SoundNarratives: Rich Auditory Scene Descriptions to Support Deaf and Hard of Hearing People

Liang-Yuan “Leo” Wu and Dhruv Jain | University of Michigan

Prior work
“humming” | “computer keyboard” | “speech”

Our approach
Soft **conversation** mixed with **keyboard typing**, while the coffee machine **hums and beeps**.

Prior work
“car honks” | “footsteps” | “traffic noise”

Our approach
Footsteps approaching steadily with distant **traffic noise**. Suddenly, a car **honks** loudly.

Prior work
”laughter” | “microwave beep” | “cutlery”

Our approach
Child **laughs** as **cutlery clinks** against plates, and then the **microwave timer beeps** twice.

Motivation

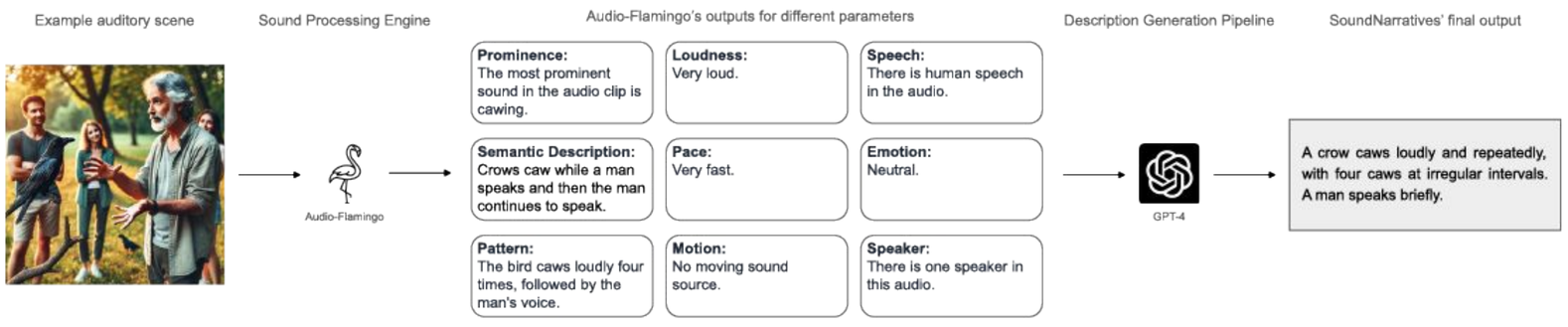
Existing sound awareness technologies classify sounds into fixed categories, which fail to capture the full complexity of real-world auditory scenes. In this work, we introduce ***SoundNarratives***, a real-time system based on an audio-language model that generates rich, contextual auditory scene descriptions tailored to DHH users.

Research Questions

1. What auditory scene information do DHH individuals need beyond basic sound classification?
2. Can state-of-the-art audio-language models capture and structure these auditory parameters effectively?
3. How do DHH users perceive and interact with AI-generated auditory scene descriptions?

Methods

To establish the foundational elements for comprehensive auditory scene awareness, we elicited nine key sound parameters through semi-structured interviews with 10 DHH individuals and a thorough literature review. Subsequently, we employed prompt engineering with *AudioFlamingo*, a state-of-the-art audio-language model, to assess its capacity to extract these critical parameters. Building upon these insights, we developed and evaluated an end-to-end system designed to provide rich, real-time auditory scene understanding for DHH individuals, with a user study involving an additional 10 DHH participants.



Results

1. DHH participants shared excitement about this novel idea of holistic auditory scene understanding, and we identified nine key sound parameters.
2. AudioFlamingo performed well (75%-90%) on well-defined perceptual properties such as sound class and loudness while achieved lower accuracy (50%-65%) on more abstract attributes like spatial dynamics and sound patterns.
3. Participants showed a significant preference for SoundNarratives over the freeform descriptions, underscoring the value of our structured prompting pipeline. Subjective insights indicated high user satisfaction with participants reporting increased confidence and situational awareness when accessing auditory information.