Neural Encoding of Acoustic Features Across Speech and Music in the Human Brain

The University of Texas at Austin
Speech, Language, and Hearing Sciences
Moody College of Communication

Rajvi Agravat[1,2], Maansi Desai[2], Gabrielle Foox[3], Alyssa Field[2], Anne Anderson[3], Dave Clarke[4], Elizabeth Tyler-Kabara[4], Howard Weiner[3], Liberty Hamilton[1,4]

[1]UT Austin Institute for Neuroscience, [2]UT Austin Moody College of Communication, [3]Texas Children's Hospital/Baylor College of Medicine, [4]Dell Children's Medical Center/Dell Medical School

Introduction

In everyday life, our brains constantly separate overlapping sounds, such as speech and music, to focus on important information—a process called **auditory streaming**. This is essential for understanding conversations in noisy environments and distinguishing between different sounds.

To investigate how the brain performs this, we presented **movie trailers** (with both speech and music) as stimuli. After collecting **intracranial EEG** (sEEG) data from epilepsy patients, we used **Moises**, a deep neural network (DNN), to split the original audio into **speech-only** and **music-only** tracks, mimicking the brain's ability to separate complex sounds into individual sources.

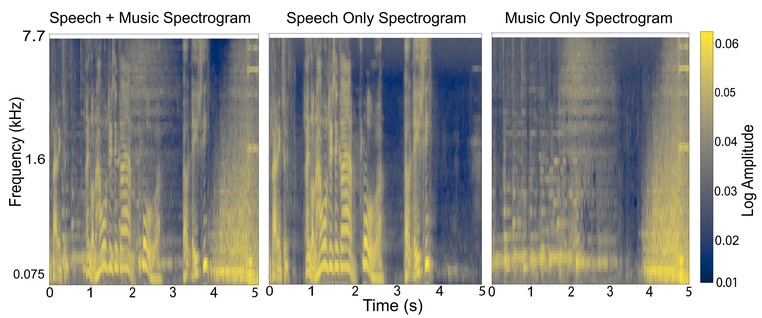
This study aims to reveal **how different brain regions encode speech and music using separated auditory streams.**

sEEG Data Collection & Methods

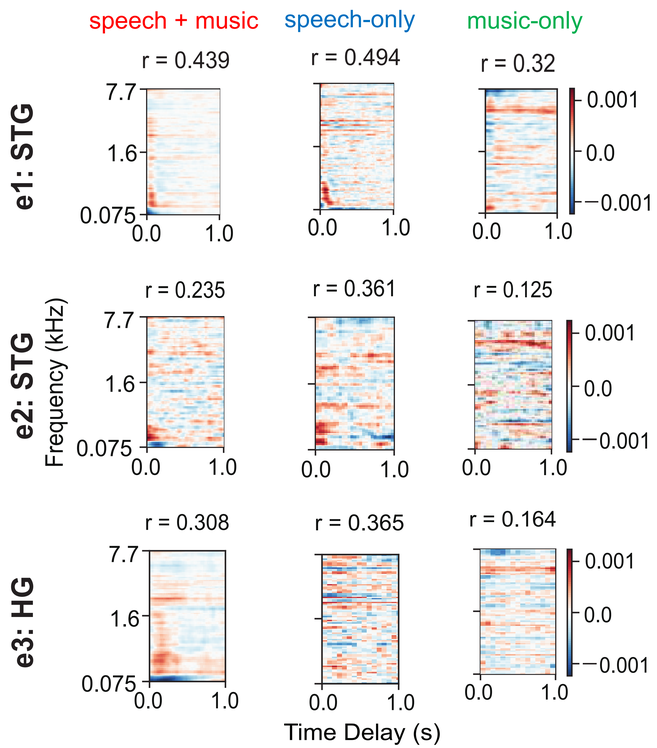
- Recorded brain activity from 26 intractable epilepsy patients (ages 4 to 21, 16M/10F) while they **listened and watched movie trailers** as stimuli.
- Measured **local field potentials** from the high-gamma band (70-150 Hz) using depth electrodes in bilateral auditory-related brain regions.
- Fit **Spectro-Temporal Receptive Field (STRF)** models to the original, speech-only, and music-only stimuli to explore how different brain regions encode these separated streams.

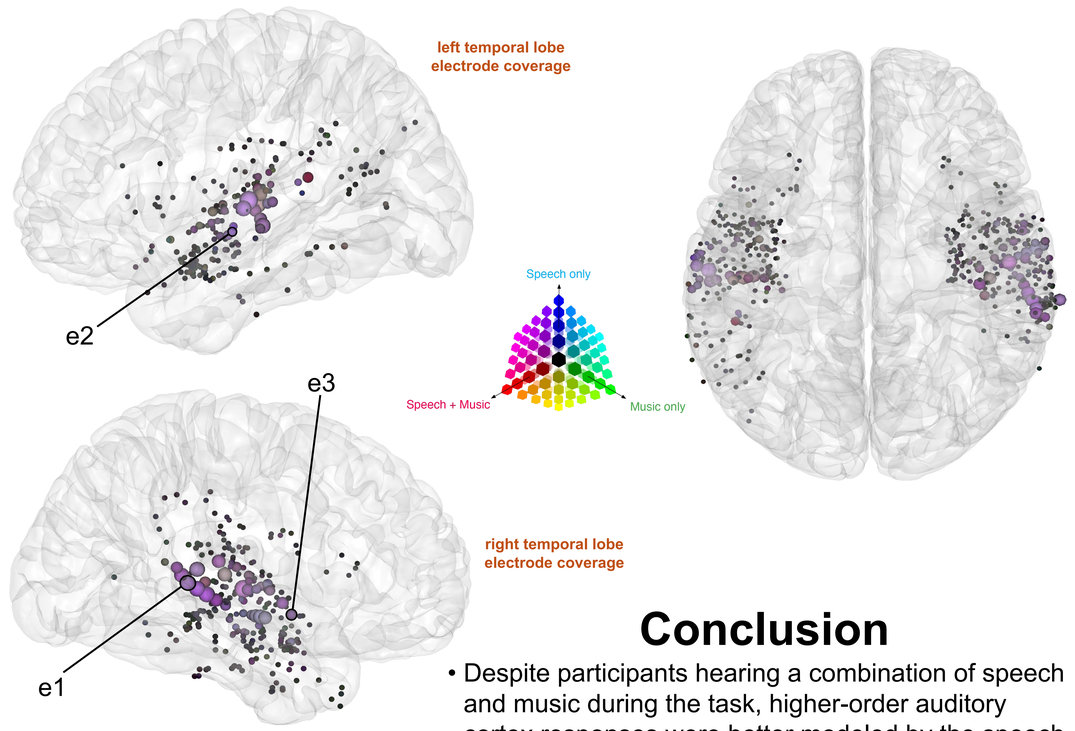
$$sEEG(t,n) = \sum_f \sum_\tau w(f,\tau,n)\, s(f,t-\tau) + \varepsilon(t,n)$$

predicted electrode activity = receptive field × stimulus representation + residuals

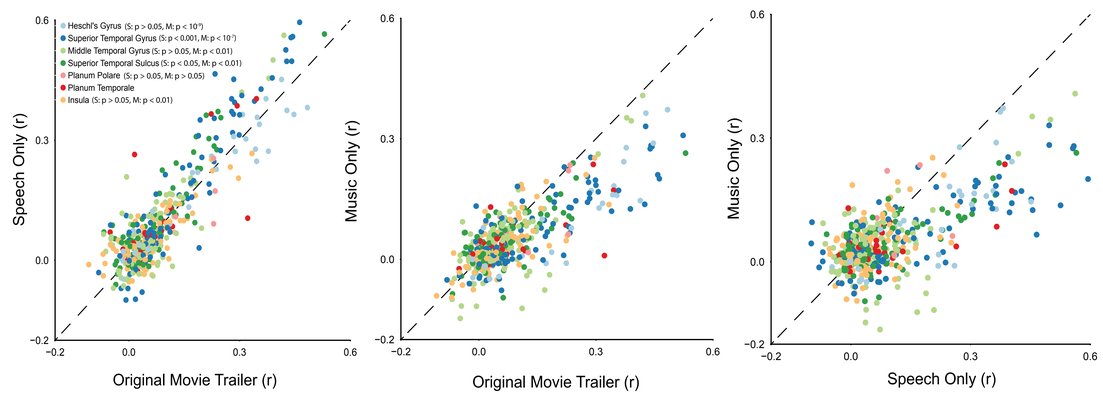


Higher Order Auditory Cortex Encodes Speech Over Combined Audio Stimuli





Even though participants listened to mixtures of speech and music, models using the speech spectrograms from the DNN-separated audio performed better at predicting brain activity compared to using the original combined audio or music-only spectrograms.



Conclusion

- Despite participants hearing a combination of speech and music during the task, higher-order auditory cortex responses were better modeled by the speech-separated audio, indicating **preferential encoding of speech** over music in these regions.
- Neural responses show **strong correlations in STG and HG**, both known for speech sound encoding.
- The variability in music encoding across different regions suggests that **music may engage more diverse neural networks** compared to speech, which shows a more localized and robust representation.

References & Funding

1. Desai, M et al. (2021) J Neuro. 41(43): 8946-8962
2. Holdgraf et al. (2017) Frontiers in Systems Neuroscience 11,61.
3. The Musician's App. https://moises.ai/

This study is funded under a grant by the National Institutes of Health to LSH (1R01DC018579, NIDCD)