Advancing Expert-Level Reasoning and Understanding in Large Audio Language Models

Ramani Duraiswami[♦] Santosh Kesiraju[♠] Dinesh Manocha[♦] Sreyan Ghosh[♠] Sonal Kumar[♦]

[♦]University of Maryland, College Park, MD, USA [♠] Technical University, Brno, Czechia
{ramanid, dmanocha, sreyang, sonalkum}@umd.edu kesiraju@fit.vut.cz

I. Summary

To exhibit intelligence in the physical world, both AI agents and humans must comprehend and then reason about sound (including speech, non-speech sounds, and music). However, research in complex reasoning with audio has lagged behind modalities such as language and vision. This discrepancy is due to several challenges, the capabilities of algorithms for audio understanding, scarcity of large-scale training datasets, architectures, and, the lack of comprehensive benchmarks for assessing advanced audio processing capabilities. The proposed JSALT effort will allow us to create an intensive project to address critical limitations in Large Audio Language Models (LALMs). Our recent open-source MMAU benchmark [6] has revealed that even state-of-the-art LALMs, including proprietary ones, achieve only 53% accuracy on complex audio reasoning tasks. This deficiency represents a crucial bottleneck in the development of multimodal AI systems and the progression toward AGI. The JSALT project aims to improve LALM architectures and evaluation methodologies, building upon baseline open source materials we have contributed: the GAMA architecture [2](EMNLP 2024) and the MMAU benchmark, under review at ICLR.

II. Current Limitations in LALMs

The emergence of Large Language Models (LLMs) agents with their understanding of human language, and their ability to reason and plan, presents an opportunity for speech and audio processing. Large Audio-Language Models (LALMs), a subclass of Multi-Modal Large Language Models (MLLMs), are designed to process audio inputs alongside language, and possibly vision and robotics. Recent developments in LALMs [2], [3] have shown advancements in foundational processing tasks, such as Automatic Speech Recognition (ASR), Acoustic Scene Classification, and Music Genre Classification. However, LALMs fall far short of the AGI benchmark proposed by Morris *et al.* [5], who define it as a system performing at the "90th percentile of skilled adults" across a broad spectrum of tasks. Tasks like speech and emotion recognition, are often manageable even for young children [1], [4], and, while valuable for basic audio comprehension, do not require the kind of deliberate and complex reasoning for AGI.

To bring current LALMs to this standard, this JSALT workshop is focused on advancing expert-level understanding and complex reasoning in audio-language models. The team, drawn from several universities and industry in the US, Europe and Asia, and with students and senior professionals from various disciplines, will allow us to advance architectures, training approaches, and data resources to enhance these models' capabilities. The aim is to push the field forward, achieving a more sophisticated level of reasoning and comprehension in audio processing that parallels the complexity of human intelligence. Our recent work MMAU [6], the first comprehensive benchmark tailored for multimodal audio understanding and reasoning, will be improved and made more aligned with human reasoning. MMAU features over 10,000 expertly annotated audio-question-response pairs evenly distributed across speech, sound, and music domains; it already achieves significant **breadth** by testing on 27 distinct tasks (16 reasoning and 11 information extraction. The tasks have **depth** by requiring advanced reasoning (e.g., multi-speaker role mapping, emotional shift detection, and temporal acoustic event analysis). Each question tests the model's ability to infer, reason, recall relevant knowledge, and understand, and not just transcribe or describe content, (Fig. 1).

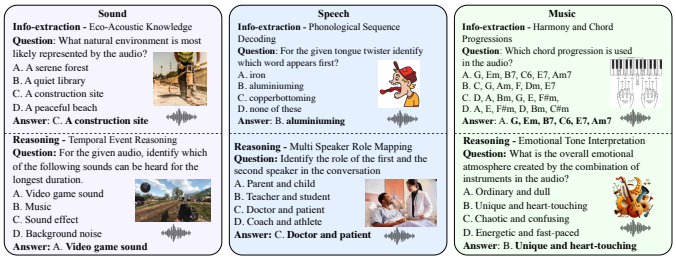

Fig. 1. The MMAU benchmark includes diverse reasoning and information extraction tasks across sound, speech, and music. Each is rich, context-specific with human-annotated expert-level QA pairs.

We evaluated multiple Large Audio-Language Models (LALMs) on the MMAU benchmark. The best closed source performer - Gemini Pro v1.5 achieves only 52.97% accuracy, while the top-open-source model, Qwen2-Audio, reaches 52.50%. *These results underscore both the substantial room for improvement in current models and the minimal performance gap between open-source and proprietary solutions—highlighting a valuable opportunity for JSALT to advance the field.* Analysis in our paper suggests that models face challenges in audio perception and alignment, which requires improvements in the LALM technologies, using improvements from the literature, and from work ongoing at Brno.

III. GAMA Architecture and Improvements

Models capable of accurately responding to such questions will be developed using starter code and datasets, based on our previous work [2]. All data, models, and training code

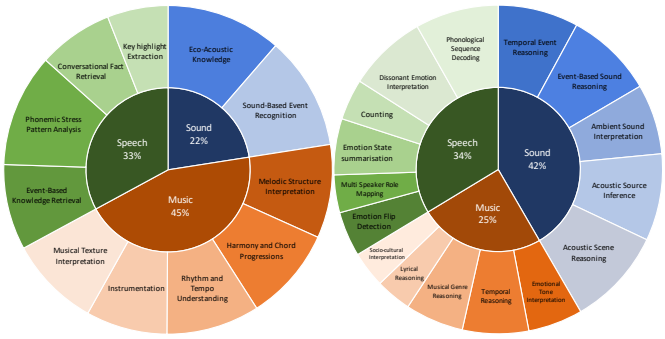

Fig. 2. Distribution of skills required for **(Left)** information extraction and **(Right)** reasoning questions, in the MMAU benchmark.

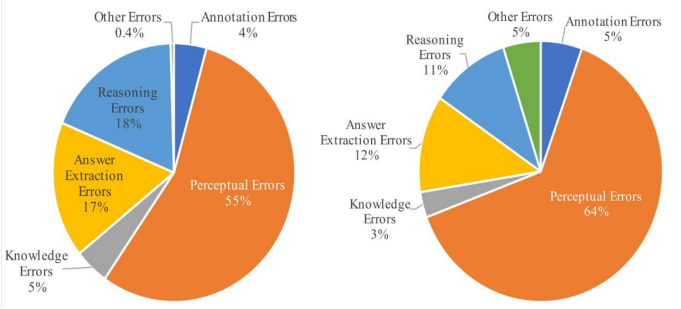

Fig. 3. Distribution of error types across 500 instances for top performing LALMs QWEN2-Audio-Instruct and Gemini Pro.

developed during the workshop will be open-sourced. The GAMA architecture represents a significant step forward in LALM design, but MMAU analysis suggests several critical areas for improvement: adding speech processing (work has begun); addressing longer audio clips; advanced and efficient connector architecture using discrete latent variables; and agentic planning-based inference for complex reasoning. These all will be worked on before and during the workshop.

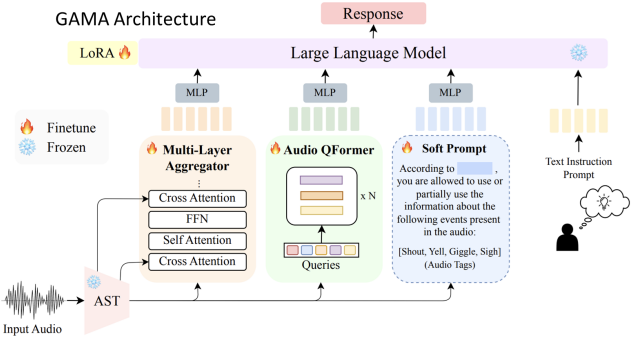

Fig. 4. The GAMA architecture. We will explore alternatives for the choices made to see if performance of the LALM can be improved. We are already working on adding speech to GAMA.

IV. Synergies and Proposed Team

The JSALT community is very interested in the project, and a number of workshop attendees and members of their networks have expressed interest. The proposed project will be lead by Prof. Ramani Duraiswami (UMD) and students who worked on the papers. Three of the students from his team would join (Sreyan Ghosh, Sonal Kumar, S. Sakshi, Nitish Anand, Utkarsh Tyagi, S. Ramaneshwaran, Bowen Zhi and Armin Gerami); while the remaining would be heavily involved, but potentially be at internships. Dr. Santosh Kesiraju (BUT) will co-lead the project and be joined by PhD students Simon Sedlacek and Katia Vendrame. Other students from other universities will also be welcomed. Prof. Har at (BUT) would also be involved. We have potential interest from Profs. David Harwath at UT Austin, Chao Zhang and Wenyi Zhu at Tsinghua, Shinji Watanabe at CMU, and Xavier Serra at UPF Barcelona. Shijia Liao (Fish Audio), Dr. Emre Eskimez (Microsoft), Dr. Murali Karthick Baskar (Google), Dr. Shankar Kumar (Google), and Dr. Viktor Rozgic (Amazon) have also expressed interest. GPU resources have been promised from Fish Audio, LUMI, BSC, Karolina supercomputing clusters and from the UMD Computer Science department.

V. Project Outcomes and Impact

This project represents a crucial step toward advancing multimodal foundational models and progressing toward AGI. Audio understanding represents a critical but currently underperforming component in multimodal AI systems. Fundamental challenges in audio AI will be addressed through three key research directions: improved learning algorithms and architectures; improved training methodologies including methods to create augmented data sets; and benchmark and metric advancement. Significant improvement in LALM performance and building new research collaborations are expected to be achieved. The community will benefit from open-source implementations and public datasets and benchmarks. The researchers involved have a history of high productivity, and delivering high quality papers.

VI. Conclusion

This six-week project represents a focused effort to advance audio AI through improvements in both model architectures and evaluation methodologies. The project addresses critical limitations in current LALMs while pushing toward better multimodal AI systems and AGI capabilities. The combination of strong technical foundations, clear objectives, and a collaborative approach positions the project for significant impact in advancing audio AI technology.

References

[1] Peter Gerhardstein and Carolyn Rovee-Collier. The development of visual search in infants and very young children. *Journal of Experimental Child Psychology*, 81(2):194–215, 2002.

[2] Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. GAMA: a large audio-language model with advanced audio understanding and complex reasoning abilities. *arXiv preprint arXiv:2406.11768*, 2024.

[3] Yuan Gong et al. Listen, think, and understand. In *ICLR 2024*.

[4] Richard P. Lippmann. Speech recognition by machines and humans. *Speech Communication*, 22(1):1–15, 1997.

[5] Meredith Ringel Morris et al. Position: Levels of AGI for operationalizing progress on the path to AGI. In *ICML 2024*.

[6] S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. MMAU: A massive multi-task audio understanding and reasoning benchmark. In *Submitted to ICLR 2024*. under review.