Documentation Output

Summary:

Invention Disclosure Form

1. Title of the Invention

AHNA (Audio Hearing and Neural Analysis)- Intelligent Audio Understanding Model Integra8ng ASR, Diariza8on, Event Detec8on, Paralinguis8cs, and Reasoning for Edge and SoCware 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 mul8modal language models. Despite these advancements, most exis8ng approaches focus on individual audio subtasks in

isola8on, such as Automa8c Speech Recogni8on (ASR), speaker diariza8on, audio event detec8on, or paralinguis8c/emo8on analysis, without unified contextual reasoning or efficient deployment on edge devices.

Structure:

They lack:

audio understanding system. Such a system would:

that dis8nguish it from exis8ng audio understanding technologies:

Unified Mul.task Audio Understanding:

Context-Aware Audio Reasoning :

Parallel Mul.modal Processing Architecture :

Edge-Op.mized Deployment :

Adap.ve and Explainable Outputs :

Scalable Offline and Real-Time Opera3on :

Comprehensive Fusion of Audio Modali.es :

4.1. Audio Chunking Module :

4.2. Parallel Processing Units :

The architecture employs four primary, concurrent processors:

ASR Processor :

Example output:

Diariza8on Processor :

Event Detec8on Processor :

Paralinguis8c Processor :

4.3. Fusion Engine :

Fusion methods may include:

4.4. Reasoner Module :

Finaliza8on is triggered by :

5. Workflow Summary :

6. Implementa.on and Deployment :

Easy Text:

Invention Disclosure Form

1. Title of the Invention

AHNA (Audio Hearing and Neural Analysis)- Intelligent Audio Understanding Model Integra8ng ASR, Diariza8on, Event Detec8on, Paralinguis8cs, and Reasoning for Edge and SoCware 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 mul8modal language models.

Despite these advancements, most exis8ng approaches focus on individual audio subtasks in isola8on, such as Automa8c Speech Recogni8on (ASR), speaker diariza8on, audio event detec8on, or paralinguis8c/emo8on analysis, without unified contextual reasoning or efficient deployment on edge devices.

Prior Arts and Limita.ons

(a) Automa.c Speech Recogni.on (ASR) Systems

Models such as Whisper (OpenAI, 2022), DeepSpeech (Mozilla, 2017), and Wav2Vec2.0 (Meta AI, 2020) achieve high transcrip8on accuracy across mul8ple languages and acous8c condi8ons.

Limita8ons: While robust in speech-to-text conversion, these models are task-specific.

They lack:

Integra8on with event or speaker context.

Emo8onal and paralinguis8c understanding.

Reasoning over non-speech audio.

(b) Speaker Diariza.on and Iden.fica.on Models

Conven8onal systems (x-vectors, ECAPA-TDNN, pyannote.audio) excel at segmen8ng and labeling "who spoke when."

Limita8ons: These models operate independently of the seman8c or emo8onal content of

speech and do not integrate with other auditory modalties for unified reasoning.

(c) Audio Event and Scene Detecton

Models like AudioSet, YAMNet, and PANNs (Pretrained Audio Neural Networks) have advanced environmental sound classificaton and event tagging.

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

(d) Paralinguistc and Emotonal Speech Models

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

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

(e) Multmodal Audio-Language Models (LALMs)

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

Limitatons: These models are computatonally intensive, unsuitable for real-tme or edge applicatons, do not fully integrate ASR, diarizaton, and paralinguistcs into a single unified pipeline, and cannot reason over overlapping or concurrent sound sources effectvely.

Identfied Gaps in Prior Arts

Fragmentaton of Audio Subtasks: Most systems process ASR, speaker recogniton, event detecton, and paralinguistcs independently, without a unified reasoning framework.

Lack of Contextual Reasoning: Existng models perform classificaton or transcripton but do not understand relatonships among multple auditory cues (e.g., linking tone, speaker identty, 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 Multtask Intelligence: No existng soluton fully integrates speech, event, emoton, and reasoning into a cohesive pipeline that can scale from edge devices to soCware placorms.

Need for the Inventon

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

(a) Combine speech recogniton, speaker tracking, event and emoton detecton, and reasoning

in a unified neural architecture.

(b) Achieve real-8me performance with low computa8onal overhead suitable for embedded devices.

(c) Provide explainable outputs and adapt to diverse auditory scenes and mul8-speaker interac8ons.

3. Novelty of the Invention

The AHNA (Audio Hearing and Neural Analysis) system introduces several novel contribu8ons that dis8nguish it from exis8ng audio understanding technologies:

Unified Mul.task Audio Understanding:

Unlike prior systems that handle Automa8c Speech Recogni8on (ASR), speaker diariza8on, audio event detec8on, and paralinguis8cs in isola8on, AHNA integrates all these modali8es within a single unified neural pipeline. This enables simultaneous reasoning across speech, speaker iden8ty, environmental sounds, and emo8onal 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 rela8onships among mul8ple auditory cues, such as linking tone, speaker intent, and environmental events - a capability largely absent in prior arts.

Parallel Mul.modal Processing Architecture :

The system uses parallel processors for ASR, diariza8on, event detec8on, and paralinguis8cs, which run simultaneously on raw audio input. This parallelism enables high efficiency and reduces latency, allowing near-real-8me processing suitable for edge devices - an improvement over sequen8al or cloud-dependent architectures in exis8ng models.

Edge-Op.mized Deployment :

AHNA employs lightweight transformer compression, hierarchical aaen8on, and quan8za8on 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.

Adap.ve and Explainable Outputs :

The system produces interpretable reasoning outputs for mul8-speaker dialogues, overlapping events, and paralinguis8c cues, enhancing transparency and adaptability in real-world applica8ons. Exis8ng systems generally lack explainability or adaptability for dynamic auditory

environments.

Scalable Offline and Real-Time Opera3on :

AHNA is designed for both offline and real-8me inference, allowing seamless integra8on into soCware placorms, IoT systems, and assis8ve AI devices. This dual-mode opera8on is a unique feature, as most high-performing audio-language models are constrained to offline or cloud-only deployment.

Comprehensive Fusion of Audio Modali.es :

Through a mul8modal fusion mechanism, AHNA effec8vely combines diverse auditory inputs into a cohesive feature space, enabling more accurate predic8ons and higher-level reasoning compared to single-task models or LALMs that fuse only limited modali8es.

In summary, the novelty of AHNA lies in its ability to unify mul8ple audio processing tasks, reason contextually over overlapping and complex auditory inputs, operate efficiently on edge devices, provide explainable outputs, and support scalable deployment - capabili8es that are not collec8vely addressed by any known prior art.

4. Detailed Description of the Invention

Workflow Diagram

4.1. Audio Chunking Module :

The system captures a con8nuous 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 8me-stamped rela8ve 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 8mestamps.

Example output:

{

"processor": "ASRProcessor",

"start_8me": 12.45,

"end_8me": 12.75,

"payload": {"text": "Turn on the light", "tokens": [...]},

"confidence": 0.92

}

Diariza8on Processor :

Detects speaker iden88es and boundaries across overlapping speech segments.

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

Event Detec8on Processor :

Recognizes non-speech acous8c events such as door knocks, typing, or ambient signals.

{"events": [{"label": "door_open", "8me": 12.62, "score": 0.91}]}

Paralinguis8c Processor :

Analyzes voice aaributes like pitch, energy, and emo8on to infer speaker affect and tone.

{"pitch": 120, "energy": 0.34, "emo8on": "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 8me-synchronized integra8on layer that aggregates and aligns outputs from all processors using a common 8me base.

It merges text tokens, speaker informa8on, detected events, and paralinguis8c cues into a unified mul8modal data structure.

Fusion methods may include:

(a) Token-level enrichment: augmen8ng ASR tokens with speaker, event, and emo8on metadata.

(b) Frame-level concatena8on: combining numeric feature vectors for temporal correla8on.

(c) Confidence-weighted merging: priori8zing high-confidence processor outputs.

The fused output is formaaed for downstream reasoning and contains seman8cally 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 par8al summaries during speech and final summaries at uaerance comple8on.

Finaliza8on is triggered by :

Detec8on of speech silence exceeding a threshold (e.g., 500-800 ms).

Explicit end-of-uaerance markers.

The Reasoner's output evolves dynamically, revising par8al hypotheses into stable, high-confidence summaries once finaliza8on occurs.

5. Workflow Summary :

Audio input is con8nuously captured and segmented into overlapping chunks with lookahead.

Each chunk is processed concurrently by ASR, Diariza8on, Event, and Paralinguis8c modules.

All processor outputs are 8mestamped and sent to the Fusion Engine.

The Fusion Engine synchronizes outputs and produces enriched representa8ons.

The Reasoner consumes fused data incrementally to generate par8al summaries.

Upon uaerance comple8on, the Reasoner generates a comprehensive final summary.

This workflow ensures real-8me adaptability, temporal precision, and context-aware mul8modal understanding.

6. Implementa.on and Deployment :

The inven8on 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 con8nuous audio handling, bounded queues for backpressure management, and state persistence for con8nuity across chunks.

5. Advantages of the Invention

(a) Low Latency: Lookahead buffering enables fast, context-aware inference without wai8ng for full uaerance comple8on.

(b) Parallel Modularity: Each processor operates independently, allowing scalability and hardware op8miza8on.

(c) Temporal Fusion: Consistent 8mestamping ensures accurate alignment of mul8modal signals.

(d) Incremental Reasoning: The Reasoner provides both immediate feedback (par8al summaries) and refined outputs (final summaries).

(e) Versa8lity: The framework supports both on-device and cloud-deployed models, enabling flexible deployment on edge systems.

(f) Robustness: Confidence-based fusion mi8gates 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-8me understanding of commands and emo8ons for responsive AI interac8on.

2. Mee8ng & Call Analy8cs: Speaker-aware transcrip8ons and automa8c summaries of conversa8ons.

3. Surveillance & Security: Detect cri8cal acous8c events and monitor environments in real 8me.

4. Healthcare & Therapy: Track speech paaerns and emo8onal tone for therapy and health monitoring.

5. Human-Robot Interac8on: Enable robots to understand commands, context, and emo8ons instantly.

6. Media Monitoring & Broadcast Analysis: Analyze live audio streams for speakers, events, and emo8onal cues.

7.Customer Support and Call Centers: The system can analyze customer speech, detect emo8ons such as frustra8on or sa8sfac8on, and guide agents or automated bots to provide more effec8ve and empathe8c responses.

8.Legal and Forensic Audio Analysis: The inven8on can aid in analyzing recorded speech for stress, decep8on, or emo8onal indicators, suppor8ng inves8ga8ons or courtroom analysis.

9.Emergency Response and Disaster Management: The inven8on can assist in analyzing distress calls or field communica8ons in real 8me, iden8fying emo8onal urgency and guiding rapid, informed decision-making.