PREPRINT VERSION 1.1

A Primer on the Challenges of Audio Latency in Artificial Intelligence Systems

Noah M. Kenney[1*]
[1]*Georgia Institute of Technology, Atlanta, GA*

ABSTRACT
Audio latency in Artificial Intelligence (AI) systems poses significant challenges, especially in applications requiring real-time processing and interaction, such as the use of AI in call centers, translation, and live audio processing. This paper explores the technical complexities and mathematical frameworks underlying audio latency, including a brief analysis of its causes, impacts, and potential mitigation strategies. It aims to provide a comprehensive understanding of the challenges faced by AI systems in managing audio latency by looking at signal processing, neural network inference, and hardware-software co-design.

Key words: Audio Latency – Neural Network Interface – Latency Mitigation

1 INTRODUCTION

For purposes of this paper, we define audio latency as the delay between input audio signal and the corresponding output. Audio latency is a critical factor in the performance of nearly any audio-based AI system, particularly in speech recognition, real-time audio synthesis, and interactive voice response systems.

The increasing integration of AI in these domains necessitates a deeper understanding of the sources and implications of latency. This paper aims to analyze the challenge of audio latency, analyze its effects on system performance, and explore advanced strategies to minimize it.

2 SOURCES OF AUDIO LATENCY

There are many possible sources for audio latency, which we've broadly broken up into four categories:

1) Signal Acquisition and Preprocessing
   - Analog-to-Digital Conversion (ADC): The process of converting analog signals to digital form introduces inherent delays
   - Preprocessing Algorithms: noise reduction, echo cancellation, and other preprocessing steps add to the latency.
2) Data Transmission
   - Network Latency: In distributed systems, data transmission over networks can introduce significant delays
   - Buffering: Buffers used to manage data flow can introduce additional latency.
   - Server Response: Limits of server-side computational resources can increase latency.
3) Computational Delays
   - Neural Network Inference: The time taken to process data through Deep Neural Networks (DNNs) or Convolutional Neural Networks (CNNs) can be significant, especially in cases where the network is complex, with multiple layers.
   - Algorithmic Complexity: Complex algorithms for feature extraction and pattern recognition contribute to latency.
4) System Integration
   - Hardware-Software Interfacing: Communication between hardware components and software layers can introduce delays.
   - Operating System Scheduling: The method an operating system schedules processes can affect the timing of audio processing tasks.

These sources of latency have varying levels of impact, with the size of the audio file, bit rate of the audio, and acceptable level of compression all having a significant impact.

3 MATHEMATICAL MODELING OF LATENCY

To quantitative analyze audio latency, we model it as a series of delays D_i for each processing stage i:

$$L = \sum_{i=1}^{n} D_i$$

where L is the total latency and D_i represents the delay at the i-th stage.

We can represent ADC latency by the following equation:

$$D_{ADC} = \frac{1}{f_s}$$

where f_s is the sampling frequency of the audio.

We can represent preprocessing latency by the following equation:

$$D_{prep} = \sum_{j=1}^{m} T_{prepj}$$

where T_{prepj} is the time taken by the j-th preprocessing step.

We can represent network latency by the following equation:

$$D_{net} = \frac{P_{trans}}{B}$$

where P_{trans} is the packet size and B is the bandwidth.

We can represent inference latency by the following equation:

$$D_{inf} = \sum_{k=1}^{p} T_{inf_k}$$

where T_{inf_k} is the time taken by the k-th layer of the neural network.

We can represent buffering latency by the following equation:

$$D_{buff} = N_{buff} * T_{buff}$$

where N_{buff} is the number of buffers and T_{buff} is the time per buffer cycle.

4 MITIGATION STRATEGIES

Addressing audio latency requires a multi-faceted approach, often combining both hardware and software.

In terms of hardware optimization, primary advantages come from High-Speed ADCs, which can reduce the conversion latency. Dedicated processing units such as Graphics Processing Units (GPUs) and Tensor Processing Units (TPUs) can accelerate neural network interface.

In terms of software optimization, primary advantages come from algorithmic efficiency, optimized for preprocessing and interface to reduce computational delays. Additionally, parallel processing can be utilized to distribute tasks across multiple processors to reduce processing time. This is often done automatically when utilizing cloud compute.

In terms of network optimization, primary advantages come from low-latency protocols and utilization of edge computing, which processes data closer to the source (at the network edge) to reduce transmission delays.

In terms of system-level enhancements, primary advantages come from utilization of Real-Time Operating Systems (RTOS), which can improve scheduling and reduce latency. Additionally, it can be beneficial to utilize latency-aware scheduling, which prioritizes latency-sensitive tasks, such as real-time audio processing or processing of key audio tracks.

5 CONCLUSION

Audio latency remains a critical challenge in AI systems, especially in applications requiring real-time processing and interaction. Mathematically, achieving zero latency is not feasible; however, achieving a lower level of latency through mitigation strategies (such as those outlined previously) can improve performance metrics. This can be accomplished by optimizing the mathematical equations provided based on the category of optimization with the largest latency improvement (i.e., network latency, inference latency, etc.).

There is future research focusing on developing novel hardware architectures, optimizing neural network designs for low latency, and developing new communication protocols tailored for real-time AI applications. There is also potential found in cross-disciplinary approaches which integrate insights from signal processing, computer science, and human-computer interaction. Though outside the scope of this paper, optimization in the AI model itself, combined with neural network layer optimization and feature extraction can provide significant improvements in the overall processing time and a reduction in latency. Similarly, a reduction in complexity of data flows from audio input to output can have a significant improvement on latency, particularly in preprocessing and processing, but also in buffering.