DOI: 10.3217/978-3-99161-050-2-238

Brain-Informed Auditory Scene Understanding: A Listener-Aware Auditory Foundation Model for Personalized Speech Processing

Xilin Jiang[1,2], Sukru Samet Dindar[1,2], Vishal Choudhari[1,2], Stephan Bickel[3,4], Ashesh D. Mehta[3,4], Catherine Schevon[5], Guy M. McKhann[5], Adeen Flinker[6], Daniel Friedman[6], Nima Mesgarani[1,2*] (nima@ee.columbia.edu)

[1]Columbia University, [2]Mortimer B. Zuckerman Mind Brain Behavior Institute, [3]Hofstra Northwell School of Medicine, [4]The Feinstein Institutes for Medical Research, [5]Department of Neurology, Columbia University, [6]New York University School of Medicine

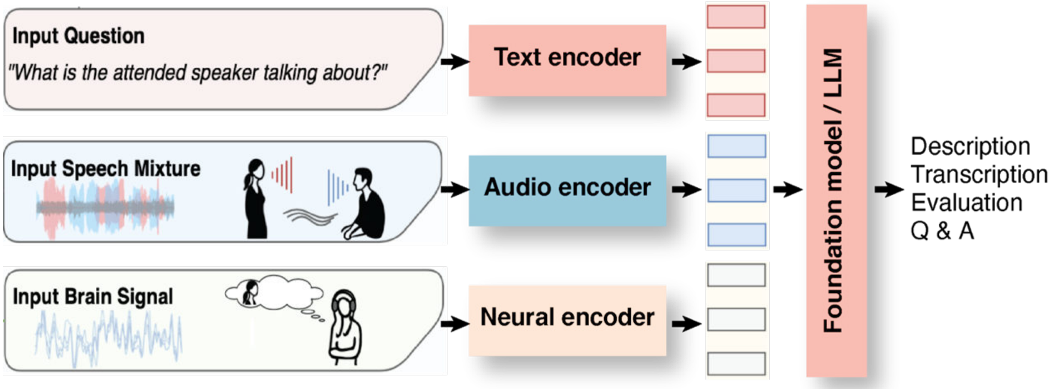
Introduction: Auditory foundation models represented by auditory large language models (LLMs) process speech and text inputs to analyze speech mixtures, recognize speakers, and transcribe content. However, they operate independently of the listener's perceptual experience, treating all auditory inputs equally. In reality, auditory perception is inherently selective—different listeners focus on different elements of the same acoustic scene at different times. We introduce a novel foundation model that extends auditory LLMs beyond traditional multimodal inputs by incorporating neural data from the listener, enabling personalized auditory scene understanding.

Gap: While multimodal foundation models integrate text and audio, they lack a mechanism to account for individual auditory attention. Existing models cannot distinguish between foreground and background speech based on the listener's intent, limiting their ability to provide personalized responses about the scene. Prior research in auditory attention decoding (AAD) has demonstrated that neural signals can reveal which speaker a person is focusing on, but this has not been integrated into auditory LLMs for selective response generation.

Methods: We introduce an attention-aware framework that integrates neural recordings into an auditory LLM, enabling selective speech processing. Our approach consists of two key steps. First, using intracranial EEG (iEEG) data, we decode the listener's auditory focus, predicting an attention token that represents the attended speaker. This token is derived from speaker-specific neural features, allowing the model to differentiate between attended and ignored speech sources. Next, the predicted attention token is incorporated into an auditory LLM's input, allowing it to generate responses that align with the listener's perceptual focus. A chain-of-thought reasoning mechanism ensures that transcriptions, summaries, and speech enhancements prioritize the listener's intended speaker while also enabling background summarization if needed.

Results: Our model enables a range of listener-personalized tasks, including speaker-aware transcription, selective speech enhancement, and customized background summarization. Experimental results demonstrate that the model can accurately extract and process speech from the attended source while filtering out distractions, outperforming traditional auditory foundation models that lack neural integration.

Significance: This work represents the first integration of listener-specific neural signals into a foundation model for auditory scene understanding, bridging the gap between passive speech processing and human-centered auditory AI. By extending auditory LLMs beyond text and audio to include neural data, we move toward AI systems that dynamically adapt to human perception. This advancement has significant implications for assistive hearing technologies, human-computer interaction, and personalized AI-driven auditory experiences.



Published by Verlag der Technischen Universität Graz