	4E-9	0.784	0.009	0.044	0.985
P6	2E-17	0.089	8E-6	1E-5	0.678
P7	7E-9	0.546	0.007	0.001	0.095
P8	7E-16	0.001	2E-5	5E-7	0.931

Table S2. Statistical *p*-values associated with the sound category $\times$ fROI interaction, for each participant and sound type (natural sounds and synthetic sounds from each DNN layer). *P*-values for synthetic sounds were Bonferroni-corrected across layers.

	natural		c1		c3		c4		f1	
	speech fROI	music fROI	speech fROI	music fROI	speech fROI	music fROI	speech fROI	music fROI	speech fROI	music fROI
P1	1.09 (2E-14)	0.39 (1E-4)	0.16 (0.044)	0.02 (0.818)	0.33 (0.006)	0.12 (0.076)	0.44 (2E-5)	0.10 (0.073)	0.13 (0.102)	0.28 (0.000)
P2	1.23 (4E-13)	1.14 (9E-13)	0.03 (0.660)	0.26 (0.000)	0.36 (3E-5)	0.35 (1E-5)	0.38 (1E-4)	0.38 (2E-6)	0.30 (0.001)	0.05 (0.500)
P3	0.97 (3E-11)	0.34 (2E-5)	0.18 (0.022)	0.13 (0.015)	0.28 (0.002)	0.30 (2E-5)	0.22 (0.020)	0.10 (0.143)	0.03 (0.708)	0.15 (0.022)
P4	1.19 (2E-7)	0.13 (0.208)	0.18 (0.079)	0.08 (0.331)	0.36 (0.026)	0.10 (0.367)	0.33 (0.002)	0.07 (0.332)	0.09 (0.529)	0.01 (0.900)
P5	1.31 (3E-5)	0.18 (0.349)	-0.16 (0.214)	0.25 (0.024)	0.24 (0.115)	0.03 (0.785)	0.16 (0.182)	0.35 (0.002)	0.06 (0.481)	-0.00 (0.986)
P6	1.15 (2E-10)	0.39 (2E-4)	0.06 (0.412)	0.10 (0.137)	0.41 (0.002)	0.00 (0.986)	0.24 (0.025)	0.22 (0.022)	0.05 (0.629)	0.04 (0.531)
P7	0.80 (3E-8)	0.32 (0.009)	-0.06 (0.491)	0.09 (0.285)	0.123 (0.203)	0.18 (0.108)	0.04 (0.618)	0.23 (0.009)	0.04 (0.634)	0.90 (0.364)
P8	1.03 (4E-8)	0.26 (0.062)	0.22 (0.061)	0.10 (0.235)	0.23 (0.096)	0.23 (0.022)	0.49 (0.000)	0.09 (0.302)	-0.04 (0.766)	0.12 (0.234)

Table S3. Statistical *p*-values associated with simple effects of sound category (in speech fROI: speech activator > music activator; in music fROI: music activator > speech activator), for each participant and sound type (natural sounds and synthetic sounds from each DNN layer). *P*-values were Bonferroni-corrected across layers, the uncorrected *p*-values are shown in parentheses.

Chapter 4

Deep Sound Synthesis Reveals Sound Features Underlying Functional Asymmetry of the Auditory Cortices

Based on

Xing, L., Formisano, E., Riecke, L. (in preparation). Deep sound synthesis reveals sound features underlying auditory cortical asymmetry.

Abstract

Lateralization of auditory processing is a hallmark of human brain organization, yet the acoustic features driving hemispheric specialization in the auditory cortex remain debated. Prevailing models emphasize spectrotemporal modulation (STM) sensitivity—to rapid temporal cues in the left hemisphere (LH) and to fine spectral cues in the right hemisphere (RH)—but fail to fully explain lateralization of non-primary auditory cortical responses. Using fMRI, natural-sound analysis, and deep neural network (DNN)-guided sound synthesis, we show that spectral cross-band correlations in acoustic input are a dominant predictor of hemispheric asymmetry in the planum polare (PP), planum temporale (PT) and superior temporal gyrus (STG), whereas STM features of acoustic input better account for hemispheric differences in the superior temporal sulcus (STS). “Left-” and “right-activator” sounds (auditory stimuli synthesized in a DNN-based, hypothesis-free procedure to selectively activate a given hemisphere) were found to selectively engage their respective target hemisphere in line with the DNN’s predictions. More importantly, they were found to replicate the strong cross-band correlations observed in natural sounds evoking strongly lateralized responses. These findings provide converging evidence that higher-order spectral statistics, not STM alone, underlie auditory cortical asymmetry and establish DNN-based stimulus synthesis as a powerful tool for investigating neural feature selectivity.

Keywords

Auditory cortex; Hemispheric asymmetry; Cross-band correlations; Spectrotemporal modulation; Deep neural networks

1 Introduction

Functional asymmetry of the human brain is a central topic in cognitive neuroscience. While both hemispheres are involved in the processing of speech and music (Norman-Haignere et al., 2015; Overath et al., 2015), evidence also points to a hemispheric preference: the left hemisphere (LH) appears to be more specialized for speech processing, whereas the right hemisphere (RH) shows preferential engagement in music processing (Zatorre et al., 2002; Poeppel, 2003; McGettigan and Scott, 2012). This functional preference is typically attributed to the distinct acoustic demands of speech and music processing—speech requiring high temporal resolution for phoneme discrimination, and music relying on fine spectral resolution for pitch and timbre analysis (Zatorre et al., 2002). A prominent theoretical framework for auditory processing posits that auditory cortical neurons function as spectrotemporal modulation (STM) filters that respond selectively to distinct combinations of temporal and spectral modulations (Chi et al., 2005). This modulation-based model has also been proposed to explain hemispheric specialization. More specifically, the left auditory cortex is thought to be more sensitive to temporal modulations supporting the decoding of rapid phonemic transitions, while the right auditory cortex favors spectral modulations facilitating the perception of melodic contours and prosodic features (Flinker et al., 2019).

Recent evidence found that STM model-matched artificial sounds may fail to replicate the responses that their natural counterpart sounds elicit in non-primary auditory regions, suggesting that the STM model may not fully capture the complexity of auditory cortical processing (Norman-Haignere and McDermott, 2018). Instead, higher-order representations—such as those learned by deep neural networks (DNNs)—may better characterize these computations (Kell et al., 2018; Giordano et al., 2023; Tuckute et al., 2023). These findings raise a critical question: is the STM framework sufficient to explain hemispheric specialization across human auditory cortex? In this study, we address this question using functional magnetic resonance imaging (fMRI), natural auditory stimuli, and a hypothesis-free approach exploiting a DNN-based sound synthesis.

In the first part of the study, unlike prior studies that manipulated sound features based on pre-defined hypotheses, we presented eight human participants with a diverse set of everyday sounds, recorded the sound-evoked responses with fMRI, identified the sounds that elicited the strongest hemispheric lateralization, and then analyzed the distinct acoustic features of these most-lateralizing sounds. Besides STM, we also probed more time-averaged acoustic features in a sound texture analysis. Sounds synthesized to match the sound texture of natural sounds have been found to elicit realistic percepts, suggesting that texture analysis is psychophysically meaningful for the human auditory system. The perception of these textures has been shown to depend on the statistical structure of early auditory cortical representations—particularly cross-band correlations—that might be computed by downstream cortical populations (McDermott and Simoncelli, 2011).

In the second, confirmatory part of the study, we employed a sound synthesis approach guided by a DNN trained for sound classification. DNNs, which encode auditory features in a high-dimensional hierarchical space, have been shown to outperform conventional acoustic models in predicting auditory cortical responses (Kell et al., 2018; Giordano et al., 2023; Tuckute et al.,

2023). Sound representations in the DNN remain difficult to interpret due to the "black box" nature of the DNN, yet the DNN allows for an efficient feature-identification strategy: utilizing the model's predictions, it is feasible to synthesize stimuli composed of features that elicit neural targets, i.e., specific neural response patterns. This feature-identification approach has been used to visualize or sonify the abstract features embedded in DNNs that underlie stimulus processing in the visual cortex (Bashivan et al., 2019; Ponce et al., 2019; Ratan Murty et al., 2021; Tuckute et al., 2023) and auditory cortex (Xing et al., 2025). Using this framework, we synthesized sounds predicted to preferentially activate either the left or right auditory cortex based on DNN outputs. We then compared the STM and higher-order acoustic features of these "left-activator" and "right-activator" sounds to identify the sounds feature that may drive the functional asymmetry of the auditory cortex.

2 Methods

The fMRI data were acquired in a previous study, for a full description of the methods used, see Xing et al. (2025).

2.1 Participants

Eight healthy participants (median age = 29, two males) participated in the fMRI experiment. All participants reported normal hearing and no history of a hearing disorder or neurological disease and gave informed consent before undergoing the experimental procedures. The experimental procedure was approved by the local ethics committee of the Faculty of Psychology and Neuroscience, Maastricht University (ERCPN #_233_09_02_2021).

2.2 Extraction of ROI Responses to Natural Sounds

2.2.1 Auditory stimuli

The auditory stimulation consisted of 288 recordings of natural sounds (Santoro et al., 2017), with 48 sounds from each of the following six semantic categories: speech, (nonspeech) voice, animal cries, musical instruments, nature scenes, and tools. All sounds had a 1-s duration including 10-ms ramps at their onset and offset. They were sampled at 16 kHz, matched for RMS, filtered to correct for non-flatness of the frequency response of the MR-compatible in-ear system (Sensimetrics S14), and presented at a comfortable sound level.

2.2.2 fMRI Procedure

To establish six unique fMRI runs, the 288 sounds were randomly split into six sound sets with equal, uniform category distributions. Within each set, the sounds were randomly ordered, with the constraint that a given category was never repeated immediately. Sounds were presented in blocks of five, so that each block included five presentations of the same sound. To encourage participants to focus on the sounds, they were instructed to press a button when

they detected the target. The latter was defined as a 7-dB level decrease applied to a single, randomly chosen sound repetition in every block. Loudness-change detection performance ranged from 72.5% to 96.1% (mean $\pm$ SEM: $84.9 \pm 0.03\%$ across participants) given a chance level of 25%, suggesting that all participants paid attention to the auditory stimulation. After every fourth block (or “trial”) a baseline block was presented that was identical to the aforementioned stimulation blocks, except that no sound or task was presented. For participant P1, the stimulation protocol included a larger number of blocks with shorter durations (blocks of four sounds alternated with baseline blocks of half the duration of sound blocks). The aforementioned procedure was conducted twice (i.e., in two fMRI sessions) with a gap of 1–7 days in-between.

The MRI data were acquired with a 3T scanner (Siemens) at Scannexus (<https://scannexus.nl/>). Anatomical T1-weighted images were acquired using a modified magnetization-prepared rapid-acquisition gradient-echo sequence (number of slices = 192, TR = 2250 ms, TE = 2.21 ms, flip angle = 9° , voxel size = 1 mm isotropic, GRAPPA acceleration factor = 3). BOLD signal changes were recorded using a modified gradient-echo echo-planar imaging (EPI) sequence (number of slices = 33 without gap, TR = 2400 ms, TA = 1000 ms, TE = 30 ms, flip angle = 82° , voxel size = 2 mm isotropic, GRAPPA acceleration factor = 3). Functional image acquisition and sound presentation alternated to ensure clear audibility of the sounds in the absence of scanner noise. Functional slices were centered on the Sylvian fissures to cover the temporal lobes. All functional and anatomical images were acquired in anterior-posterior phase encoding direction. To enable subsequent correction for EPI distortions, an additional pair of EPI volumes with reversed phase-encoding direction was acquired before the first functional run (Andersson et al., 2003).

2.2.3 fMRI Data Preprocessing

Imaging data were preprocessed with BrianVoyager 22.4 (Brain Innovation). Functional data were processed in four steps including slice scan-time correction (using cubic spline interpolation), 3D motion correction (using trilinear/sinc interpolation), temporal high-pass filtering (cutoff frequency: 7 cycles per run), and EPI distortion correction (using the TOPUP method). The anatomical and functional data from all fMRI sessions were spatially coregistered and normalized to Talairach space.

The fMRI signal of each voxel was z-scored along the time course within each run and delayed by one TR relative to the stimulation protocol to compensate for the hemodynamic delay. Functional volumes acquired in response to the four sound repetitions within a given block were averaged to obtain a single, time-averaged response measure for that block (the volume acquired immediately after the first sound was rejected to reduce carryover effects from the preceding block). The same approach was applied to baseline blocks to obtain a time-averaged baseline response.

2.2.4 Regions of Interest (ROIs) Definition

We first defined four distinct anatomical ROIs in each hemisphere of each participant (Figure 1): Heschl's gyrus (HG), planum polare and planum temporale (PP/PT), superior temporal gyrus (STG), and superior temporal sulcus (STS). These regions span hierarchical stages from primary auditory cortex to higher-order associative areas. We defined a combined ROI encompassing both the PP and PT. This decision was motivated by several factors. First, neuroanatomically, these regions form a continuous belt of secondary auditory cortex surrounding Heschl's gyrus, jointly serving as the primary substrate for complex sound analysis (Kaas and Hackett, 2000). Second, rather than representing discrete functional modules, PP and PT are thought to lie along a spectrotemporal processing gradient (Woods et al., 2010); for stimuli requiring integrated spectral and temporal analysis, separating them may artificially disrupt a unified computational unit. Third, given the well-documented anatomical variability in the boundaries of Heschl's sulcus, a combined ROI provides a more robust and reliable anatomical definition across participants.

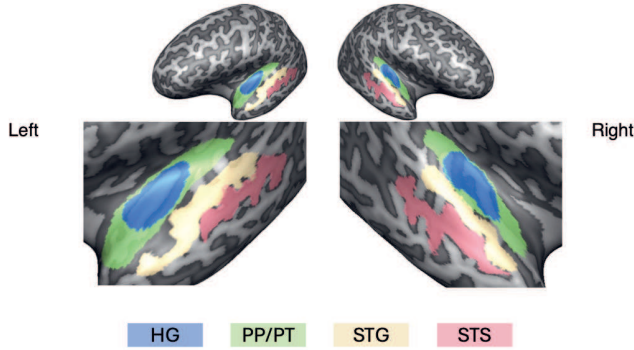


Figure 1. Anatomical regions of interest in the temporal cortex. Cortical map of the individual ROIs in the left and right hemispheres, illustrated for participant 1. ROIs are color-coded: HG (blue), surrounding PP/PT areas (green), STG (yellow), and STS (red). These subdivisions were used for region-specific analyses of acoustic feature selectivity underlying hemispheric asymmetry.

The ROIs were manually labeled on each subject's cortical surface mesh representation according to the criteria defined by Kim et al. (Kim et al., 2000). The HG is the most anterior transverse gyrus on the superior temporal plane. Its anteromedial border was defined by the first transverse sulcus (FTS), and its posterolateral border was defined by the Heschl's sulcus (HS), or by the sulcus intermedius (SI) in case it was present. Medially, the HG was confined by the circular sulcus of the insula (CSI). The PP/PT was defined as PP and PT. PP is a region on the superior temporal plane anterior to the HG. Medially, the PP was confined by the CSI. The lateral border was defined following the FTS until the anterior end of the HS or the SI; here, the lateral border was changed to the lateral rim of the superior temporal plane. The PT is a triangular region posterior to the HG on the superior temporal plane. Its anteromedial border

was defined by the HS, or by the SI when it was present. The lateral rim of the superior temporal plane confined the PT laterally; the medial border was defined as the deepest point of the Sylvian fissure from the medial origin of the HS until the posterior point of the STG at the temporal-parietal junction. The STG constitutes the lower bank of the Sylvian fissure and runs parallel to the STG. It was traced from its anterior border at the temporal pole to the posterior end of the Sylvian fissure at the temporal parietal junction. The STS is a prominent sulcus running parallel and inferior to the STG. We obtained volumetric ROIs by projecting the vertex patches defined on the surface meshes into the volume space of the same subject.

2.3 Response Asymmetry to Natural Sounds

To quantify the response asymmetry to each natural sound, we computed the softmax score as follows:

$$\text{Response asymmetry} = \frac{e^{RH}}{e^{RH} + e^{LH}},$$

Where e denotes the base of the natural logarithm, and RH and LH represent the averaged BOLD responses in the right and left hemispheres, respectively. A score above 0.5 indicated stronger activation in the right hemisphere, while a score below 0.5 indicated stronger activation in the left.

The softmax score was used due to its ideal gradient properties for our gradient-based sound synthesis approach (see Section 2.6). Specifically, it generates larger gradients for balanced hemispheric responses, which drives effective optimization, and smaller gradients as the response becomes more lateralized, ensuring stable convergence.

2.4 Acoustical Models

The modulation transfer function (MTF) was computed using NSLtools (Chi et al., 2005) as follows: first, the sound cochleagram was computed, which consisted of the time-varying output (temporal resolution = 8 ms) of 128 cochlear filters with $\log_2$ spaced frequencies (179–7246 Hz). Then, the magnitude of the spectrotemporal modulations (scale from 0.25 to 8 cycles/octave; unsigned rate from 2 to 32 Hz) in each channel of the cochleagram representation was assessed and averaged across time points. The sound texture was analyzed using the Sound Texture Synthesis Toolbox v.1.7 (McDermott and Simoncelli, 2011). The texture model considered here includes four separate components measuring different summary statistics of the time-varying amplitude envelopes of the cochleagram (32 frequency bands evenly spaced on an ERB-rate scale from 20 Hz to 10 kHz), or of the modulation analysis specific to each of the frequency bands (six modulation filters with center frequencies log-spaced between 30 and 100 Hz): (1) the marginal statistics of the band-specific envelopes (mean, EM; variance, EV; skewness, ES; kurtosis, EK); (2) the pairwise correlations C between band-specific amplitude envelopes (EC); (3) the pairwise correlation between the modulation analysis of each frequency band at a constant modulation filter frequency ($C1$ correlations); and (4) the correlation between the modulation analysis of the same frequency band at adjacent modulation filter frequencies

(C2 correlations). For each type of representation, between-stimulus cosine distances were computed, and all parameters were given equal weight.

2.5 Support Vector Machine (SVM) Analysis for Natural Sound

We fitted the SVMs with a linear kernel function and set the regularization parameter to 1. When fitting the SVM, the sounds were randomly split into a training set and a test set, each including half of the preferred sounds of each hemisphere. This cross-validation was repeated 30 times and resulted in a distribution of the classification accuracy. Besides, the labels of the natural sounds were shuffled 1000 times to obtain a null distribution.

Statistical significance was assessed using a one-sided permutation test, in which the observed classification accuracies were compared against permutation-based null distributions. Multiple comparisons were controlled using a maximum-statistic correction.

2.6 Validation via Deep Sound Synthesis

We applied a sound synthesis approach that allowed us to identify the features preferred by each hemisphere in a hypothesis-free way, i.e., within the vast feature space of a DNN (Figure 2).

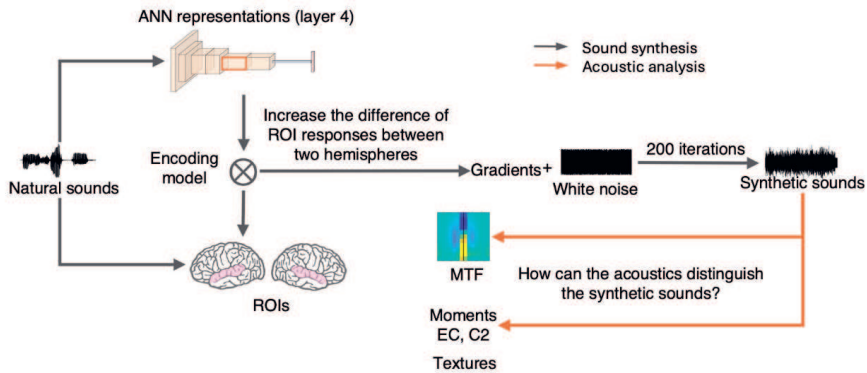


Figure 2. Sound synthesis-based feature-identification approach. Black arrows indicate the steps involved in the DNN-based sound feature identification. Orange arrows indicate steps involved in the characterization of the features identified to preferentially activate either the left or right auditory cortex. In the first part, natural categorical sounds were presented diotically to both DNN and participants, and the DNN and fMRI responses to these sounds were recorded. A linear encoding model was fitted that mapped the DNN responses onto the fMRI responses within the anatomically-defined ROIs in the left and right auditory cortex (see Figure 1). Utilizing this mapping and a gradient-based approach, features embedded in the DNN that drove the predicted ROI response in a given hemisphere, while suppressing that in the contralateral hemisphere, were sonified, resulting in novel synthetic, hemisphere-specific activator sounds. Subsequently, the left- and right hemisphere activator sounds were analyzed with various acoustical models to obtain their representations in different sound-feature spaces. Linear

SVMs were used to assess the discriminability of the representations of the left- vs. right-hemisphere activator sounds.

2.6.1 Extraction of DNN Responses to Sounds

The DNN used here was the VGGish (Hershey et al., 2017), a DNN that has been pre-trained to classify 10-s YouTube audio tracks with a set of 527 labels from the Audioset ontology (Gemmeke et al., 2017). The VGGish takes audio segments of 0.975-s duration (converted to a stabilized log-mel spectrogram) as input and feeds it through an architecture consisting of an input layer, four convolutional blocks (convolution, ReLU, MaxPooling), and three fully connected blocks (fully connected, ReLU) that progressively reduce dimensionality, resulting in a 128-dimensional semantic embedding (output layer). We fed the 288 natural sounds into the VGGish and extracted representations of these sounds from its convolutional layers and fully connected layers as follows: after generating a cochleagram of each sound, we passed it through the VGGish, and extracted the response of each model unit, from each layer. The convolutional layers represented the sound along three dimensions: time, frequency, and kernel. To facilitate prediction of the time-averaged ROI responses, we averaged the time dimension yielding a time-averaged vector whose length (4096 points) equaled the product of the number of frequency points and number of kernels. A vector of equal length was directly extracted from the fully connected layers (i.e., without time dimension). Thus, this procedure resulted in eight sets of fixed-size (4096 points) sound representations (conv1, conv2, conv3_1, conv3_2, conv4_1, conv4_2, fc1_1, fc1_2). Considering the computational load, we limited the study to the fourth layer, which has been shown to accurately predict human auditory cortical fMRI responses to natural sounds (Giordano et al., 2023).

2.6.2 Encoding Models: Mapping DNN Responses onto ROI Responses

We next trained an encoding model to map the DNN layer 4 representations of the sounds onto the measured ROI responses to these sounds, separately for each ROI. We modeled each ROI's time-averaged response to a given sound as a linear combination of the 4096 points defining that sound's time-averaged representation in layer 4. The model was fitted using regularized linear regression and a six-fold cross-validation procedure. The six unique sound sets from the six fMRI runs were split into five training sets (used to fit the model) and one test set (used to evaluate the accuracy of the prediction of the fitted model). Each set contained responses to 48 unique sounds and was z-scored independently. On each iteration of the procedure, a different sound set was defined as test set. Given that the number of regressors (4096) exceeded the number of training sounds (240), L2-regularized regression was used. Introducing the L2-penalty on the weights results in a closed-form solution to the regression problem, which is similar with the least-squares regression normal equation:

$$\mathbf{w} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y},$$

where $\mathbf{w}$ was a 4096-point column vector containing the regression weights, $\mathbf{y}$ was a 240-point column vector representing the response variable (i.e., the ROI's response to each sound), $\mathbf{X}$

was a 240×4096 matrix representing the predictor variable (i.e., the layer 4 representation of each sound), and I was the 4096×4096 identity matrix. The ridge parameter was optimized within the range $\lambda = 10^2, 10^1, \dots, 10^{-6}, 10^{-7}$. The value yielding the most accurate predictions (maximum mean predictive variance) was retained. The predictive variance R^2 was estimated as $1 - \text{SSE}_{\text{test}} / \text{SST}_{\text{test}}$, where SSE_{test} represents the sum of the squared error of the trained model's prediction of the test-set data and SST_{test} represents the total sum of squares of the test-set data. To obtain an empirical chance level, we computed 1000 permutations across the test-set sounds for each split. The same permutations were applied to the corresponding sounds in the training set. The aforementioned procedure was conducted separately for each ROI $\times$ hemisphere (4×2).

2.6.3 Sound Synthesis

We utilized the VGGish and the fitted encoding model to synthesize sounds via a gradient-ascent optimization approach. The procedure started with 30 white noise samples, the RMS level of which was matched to the natural sounds. The noise was fed into the VGGish, its representation was extracted from layer 4, and the ROI response to this noise was predicted using the fitted encoding model. The value of each time point of the noise waveform was adjusted in the direction that increased the hemispheric difference, using a cost function based on the softmax score of the predicted ROI response and the total variation between directly neighboring time points (the latter served to reduce acoustic transients). The learning rate was set to a value of 0.001. The gradient of the cost function was calculated across 200 iterations using the function built in TensorFlow 2.13 (www.tensorflow.org), after which the optimized sounds showing the minimum costs were chosen. In each iteration, the novel sound from the previous iteration was temporally shifted in random direction (forward or backward) with a maximum of 1 ms to generate the adversarial component; the latter served to promote the synthesis of sounds that elicit predicted ROI response levels unaffected by such small shifts. The above procedure was conducted separately for each ROI $\times$ hemisphere (4×2), each time starting with the same set of 30 (frozen) white noise samples. Overall, this resulted in a set of 240 ($4 \times 2 \times 30$) unique synthetic sounds.

2.6.4 SVM Analysis for Synthesized Sound

For the SVM analysis of the synthesized sounds, we employed a cross-validation procedure that explicitly accounted for the fact that the sound-synthesis process was performed separately for each participant. Specifically, the participant sample was evenly split into a training set and a test set. All possible combinations of training–test splits were considered, resulting in a total of 70 cross-validation folds. The SVM classification procedure for synthesized sounds followed the same settings used for the natural sounds. Likewise, the parameter configuration,

permutation procedure, and statistical evaluation were identical to those applied in the natural-sound analysis.

3 Results

3.1 SVM-Based Classification of Lateralizing Sounds in Various Acoustic Spaces

For each anatomical ROI, we quantified response asymmetry to the diotically presented natural sounds using a softmax-based score. A score above 0.5 indicates stronger activation in the right hemisphere, while a score below 0.5 indicates stronger activation in the left. Based on this score, we then selected the 30 most “left-lateralizing” sounds (LLS) and the 30 most “right-lateralizing” sounds (RLS) per ROI. Next, we transformed these sounds into various acoustic representations (see *Methods*) to identify the feature space in which LLS and RLS could be discriminated best. The acoustical models used to transform the sounds included the MTF model and the sound texture model, the latter includes multiple time-averaged statistical indices of its acoustic input.

To assess the separability of LLS and RLS in each feature space, we trained a SVM to classify the model representations of LLS and RLS. To visualize the results, we computed representational dissimilarity matrices (RDMs) and projected them onto a two-dimensional space using multidimensional scaling (MDS) (Figure 3B). The results demonstrate that the two sets of natural sounds (LLS vs. RLS) are linearly separable within each feature space tested, with the strongest separability observed in the cross-band correlation feature space (Figure 3B; see also Figure S1–4). To quantify classification performance, we conducted cross-validation and compared the observed accuracy against a null distribution (obtained from 1,000 permutations). As shown in Figure 3C, the observed classification accuracy was significantly above chance for most models in HG, PP/PT and STG ($p < 0.013$, corrected across models, LLS and LRS are selected according to averaged responses across participants), except the single band statistics ($p > 0.093$, corrected). In STS, only the MTF was found to discriminate

the two sets of sound ($p = 0.005$, corrected). These results indicate that LLS and RLS can be reliably distinguished throughout the auditory cortex based on multiple sound features.

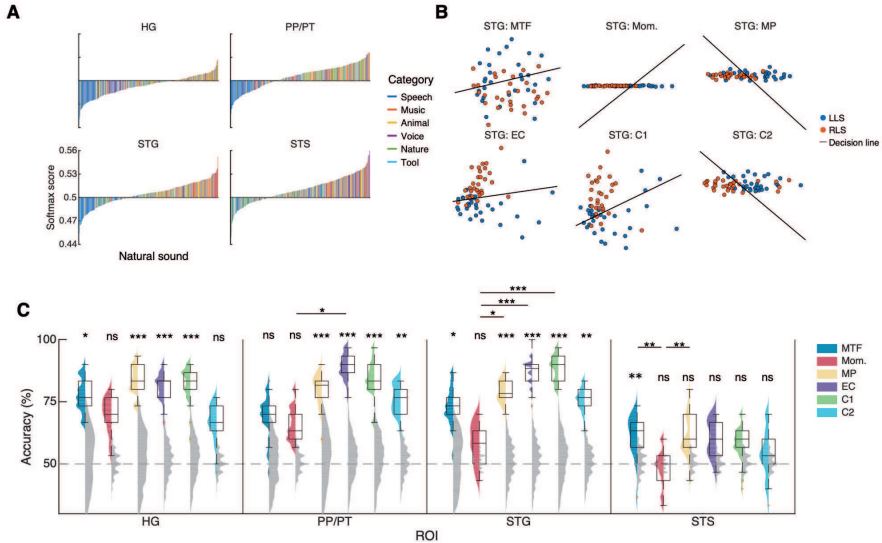


Figure 3. Response asymmetry to natural sounds. **A.** Lateralization of fMRI responses to natural sounds (averaged across participants) for each ROI. Each colored vertical line represents a single natural sound, with sound categories color-coded. The softmax score is computed as the response of the left hemisphere divided by the summed responses of both hemispheres. A score greater than 0.5 indicates a stronger activation in the right hemisphere, whereas a score below 0.5 indicates stronger activation in the left hemisphere. **B.** Two-dimensional (MDS-based) visualization of the representational similarity of the natural sounds that showed the most lateralized responses in STG, for each sound-feature space (MTF, modulation transfer function; Mom., marginal moments; EC, envelope correlation; C1 and C2, modulation correlations). Each dot represents the model representation of a given sound, and the distance between dots represents the representational dissimilarity of the sounds in the given feature space. The lines depict the decision lines of SVMs trained to discriminate the representations of the LLS (blue dots) and RLS (red dots). **C.** SVM classification accuracy within each sound-feature space (color-coded; gray distributions represent the empirical chance level obtained from 1000 permutations) is shown for each ROI. It can be seen that the left-hemisphere and right-hemisphere activator representations were discriminable in multiple sound-feature spaces and ROIs. Moreover, discriminability varied across sound-feature spaces in a subset of ROIs. Summary statistics (across validation folds): center, median; lower/upper box limits, first/third quartile; bottom/ top whisker, data within 1.5 interquartile ranges from first and third quartiles, respectively. ns, non-significant; *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

To identify the most accurate model, we further compared classification performance within each possible pair of models. We found no significant difference across models in HG ($p >$

0.092, corrected). However, models based on cross-band correlations (EC and C1) outperformed models based on individual band statistics (e.g., Moments; $p < 0.05$, corrected) in STG and PP/PT. Moreover, modulation-based models (MTF and MP) outperformed individual band statistics in STS (Mom., $p < 0.002$, corrected). These findings suggest that lateralized neural responses in PP/PT or STG are best explained by cross-band correlations, while in STS they are better explained by spectrotemporal modulations.

To gain further insight into the relation between the aforementioned sound features and auditory hemispheric lateralization, we visualized the representational difference between LLS and RLS in each feature space and ROI (Figures S1–4). Notably, the left STG showed a preference for high cross-band correlations along the diagonal, while the right STG favored high correlations away from the diagonal. High values near the diagonal represent correlations between pairs of frequency channels that are adjacent to each other. Additionally, the right STS showed more sensitivity to spectral than temporal modulation, consistent with previous literature (Obleser et al., 2008; Flinker et al., 2019).

3.2 SVM-Based Classification of the Synthesized Sounds in Various Acoustic Spaces

To further validate these results, we submitted the synthesized sounds (see *Methods*) to the same SVM analysis as the natural sounds. We refer to sounds synthesized to activate the different hemispheres respectively as “left activators” and “right activators”. Figure 4A shows the spectrogram of an exemplary synthesized sound for participant P1. Audio files are provided in Supplementary material.

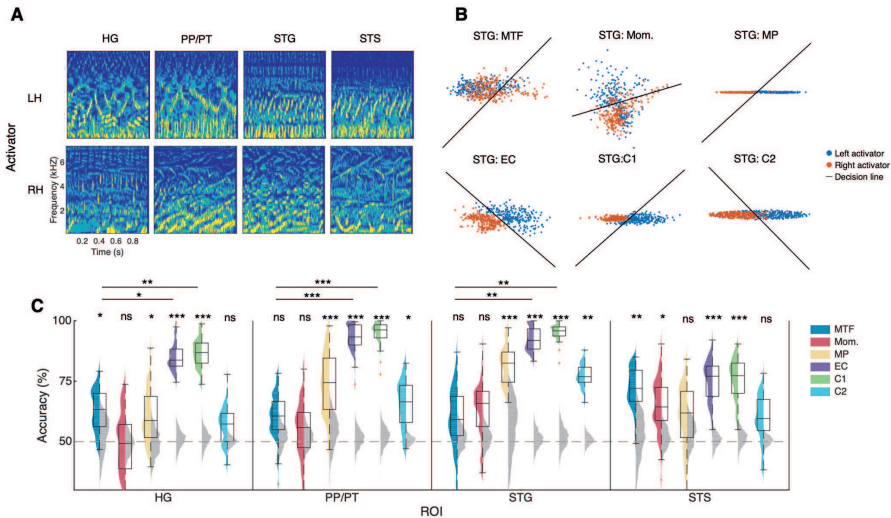


Figure 4. Representational discriminability of the hemisphere-specific activators. A. Examples of the DNN-based synthetic sounds for each ROI of participant P1. A frozen sample of white noise was iteratively fine-tuned to synthesize a sound that increased (“left activator”,

upper row) or decreased (“right activator”, lower row) the predicted softmax score of a given ROI. **B.** Two-dimensional (MDS-based) visualization of the representational similarity of the sounds synthesized to activate the left STG (blue dots) or right STG (red dots) of all participants, for each sound-feature space. Each dot represents the representation of a given sound, and the distance between dots represents the representational dissimilarity of the sounds in the given feature space. The lines depict the decision lines of SVMs trained to discriminate the representations of the left-hemisphere and right-hemisphere synthetic activators. For details, see Figure 3B. **C.** SVM classification accuracy within each sound-feature space is shown for each ROI. It can be seen that the left-hemisphere and right-hemisphere activator representations were discriminable in multiple sound-feature spaces and ROIs. Moreover, discriminability varied across sound-feature spaces in a subset of ROIs. Summary statistics (across cross-validation folds): center, median; lower/upper box limits, first/third quartile; bottom/ top whisker, data within 1.5 interquartile ranges from first and third quartiles, respectively. Some observations surpass 100% accuracy because the distributions were smoothed for visualization purpose only. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; For details, see Figure 3C.

As for the natural sounds, we found that the left and right activators could be distinguished by most acoustical models in each ROI, indicating that the synthesis procedure transformed the initially random waveforms into two distinct sets of sounds: one driving preferentially the left auditory cortex, and the other driving preferentially the right auditory cortex. Moreover, we found again that cross-band correlations are the most discriminating feature between the two sets of sound (Figures 4B and S5–8), as they significantly outperformed individual band moments in HG, PP/PT and STG ($p < 0.008$, corrected). We also found that in PP/PT and STG, cross-band correlations outperformed the MTF model ($p < 0.015$, corrected). However, unlike for the natural sounds, we did not find any acoustic model effect in STS ($p > 0.411$, corrected).

We then computed the representational difference between the two sets of activators as before for the natural sounds (Figures S5–8). Visual inspection showed that left activators had higher cross-band correlations along the diagonal than right activators in HG, PP/PT, and STG, consistent with the results from natural sounds. For STS, the left vs. right activators showed higher spectral modulation above 0.5 cycle/oct, which partially aligns with the results from the natural sounds.

4 Discussion

The present study investigated the acoustic features underlying functional hemispheric specialization in the human auditory cortex. Combining fMRI with a natural sound analysis and a DNN-guided sound synthesis, we found strong evidence that higher-order acoustic features—particularly cross-band spectral correlations—are linked to auditory cortical lateralization, especially in non-primary regions. This finding provides new insights into the neural basis of

hemispheric specialization and enriches theories on lateralized auditory processing, which so far have considered STM as a primary factor (Flinker et al., 2019).

4.1 Revisiting the STM Hypothesis

Functional asymmetries in the auditory cortex have long attracted scientific interest. Three major theoretical accounts have been proposed to explain auditory lateralization: the temporal versus spectral processing hypothesis (Zatorre et al., 2002), the asymmetric sampling in time theory (AST) (Poeppel, 2003), and, more recently, the spectrotemporal modulation framework (Flinker et al., 2019), which has been introduced as a unifying model.

Consistent with the STM framework, we found that STM-based models could significantly distinguish left-lateralizing from right-lateralizing sounds, particularly in the STS (Zatorre, 2001; Boemio et al., 2005; Obleser et al., 2008). Thus, our results support the long-standing hypothesis that the left auditory cortex is more attuned to rapid temporal modulations (relevant for speech processing), while the right auditory cortex favors spectral modulations (important for music and prosody). Importantly, we further found that STM models alone failed to fully account for the observed lateralization patterns in other auditory cortical regions, such as the PP/PT and STG, in which lateralization patterns were better explained by a texture model.

This pattern is partially consistent with prior findings: Boemio et al. (2005) reported lateralization of temporal integration windows in the STS, but not in the STG. Similarly, in our data, the MTF model best explained the asymmetry in the STS, whereas it was outperformed by cross-band correlation (texture) features in the STG. Together, these results indicate that hemispheric asymmetry exists in multiple cortical regions that show lateralized processing of multiple, distinct stimulus features.

4.2 Cross-Band Correlations as a Key Feature for Hemispheric Specialization in Auditory Cortex

The main contribution of this study is the identification of cross-band spectral correlations as a dominant feature driving hemispheric asymmetries in PP/PT and STG. These correlations, which reflect how spectral bands co-vary over time, capture higher-order statistical regularities in complex, e.g., natural, sounds. Our classification analyses consistently show that models incorporating cross-band correlations outperform both STM and simpler statistical models (e.g., moments of individual bands) in predicting lateralized responses in the aforementioned regions. Importantly, this observation held across both natural sounds and artificial DNN-based sounds (synthesized to induce lateralized responses), reinforcing the robustness of our finding. It aligns with previous work suggesting that sound texture perception relies on statistical regularities in spectral correlations (McDermott and Simoncelli, 2011; McDermott et al., 2013).

Why might cross-band correlations drive hemispheric specialization? Part of the effect might be attributed to a putative relation between cross-band correlations and STM-based acoustics. The observation that left-STG activators are characterized by high values along the diagonal of the EC space indicates that neighboring frequency bands are strongly correlated in these activator sounds, suggesting a low spectral resolution in left STG. It is possible that the left

auditory cortex, traditionally associated with speech processing (Scott et al., 2000), can utilize these more local spectral correlations to facilitate the extraction of phonetic and syllabic structure, even when the speech signal has been spectrally degraded (Shannon et al., 1995; Elhilali et al., 2003; Elliott and Theunissen, 2009). Conversely, the right auditory cortex, often linked to music (Samson and Zatorre, 1994) or syllable pattern perception (Abrams et al., 2008), might prioritize more widespread, possibly harmonically spaced spectral correlations, aiding in melodic contour and timbre discrimination.

An alternative interpretation is that cross-band correlations may largely arise from broadband transient events, such as abrupt amplitude onsets. When a transient occurs, it produces simultaneous increases in energy across many cochlear channels, leading to highly correlated envelope fluctuations across frequency bands. In this case, cross-band correlations would effectively serve as a statistical descriptor of broadband temporal events, rather than representing an independent acoustic dimension. Because such events are closely related to rapid temporal fluctuations, they overlap conceptually with acoustic cues often associated with left-hemisphere sensitivity to fast temporal information in the auditory cortex. Under this view, the apparent explanatory power of cross-band correlations might partly reflect their ability to summarize the temporal alignment of energy changes across frequency channels, which is already implicit in representations of temporal modulation or onset structure. Consequently, the observed neural effects could reflect sensitivity to broadband temporal events, with cross-band correlations acting as a convenient statistical measure of this phenomenon.

Although these interpretation align with the classic temporal vs. spectral dichotomy in left vs. right auditory cortex (Zatorre et al., 2002; Albouy et al., 2020), our result showing that functional asymmetry in STG is explained significantly better by EC than MTF extends this dichotomy to additional higher-order statistical structures as a critical contributor to hemispheric differences. STM representation characterizes sound in terms of modulation energy distributed across spectral and temporal modulation frequencies. In practice, this is often obtained by filtering a cochleagram with a bank of spectrotemporal modulation filters. Cross-band correlations, in contrast, measure the statistical dependencies between the envelopes of different frequency channels after cochlear filtering. From this viewpoint, STM describes local modulation structure in time–frequency space, while cross-band correlations describe global coordination across frequency channels. Therefore, our results suggest that the lateralization observed in the STG may be better explained by global acoustic features.

4.3 Natural Sounds and DNN-Based Synthetic Sounds

Unlike most previous studies using highly controlled artificial sounds, we employed a large number of natural sounds that may elicit stronger and more ecologically valid neural responses than artificial sounds (Leaver and Rauschecker, 2010; Norman-Haignere and McDermott, 2018). Meanwhile, the higher acoustical complexity of these natural sounds can facilitate the exploration of a larger variety of acoustical features (Norman-Haignere et al., 2015). In addition, the DNN-driven sound synthesis enabled us to move beyond correlational inference and approach a more causal understanding of feature selectivity (Bashivan et al., 2019; Tuckute et al., 2024). By generating stimuli designed to maximize differential activation of the

two hemispheres, we obtained DNN-derived “left activators” and “right activators” that we found to exhibit clearly distinguishable acoustic properties, primarily in terms of cross-band correlations.

Our observation that acoustic features that we found to be predictive of neural lateralization in natural sounds emerged also in the synthesis of activator sounds corroborates the validity of our feature identification framework. The converging results support that these features reflect fundamental properties of auditory cortical processing asymmetry, rather than idiosyncrasies of specific sound sets (Ratan Murty et al., 2021).

4.4 Implications for Models of Auditory Processing

Our findings challenge the sufficiency of low-level STM models for explaining functional hemispheric specialization in auditory cortex. They support a hybrid framework, in which different acoustic features drive asymmetry in different auditory regions. Specifically, the PP/PT and STG appear to be sensitive to time-averaged features (e.g., cross-band correlations), while the STS seems to increasingly integrate these local statistical features within time windows of different durations, possibly into higher-order features relevant to natural sound categories.

Moreover, the success of the DNN in modeling auditory cortical responses and in guiding the synthesis of novel stimuli underscores its potential as a tool for predicting neural responses and identifying stimulus features in a hypothesis-free manner. DNN models may allow simulating the abstraction processes of the human brain, especially in regions where classic analytical models fail.

4.5 Limitations and Future Directions

While our study highlights cross-band correlations as a critical feature for hemispheric asymmetry in auditory cortical processing, several questions remain. First, the exact computational role of these correlations—whether they reflect integrative mechanisms (e.g., integrating the co-varying frequencies) or predictive coding (e.g., predicting the following spectral structure)—warrants further investigation. Second, our synthesis approach relied on a single DNN architecture (VGGish). Future work could test the generalizability of our findings across different DNN models. Finally, as discussed above, there is a potential correlation between cross-band covariation and STM-based features. The isolation of these features is necessary to disambiguate their individual contributions in a future study.

4.6 Conclusion

By combining natural sound analysis, DNN-based synthesis, and fMRI, we have demonstrated that functional hemispheric specialization in early and intermediate processing stages of the human auditory cortex is strongly linked to cross-band spectral correlations in the acoustic input, rather than to modulation-based acoustic features alone. These findings refine existing models of auditory lateralization and underscore the importance of higher-order acoustic statistics in

shaping cortical organization. Future research should explore how these representations evolve across the auditory hierarchy and their role in real-world listening tasks.

5 Supplementary Materials

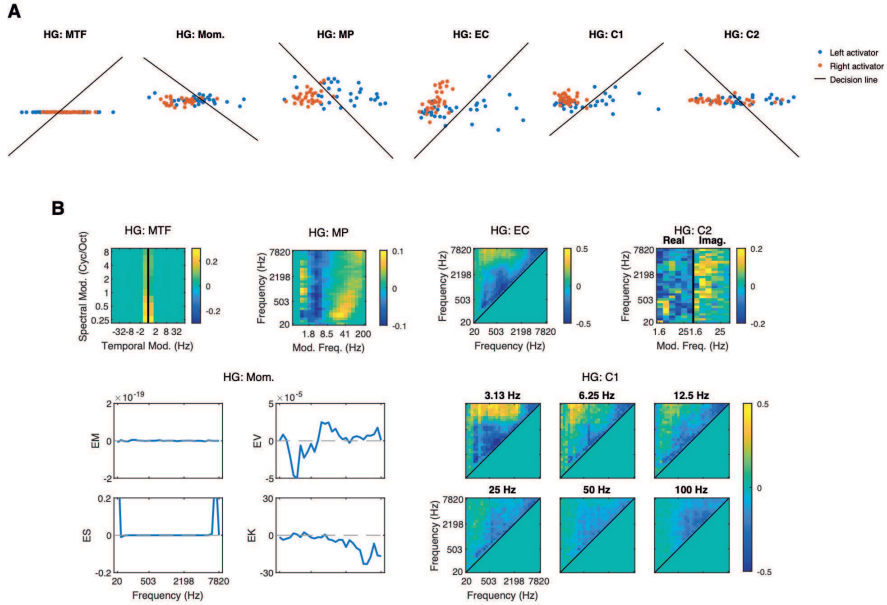


Figure S1. The separability of LLS and RLS in HG. **A.** Two-dimensional (MDS-based) visualization of the representational similarity of LLS and RLS. For details, see Figure 3B. **B.** The average acoustic difference between LLS and RLS (RLS minus LLS) is shown for each model. For details of each model, see Methods. In MTF space, the RLS have stronger spectral modulations. In MP space, they have stronger high-frequency (~ 40 Hz) power modulations and weaker low-frequency ($\sim 1\text{--}8$ Hz) power modulations. In moments space, differences exist mainly in the EV, where RLS have more variance in high frequency bands (> 500 Hz) and less variance in low frequency bands (< 500 Hz). In EC space, RLS have higher cross-band envelope correlations far off the diagonal and lower correlations near the diagonal. In C1 space, they have higher cross-band modulation correlations far off the diagonal and lower correlations near the diagonal, particularly in low modulation frequency bands (< 25 Hz). In C2 space, RLS have stronger within-band modulation correlations in the imaginary part and weaker modulation correlations in the real part.

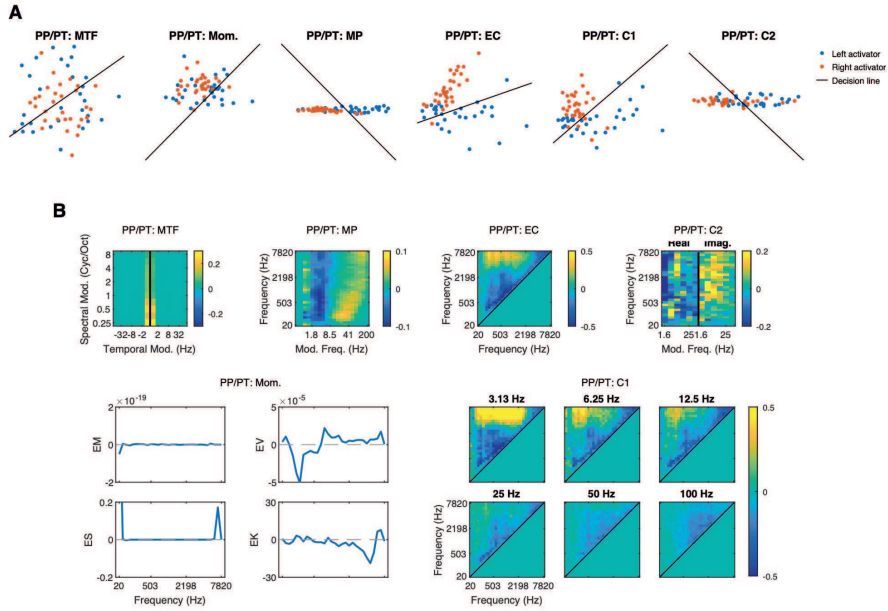


Figure S2. The separability of LLS and RLS in PP/PT. A. Two-dimensional (MDS-based) visualization of the representational similarity of LLS and RLS. For details, see Figure 3B. **B.** The average acoustic difference between LLS and RLS (RLS minus LLS) is shown for each model. For details of each model, see Methods. Results in PP/PT are qualitatively similar to those in HG.

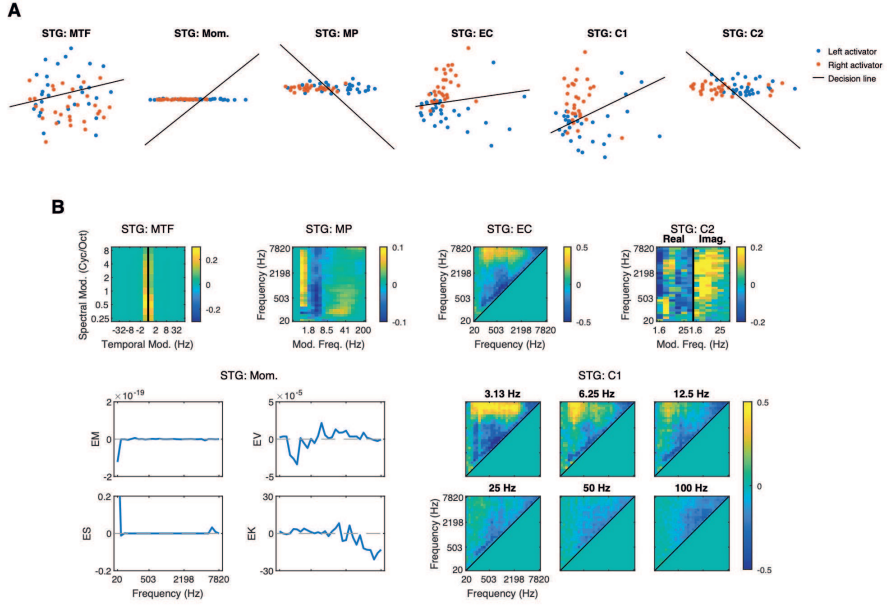


Figure S3. The separability of LLS and RLS in STG. A. Two-dimensional (MDS-based) visualization of the representational similarity of LLS and RLS. For details, see Figure 3B. **B.** The average acoustic difference between LLS and RLS (RLS minus LLS) is shown for each model. For details of each model, see Methods. Results in PP/PT are qualitatively similar to those in HG, except that in MP space, the RLS have stronger high-frequency ($\sim 2\text{--}8$ Hz) power modulations and weaker low-frequency (~ 1 Hz) power modulations.

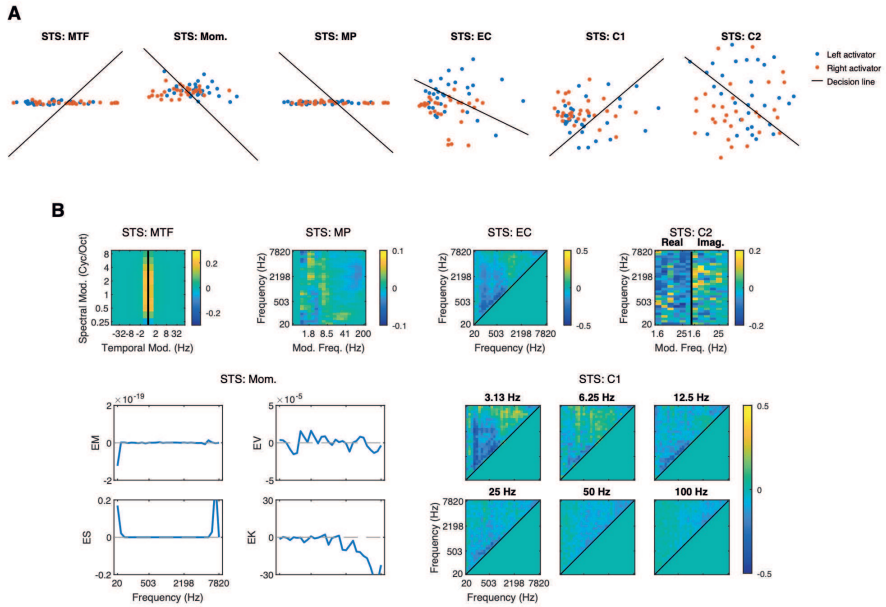


Figure S4. The separability of LLS and RLS in STS. A. Two-dimensional (MDS-based) visualization of the representational similarity of LLS and RLS. For details, see Figure 3B. **B.** The average acoustic difference between LLS and RLS (RLS minus LLS) is shown for each model. For details of each model, see Method. In MTF space, the RLS have stronger spectral modulations. In MP space, no clear difference pattern was observed. In moments space, the difference mainly comes from the EK, where RLS have smaller kurtosis in high frequency bands (> 2000 Hz). In EC space, RLS have lower cross-band envelope correlations in the low frequency range (< 1000 Hz). In C1 space, RLS have lower cross-band modulation correlations at low frequency bands (< 1000 Hz), particularly in low modulation frequency bands (< 25 Hz). In C2 space, RLS have stronger within-band modulation correlations in the imaginary part and weaker modulation correlations in the real part.

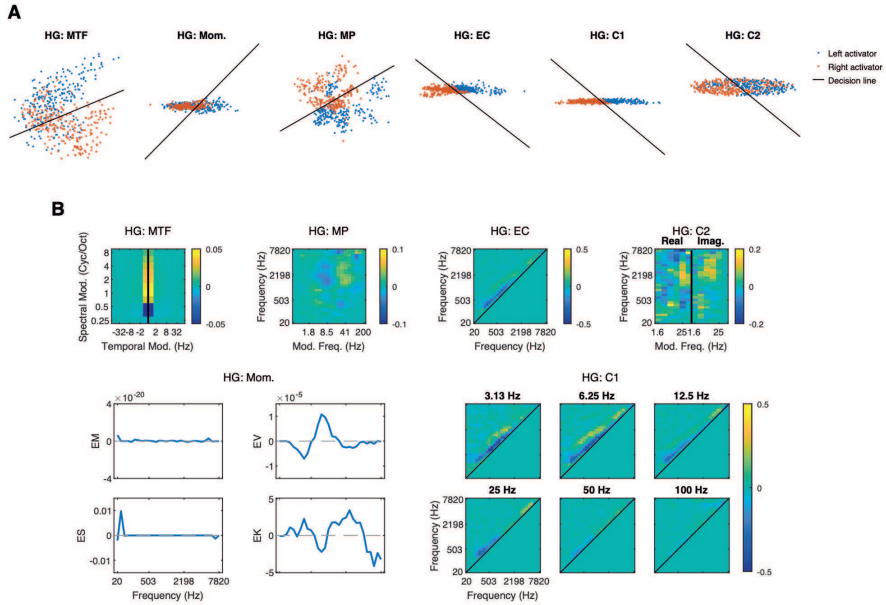


Figure S5. The separability of left activators and right activators in HG. **A.** Two-dimensional (MDS-based) visualization of the representational similarity of left activators and right activators. For details, see Figure 3B. **B.** The average acoustic difference between left activators and right activators (right activators minus left activators) is shown for each model. For details of each model, see Methods. In MTF space, the right activators have stronger fast spectral modulations (> 0.5 Cyc./Oct.) and weaker slow spectral modulations (< 0.5 Cyc./Oct.). In MP space, no clear difference pattern was observed. In moments space, the difference mainly comes from the EV, where right activators have more variance between 500–2000 Hz. In EC space, right activators have lower cross-band envelope correlations near the diagonal. In C1 space, right activators have lower cross-band modulation correlations near the diagonal, particularly in low modulation frequency bands (< 50 Hz). In C2 space, right activators have stronger within-band modulation correlations in the imaginary part.

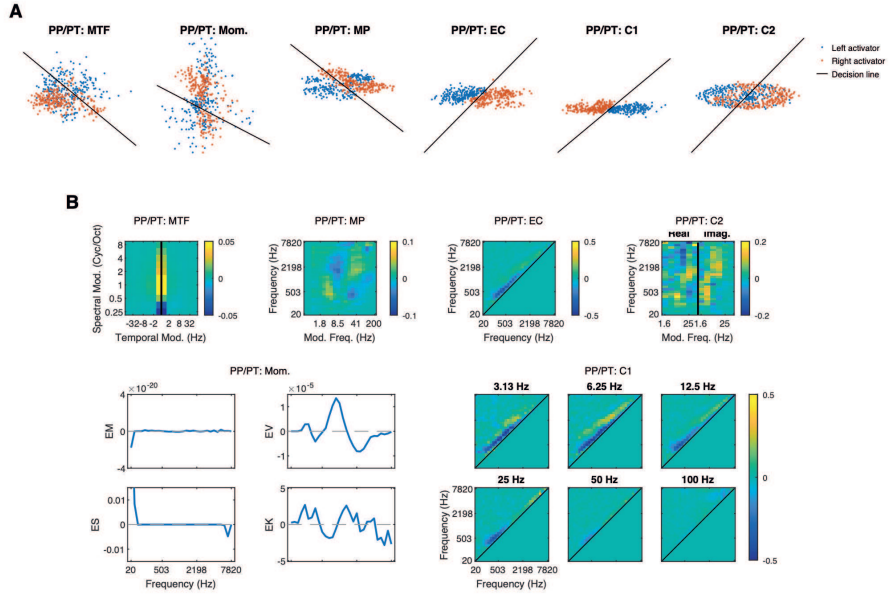


Figure S6. The separability of left activators and right activators in PP/PT. A. Two-dimensional (MDS-based) visualization of the representational similarity of left activators and right activators. For details, see Figure 3B. **B.** The average acoustic difference between left activators and right activators (right activators minus left activators) is shown for each model. For details of each model, see Methods. Results in PP/PT are qualitatively similar to those in HG.

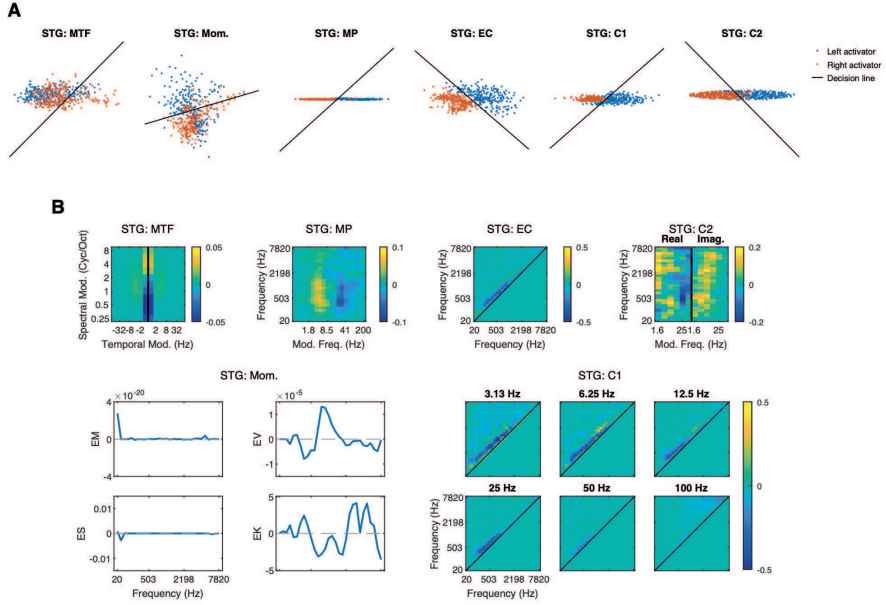


Figure S7. The separability of left activators and right activators in STG. A. Two-dimensional (MDS-based) visualization of the representational similarity of left activators and right activators. For details, see Figure 3B. **B.** The average acoustic difference between left activators and right activators (right activators minus left activators) is shown for each model. For details of each model, see Methods. Results in PP/PT are qualitatively similar to those in HG, except that in MTF space, the right activators have stronger fast spectral modulations (> 2 Cyc./Oct.) and weaker slow spectral modulations (< 2 Cyc./Oct.). In MP space, the right activators have stronger low-frequency (~ 4 – 8 Hz) power modulations and weaker high-frequency (~ 40 Hz) power modulations.

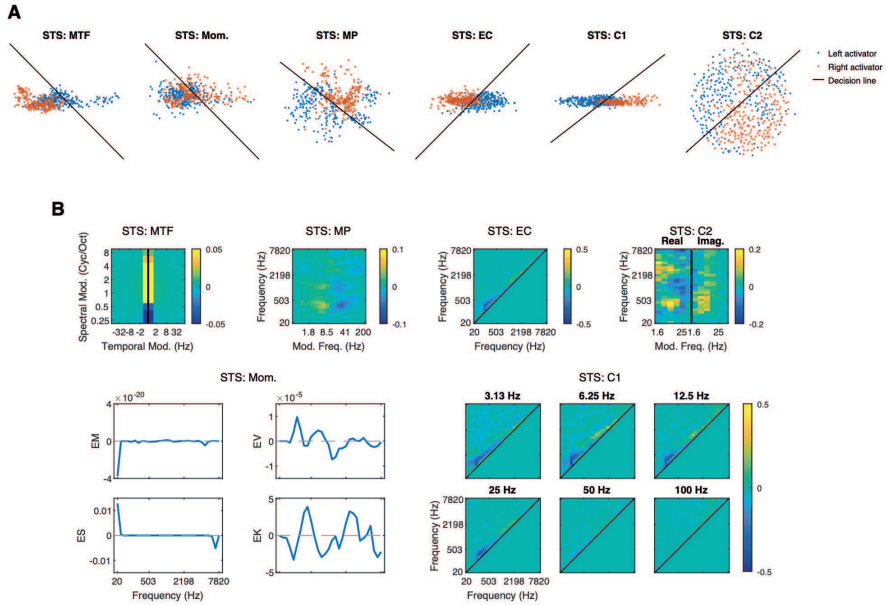


Figure S8. The separability of left activators and right activators in STG. A. Two-dimensional (MDS-based) visualization of the representational similarity of left activators and right activators. For details, see Figure 3B. **B.** The average acoustic difference between left activators and right activators (right activators minus left activators) is shown for each model. For details of each model, see Methods. Results in PP/PT are qualitatively similar to those in HG, except that in MP space, the right activators have stronger low-frequency ($\sim 4\text{--}8\text{ Hz}$) power modulations and weaker high-frequency ($\sim 40\text{ Hz}$) power modulations. In moments space, no clear difference pattern was observed.

Chapter 5

General Discussion

The central aim of this dissertation is to evaluate the feasibility of a feature identification approach in auditory neuroscience and further to assess the perceptual and neural significance of the features identified through this framework. To this end, we conducted three empirical studies designed to measure cortical and behavioral responses. We validated this approach across examined three relevant dimensions of auditory cortical processing—temporal dynamics (Chapter 2), functional spatial patterns (Chapter 3), and anatomical spatial organization (Chapter 4).

This chapter summarizes the key findings and their direct implications, as well as the implications in the broader picture. Additionally, we propose prospective avenues for applying the approach in future research.

1 Summary and Implications of the Findings

1.1 Encoding Dynamic Loudness Patterns

1.1.1 Summary of the Findings

In Chapter 2, we aimed to reconstruct periodic amplitude fluctuations from neural responses and validate their psychophysical relevance. Participants were presented with rhythmic tone sequences (e.g., ascending/descending loudness patterns) while neural activity was recorded via EEG. Using a linear TRF model, we mapped EEG responses onto the envelope of the auditory stimuli. This enabled a backward reconstruction of the original input's amplitude modulation. While prior studies successfully reconstructed speech envelopes and quantified reconstruction accuracy using stimulus-reconstruction correlations, the perceptual recognizability of such reconstructions remained untested. To address this gap, we created two types of mixtures: one combining the original tone sequences with the reconstructed loudness modulations (support mixture), and another combining them with the inverted reconstructions (reduction mixture). These mixtures were then presented directly to participants. We then evaluated whether these reconstructions elicited significant perceptual and neural responses. The behavioral results show that participants judged reconstructions from the support mixture as more perceptually similar to the original sequences than those from a random control mixture, which validates the reconstructions at the behavioral level. Consistently, participants perceived reduction mixture reconstructions as less similar to the original sequences than the control mixture. We found similar outcomes at the neural level: neural responses to the support-mixture reconstructions resembled more closely responses to the original sequences than did responses to the control mixture. Responses to the control mixture, in turn, showed more similarity to original responses than those evoked by the reduction mixture. In sum, these perceptual and neural results demonstrate consistently that the reconstructions influenced cognition and neural processing as expected, confirming their validity and potential utility for neural control.

1.1.2 Implications of the Findings

Our approach advances the evaluation of reconstruction fidelity beyond traditional correlation-based metrics by incorporating direct perceptual judgments and brain-based measurements. The results show that the reconstructed amplitude fluctuations were not merely superficial approximations that correlate with the original stimulus but were sufficiently faithful to modulate both perception and neural responses.

1.2 Encoding of Semantic Sound Features in Regional Activity

1.2.1 Summary of the Findings

In Chapter 3, we shifted focus to two human-unique sound categories: speech and music. Our goal was to identify the intermediate categorical sound features that underlie internal neural representations. Participants listened to a set of naturalistic sounds while we recorded their neural responses via fMRI. Concurrently, we extracted layer-wise activations from a DNN trained to map acoustic inputs to human-defined semantic labels, capturing computationally derived intermediate sound features. Using ridge regression, we mapped these DNN layer features onto fMRI responses within functionally defined ROIs, specifically targeting speech-preferring and music-preferring cortices. Critically, we then sonified the internal neural representation of the sound categories in the fROIs as captured by the DNN using a stimulus synthesis approach. This method generated novel sounds ("speech activators" and "music activators") designed to maximally drive the predicted fMRI responses in their respective target regions, providing an audible instantiation of these abstract neural representations. Following the approach of Chapter 2, we then presented these synthesized sounds to participants to validate their causal role for categorization at the perceptual and neural level. The behavioral results show that participants reliably categorized the speech activators as speech-like and the music activators as music-like, despite their unnatural acoustic properties. At the neural level, the speech-preferring regions were found to respond most strongly to speech activators, while music-preferring regions responded most strongly to music activators. Furthermore, a RSA revealed that these intermediate features partially overlap with known modulation- or texture-based acoustical cues. In sum, the results demonstrate that the sound features embedded in the DNN are causally relevant for categorical sound perception and for cortical category selectivity in the human auditory system. These findings not only validate the feasibility of the forward feature-identification approach in auditory neuroscience but also reveal categorical sound features that extend beyond those identified by conventional methods.

1.2.2 Implications of the Findings

Studies on the functional organization of the human auditory cortex have often relied on contrasting fundamental sound categories, particularly those unique to humans, like speech (Scott et al., 2000; Mesgarani et al., 2014; Norman-Haignere et al., 2015; Overath et al., 2015) and music (Leaver and Rauschecker, 2010; Tierney et al., 2013; Norman-Haignere et al., 2015). Based on observed dissociations in neural responses, several influential studies have proposed

distinct, specialized neural pathways dedicated to processing these categories (Peretz and Coltheart, 2003; McGettigan and Scott, 2012; Norman-Haignere et al., 2015). In Chapter 3, we adopted the term fROI to refer to the auditory cortical stage of these hypothesized category-selective pathways. However, an inherent limitation of the definition of an fROI is imposed by its reliance on human-defined labels. For example, we defined a speech fROI based on the observation that it responded preferentially to sounds that human listeners label as 'speech'. Therefore, many researchers argued that the preferential responses in fact originate from a distinct sensitiveness to acoustical cues (Albouy et al., 2020; Giordano et al., 2023). Despite these arguments, the precise computational role of these non-primary auditory fROIs in transforming raw acoustic inputs into meaningful semantic representations has remained largely unresolved (Giordano et al., 2023).

Our study directly addressed this gap by leveraging DNNs as computational models, which map acoustic signals onto their appropriate human-defined semantic labels (Hershey et al., 2017). We generated synthetic sounds by optimizing inputs to match activation patterns within these fROIs across different DNN layers. Critically, this approach provided a characterization of the abstract intermediate features encoded by the fROIs. We found that sounds synthesized from middle DNN layers consistently elicited both the categorical percepts (e.g., being identified as 'speech' or 'music') and the preferential neural responses associated with the targeted fROIs. This convergence strongly supports the hypothesis, aligned with frameworks like Giordano et al.'s, that these fROIs act as crucial intermediate processing stages situated between low-level acoustic analysis and high-level semantic interpretation (Giordano et al., 2023).

1.3 Asymmetric Encoding of Sound Features

1.3.1 Summary of the Findings

In Chapter 4, we revisited a long-known phenomenon in auditory neuroscience: the functional hemispheric asymmetry in auditory processing. Participants listened to natural sounds while we recorded neural responses using fMRI. We analyzed responses within four anatomically defined ROIs pairs per hemisphere (HG, PP/PT, STG, STS), covering primary and non-primary auditory cortex. To characterize hemispheric specialization for each ROI, we first identified the natural sounds that preferentially activated the left or right hemisphere. We extracted acoustic representations of these sounds using various computational models. We then trained SVMs to classify sounds based on their 'hemispheric preference' (i.e., the specific hemisphere they activated more strongly), assessing classification accuracy via cross-validation. This accuracy quantified how correctly each acoustic model explained the hemispheric specialization. Consistent with previous hypotheses, a spectrotemporal modulation-based model (e.g., MTF) could accurately explain the observed hemispheric differences. Crucially, however, we discovered that models capturing cross-band correlations (specifically envelope correlations and cross-band modulation correlations) provided a significantly better account of these differences. To validate this observation, we synthesized novel sounds using the same DNN as in Chapter 3. We focused on features from the DNN's middle layer (c4), given its high correlation with neural responses. For each ROI, we built ridge regression encoding models to predict neural responses from these DNN features. Using these models, we synthesized two

novel sound sets per ROI: one designed to maximally drive responses in the left hemisphere while suppressing the right, and vice versa. Applying the same SVM classification analysis to these synthesized sounds yielded results consistent with the natural sound analysis: sounds synthesized to activate a given hemisphere could be robustly discriminated from sounds synthesized to activate the opposite hemisphere. Critically, classification based on the cross-band correlation features (e.g., EC/C1) remained highly effective. This convergence between natural and synthesized sounds strongly implicates cross-band correlations as a key acoustic factor underlying hemispheric specialization.

1.3.2 Implications of the Findings

The notion of a functional asymmetry of the two brain hemispheres originated from Broca's observation that left-hemisphere lesions selectively impair speech production (Broca, 1861). This early evidence led researchers to propose that the brain evolved partially specialized pathways for ecologically salient signals, with the left hemisphere supporting speech and the right supporting music (Peretz and Coltheart, 2003; McGettigan and Scott, 2012). In parallel, other researchers sought to uncover the acoustic principles underlying this asymmetry. Three major theoretical accounts have been proposed: the spectrotemporal processing hypothesis (Zatorre et al., 2002), the AST theory (Poeppel, 2003), and the spectrotemporal modulation framework (Flinker et al., 2019). Although these theories have guided research for decades, it remains uncertain whether the frameworks fully capture the dynamic and context-dependent nature of hemispheric lateralization during naturalistic listening.

In Chapter 4, we proposed a novel account grounded in cross-band correlation statistics, as our findings demonstrate that these correlations can explain hemispheric asymmetry more robustly than domain- or modulation-based models. Specifically, the left STG showed preferential encoding of cross-band correlations among neighboring frequency bands reflecting a rough spectral structure, while the right STG showed this among bands that are more distant from each other along the frequency axis. This dissociation was consistent across naturalistic and synthesized sounds, confirming its generality. Notably, we also identified region-specific asymmetry patterns: the STS, unlike the STG, exhibited lateralization for modulation-based features. In sum, these results argue for a re-evaluation of classical theories of functional hemispheric asymmetry and introduce cross-band correlations as an important, abstract acoustic feature.

Having summarized the empirical findings and their immediate implications, we now turn to the broader methodological and conceptual implications of the feature identification framework.

2 The Implications of The Feature Identification Approach

This thesis examined the feasibility of a feature identification approach that leverages forward and backward models to link stimulus features with brain responses. The predictive accuracy of such models—either in estimating neural responses from stimulus features (forward modeling) or in reconstructing stimulus features from neural responses (backward modeling)—

is typically taken as an index of how faithfully those features are represented in the brain. Accordingly, this framework can be used to infer which aspects of the stimulus are encoded in neural activity. In this thesis, we refer to these two model directions as forward synthesis and backward reconstruction, respectively. In the following sections, we discuss the implications of forward synthesis and backward reconstruction separately.

2.1 Forward Synthesis

The forward-synthesis approach begins with a target neural activation pattern and iteratively adjusts an initial sound waveform in the direction that increases the model-predicted similarity to that target pattern. This procedure therefore requires an encoding model capable of predicting neural responses to inputs with very high accuracy. To date, all studies employing forward synthesis have relied on DNN models (Bashivan et al., 2019; Ponce et al., 2019; Ratan Murty et al., 2021; Tuckute et al., 2024), as these models consistently outperform traditional models in their ability to capture neural response patterns. In this thesis, we extend this approach to the auditory domain for the first time. From a theoretical perspective, this enables new insights into the intermediate stages of auditory processing that transform acoustic input into semantic representations. From a methodological perspective, the approach provides a hypothesis-free means of probing neural representations within the feature space of a DNN.

2.1.1 Intermediate Sound Features

The intermediate sound features embedded in DNNs have been proposed to characterize how the human auditory system transforms acoustic waveforms into meaningful semantics, which is one of the least understood aspects of sensory processing (Giordano et al., 2023). A primary constraint in many previous studies is the reliance on relatively small and simplified stimulus sets. These limited sets fail to capture the immense complexity and diversity of natural sounds encountered in the real world, potentially biasing results and obscuring the true functional topography of the cortex. In our work, we investigated the intermediate sound features by extracting them from a pre-trained DNN and sonifying them to make the learned representations perceptually accessible.

A closer look at the layer-specific synthetic sound spectrogram (Chapter 3, Figure 2) suggests that early DNN layers retained the basic spectral and temporal structure akin to peripheral and early cortical auditory encoding, whereas deeper layers emphasized category-relevant, abstract spectrotemporal patterns—paralleling previous fMRI-based results. Conversely, sounds synthesized from late layers appeared noisier and structurally ambiguous, suggesting a potential divergence between the DNN's highest-level representations and the human neural system's semantic encoding mechanisms, a point corroborated by previous findings (Feather et al., 2023). The sonification in our work not only aided interpretability but also provided a bridge between machine-learned features and human perceptual experience, offering a new experimental tool for probing intermediate auditory representations.

2.1.2 Hypothesis-Free Exploration of Sound Features

A fundamental advantage of our approach lies in its ability to investigate auditory representations without being constrained by predetermined feature assumptions. The layer-by-layer abstraction inherent in DNN architectures enables systematic exploration of novel auditory features that conventional analytical models might miss, particularly those involving complex nonlinear interactions or higher-order statistical regularities in natural sounds. This data-driven approach is more efficient than traditional hypothesis-driven approaches that must test individual features sequentially through separate experimental designs (Brown and Cooke, 1994; Patterson, 2000; Chi et al., 2005; McDermott and Simoncelli, 2011).

Auditory processing in humans and animals has been explained primarily by modulation-based frameworks including STRF models. These models typically incorporate biologically inspired but potentially restrictive assumptions that may exclude important aspects of auditory processing, particularly for complex natural sounds. Recent evidence further suggests that low-level acoustic features become increasingly inadequate for explaining neural responses beyond primary auditory areas, where representations appear to reflect more abstract, behaviorally-relevant sound attributes (Norman-Haignere and McDermott, 2018; Giordano et al., 2023). These constraints have substantially limited our comprehension of the human auditory system's computations particularly for natural listening scenarios.

The data-driven features uncovered through our approach create interesting opportunities for designing novel experiments that can deepen our understanding of auditory processing. By utilizing synthetic stimuli optimized to selectively induce target neural response patterns, researchers can more directly probe the causal relationship between specific acoustic features and a given neural activity pattern. This paradigm, where the feature-identification approach directly informs experimental design, represents a significant advance over traditional approaches, as it allows the nervous system itself to reveal the stimulus features it primarily encodes, rather than relying on researchers' intuition about potentially relevant acoustic dimensions (Doerig et al., 2023). The resulting experiments are both more efficient, as they focus on biologically validated features, and more likely to uncover novel principles of auditory processing beyond existing theoretical frameworks. For example, Bashivan et al. achieved this in monkey's visual cortex by synthesizing novel images that maximally drive the response of V4 neurons, which had been originally thought to be tuned to curvatures (Bashivan et al., 2019).

Similarly leveraging this approach, this thesis revealed that the DNN prioritized modulation transfer function (MTF) (Elliott and Theunissen, 2009) or texture-based (McDermott and Simoncelli, 2011) sound features during the synthesis process according to different target neural response patterns (RSA results in Chapter 3 and sound synthesis results in chapter 4). The results reinforced previous hypotheses about the acoustic basis underlying specialization for speech and music (Ding et al., 2017; Albouy et al., 2020; Chang et al., 2024) and provided novel insights into cortical asymmetry.

2.2 Backward Reconstruction: The Assessment of the Neural Fidelity

Neural fidelity to the temporal dynamics of sounds has been extensively studied (Ding and Simon, 2014; Lakatos et al., 2019), typically by measuring reconstruction accuracy—the correlation between stimuli reconstructed from neural signals and the original acoustic stimuli. This approach offers several advantages compared to forward modeling, particularly when analyzing population-level responses with multi-channel techniques such as EEG. First, because stimulus reconstruction projects the neural data back into the stimulus domain, it eliminates the need to preselect response channels. Including all channels is beneficial, as the reconstruction framework naturally down-weights irrelevant ones (Pasley et al., 2012). In contrast, forward models typically treat each channel independently, more suitable for predicting the response of a single region (e.g., the averaged response of a speech- or music-selective cortical region). Second, backward modeling is more sensitive to subtle but important differences between response channels, even when those channels are highly correlated, because the method considers the multivariate structure of the data and can account for inter-channel correlations (Mesgarani et al., 2009).

While this approach has provided foundational insights into how the brain tracks rhythmic patterns (Poeppel and Teng, 2020; Duecker et al., 2024), its reliance on correlation-based metrics imposes significant limitations. Specifically, such methods only assess phase-locking between neural activity and external sounds (Luo and Poeppel, 2007; Peelle et al., 2013), reducing the rich, multidimensional nature of neural processing to a single statistical measure. As a result, critical questions remain unanswered regarding the perceptual relevance and functional consequences of these reconstructed patterns. For example, in Chapter 2, the conventional correlation-based approach enabled us to successfully reconstruct each 1-Hz and 4-Hz components of basic tone sequences. However, it remained unclear whether the perceptual properties of the temporal pattern—such as ascending or descending loudness—were accurately preserved.

Our work (Chapter 2) addressed this gap by assessing the neural fidelity and perceptual relevance of the stimulus reconstructions, thereby moving beyond traditional approaches that treat reconstruction accuracy as an isolated metric. Notably, we found that the reconstructions reproduced both neural and perceptual responses, indicating their high fidelity.

3 Potential Applications of the Feature Identification Approach

The establishment of precise mappings between specific stimulus features—such as spectral or temporal patterns in auditory signals—and their corresponding neural responses, opens new possibilities for developing new non-invasive methods for decoding or modulating auditory brain activity. This approach represents an alternative to conventional non-sensory neural modulation techniques that rely on electric or magnetic stimulation (Paulus, 2014). Transcranial stimulation techniques, such as transcranial magnetic stimulation (TMS) and transcranial electrical stimulation (TES), have been widely used in both research and clinical settings. However, they continue to suffer from limited spatial resolution due to signal dispersion through the skull and surrounding tissues (Paulus, 2014; Kozyrev et al., 2018). Their neuronavigation requires a

model of the conductive properties of the head/brain resembling the signal pathway from stimulating apparatus (e.g., TMS coil or TES electrodes) to ROI. Intracranial techniques, on the other hand offer greater precision, but require invasive surgical implantation with associated risks of tissue damage and long-term complications (Histed et al., 2009; Normann and Fernandez, 2016). The sensory-based technique requires a model of the entire sensory pathway of the brain which is obviously more complex and therefore requires a more appropriate model architecture in the form of a DNN. It provides a complementary technique for non-invasive neural modulation, leveraging the brain's intrinsic sensory processing pathways.

Our results suggest that the feature identification approach might be utilized as a neuromodulation technique for promoting specific target response patterns in the cortex and perception. The advantages of this technique are multifaceted. First, controlled perceptual stimuli can be delivered safely and non-invasively, engaging neural circuits through the brain's natural processing pathways rather than through external "brute-force" perturbation. Second, the resulting neural activity is highly functionally specific—for example, selectively activating speech- or music-responsive fROIs—by leveraging the brain's intrinsic topographic organization. Third, perceptual stimulation is easy and flexible to administer and can be applied in practical settings without the complex equipment required for techniques such as TMS or TES. This simplicity makes the sensory-based approach well suited for both laboratory research and potential clinical deployment. These characteristics make the technique particularly valuable for applications demanding both precision and safety, such as targeted modulation of auditory cortex regions in tinnitus treatment (Eggermont and Roberts, 2004; Peter and Kleinjung, 2019), or rehabilitation of speech processing pathways in aphasia patients (Kendall et al., 2008). Furthermore, the technique is not necessarily limited to sensory stimulation, but could also be applied to electric (auditory nerve) stimulation. For example, auditory prostheses such as cochlear implants could possibly integrate feature-enhancement algorithms that amplify speech-relevant cues to improve intelligibility, or selectively boost music-relevant features to enhance musical perception.

Despite its promise, sensory-based neural control has important constraints: First, response mappings require extensive data from each individual or population, limiting rapid clinical deployment. Second, because sensory-based methods operate through peripheral pathways, they cannot bypass degradation in the cochlea or auditory nerve. While this sensory stimulus-driven control technique represents a meaningful evolution in neural intervention strategies, it cannot replace these electromagnetic methods. Looking ahead, several research directions promise to further improve this technique. These include optimizing feature-response mappings across individuals, developing robust real-time closed-loop implementations, exploring synergistic effects of multimodal stimulus combinations, and investigating long-term plasticity effects induced by repeated stimulus-based modulation.

4 Limitations and Future Directions

While this dissertation has systematically investigated the feasibility of extracting information from neural activities through both temporal and spatial patterns, several limitations should be

acknowledged that point to important directions for future research. A key constraint of our approach was the artificial separation between temporal and spatial pattern analyses. In Chapter 2, we focused exclusively on temporal dynamics using simple pure tone sequences, while in Chapters 3 and 4, we examined spatial patterns with rather short 1-s natural sound segments. This compartmentalized approach, while methodologically necessary, may have overlooked important interactions between temporal and spatial processing in the auditory system.

The stimulus choices in our experiments also imposed ecological limitations. The pure tone sequences used in Chapter 2 represent highly controlled but unnatural auditory stimuli that lack the spectrotemporal complexity of real-world sounds. Conversely, while the natural sounds in later chapters provided more ecological validity, the 1-s duration restriction prevented us from capturing longer-term temporal modulations that are characteristic of continuous speech, music, and environmental sounds. This duration limitation means we may have missed important neural coding principles that emerge only during the processing of more sustained auditory stimulus features.

These methodological constraints suggest several directions for future work. Most importantly, there is a need to bridge the gap between controlled laboratory stimuli and ecologically valid sounds by incorporating more naturalistic auditory materials with longer durations. Future experiments could utilize continuous speech narratives (Ding and Simon, 2012b, 2014; Zion Golumbic et al., 2013), musical pieces, or environmental soundscapes to better capture how the auditory system processes real-world acoustic signals. Additionally, developing analysis methods that can simultaneously track both spatial and temporal patterns would provide a more integrated understanding of neural processing.

A particularly promising direction for future research lies in systematically investigating how attentional states and behavioral task demands influence the neural representations we have identified. While our current findings provide insights into bottom-up auditory processing under passive listening conditions, they necessarily present an incomplete picture of real-world auditory perception, which invariably involves dynamic interactions between sensory input and cognitive processes. Future studies could employ carefully designed active listening paradigms where participants engage in tasks such as auditory stream segregation, target detection in noise, or semantic categorization of speech sounds while neural activity is recorded. Such approaches would allow us to examine how top-down attentional mechanisms selectively enhance or suppress different aspects of the bottom-up feature representations we have characterized.

This line of investigation would be especially valuable for understanding the neural basis of the "cocktail party effect" and other challenging listening situations where listeners must extract meaningful information from complex auditory scenes. By comparing neural responses to identical acoustic stimuli under different attentional sets or task requirements, we could identify how cognitive factors reshape the sensory representation of acoustic features at various stages of the auditory processing hierarchy. For instance, we might discover that attention selectively enhances the neural representation of task-relevant acoustic features while suppressing irrelevant ones (Fritz et al., 2003; Elhilali et al., 2009; Riecke et al., 2016), or that it alters the

temporal dynamics of neural responses to optimize processing of behaviorally significant sounds (Lakatos et al., 2008).

5 Conclusion

By integrating natural sound analysis, DNN-based feature discovery, and rigorous neuroimaging validation, this thesis refines our understanding of auditory cortical function and offers potentially practical tools for neural control. We move from descriptive accounts of cortical responses toward DNN-based models. The identification of cross-band correlations as a key driver of hemispheric specialization, the demonstration that mid-level DNN features carry categorical perceptual weight, and the extensive validation of EEG-based temporal stimulus reconstruction collectively mark a substantive step forward in auditory science—with implications for both theory and application. Future work may build on this framework to develop closed-loop neuromodulation systems or to further unravel the hierarchical transformation of sound into meaning

References

- Abrams DA, Nicol T, Zecker S, Kraus N (2008) Right-Hemisphere Auditory Cortex Is Dominant for Coding Syllable Patterns in Speech. *J Neurosci* 28:3958–3965.
- Albouy P, Benjamin L, Morillon B, Zatorre RJ (2020) Distinct sensitivity to spectrotemporal modulation supports brain asymmetry for speech and melody. *Science* 367:1043–1047.
- Andersson JLR, Skare S, Ashburner J (2003) How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. *NeuroImage* 20:870–888.
- Angrick M, Luo S, Rabbani Q, Candrea DN, Shah S, Milsap GW, Anderson WS, Gordon CR, Rosenblatt KR, Clawson L, Tippet DC, Maragakis N, Tenore FV, Fifer MS, Hermansky H, Ramsey NF, Crone NE (2024) Online speech synthesis using a chronically implanted brain–computer interface in an individual with ALS. *Sci Rep* 14:9617.
- Bartlett AM, Ovaysikia S, Logothetis NK, Hoffman KL (2011) Saccades during Object Viewing Modulate Oscillatory Phase in the Superior Temporal Sulcus. *J Neurosci* 31:18423–18432.
- Bashivan P, Kar K, DiCarlo JJ (2019) Neural population control via deep image synthesis. *Science* 364:eaav9436.
- Belin P, Zatorre RJ (2003) Adaptation to speaker's voice in right anterior temporal lobe. *Neuroreport* 14:2105–2109.
- Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57:289–300.
- Berezutskaya J, Freudenburg ZV, Vansteensel MJ, Aarnoutse EJ, Ramsey NF, van Gerven MAJ (2023) Direct speech reconstruction from sensorimotor brain activity with optimized deep learning models. *J Neural Eng* 20:056010.
- Bialek W, Rieke F, de Ruyter van Steveninck RR, Warland D (1991) Reading a Neural Code. *Science* 252:1854–1857.
- Bizley JK, Cohen YE (2013) The what, where and how of auditory-object perception. *Nat Rev Neurosci* 14:693–707.
- Boemio A, Fromm S, Braun A, Poeppel D (2005) Hierarchical and asymmetric temporal sensitivity in human auditory cortices. *Nat Neurosci* 8:389–395.
- Boynton GM, Engel SA, Glover GH, Heeger DJ (1996) Linear systems analysis of functional magnetic resonance imaging in human V1. *J Neurosci* 16:4207–4221.
- Broca MP (1861) REMARQUES SUR LE SIÈGE DE LA FACULTÉ DU LANGAGE ARTICULÉ,

- SUIVIES D'UNE OBSERVATION D'APHÉMIE (PERTE DE LA PAROLE). *Bull et mémoires de la Société Anatomique de Paris* 36:330–356.
- Brown GJ, Cooke M (1994) Computational auditory scene analysis. *Computer Speech & Language* 8:297–336.
- Brumberg JS, Pitt KM, Burnison JD (2018) A Non-Invasive Brain-Computer Interface for Real-Time Speech Synthesis: The Importance of Multimodal Feedback. *IEEE Trans Neural Syst Rehabil Eng* 26:874–881.
- Chang A, Teng X, Assaneo MF, Poeppel D (2024) The human auditory system uses amplitude modulation to distinguish music from speech Malmierca MS, ed. *PLoS Biol* 22:e3002631.
- Chang JE, Zeng F-G (2012) Tinnitus suppression by electric stimulation of the auditory nerve. *Front Syst Neurosci* 6 Available at: <https://www.frontiersin.org/journals/systems-neuroscience/articles/10.3389/fnsys.2012.00019/full> [Accessed August 11, 2025].
- Chi T, Gao Y, Guyton MC, Ru P, Shamma S (1999) Spectro-temporal modulation transfer functions and speech intelligibility. *The Journal of the Acoustical Society of America* 106:2719–2732.
- Chi T, Ru P, Shamma SA (2005) Multiresolution spectrotemporal analysis of complex sounds. *J Acoust Soc Am* 118:887–906.
- Crosse MJ, Di Liberto GM, Bednar A, Lalor EC (2016) The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli. *Front Hum Neurosci* 10 Available at: <https://www.frontiersin.org/journals/human-neuroscience/articles/10.3389/fnhum.2016.00604/full> [Accessed August 8, 2025].
- Deneux T, Kempf A, Daret A, Ponsot E, Bathellier B (2016) Temporal asymmetries in auditory coding and perception reflect multi-layered nonlinearities. *Nat Commun* 7:12682.
- Ding N, Chatterjee M, Simon JZ (2014) Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. *NeuroImage* 88:41–46.
- Ding N, Patel AD, Chen L, Butler H, Luo C, Poeppel D (2017) Temporal modulations in speech and music. *Neuroscience & Biobehavioral Reviews* 81:181–187.
- Ding N, Simon JZ (2012a) Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences* 109:11854–11859.
- Ding N, Simon JZ (2012b) Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of Neurophysiology* 107:78–89.

- Ding N, Simon JZ (2014) Cortical entrainment to continuous speech: functional roles and interpretations. *Front Hum Neurosci* 8 Available at: <http://journal.frontiersin.org/article/10.3389/fnhum.2014.00311/abstract> [Accessed June 25, 2025].
- Doerig A, Sommers RP, Seeliger K, Richards B, Ismael J, Lindsay GW, Kording KP, Konkle T, van Gerven MAJ, Kriegeskorte N, Kietzmann TC (2023) The neuroconnectionist research programme. *Nat Rev Neurosci* 24:431–450.
- Duecker K, Doelling KB, Breska A, Coffey EBJ, Sivarao DV, Zoefel B (2024) Challenges and Approaches in the Study of Neural Entrainment. *J Neurosci* 44:e1234242024.
- Eggermont JJ, Roberts LE (2004) The neuroscience of tinnitus. *Trends in Neurosciences* 27:676–682.
- Ehrlé N, Samson S, Baulac M (2001) Processing of rapid auditory information in epileptic patients with left temporal lobe damage. *Neuropsychologia* 39:525–531.
- Elhilali M, Chi T, Shamma SA (2003) A spectro-temporal modulation index (STMI) for assessment of speech intelligibility. *Speech Communication* 41:331–348.
- Elhilali M, Xiang J, Shamma SA, Simon JZ (2009) Interaction between Attention and Bottom-Up Saliency Mediates the Representation of Foreground and Background in an Auditory Scene. *PLOS Biology* 7:e1000129.
- Elliott TM, Hamilton LS, Theunissen FE (2013) Acoustic structure of the five perceptual dimensions of timbre in orchestral instrument tones. *J Acoust Soc Am* 133:389–404.
- Elliott TM, Theunissen FE (2009) The Modulation Transfer Function for Speech Intelligibility Friston KJ, ed. *PLoS Comput Biol* 5:e1000302.
- Feather J, Leclerc G, Mađry A, McDermott JH (2023) Model metamers reveal divergent invariances between biological and artificial neural networks. *Nat Neurosci* 26:2017–2034.
- Flinker A, Doyle WK, Mehta AD, Devinsky O, Poeppel D (2019) Spectrotemporal modulation provides a unifying framework for auditory cortical asymmetries. *Nat Hum Behav* 3:393–405.
- Formisano E, Kim D-S, Di Salle F, van de Moortele P-F, Ugurbil K, Goebel R (2003) Mirror-Symmetric Tonotopic Maps in Human Primary Auditory Cortex. *Neuron* 40:859–869.
- Fritz J, Shamma S, Elhilali M, Klein D (2003) Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nature neuroscience* 6:1216–1223.
- Gaver WW (1993) What in the World Do We Hear?: An Ecological Approach to Auditory Event Perception. *Ecological Psychology* 5:1–29.

- Gemmeke JF, Ellis DPW, Freedman D, Jansen A, Lawrence W, Moore RC, Plakal M, Ritter M (2017) Audio Set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 776–780. New Orleans, LA: IEEE. Available at: <http://ieeexplore.ieee.org/document/7952261/> [Accessed April 26, 2024].
- Giordano BL, Esposito M, Valente G, Formisano E (2023) Intermediate acoustic-to-semantic representations link behavioral and neural responses to natural sounds. *Nat Neurosci* 26:664–672.
- Giordano BL, McAdams S, Zatorre RJ, Kriegeskorte N, Belin P (2013) Abstract Encoding of Auditory Objects in Cortical Activity Patterns. *Cerebral Cortex* 23:2025–2037.
- Giraud A-L, Kleinschmidt A, Poeppel D, Lund TE, Frackowiak RSJ, Laufs H (2007) Endogenous cortical rhythms determine cerebral specialization for speech perception and production. *Neuron* 56:1127–1134.
- Giraud A-L, Poeppel D (2012) Cortical oscillations and speech processing: emerging computational principles and operations. *Nat Neurosci* 15:511–517.
- Griffiths TD, Warren JD (2002) The planum temporale as a computational hub. *Trends Neurosci* 25:348–353.
- Guo W, Clause AR, Barth-Maroon A, Polley DB (2017) A Corticothalamic Circuit for Dynamic Switching between Feature Detection and Discrimination. *Neuron* 95:180–194.e5.
- Hamilton LS, Oganian Y, Hall J, Chang EF (2021) Parallel and distributed encoding of speech across human auditory cortex. *Cell* 184:4626–4639.e13.
- Handy TC ed. (2004) *Event-Related Potentials: A Methods Handbook*. Cambridge, MA, USA: MIT Press.
- Herff C, Heger D, de Pestiers A, Telaar D, Brunner P, Schalk G, Schultz T (2015) Brain-to-text: decoding spoken phrases from phone representations in the brain. *Front Neurosci* 8 Available at: <https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2015.00217/full> [Accessed January 9, 2026].
- Herff C, Krusienski DJ, Kubben P (2020) The Potential of Stereotactic-EEG for Brain-Computer Interfaces: Current Progress and Future Directions. *Front Neurosci* 14 Available at: <https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2020.00123/full> [Accessed January 9, 2026].
- Hershey S, Chaudhuri S, Ellis DPW, Gemmeke JF, Jansen A, Moore RC, Plakal M, Platt D, Saurous RA, Seybold B, Slaney M, Weiss RJ, Wilson K (2017) CNN Architectures for Large-Scale Audio Classification. Available at: <http://arxiv.org/abs/1609.09430> [Accessed April 25, 2024].

- Hickok G, Poeppel D (2007) The cortical organization of speech processing. *Nat Rev Neurosci* 8:393–402.
- Histed MH, Bonin V, Reid RC (2009) Direct Activation of Sparse, Distributed Populations of Cortical Neurons by Electrical Microstimulation. *Neuron* 63:508–522.
- Irsik VC, Almanaseer A, Johnsrude IS, Herrmann B (2021) Cortical Responses to the Amplitude Envelopes of Sounds Change with Age. *J Neurosci* 41:5045–5055.
- Ito J, Maldonado P, Singer W, Grün S (2011) Saccade-Related Modulations of Neuronal Excitability Support Synchrony of Visually Elicited Spikes. *Cerebral Cortex* 21:2482–2497.
- Jamison HL, Watkins KE, Bishop DVM, Matthews PM (2006) Hemispheric specialization for processing auditory nonspeech stimuli. *Cereb Cortex* 16:1266–1275.
- Kaas JH, Hackett TA (2000) Subdivisions of auditory cortex and processing streams in primates. *Proc Natl Acad Sci USA* 97:11793–11799.
- Kay KN, Naselaris T, Prenger RJ, Gallant JL (2008) Identifying natural images from human brain activity. *Nature* 452:352–355.
- Kell AJE, Yamins DLK, Shook EN, Norman-Haignere SV, McDermott JH (2018) A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron* 98:630–644.e16.
- Kendall DL, Rosenbek JC, Heilman KM, Conway T, Klenberg K, Gonzalez Rothi LJ, Nadeau SE (2008) Phoneme-based rehabilitation of anomia in aphasia. *Brain and Language* 105:1–17.
- Keshishian M, Akkol S, Herrero J, Bickel S, Mehta AD, Mesgarani N (2023) Joint, distributed and hierarchically organized encoding of linguistic features in the human auditory cortex. *Nat Hum Behav* 7:740–753.
- Kim JJ, Crespo-Facorro B, Andreasen NC, O'Leary DS, Zhang B, Harris G, Magnotta VA (2000) An MRI-based parcellation method for the temporal lobe. *Neuroimage* 11:271–288.
- King AJ, Teki S, Willmore BDB (2018) Recent advances in understanding the auditory cortex. *F1000Res* 7:F1000 Faculty Rev-1555.
- Koelsch S (2011) Toward a neural basis of music perception - a review and updated model. *Front Psychol* 2:110.
- Kozyrev V, Staadt R, Eysel UT, Jancke D (2018) TMS-induced neuronal plasticity enables targeted remodeling of visual cortical maps. *Proc Natl Acad Sci U S A* 115:6476–6481.
- Kriegeskorte N, Mur M, Bandettini P (2008) Representational Similarity Analysis – Connecting the Branches of Systems Neuroscience. *Front Syst Neurosci* 2:4.

- Lakatos P, Chen C-M, O'Connell MN, Mills A, Schroeder CE (2007) Neuronal Oscillations and Multisensory Interaction in Primary Auditory Cortex. *Neuron* 53:279–292.
- Lakatos P, Gross J, Thut G (2019) A New Unifying Account of the Roles of Neuronal Entrainment. *Current Biology* 29:R890–R905.
- Lakatos P, Karmos G, Mehta AD, Ulbert I, Schroeder CE (2008) Entrainment of Neuronal Oscillations as a Mechanism of Attentional Selection. *Science* 320:110–113.
- Lakatos P, Shah AS, Knuth KH, Ulbert I, Karmos G, Schroeder CE (2005) An Oscillatory Hierarchy Controlling Neuronal Excitability and Stimulus Processing in the Auditory Cortex. *Journal of Neurophysiology* 94:1904–1911.
- Leaver AM, Rauschecker JP (2010) Cortical Representation of Natural Complex Sounds: Effects of Acoustic Features and Auditory Object Category. *J Neurosci* 30:7604–7612.
- Lewis JW, Talkington WJ, Puce A, Engel LR, Frum C (2011) Cortical networks representing object categories and high-level attributes of familiar real-world action sounds. *J Cognitive Neuroscience* 23:2079–2101.
- Liégeois-Chauvel C, Peretz I, Babai M, Laguitton V, Chauvel P (1998) Contribution of different cortical areas in the temporal lobes to music processing. *Brain* 121 (Pt 10):1853–1867.
- Luck SJ (2014) *An Introduction to the Event-Related Potential Technique*, 2nd ed. Cambridge, MA, USA: MIT Press.
- Luo H, Poeppel D (2007) Phase Patterns of Neuronal Responses Reliably Discriminate Speech in Human Auditory Cortex. *Neuron* 54:1001–1010.
- Luo S, Rabbani Q, Crone NE (2022) Brain-Computer Interface: Applications to Speech Decoding and Synthesis to Augment Communication. *Neurotherapeutics* 19:263–273.
- Maris E, Oostenveld R (2007) Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods* 164:177–190.
- Marmarelis PVZ (2004) *Nonlinear Dynamic Modeling of Physiological Systems*. John Wiley & Sons.
- McClelland JL, Elman JL (1986) The TRACE model of speech perception. *Cogn Psychol* 18:1–86.
- McDermott JH, Schemitsch M, Simoncelli EP (2013) Summary statistics in auditory perception. *Nat Neurosci* 16:493–498.
- McDermott JH, Simoncelli EP (2011) Sound Texture Perception via Statistics of the Auditory Periphery: Evidence from Sound Synthesis. *Neuron* 71:926–940.
- McGettigan C, Scott SK (2012) Cortical asymmetries in speech perception: what's wrong,

- what's right and what's left? *Trends in Cognitive Sciences* 16:269–276.
- Mesgarani N, Cheung C, Johnson K, Chang EF (2014) Phonetic Feature Encoding in Human Superior Temporal Gyrus. *Science* 343:1006–1010.
- Mesgarani N, David SV, Fritz JB, Shamma SA (2009) Influence of Context and Behavior on Stimulus Reconstruction From Neural Activity in Primary Auditory Cortex. *J Neurophysiol* 102:3329–3339.
- Miyawaki Y, Uchida H, Yamashita O, Sato M, Morito Y, Tanabe HC, Sadato N, Kamitani Y (2008) Visual Image Reconstruction from Human Brain Activity using a Combination of Multiscale Local Image Decoders. *Neuron* 60:915–929.
- Moerel M, De Martino F, Formisano E (2014) An anatomical and functional topography of human auditory cortical areas. *Front Neurosci* 8 Available at: <https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2014.00225/full> [Accessed August 3, 2025].
- Morillon B, Lehongre K, Frackowiak RSJ, Ducorps A, Kleinschmidt A, Poeppel D, Giraud A-L (2010) Neurophysiological origin of human brain asymmetry for speech and language. *Proc Natl Acad Sci USA* 107:18688–18693.
- Naselaris T, Prenger RJ, Kay KN, Oliver M, Gallant JL (2009) Bayesian Reconstruction of Natural Images from Human Brain Activity. *Neuron* 63:902–915.
- Neuhoff JG (1999) Perception of changes in loudness. *Nature* 398:673–674.
- Norman-Haignere S, Kanwisher NG, McDermott JH (2015) Distinct Cortical Pathways for Music and Speech Revealed by Hypothesis-Free Voxel Decomposition. *Neuron* 88:1281–1296.
- Norman-Haignere SV, Feather J, Boebinger D, Brunner P, Ritaccio A, McDermott JH, Schalk G, Kanwisher N (2022) A neural population selective for song in human auditory cortex. *Current Biology* 32:1470–1484.e12.
- Norman-Haignere SV, McDermott JH (2018) Neural responses to natural and model-matched stimuli reveal distinct computations in primary and nonprimary auditory cortex. *PLOS Biology* 16:e2005127.
- Normann RA, Fernandez E (2016) Clinical applications of penetrating neural interfaces and Utah Electrode Array technologies. *J Neural Eng* 13:061003.
- Obleser J, Eisner F, Kotz SA (2008) Bilateral Speech Comprehension Reflects Differential Sensitivity to Spectral and Temporal Features. *J Neurosci* 28:8116–8123.
- Overath T, McDermott JH, Zarate JM, Poeppel D (2015) The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nat Neurosci*

18:903–911.

- Pasley BN, David SV, Mesgarani N, Flinker A, Shamma SA, Crone NE, Knight RT, Chang EF (2012) Reconstructing Speech from Human Auditory Cortex. *PLOS Biology* 10:e1001251.
- Patel AD (2008) Music, language, and the brain. New York, NY, US: Oxford University Press.
- Patel AD, Peretz I, Tramo M, Labreque R (1998) Processing prosodic and musical patterns: a neuropsychological investigation. *Brain Lang* 61:123–144.
- Patterson RD (2000) Auditory images. How complex sounds are represented in the auditory system. *J Acoust Soc Jpn (E), J Acoust Soc Jpn E* 21:183–190.
- Paulus W (2014) Transcranial brain stimulation: potential and limitations. *e-Neuroforum* 5:29–36.
- Peelle JE, Gross J, Davis MH (2013) Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cereb Cortex* 23:1378–1387.
- Penhune VB, Zatorre RJ, Feindel WH (1999) The role of auditory cortex in retention of rhythmic patterns as studied in patients with temporal lobe removals including Heschl's gyrus. *Neuropsychologia* 37:315–331.
- Peretz I, Coltheart M (2003) Modularity of music processing. *Nat Neurosci* 6:688–691.
- Peter N, Kleinjung T (2019) Neuromodulation for tinnitus treatment: an overview of invasive and non-invasive techniques. *J Zhejiang Univ Sci B* 20:116–130.
- Pickles J (2013) An Introduction to the Physiology of Hearing. In: Brill. Available at: <https://brill.com/display/title/24209> [Accessed August 3, 2025].
- Poeppel D (2003) The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time.' *Speech Communication* 41:245–255.
- Poeppel D, Teng X (2020) Entrainment in Human Auditory Cortex: Mechanism and Functions. In: *The Senses: A Comprehensive Reference*, pp 63–76. Elsevier. Available at: <https://linkinghub.elsevier.com/retrieve/pii/B978012805408600018X> [Accessed July 3, 2025].
- Ponce CR, Xiao W, Schade PF, Hartmann TS, Kreiman G, Livingstone MS (2019) Evolving Images for Visual Neurons Using a Deep Generative Network Reveals Coding Principles and Neuronal Preferences. *Cell* 177:999–1009.e10.
- Ratan Murty NA, Bashivan P, Abate A, DiCarlo JJ, Kanwisher N (2021) Computational models of category-selective brain regions enable high-throughput tests of selectivity. *Nat Commun* 12:5540.

- Rauschecker JP, Scott SK (2009) Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nat Neurosci* 12:718–724.
- Rauschecker JP, Tian B (2000) Mechanisms and streams for processing of “what” and “where” in auditory cortex. *Proceedings of the National Academy of Sciences* 97:11800–11806.
- Riecke L, Formisano E, Sorger B, Başkent D, Gaudrain E (2018) Neural Entrainment to Speech Modulates Speech Intelligibility. *Current Biology* 28:161–169.e5.
- Riecke L, Peters JC, Valente G, Kemper VG, Formisano E, Sorger B (2016) Frequency-Selective Attention in Auditory Scenes Recruits Frequency Representations Throughout Human Superior Temporal Cortex. *Cereb Cortex*:bhw160.
- Ringach D, Shapley R (2004) Reverse correlation in neurophysiology. *Cognitive Science* 28:147–166.
- Robin DA, Tranel D, Damasio H (1990) Auditory perception of temporal and spectral events in patients with focal left and right cerebral lesions. *Brain Lang* 39:539–555.
- Samson S, Zatorre RJ (1994) Contribution of the right temporal lobe to musical timbre discrimination. *Neuropsychologia* 32:231–240.
- Sankaran N, Leonard MK, Theunissen F, Chang EF (2024) Encoding of melody in the human auditory cortex. *Science Advances*.
- Santoro R, Moerel M, De Martino F, Valente G, Ugurbil K, Yacoub E, Formisano E (2017) Reconstructing the spectrotemporal modulations of real-life sounds from fMRI response patterns. *Proceedings of the National Academy of Sciences* 114:4799–4804.
- Santoro R, Moerel M, Martino FD, Goebel R, Ugurbil K, Yacoub E, Formisano E (2014) Encoding of Natural Sounds at Multiple Spectral and Temporal Resolutions in the Human Auditory Cortex. *PLOS Computational Biology* 10:e1003412.
- Scott SK, Blank CC, Rosen S, Wise RJS (2000) Identification of a pathway for intelligible speech in the left temporal lobe. *Brain* 123:2400–2406.
- Shamma S (2001) On the role of space and time in auditory processing. *Trends in Cognitive Sciences* 5:340–348.
- Shannon RV, Zeng F-G, Kamath V, Wygonski J, Ekelid M (1995) Speech Recognition with Primarily Temporal Cues. *Science* 270:303–304.
- Staeren N, Renvall H, De Martino F, Goebel R, Formisano E (2009) Sound Categories Are Represented as Distributed Patterns in the Human Auditory Cortex. *Current Biology* 19:498–502.
- Su K, Tian L (2025) Systematic review: progress in EEG-based speech imagery brain-computer interface decoding and encoding research. *PeerJ Comput Sci* 11:e2938.

- Talavage TM, Ledden PJ, Benson RR, Rosen BR, Melcher JR (2000) Frequency-dependent responses exhibited by multiple regions in human auditory cortex. *Hear Res* 150:225–244.
- Telkemeyer S, Rossi S, Koch SP, Nierhaus T, Steinbrink J, Poeppel D, Obrig H, Wartenburger I (2009) Sensitivity of Newborn Auditory Cortex to the Temporal Structure of Sounds. *J Neurosci* 29:14726–14733.
- Theunissen FE, Elie JE (2014) Neural processing of natural sounds. *Nat Rev Neurosci* 15:355–366.
- Tierney A, Dick F, Deutsch D, Sereno M (2013) Speech versus Song: Multiple Pitch-Sensitive Areas Revealed by a Naturally Occurring Musical Illusion. *Cereb Cortex* 23:249–254.
- Tuckute G, Feather J, Boebinger D, McDermott JH (2023) Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions Poeppel D, ed. *PLoS Biol* 21:e3002366.
- Tuckute G, Sathe A, Srikant S, Taliaferro M, Wang M, Schrimpf M, Kay K, Fedorenko E (2024) Driving and suppressing the human language network using large language models. *Nat Hum Behav* 8:544–561.
- Vos PG, Troost JM (1989) Ascending and Descending Melodic Intervals: Statistical Findings and Their Perceptual Relevance. *Music Perception: An Interdisciplinary Journal* 6:383–396.
- Wang Y, Zhang J, Zou J, Luo H, Ding N (2019) Prior Knowledge Guides Speech Segregation in Human Auditory Cortex. *Cerebral Cortex* 29:1561–1571.
- Warren RM, Gardner DA, Brubaker BS, Bashford JA (1991) Melodic and Nonmelodic Sequences of Tones: Effects of Duration on Perception. *Music Perception: An Interdisciplinary Journal* 8:277–289.
- Wernicke C (1874) Der aphasische Symptomencomplex; eine psychologische Studie auf anatomischer Basis (Cohn & Weigert, Breslau, 1874). Available at: https://www.deutschestextarchiv.de/book/view/wernicke_symptomencomplex_1874?p=5 [Accessed August 1, 2025].
- Winer JA, Schreiner CE eds. (2011) *The Auditory Cortex*. Boston, MA: Springer US. Available at: <https://link.springer.com/10.1007/978-1-4419-0074-6> [Accessed August 3, 2025].
- Woods DL, Herron TJ, Cate AD, Yund EW, Et. A (2010) Functional Properties of Human Auditory Cortical Fields. *Frontiers in Systems Neuroscience* Available at: <https://scihub.su/10.3389/fnsys.2010.00155> [Accessed March 12, 2026].
- Wu MC-K, David SV, Gallant JL (2006) Complete functional characterization of sensory neurons by system identification. *Annu Rev Neurosci* 29:477–505.

- Xing L, Formisano E, Riecke L (2025) Deep Sound Synthesis Reveals Novel Category-Defining Sound Features in the Human Auditory Cortex. :2025.02.14.638230 Available at: <https://www.biorxiv.org/content/10.1101/2025.02.14.638230v1> [Accessed August 9, 2025].
- Zatorre RJ (2001) Spectral and Temporal Processing in Human Auditory Cortex. *Cerebral Cortex* 11:946–953.
- Zatorre RJ, Belin P, Penhune VB (2002) Structure and function of auditory cortex: music and speech. *Trends in Cognitive Sciences* 6:37–46.
- Zion Golumbic EM, Ding N, Bickel S, Lakatos P, Schevon CA, McKhann GM, Goodman RR, Emerson R, Mehta AD, Simon JZ, Poeppel D, Schroeder CE (2013) Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party.” *Neuron* 77:980–991.

Appendix

Impact Paragraph

The work presented in this thesis contributes to auditory neuroscience by introducing and evaluating a feature identification framework for relating acoustic structure to neural activity. The three empirical studies addressed distinct aspects of auditory processing, but together outline a coherent approach for characterizing auditory representations. The impact of these findings is primarily of theoretical nature, influencing how auditory features can be discovered, interpreted, and experimentally manipulated. In addition, the framework offers potential future relevance for neuromodulation, although these envisioned applications remain preliminary.

1. Scientific Impact

The EEG study demonstrates that cortical temporal response patterns can be reconstructed from scalp recordings with high fidelity. The sonification approach could be generalized to other sensory modalities. Researchers can design visual or tactile stimuli that preferentially induce specific oscillatory or temporal neural patterns. This capability provides a new avenue for testing hypotheses about temporal coding.

The two fMRI studies show that applying the feature identification framework can unravel novel features accounting for speech- and music-preferring fROIs and auditory cortical asymmetry. For the scientific community, these insights refine our knowledge about the functional organization of human auditory cortex. The first fMRI study opens opportunities for future experiments to further investigate how exactly the features we identified contribute to speech or music processing. The second fMRI study suggests other researchers can revise/extend theories of brain asymmetry based on our findings. Collectively, future work can try to isolate the novel features uncovered in this thesis and present them as controlled stimuli to investigate their perceptual or learning-related effects.

2. Methodological Impact

Across all three studies, the feature identification framework provides a general strategy for identifying the acoustic features that drive perceptual and neural responses. This approach is one step forward from regression-based modeling which can either forward predict neural responses or backward reconstruct the acoustic input. Future research can extend this strategy to identify additional feature dimensions and evaluate their biological significance. For instance, forward synthesis may be leveraged to modulate neural responses with greater spatial specificity—for example, targeting voxel-wise response patterns rather than overall ROI-wise activity. Meanwhile, backward reconstruction may enable the recovery of stimulus features from internally generated neural activity, such as during imagery (e.g., imagined speech).

Furthermore, instead of testing one candidate feature at a time, the DNN-based approach allows hypothesis-free exploration of the rich feature spaces that the DNN has acquired during its training. This opens the door for future work to examine a variety of neural network architectures, training objectives, and datasets, each of which may capture distinct and complementary aspects of sound processing.

3. Broader Applications: Toward Neuromodulation

The feature identification framework opens promising directions for sensory-based neuromodulation. Because the approach allows to identify features that selectively activate

specific auditory regions, it further enables feature-targeted sensory stimulation. While we succeeded in selectively activating speech- and music-preferring cortical regions, we failed to drive these regions to supernatural activation levels. The latter might be useful for a more powerful neural modulation. Future work needs to further fine-tune our approach to further boost the activation level. For example, invasive electrophysiologic recordings with more detailed spatiotemporal neural response patterns could potentially reveal additional features of the internal representation that contribute to cortical sound categorization. Furthermore, signal processing in auditory prostheses (e.g., cochlear implants) could be fine-tuned to enhance the DNN-derived features, as this could improve speech comprehension or music appreciation in the prosthesis user.

Summary

This doctoral thesis develops and validates a feature identification approach for extracting sound features from neural activity, with the goal of enabling the modulation of cortical and behavioral responses toward designed patterns. By combining high temporal resolution electroencephalography (EEG) and high spatial resolution functional magnetic resonance imaging (fMRI), this research establishes a methodological framework that integrates backward reconstruction and forward synthesis to link neural responses to acoustic stimulus features across temporal dynamics, semantic categories, and hemispheric asymmetries.

In this dissertation, we conducted three empirical studies to test the feasibility and generalizability of this feature identification approach across different dimensions of auditory processing. In Chapter 2, we applied a backward reconstruction method using a linear temporal response function (TRF) to reconstruct dynamic loudness patterns from EEG responses to basic tone sequences. We then developed a behavioral and neural validation paradigm to assess whether these reconstructions—originally evaluated only by correlation metrics—hold perceptual and neural significance for listeners. This established a more rigorous framework for evaluating reconstruction fidelity beyond conventional acoustic-domain metrics. In Chapter 3, we implemented a forward synthesis method using a deep neural network (DNN) to sonify layer-wise categorical sound representations embedded in the model. We then designed perceptual and neural validation experiments to test whether the synthesized sounds—though acoustically unnatural—could serve as effective probes for investigating causal relationships between DNN-derived features and auditory cortical processing. In Chapter 4, we extended this forward synthesis framework to target hemispheric differences, synthesizing sounds that maximized the predicted response difference between the left and right auditory cortices. Subsequently, we developed an acoustic feature analysis pipeline to characterize the distinguishing features of these left- and right-activating sounds using multiple acoustic models (e.g., spectrotemporal modulation and sound texture statistics). The methodological contributions and generalizability of this feature identification approach are discussed in the final chapter.

Samenvatting

Dit proefschrift ontwikkelt en valideert een benadering voor kenmerkidentificatie voor het extraheren van geluidskenmerken uit neurale activiteit, met als doel de modulatie van corticale en gedragsresponsen naar ontworpen patronen mogelijk te maken. Door het combineren van elektro-encefalografie (EEG) met hoge temporele resolutie en functionele magnetische resonantiebeeldvorming (fMRI) met hoge ruimtelijke resolutie, wordt in dit onderzoek een methodologisch raamwerk gevestigd dat achterwaartse reconstructie en voorwaartse synthese integreert om neurale responsen te koppelen aan akoestische stimuluskenmerken op het gebied van temporele dynamiek, semantische categorieën en hemispherische asymmetrieën.

In dit proefschrift hebben we drie empirische studies uitgevoerd om de haalbaarheid en generaliseerbaarheid van deze benadering voor kenmerkidentificatie te toetsen over verschillende dimensies van auditieve verwerking. In Hoofdstuk 2 pasten we een methode voor achterwaartse reconstructie toe, gebruikmakend van een lineaire temporele responsfunctie (TRF), om dynamische luidheidspatronen te reconstrueren uit EEG-responsen op basische toonreeksen. Vervolgens ontwikkelden we een gedrags- en neurale validatieparadigma om te beoordelen of deze reconstructies – die oorspronkelijk alleen werden geëvalueerd met correlatiemetingen – perceptuele en neurale betekenis hebben voor luisteraars. Dit vestigde een meer rigoreus raamwerk voor het evalueren van reconstructiegetrouwheid, verdergaand dan conventionele akoestisch-domein metrieken. In Hoofdstuk 3 implementeerden we een methode voor voorwaartse synthese met behulp van een diep neurale netwerk (DNN) om laagsgewijze categorische geluidsrepresentaties, ingebed in het model, te verklankijken. Vervolgens ontwierpen we perceptuele en neurale validatie-experimenten om te toetsen of de gesynthetiseerde geluiden – hoewel akoestisch onnatuurlijk – konden dienen als effectieve probes voor het onderzoeken van causale relaties tussen DNN-afgeleide kenmerken en auditieve corticale verwerking. In Hoofdstuk 4 breidden we dit raamwerk voor voorwaartse synthese uit om hemispherische verschillen te targeten, waarbij we geluiden synthetiseerden die het voorspelde responsverschil tussen de linker- en rechterauditieve schors maximaliseerden. Vervolgens ontwikkelden we een analysepijplijn voor akoestische kenmerken om de onderscheidende eigenschappen van deze links- en rechtsactiverende geluiden te karakteriseren met behulp van meerdere akoestische modellen (bijv. spectrotemporele modulatie en geluidstextuurstatistieken). De methodologische bijdragen en de generaliseerbaarheid van deze benadering voor kenmerkidentificatie worden besproken in het slothoofdstuk.

Acknowledgement

The completion of this PhD marks the culmination of several years of intense learning, exploration, and perseverance. This journey has been far more than a research endeavor; it has been a transformative chapter of my life. Along the way, I have been supported, guided, challenged, and encouraged by many people, each of whom has contributed in their own unique way to this work.

First and foremost, I would like to express my deepest gratitude to my supervisor, Lars Riecke. I still remember our very first online meeting, when I nervously introduced myself and my research experience. Your warm and positive response immediately put me at ease and revealed how approachable you truly are. Your intellectual guidance, high standards, and genuine dedication to mentoring have been fundamental to my development as a researcher. You created an environment in which I always felt comfortable discussing ideas, asking questions, and sharing concerns. Your continuous encouragement and constructive feedback strengthened my confidence and supported my growth both scientifically and personally. I am especially grateful for the thoughtful balance of independence and guidance you provided throughout this journey. Beyond research, your impact on my ability to communicate has been invaluable. Your patience, attentive listening, and exceptional writing skills have helped me find my voice and express myself as a true scientist.

I would also like to express my heartfelt gratitude to Elia Formisano. Your profound intelligence, sharp insights, and extensive expertise have been a constant source of inspiration. Your ability to provide clear, thoughtful, and highly insightful feedback has shaped the direction of my project in meaningful ways. I still vividly remember the day I presented our deep synthesized sounds to you—you showed great interest and asked how I had achieved it. That moment was incredibly encouraging and has stayed with me ever since. Your encouragement throughout this journey has meant a great deal, and your vision and guidance were instrumental in shaping the most important study of my PhD. I am especially grateful for the trust you have placed in me. Your belief in my potential gave me the confidence to persevere through the most challenging moments. For your trust, your inspiration, and your leadership, I am deeply thankful.

I would like to express my sincere appreciation to my thesis committee for the time and effort they devoted to carefully reviewing my thesis. Your thoughtful feedback, insightful comments, and constructive critiques have significantly improved the quality of this work and provided valuable guidance during the final stage of my PhD. I am truly grateful for the expertise, care, and attention you invested in this process. I would also like to gratefully acknowledge the China Scholarship Council (CSC 202109110036) for their support, which made my doctoral studies abroad possible.

I would also like to thank all members of the auditory group for creating such a supportive and inspiring research environment. I am especially grateful to Lars H and Fren for their valuable input and helpful suggestions on my projects. Federico, thank you for your guidance and support during the early stages of setting up my fMRI study. I would also like to thank Giancarlo for sharing his expertise in advanced statistical methods, which has been highly valuable for my work. It was a great pleasure to share an office with Michele and Maria—your presence

made my PhD journey both enjoyable and memorable. Many thanks to Lonike for organizing MBIC and CN activities, which contributed greatly to the group spirit. Finally, I would like to thank Isma, Agustin, Miriam, and Gijs for fostering such a friendly, collaborative, and supportive atmosphere within the auditory group.

Regarding my colleagues at CN, there are too many to mention individually. I have shared many memorable moments with all of you, including the CN Christmas parties, lunches, and retreats, which made my time in the group truly enjoyable. I would also like to sincerely thank our secretarial team, Christl, Eva, and Andra. Your constant support has been invaluable and has greatly facilitated my work throughout this journey.

I would also like to express my sincere gratitude to Bettina for her invaluable support during the initial stages of setting up my fMRI study. Your guidance made this challenging process much more manageable. Bettina, I am also deeply thankful for the lovely dinner you prepared at your and Lars's home. I truly appreciated your warm hospitality and kindness.

I would also like to express my sincere gratitude to Prof. Du and Dr. Ding. Your invaluable guidance in experimental design and manuscript preparation has contributed greatly to this work. I am truly thankful for your patience, expertise, and continued encouragement, which helped me overcome many challenges along the way. This work would not have been possible without your generous support.

Beyond academia, I have been incredibly fortunate to be surrounded by friends who have supported me unconditionally. Thank you for listening when I needed to vent, celebrating the small victories, and reminding me to find joy beyond research. Whether through a coffee break, a shared meal, or a simple message, your presence has helped me maintain balance and perspective.

最后，我要衷心感谢我的家人。感谢我的父母，谢谢你们无条件的爱、一如既往的支持和始终如一的鼓励。这些年来，你们始终相信我，并给予我追求梦想的自由。感谢我的挚爱，谢谢你在最艰难的时光里给予的耐心、理解和陪伴。你的存在是我坚持下去的重要力量。感谢所有家人，你们的关怀与支持是我一路走来最温暖的依靠。

Curriculum Vitae

Lidongsheng Xing (邢立冬生) was born on 22nd November, 1990 in Tai'an, Shandong Province, China. He completed his secondary education in his hometown before pursuing higher education in the field of auditory science.

In 2009, he enrolled at Zhejiang Chinese Medical University, where he pursued his undergraduate studies in Audiology. During this period, he developed a solid foundation in hearing science, speech perception, and clinical audiology, which sparked his growing interest in the cognitive and neural mechanisms underlying auditory processing. He obtained his Bachelor's degree in Audiology upon graduation.

Motivated by an increasing interest in the intersection between auditory perception and brain function, Lidongsheng Xing continued his academic training at the Institute of Psychology, Chinese Academy of Sciences. Under the supervision of Prof. Dr. Yi Du, he conducted research on how musical experience influences speech processing. His Master's work focused on the interaction between music and language, particularly how long-term musical training modulates speech perception and cognitive processing, laying the groundwork for his future research in auditory cognitive neuroscience.

With the support of the Chinese Scholarship Council (CSC), Lidongsheng Xing is currently pursuing a PhD in Cognitive Neuroscience at the Faculty of Psychology and Neuroscience, Maastricht University. During his doctoral studies, he works under the supervision of Dr. Lars Riecke and Prof. Dr. Elia Formisano. His current research focuses on auditory cognition, with particular emphasis on the neural modulation with auditory stimulation.

Throughout his academic journey, Lidongsheng Xing has developed a strong interdisciplinary background bridging audiology, psychology, and neuroscience. His research interests center on understanding how the human brain processes complex auditory information, especially in the domains of speech and music.

Publications

- Zhang, L., Fu, X., Luo, D., **Xing, L.**, & Du, Y. (2021). Musical Experience Offsets Age- Related Decline in Understanding Speech-in-Noise: Type of Training Does Not Matter, Working Memory Is the Key. *Ear and hearing*, 42(2), 258–270.
- Xing, L.**, Formisano, E., & Riecke, L. (2026). Deep Sound Synthesis Matched to Brain Activity Recapitulates Preferential Responses to Speech and Music. *The Journal of neuroscience*, 46(9), e1651252025. <https://doi.org/10.1523/JNEUROSCI.1651-25.2025>
- Xing, L.**, Riecke, L. (submitted). Neural and auditory validity of basic temporal stimulus patterns reconstructed from individual EEG recordings with a linear model.
- Xing, L.**, Formisano, E., Riecke, L. (in preparation). Deep sound synthesis reveals sound features underlying auditory cortical asymmetry.

2026

