

Invention Disclosure Form

1. Title of the Invention

AHNA (Audio Hearing and Neural Analysis)- Intelligent Audio Understanding Model Integrating ASR, Diarization, Event Detection, Paralinguistics, and Reasoning for Edge and Software Deployment.

2. Background of the Invention (i.e., Known Prior Arts)

Over the past decade, significant advancements have been made in audio intelligence, primarily driven by deep learning and multimodal language models.

Despite these advancements, most existing approaches focus on individual audio subtasks in isolation, such as Automatic Speech Recognition (ASR), speaker diarization, audio event detection, or paralinguistic/emotion analysis, without unified contextual reasoning or efficient deployment on edge devices.

Prior Arts and Limitations

(a) Automatic Speech Recognition (ASR) Systems

Models such as *Whisper* (OpenAI, 2022), *DeepSpeech* (Mozilla, 2017), and *Wav2Vec2.0* (Meta AI, 2020) achieve high transcription accuracy across multiple languages and acoustic conditions.

Limitations: While robust in speech-to-text conversion, these models are task-specific.

They lack:

Integration with event or speaker context.

Emotional and paralinguistic understanding.

Reasoning over non-speech audio.

(b) Speaker Diarization and Identification Models

Conventional systems (x-vectors, ECAPA-TDNN, pyannote.audio) excel at segmenting and labeling “who spoke when.”

Limitations: These models operate independently of the semantic or emotional content of speech and do not integrate with other auditory modalities for unified reasoning.

(c) Audio Event and Scene Detection

Models like *AudioSet*, *YAMNet*, and *PANs* (Pretrained Audio Neural Networks) have advanced environmental sound classification and event tagging.

Limitations: These systems are generally static classifiers, lack cross-domain reasoning (linking events to speech intent), and are typically cloud-based, making edge deployment challenging.

(d) Paralinguistic and Emotional Speech Models

Systems such as OpenSMILE, DeepSpectrum, and SERNet extract prosody, pitch, emotion, and voice-quality features.

Limitations: While effective in controlled environments, these models struggle with overlapping or noisy speech, and rarely connect to reasoning frameworks for decision-level interpretation.

(e) Multimodal Audio-Language Models (LALMs)

Recent models, including *Audio Flamingo* (NVIDIA, 2024), *Qwen-Audio* (Alibaba, 2024), *Pengi* (Google, 2023), and *LTU* (2024), attempt to integrate audio and text for tasks like captioning, question answering, and dialogue.

Limitations: These models are computationally intensive, unsuitable for real-time or edge applications, do not fully integrate ASR, diarization, and paralinguistics into a single unified pipeline, and cannot reason over overlapping or concurrent sound sources effectively.

Identified Gaps in Prior Arts

Fragmentation of Audio Subtasks: Most systems process ASR, speaker recognition, event detection, and paralinguistics independently, without a unified reasoning framework.

Lack of Contextual Reasoning: Existing models perform classification or transcription but do not understand relationships among multiple auditory cues (e.g., linking tone, speaker identity, and environmental sound).

Inefficiency for Edge Deployment: High-performing models rely on large transformer architectures, making low-latency, on-device processing infeasible.

Absence of Unified Multitask Intelligence: No existing solution fully integrates speech, event, emotion, and reasoning into a cohesive pipeline that can scale from edge devices to software platforms.

Need for the Invention

These limitations create a strong demand for an integrated, efficient, and contextually intelligent audio understanding system. Such a system would:

- (a) Combine speech recognition, speaker tracking, event and emotion detection, and reasoning in a unified neural architecture.
- (b) Achieve real-time performance with low computational overhead suitable for embedded devices.
- (c) Provide explainable outputs and adapt to diverse auditory scenes and multi-speaker interactions.

3. Novelty of the Invention

The *AHNA* (Audio Hearing and Neural Analysis) system introduces several novel contributions that distinguish it from existing audio understanding technologies:

Unified Multitask Audio Understanding:

Unlike prior systems that handle Automatic Speech Recognition (ASR), speaker diarization, audio event detection, and paralinguistics in isolation, AHNA integrates all these modalities within a single unified neural pipeline. This enables simultaneous reasoning across speech, speaker identity, environmental sounds, and emotional or prosodic cues, providing a comprehensive understanding of complex auditory scenes.

Context-Aware Audio Reasoning :

AHNA incorporates a reasoning layer that performs context-aware inference over concurrent or overlapping sound sources. This allows the system to not only classify or transcribe audio but also understand relationships among multiple auditory cues, such as linking tone, speaker intent, and environmental events — a capability largely absent in prior arts.

Parallel Multimodal Processing Architecture :

The system uses parallel processors for ASR, diarization, event detection, and paralinguistics, which run simultaneously on raw audio input. This parallelism enables high efficiency and reduces latency, allowing near-real-time processing suitable for edge devices — an improvement over sequential or cloud-dependent architectures in existing models.

Edge-Optimized Deployment :

AHNA employs lightweight transformer compression, hierarchical attention, and quantization techniques, enabling deployment on low-power edge devices while maintaining high inference accuracy. This contrasts with prior audio-language models that are resource-intensive and typically cloud-dependent.

Adaptive and Explainable Outputs :

The system produces interpretable reasoning outputs for multi-speaker dialogues, overlapping events, and paralinguistic cues, enhancing transparency and adaptability in real-world applications. Existing systems generally lack explainability or adaptability for dynamic auditory environments.

Scalable Offline and Real-Time Operation :

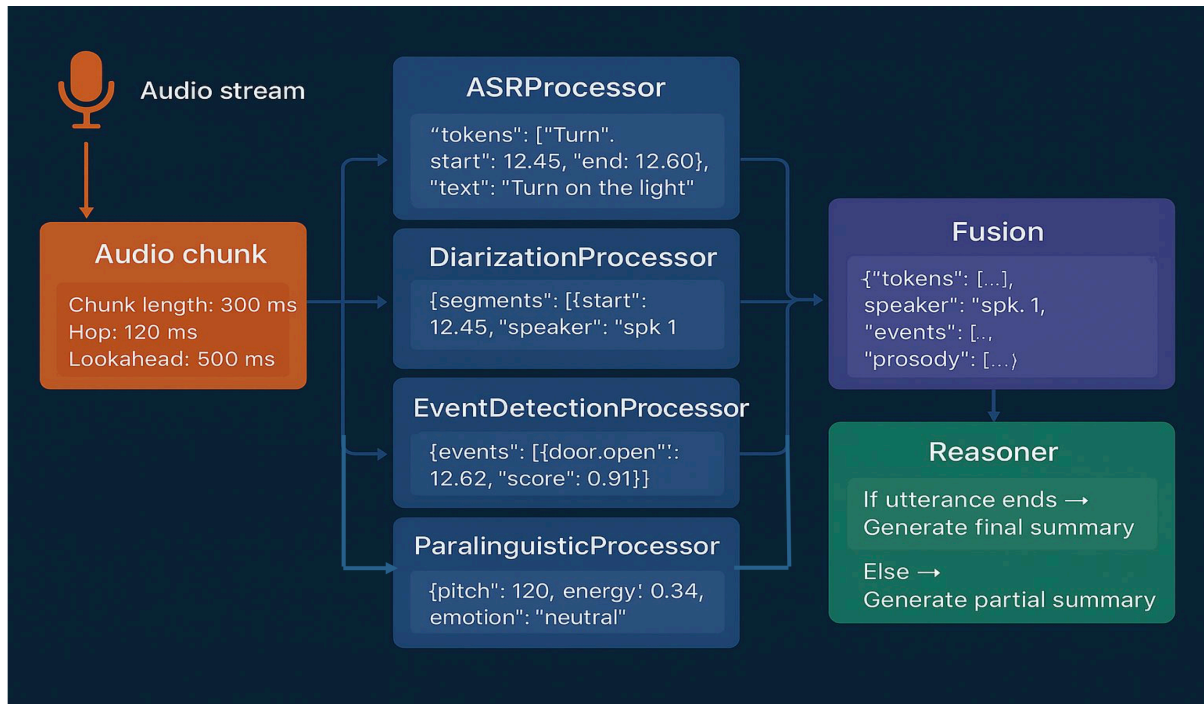
AHNA is designed for both offline and real-time inference, allowing seamless integration into software platforms, IoT systems, and assistive AI devices. This dual-mode operation is a unique feature, as most high-performing audio-language models are constrained to offline or cloud-only deployment.

Comprehensive Fusion of Audio Modalities :

Through a multimodal fusion mechanism, AHNA effectively combines diverse auditory inputs into a cohesive feature space, enabling more accurate predictions and higher-level reasoning compared to single-task models or LALMs that fuse only limited modalities.

In summary, the novelty of AHNA lies in its ability to unify multiple audio processing tasks, reason contextually over overlapping and complex auditory inputs, operate efficiently on edge devices, provide explainable outputs, and support scalable deployment — capabilities that are not collectively addressed by any known prior art.

4. Detailed Description of the Invention



Workflow Diagram

4.1. Audio Chunking Module :

The system captures a continuous audio stream at a sample rate of 16 kHz or 24 kHz.

The stream is segmented into overlapping audio chunks of 300 milliseconds each, with a hop size of 120 milliseconds and a lookahead buffer of 500 milliseconds.

The lookahead mechanism appends future samples to each chunk, improving temporal context for mid-sentence understanding without introducing excessive delay.

Each chunk is time-stamped relative to the start of the stream and dispatched to all processing units in parallel.

4.2. Parallel Processing Units :

The architecture employs four primary, concurrent processors:

ASR Processor :

Converts audio chunks into phoneme or token-level text with precise start and end timestamps.

Example output:

```
{  
  "processor": "ASRProcessor",  
  "start_time": 12.45,  
  "end_time": 12.75,  
  "payload": {"text": "Turn on the light", "tokens": [...]},  
  "confidence": 0.92  
}
```

Diarization Processor :

Detects speaker identities and boundaries across overlapping speech segments.

```
{"segments": [{"start": 12.45, "end": 12.75, "speaker": "spk_1"}]}
```

Event Detection Processor :

Recognizes non-speech acoustic events such as door knocks, typing, or ambient signals.

```
{"events": [{"label": "door_open", "time": 12.62, "score": 0.91}]}
```

Paralinguistic Processor :

Analyzes voice attributes like pitch, energy, and emotion to infer speaker affect and tone.

```
{"pitch": 120, "energy": 0.34, "emotion": "neutral"}
```

Each processor emits outputs with consistent temporal references and metadata, ensuring alignment at the fusion stage.

4.3. Fusion Engine :

The Fusion Engine is a time-synchronized integration layer that aggregates and aligns outputs from all processors using a common time base.

It merges text tokens, speaker information, detected events, and paralinguistic cues into a unified multimodal data structure.

Fusion methods may include:

- (a) Token-level enrichment: augmenting ASR tokens with speaker, event, and emotion metadata.
- (b) Frame-level concatenation: combining numeric feature vectors for temporal correlation.
- (c) Confidence-weighted merging: prioritizing high-confidence processor outputs.

The fused output is formatted for downstream reasoning and contains semantically enriched, temporally aligned frames.

4.4. Reasoner Module :

The Reasoner operates in streaming mode, maintaining internal state to track context across incoming fused segments.

It performs incremental reasoning to generate partial summaries during speech and final summaries at utterance completion.

Finalization is triggered by :

Detection of speech silence exceeding a threshold (e.g., 500–800 ms).

Explicit end-of-utterance markers.

The Reasoner's output evolves dynamically, revising partial hypotheses into stable, high-confidence summaries once finalization occurs.

5. Workflow Summary :

Audio input is continuously captured and segmented into overlapping chunks with lookahead.

Each chunk is processed concurrently by ASR, Diarization, Event, and Paralinguistic modules.

All processor outputs are timestamped and sent to the Fusion Engine.

The Fusion Engine synchronizes outputs and produces enriched representations.

The Reasoner consumes fused data incrementally to generate partial summaries.

Upon utterance completion, the Reasoner generates a comprehensive final summary.

This workflow ensures real-time adaptability, temporal precision, and context-aware multimodal understanding.

6. Implementation and Deployment :

The invention can be implemented in Python with asynchronous event-driven architectures (e.g., asyncio, message queues).

Models may be deployed via ONNX, TFLite, or TensorRT for edge devices, with cloud offloading for heavier reasoning tasks.

The system can further incorporate a circular buffer mechanism for continuous audio handling, bounded queues for backpressure management, and state persistence for continuity across chunks.

5. Advantages of the Invention

- (a) Low Latency: Lookahead buffering enables fast, context-aware inference without waiting for full utterance completion.
- (b) Parallel Modularity: Each processor operates independently, allowing scalability and hardware optimization.
- (c) Temporal Fusion: Consistent timestamping ensures accurate alignment of multimodal signals.
- (d) Incremental Reasoning: The Reasoner provides both immediate feedback (partial summaries) and refined outputs (final summaries).
- (e) Versatility: The framework supports both on-device and cloud-deployed models, enabling flexible deployment on edge systems.
- (f) Robustness: Confidence-based fusion mitigates individual processor errors, enhancing system reliability.
- (g) Emotionally aware: By embedding emotion detection, the system can adapt its tone and follow-up strategy to the user's affective state.
- (f) Acoustic context preserved: Because the system retains prosodic and acoustic cues, it can identify hesitation, sarcasm, uncertainty, etc., which purely text-based systems cannot reliably capture.
- (g) Multilingual and Code-Switching Adaptability: The invention is capable of recognizing and processing multiple languages and dialects within a single interaction, adapting automatically to language shifts and accents.
- (h) Unified End-to-End Architecture: The system integrates speech detection, emotion recognition, reasoning, question generation, and knowledge retrieval into a single end-to-end framework, reducing complexity and error propagation.
- (i) Speech and Non-Speech Differentiation: The model accurately differentiates between speech, silence, noise, and other non-speech sounds, ensuring reliable operation in real-world audio environments.

6. Application Area of the Invention

1. Smart Assistants: Real-time understanding of commands and emotions for responsive AI interaction.
2. Meeting & Call Analytics: Speaker-aware transcriptions and automatic summaries of conversations.
3. Surveillance & Security: Detect critical acoustic events and monitor environments in real time.
4. Healthcare & Therapy: Track speech patterns and emotional tone for therapy and health monitoring.

5. Human-Robot Interaction: Enable robots to understand commands, context, and emotions instantly.

6. Media Monitoring & Broadcast Analysis: Analyze live audio streams for speakers, events, and emotional cues.

7. Customer Support and Call Centers: The system can analyze customer speech, detect emotions such as frustration or satisfaction, and guide agents or automated bots to provide more effective and empathetic responses.

8. Legal and Forensic Audio Analysis: The invention can aid in analyzing recorded speech for stress, deception, or emotional indicators, supporting investigations or courtroom analysis.

9. Emergency Response and Disaster Management: The invention can assist in analyzing distress calls or field communications in real time, identifying emotional urgency and guiding rapid, informed decision-making.