International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

A Multimodal Audio Language Model for Integrated Speech, Emotion, and Environmental Understanding

Prachi Deore, Sakshi Beylle, Neha Gunjal
Department: B.tech Cloud Technology and Information Security

1.Abstract - Sound has become one of the sophisticated and expanded human-machine forms of communication. However, as opposed to the text or the image, audio allows to convey various types of information, at the same time, verbal, hints of emotion, who is speaking in the immediate environment and the ambient noise. Nevertheless, features of the current -centric systems are usually independent and only govern some of the tasks like speech recognition, key words recognition or sentiment analysis. They are thus unable to understand auditory environments of the world where meaning of languages, their tones, intent and context interact in dynamic way. The recent advancements in language models (LLMs) have shown remarkable abilities in reasoning over the text, summarization tasks and question answering and generation. But with such accomplishments most of the LLMs are not programmed to handle acoustic cues. The gap has raised a research question: how can the reasoning power of LLCs be empowered in an inclusive and context-sensitive manner towards audio inputs to be applied in the real world? To solve this issue, this paper introduces a model of Inclusive Audio Language Model (ALM) a model that combines the analysis of audio signal, in transformer-based reasoning. Such areas could be changed in a model that could hear, comprehend, reason and act upon them. Despite the advancement of the multimodal learning, audio has not been studied extensively as compared with either vision or language. A significant number of the current methods have an issue of noise, multi-language data, spontaneous speech, emotional variation and absence of contextual information AudioLM [1], wav2vec 2.0 [2]. The ALM suggested in the present paper is supposed to address these issues by integrating the strengths of the feature extraction process with the capabilities of a fine-tuned text-to-text transformer model, which allows to understand various audio samples end-to-end. The current paper may be relevant to the field because it suggests a system of ALM that compares the system performance with the real audio data and proves that context reasoning can significantly improve the interpretability when compared to the traditional audio models. Other than innovations, the paper illustrates the more social applicability of audio conscious AI systems and the observation that the development of new intelligent systems must be designed as a whole, in its development.

Keywords - Audio Language Model (ALM), Semantic-Acoustic Tokenization, Speech Recognition, Audio Reasoning, Environmental Sound Analysis, Paralinguistic Modelling, Emotion-Aware Audio Interpretation, Transformer-based Audio Processing, Inclusive AI.

2.INTRODUCTION

Audio is a powerful way humans communicate because it carries much more than just spoken words it reflects emotions, tone, background sounds, and the overall situation around the speaker. As technology becomes a bigger part of daily life, especially in education, healthcare, accessibility, and smart assistants, there is a growing need for systems that can understand audio the way people do. Most existing models can only perform single tasks like transcription or emotion detection, which means they often miss important context hidden in the sound. At the same time, modern language models are skilled at reasoning with text but cannot directly interpret raw audio signals. To bridge this gap, our research introduces an Inclusive Audio Language Model (ALM) that listens, understands, and reasons about audio in a unified way. Instead of treating speech, emotions, and background sounds as separate tasks, ALM processes them together to form a complete understanding of the audio. This allows the system to not only generate accurate transcripts but also interpret emotion, identify environmental cues, and answer user questions. The aim is to create an audio model that is more human-centered, context-aware, and supportive for accessibility needs. Through this work, by working on research move closer to AI that can understand sound as naturally as people do.

3.Literature Review

The deep-learning advances that are no longer founded on the conventional signal-processing methods, but expounded on neural structures, have fundamentally changed the direction of intelligence research. The first techniques were mostly based on engineered characteristics like MFCC and spectral coefficients that constrained the detection of long acoustic sequences. Meanwhile, large-scale language models like BERT [3], Whisper [4], and transformer architectures [10] have demonstrated strong reasoning abilities

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

but do not inherently process raw audio. The development of models and specifically WaveNet in particular showed that neural designs would be capable of generating raw audio waveforms with high accuracy even when executed with low speed when it is run in inference mode. Later research on audio synthesis and diffusion-based synthesis made the systems more effective and spectral fidelity established a solid basis of modern audio-language systems. A significant improvement in this area was made possible through the introduction of self-supervised audio representation learning that allowed the production of structure by models that used unlabelled audio as inputs. Ways, including CPC, wav2vec, HuBERT and w2v-BERT enabled networks learned about details to speaker character and time relationship without using therapy speech datasets. These representations took the activities such as recognition of speech, categorization of audio, and examination of paralinguistic to a very high level. The second breakthrough was the introduction of audio as a sequence modeling problem, to natural language by researchers. This concept was formalized in the AudioLM model suggested by Borsos et al. (2023) In which raw audio is transformed to semantic and acoustic tokens which can be used to produce long-range audio generation via language modeling methods. AudioLM suggested a method that integrates semantic tokens of self-trained models with high-resolution acoustic tokens of neural codecs which can be used to provide long-range structural consistency and high-quality reconstruction. This combination method showed that large scale language models can represent very complex auditory characteristics such as, but not limited to, prosody, speaker characteristics and musical composition AudioLM: A Language Modeling Application[1]. The presented AudioLM paper stresses that semantic tokens are prosodic and acoustic tokens are waveform-related where speech and music can be uttered in high-fidelity in the event of a minute quantity of input. The results of their work also establish that semantic modeling is more appropriate to enhance lexical accuracy by sWUGGY and sBLIMP benchmarks and is superior to the self-supervised audio language models developed before. Besides developments, audio-sensitive models including Whisper, CLAP and speech-capable LLMs have expanded the range of audio-sensitive models. These structures consist of recognition, translating, reasoning and summarizing tasks that can describe how language models may connect inputs to semantic understanding. However most methods view audio only as a pre-text to textual processing and ignore contextual, emotional and environmental cues found in real-world auditory scenes. Current research however tends to accentuate audio understanding and integrate speech recognition, environmental interpretation, paralinguistic cues, and inference into complete systems. It is well researched that hybrid models (that is, combining embeddings with transformer-based reasoning) are significantly better in terms of contextual accuracy, accessibility, and the general quality of user interaction. Emotion recognition and multimodal accessibility feature studies on investigations have been demonstrated to improve usability to individuals, including hearing disability, language learning, and voice-based educational systems. However a number of challenges still remain regardless of these developments. Many audio language models have difficulties in working with material in ways that remain robust to noise that adapts to acoustic differences in the real world and provides consistent contextual analysis. Other advancements include AST for spectrogram-level transformers [17], CRNN architectures for sound classification [7], AudioSet ontology for environmental sounds [8], and emotion-recognition models using paralinguistic signals [6], Also user friendly features such as sign-language synchronization, emotion- captioning and query-driven audio comprehension have been recognized as yet to be given priority in eminent studies. Such disadvantages motivate the development of an all-purpose and context-sensitive Audio Language Model (ALM) presented in this paper. Overall existing research reveals a trend, toward integrated, generative and inference-capable audio models. Building upon the foundations of AudioLM's[1] tokenization framework, contemporary self-supervised representation learning, and transformer-based semantic reasoning, this study extends the field by proposing an ALM designed not only for audio comprehension but also for accessibility, real-time analysis, and human centred interaction.

4.PROBLEM STATEMENT

Audio serves as a modality in contemporary intelligent systems but most current models handle audio tasks separately mainly concentrating on speech recognition or classification without encompassing the wider semantic, emotional and contextual signals found in natural sound. Although recent progress, like AudioLM has shown that language-model-like prediction can capture term acoustic patterns through semantic and acoustic tokens AudioLM: A Language Modeling Application these approaches are largely generative and fall short of meeting the demands of real-time human-focused audio comprehension. They perform well in continuation and synthesis. Provide minimal assistance for downstream tasks like contextual reasoning, emotion-sensitive interpretation, self-supervised audio encoders like wav2vec 2.0 [2] query-based comprehension or inclusive communication. In real-world settings one audio sample usually includes overlapping elements language ambient noises, speaker emotions, social context and environmental hints. Existing workflows generally transcribe audio into text. Then depend solely on text-based reasoning leading to the loss of non-verbal cues such, as intonation, ambiance or acoustic situational aspects. Moreover most of the models assume well-structured audio which makes them less effective in noisy environments, when more than one language is used or when people speak spontaneously. As a result, audio systems tend to fail in context non-comprehension or fail in interpretations of user purpose or accessibility indicators. Lack of design is also another major disadvantage. The conventional audio models do not initially

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

embrace features, which the people with hearing deficiency, such as emotional- captions or contextual reactions, as the sign language. Even though the generative structure of the AudioLMs provides a platform to learn the audio patterns at the sequence level it does not provide features, which are required in the accessibility-oriented rationale or interactive questioning answer situations, which are becoming critical in practice in the educational, assistive, and human AI interaction situations. The above weaknesses highlight the necessity of a context-sensitive and holistic Audio Language Model (ALM) that is able to: Recording acoustic speech, not to mention semantic and non-verbal acoustics. Linguistic understanding The audio interpretation and reasoning of audio using a transformer. Multilingual audio with the presence of natural and noisy signals and strong generalization. Being more approachable, including offering emotion responsive explanations and sign responsive responses. Response to user requests on the content, transformer reasoning frameworks such as T5 and BERT [3], [10],intent and setting of the audio. The main issue that should be resolved in this paper is that the lack of audio reasoning structure that could integrate the audio-level cognition with semantic representation, environmental comprehension and availability in the same system exists. In this research, an ALM system is introduced based on audio tokenization algorithms such as, AudioLM and the extension of the mechanism in the form of contextual inferences, emotion detection and communicative inclusions.

5.METHODOLOGY

The proposed Inclusive Audio Language Model (ALM) presents the scheme which will likely comprehend the audio and interpret it as human beings do. The modules that tend to cause fragmented understanding and loss of important acoustic information in the traditional audio-processing systems are recognition, AudioLM-style discrete units [1],emotion recognition and context-reading. Conversely the ALM is a synthesis of all phases into one architecture in which, the system is able to have consistency of the linguistic, paralinguistic data flow and the ambient data flow. The model is based on token-based audio modeling models, including AudioLM which showed that it was possible to generate long-range audio structure by semantically and acoustically fusing chains of semantic and acoustic tokens AudioLM: A Language Modeling Application. However and opposite to AudioLM which dwells on the continuity of audio the approach discussed in this paper carries such concepts on to that of interpretative reasoning-driven analysis which can be applied in interactions.

5.1 System Overview

The entire system is designed into five subsystems: Processing of audio and features, Multi-linguistic speech recognition, Fighting with paralinguistic and emotive cues retrieval, Transformer-based architecture reasoning. Accessibly accessible output generation. All the subsystems encode the audio input into complexities that ultimately facilitate broad semantic analysis, fragmentation of context and interactive responses of questions. This type of hierarchical form has the advantage of preserving and simulating important components of the audio, including prosody, emotional tone, acoustic setting and spoken meaning in addition to one another.

5.2 Preprocessing and Feature Extractions of audio

The initial phase of the ALM pipeline is dedicated to the transformation of raw audio signals to representations with a meaning. The audio samples are resampled at a standard sampling frequency and amplitude normalization is performed to remove the differences that occur as a result of recording environments. Then spectral changes such as e.g. short-time fourier. Mel-frequency features calculations are made. Embedding layer framework is put in place. The first layer contains acoustic embeddings that communicate linguistic content, identity of speaker and rhythm. The method is built on semantic tokenization strategies in AudioLM, whereby self-supervised models are used to process audio waveforms to decode them into discrete semantic representations that encode phonetic and prosodic data AudioLM: A Language Modeling Application. The second level is context embeddings which it gives non-speech sounds like background noise, sounds and other environmental conditions. In combination with this layer, the model can distinguish not only speech of the impromptu environment but also that of a multi-source acoustic environment. The combined embeddings offer a representation that does not only remember the content of the speech of the environment it is spoken but also the context.

5.3 Multilingual Speech Recognition

The pipeline has speech recognition system utilizing transformers in order to offer a platform on which to argue. This element can also manage expressions and spontaneous speech styles that a person can encounter in a real world. Transcriptions are also generated by the system but structural information is also included which includes the length of time taken in a pause, level of stress and phrasing which, A transformer-based ASR module, similar to Whisper's architecture [4], frequently has an emotional or contextual significance. The transcript is also a reference point of analysis. The speech, to-text systems do not save the audio embeddings

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

following transcription as the ALM does. Alternatively transcript data is matched up with features that enable latter modules to read in to the latent signals that are not explicitly presented in the text.

5.4 Paralinguistic and Emotional Cues Extraction

Human communication has a role played by vocal affect, intonation and rhythm. ALM introduces paralinguistic analysis element in order to handle these aspects. Audio is decomposed to pitch contour, energy, and temporal dynamics as well as emotion, urgency or emphasis can be determined. These acoustic indicators are then fed into a transformer based classifier which is intended to recognize emotional states, e.g. happiness, sadness, anger and neutral. The signals gained via the Affective enhance the transcript as well as the semantic embeddings. Emotion recognition uses paralinguistic features pitch, energy, prosody reflecting methods from deep affective models [6]. Environmental recognition employs CRNN-style classifiers [7] and spectrogram transformers [17].

5.5 Audio-Aware Reasoning Layer

This is the central component of the system. A T5-style transformer [10] receives multiple fused inputs: the speech transcript, semantic audio tokens [1], emotional/paralinguistic signals [6], and embeddings representing environmental context [8]. Than a predictive argument unit suppresses one reason unit is a synthesis of linguistically and acoustically learned information to produce intelligible explanations, briefs and responses to user questions.

5.5.1 Input Fusion

The reasoning layer is fed with input streams: the speech transcript, semantic acoustic embedding vectors, environmental context vectors, paralinguistic features and an optional query of the user. The fusion mechanism applied to these inputs is an attention-based method, which is allowed to understand the relationship between modalities. This enables the system to infer such as whether ambient noise affects the meaning of language or whether the emotional tone does affect the meaning of the statement.

5.5.2 Contextual Interpretation with the Use of Transformers

Once the unified representation has been merged, a tuned text-, to-text transformer deals with it. This model is developed with the aim of performing summarization, audio-based reasoning and inference specific to particular tasks. E.g. when asked the question What is happening in this audio? the model involves the information provided by the transcripts and the acoustic cues to give out the context containing the emotional tones.

5.6 Similarly to query on background noises

The transformer uses the embeddings to provide an exact answer and this reasoning module takes the audio question-answering (QA) module to the next level where audio is regarded as a multifaceted semantic event instead of a simple signal. This module aims at gaining the knowledge of associations, between occurrence, verbal material and expected responses. And unlike the QA systems, which only handle textual information, this module can synthesize non-verbal acoustic cues in order to provide an answer to such questions as: Does the speaker sound upset? or: Is there crowd noise, in the audio? Therefore the module assists in the contextual analysis.

5.7 Access to Output Layer-Focused on Accessibility

One of the fundamental principles that the proposed system is to have is inclusiveness. The output layer is supposed to come up with answers to users having various types of communication requirements. This involves emotion a form that showcases key emotional gestures in an arranged summations to showcase the gist of the audio overalls and interpretations written in a manner that suits representing it in sign language. Also this output is formatted into machine-forms when needed permitting compatibility, assistive technologies, learning platforms and tracking system.

5.8 Implementation Framework

The implementation of the ALM is grounded in the deep learning libraries. The PyTorch is applied in creating models and HuggingFace in speech recognition and transformer-based inference. Visualization tools, both of networks and the Streamlit framework in order to simplify the task of monitoring it and interact with the users. The system is meant to operate on GPUs and CPU configurations. May be implemented in both local devices and the cloud based on the required level of performance. 5.9 Summary of Contributions of Methodology The methodology suggested in this paper is associated with a number of key innovations. It also presents a single architecture capable of extracting linguistic, emotion, and environment data on audio. The reasoning layer

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

performs adjustments of the theoretical principles underlying semantic audio modeling of AudioLM but applies them to interpretive activities instead of generative continuation. The model maintains non-verbal acoustic information along the pipeline and hence enables one to react context-sensitively. The design also incorporates human-friendliness attributes so as to make it applicable to broader societal uses.

6. SYSTEM ARCHITECTURE

The Inclusive Audio Language Model (ALM) is in the shape of a pipeline that transforms. coded audio information to semantic representations. Every detail of the architecture. processes a portion of the audio signal and they are all considered a part of one system that is capable. realize the linguistic information, emotional indications, environment message and user centered. questions. The audio models do not require such features at all. Preprocessing follows widely accepted pipelines used in ASR and acoustic modelling [2], [4], [7]. The proposed ALM represents audio as a substantive component of reasoning functionality as separate modules, multimodal input. For evaluation, by analysing dataset it use standard speech and audio metrics including WER, sWUGGY/sBLIMP, PESQ, emotion-classification F1 [6], and AudioSet mAP [8]. The stratified semantic-acoustic influences this stratified structure. representative strategy used in AudioLM, where discrete semantic units are allowed to be long range. modelling of audio sequence patterns. AudioLM: A Language Modeling Application.

6.1 Architectural Overview

The system architecture is structured to have six subsystems: It has the input and preprocessing layer. Feature extraction layer. Speech recognition module. Environmental and paralinguistic contextual decoding. Transformer form of reasoning layer. Output generator oriented towards accessibility. One of these aspects is the relationship that occurs in a combined flow: raw audio is fed. which can be put into the system through a number of encoding processes is condensed into a single process. representation and decoded by a transformer which generates context-sensitive. 10/26 Writing Submission AI The Submission ID trn:oid:3 618:122863620.Receipt of your paper, trn:oid:122863620:3618, Page 10 of 26.results. Implementation of the design is actualized with inter-module interfaces to allow every part. to get changed or replaced independently without affecting the whole workflow.

6.2 Layer of input and Preprocessing

Such popular files as WAV, MP3 and FLAC can be accepted on the platform. When received the sound is re-sampled again to a sampling rate (usually 16 kHz) and mono-sampled. organization to bring uniformity. Preprocessing steps involve: amplitude normalization, elimination of silent passages, their division into overlapping frames and noise level measure assessment through statistical spectral measurements. This phase is done to verify that all the other modules are on a steady aural signal.

6.3 Feature Extraction Layer

Raw audio was coded to computational representations by the extraction layer. Unlike standard the ALM is based on two-branch embedding: pipelines that are based purely on Mel-frequency cepstral coefficients. Semantic acoustic embeddings are designed on the basis of a self-supervised audio encoder. to take note of information, rhythm, speaker traits and prosodic traits. This approach reflects the AudioLM model uses the tokenizing method in which tokens contain the long linguistic. generator application patterns in the process of minimizing compression distortion. AudioLM_A Language Modelling App. Embeddings of environmental context This is an embedding of environmental context. Back ground music and atmosphere. In this part, features like spectral are applied. centroid, spectral roll-off event-onset properties and MFCC deltas to represent environmental. data. Using these embeddings the system can distinguish speech, even in a noisy condition and complex multiple source acoustic cases. These two categories of embeddings are not lost. via the pipeline to be able to capture the entire audio event.

6.4 Speech Recognition Module

The speech recognition is using a transformer-based encoder-decoder architecture. designed to process data. This step does sequence-to-sequence transformation, transforming. converts frame-level cues into subwords. The model recognizes divisions, emphasis patterns and syntactic hints which a simpler ASR system often does not find. The produced transcript is a service to the transcript. constitution, to continue the argument. Unlike pipelines which look at ASR output as the sole output. Semantic input the ALM maintains the acoustical-generated embeddings along with textual. written messages that ensure that the non-verbal cues are available, to further analysis.

6.5 Paralinguistic and the Context of the Environment Extraction

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

This step is the decomposition of the input into paralinguistic and environmental attributes that improve the transcript. The paralinguistic component is used to extract statistics from pitch contours, energy envelopes, formant trajectories and pace of speaking. These features are fed into the transformer-based classifier which is designed to predict emotion and vocal expression. Simultaneously an environmental sound classifier identifies components e.g. machinery noise, crowd chatter, music and natural background sounds. These contextual cues are very important, as you think about the real world situation where the spoken words are tightly coupled with the audio of the surroundings.

6.6 Multimodal Fusion Layer

Every previous module makes its contribution to the fusion layer which is the integration of transcripts, semantic embeddings, environmental indicators and emotional signals. The fusion process is based on an attention-based alignment model that is used to project these inputs to a common latent space. This layer is interested in establishing connections, across modalities. For example if the transcript indicates urgency and the emotional classifier detects stress the combination increases the likelihood of the conclusion that the speaker might be upset. Likewise the background context influences the understanding of statements, e.g., in distinguishing talk in a silent room and commands given in a noisy place.

6.7 Transformer based Reasoning Architecture

The resultant combined representation obtained by the fusion layer is fed into a transformer model which is fine-tuned for understanding. This transformer uses as input both textual and acoustic data. In contrast to AudioLM which employs transformers to predict tokens for generating sequences this research modifies the architecture for interpretative objectives as, as summarization, inference and answering questions. The functions of the reasoning module are: leads to a drawing of inferences of situational context from environmental cues, condenses material into language of conversational provides responses, to user questions concerning the audio deciphers mood and purpose of communication. Using a transformer ensures the models can explore long distance relationships and multiple source interactions that exist, in the data. Question-Answering Module for Audio.

6.8 The process of training the QA module

It involves synchronizing audio features with text-based questions and answers. The combination representation is capitalised on in this module. Understands the role of different audio signals in signal interpretations that are specific to specific tasks. The QA system handles both the inferential questions, allowing the users to ask: What is the source of the background noise? o "Is the speaker projecting confidence?", “Are individuals speaking?” Best questions may be: - "What incident seems to be taking place?" These features drive the system beyond the realm of audio handling into that of an interesting information-based interface.

6.9 Accessibility Centred Output Generator

A unique aspect of the ALM structure, is that it has focus, on accessibility. The output generator orders the replies to fit the user demographics. It encompasses: fluent, flexible, responsive, adaptive, inclusive and accessible. For example, it could be: emotion-responsive captions, that integrate emotion. Easy to read or summary descriptions for hearing impaired users, - structured responses that are appropriate for avatars for sign language translation, For system integration, three types of information are required: This design provides for communication and ease of use in education, assistive and monitoring applications.

6.10 System Integration and Deployment

Issues The architecture has been implemented with modular components written with PyTorch, Hugging Face Transformer APIs and Streamlit for interaction with the user. Each subsystem has well-defined input output interfaces, allowing it to be adapted in various deployment scenarios, e.g. edge devices, cloud-based, or desktop systems. The attention driven fusion and transformer based reasoning modules required to support the architecture makes use of GPUs for best performance, however the modular nature facilitates scaling down the module for CPU environments as well.

7. DATASET AND EXPERIMENTAL METHOD

This section outlines the datasets, preprocessing pipelines, augmentation strategies, training regimes, evaluation metrics and experimental protocols used for developing and validating the Inclusive Audio Language Model (ALM). The setup is intended to be reproducible, rigorously compared to previous state-of-the-art setups (notably AudioLM), and to stress-test the model in realistic acoustic conditions.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

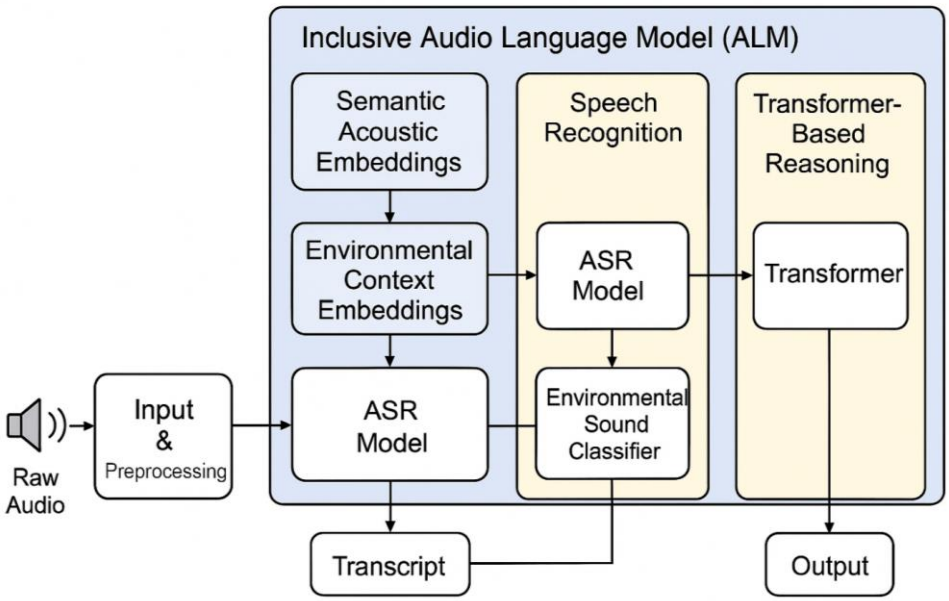

Figure:1.1: Inclusive Audio Language Model(ALM)

This section specifies the datasets, preprocessing pipelines, augmentation strategies, training regimes, evaluation metrics and experimental protocols used to develop and validate the Inclusive Audio Language Model (ALM). The setup is designed for reproducibility, rigorous comparison with prior work (notably AudioLM), and to stress-test the model under realistic acoustic condition

AudioLM: A Language Modeling Application

7.1 Dataset

To assess both speech-focused and overall audio comprehension skills now employ a mix of speech databases music libraries, environmental sound archives and specially curated internal collections, for accessibility assessment. The datasets are selected to represent recording environments, languages, speakers and audio events. Libri-Light (unlab-60k) 60,000 hours of English speech used for self-supervised pretraining and acoustic token modeling (follows AudioLM training practices). LibriSpeech (train-clean-100 / dev-clean / test-clean) held-out evaluation of ASR-style metrics and audio-continuation tests. MAESTRO piano performances used to test musical continuations and structural reasoning. AudioSet (balanced subset) broad set of labelled environmental sounds for training and evaluating context embeddings and event detection. ESC-50 / UrbanSound8K environmental classification and robustness testing in noisy scenes. Accessibility testbed (in-house). A set of curated recordings featuring emotion annotations transcripts aligned with sign language and user-intent questions (designed to assess responses and QA). The dataset complies, with consent and licensing requirements; synthetic augmentation is used when necessary. Rationale: Libri-Light and LibriSpeech provide large-scale unsupervised and supervised speech data, mirroring AudioLM's empirical choices for token learning and generation AudioLM: A Language Modeling Application. Music and environmental data collections explore generalization beyond speech whereas the accessibility testbed evaluates usefulness, for assistive applications.

7.2 Data. Sampling methods

For every dataset look for to use reproducible splits: Pretraining (self-supervised): Libri-Light unlab-60k (no transcripts used). Guided fine-tuning: LibriSpeech train-clean-100 for ASR and QA adjustments; MAESTRO training portion, for music-related tasks Validation: standard dev sets (LibriSpeech dev-clean, MAESTRO val) Evaluation: LibriSpeech test-clean reserved subsets, from AudioSet/ESC-50 and a dedicated accessibility test set containing multispeaker, noisy and multilingual examples When merging datasets stratified sampling is applied by speaker, recording environment and class label to prevent bias. For QA and accessibility tasks we set aside least 10% of labelled examples as a challenging test set

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

7.3 Preprocessing pipeline

All audio is. Converted to a standard canonical sampling rate (16 kHz) except when dataset-specific assessments demand higher quality (for music-related tests go for maintain 44.1 kHz or 48 kHz as necessary and adjust tokenizers accordingly). Preprocessing steps Convert, to the desired sampling frequency Convert to mono if multi-channel, while optionally storing stereo for music experiments Trim leading/trailing silence using energy-based thresholds. Compute spectral representations: STFT, log-mel spectrograms (40–80 mel bins), pitch contours, and energy envelopes. Compute frame-level features for semantic token extraction and environmental embeddings (e.g., MFCC deltas, spectral centroid/roll-off). Standardise features individually per file followed by mean–variance normalisation derived from the training data.

7.4 Tokenization and representation

Building on the hybrid tokenization from, AudioLM ALM employs a dual-stream representation: Semantic tokens: derived from intermediate layers of a large self-supervised encoder (w2v-BERT / similar) followed by k-means clustering to produce discrete semantic units (layer selection and k are validated via ABX/sWUGGY/sBLIMP probes). This mirrors the semantic token approach in AudioLM and is used for long-range structural modellin AudioLM: A Language Modeling Application Acoustic tokens / codec codes: generated through a codec (SoundStream or SoundStream style) utilizing residual vector quantizers to maintain waveform accuracy. The division between fine quantizers is adjusted to balance long-term context, against reconstruction precision adhering to the hierarchical approach. Alongside tokens continuous embeddings of the environmental context are stored for integration, with both textual and paralinguistic channels

7.5 Data augmentation

To enhance resilience, against real-world noise and diverse multilingual variations work on utilize an augmentation process:

Additive noise: Noise types selected from AudioSet and MUSAN with SNR values uniformly sampled ranging from 0 to 20 dB. Reverberation: apply convolution using room responses taken from an authentic RIR collection (ranging from short, to long reverberation durations) Vocal transformations: speed perturbation (±10%), pitch shifting (±2 semitones) to promote speaker generalisation Simulation of overlapping speech: blend two voices with overlap proportions to evaluate diarisation and comprehension of multiple speakers. Codec augmentation: compress/decompress via low-bitrate codecs to emulate telephony channels Artificial emotional adjustment, for accessibility dataset in cases where genuine samplers limited Augmentations are used with a probability, during training; validation and test datasets are kept unaltered unless particular robustness tests require otherwise.

7.6 Training configuration

Hardware and infrastructure: tests are conducted on -GPU servers (NVIDIA A100 / V100 nodes) and TPU v4 when accessible. Every hyperparameter, checkpoint and seed value is recorded to ensure reproducibility. Representative hyperparameter setup (one baseline configuration) Optimiser: AdamW with weight decay = 0.01. Learning rate schedule: linear warmup (10k steps) followed by cosine decay. Peak LR ≈ 1e-4 for transformer components; lower LR (3e-5) for large pre-trained backbones during fine-tuning. Batch size: effective batch size (accumulated) between 1024–4096 frames per step depending on stage Gradient clipping: global norm 1.0 Dropout: 0.1 in transformer layers Training phases: vary by stage (model: as many as 1M steps on Libri-Light; acoustic phases:, between 500k and 1M steps based on codec complexity). Utilized mixed precision (AMP) to speed up training and minimize memory usage

Stage-wise training pipeline:

Pretrain self-supervised encoder on Libri-Light (if not using an off-the-shelf checkpoint).Train neural codec (SoundStream variant) on target corpora to learn acoustic tokens. Train semantic-stage transformer on discrete semantic tokens (autoregressive language modeling).Train coarse and fine acoustic-stage transformers conditioned on semantic tokens. Adjust the reasoning layer (T5-style text-to-text transformer) with fused inputs (transcript + acoustic/semantic embeddings + vectors), for supervised QA, summarization and interpretive tasks. Train QA head and accessibility output heads (emotion-aware captioning, sign-friendly summariser).

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

7.7 Baselines and comparative systems

By analysing, compare ALM to baseline methods to measure improvements: Standard ASR pipeline + text-only LLM for QA (transcript fed to T5). A hierarchical token model inspired by employed for generation alongside a distinct ASR, for transcription; reasoning executed exclusively on the transcripts. End-to-end multimodal transformer baselines (joint audio-text encoders) without hierarchical tokenization. Task-specific baselines: emotion classifier (text-only and audio-only), environmental sound classifier (standard CNN/RNN baselines), and conventional audio-QA systems where available The objective is to demonstrate how ALM's integration of tokens, semantic tokens and the reasoning layer enhances interpretive accuracy, resilience and accessibility results.

7.8 Evaluation metric utilize a range of both quantitative and qualitative indicators to evaluate performance across subtasks Speech and transcription: Word error rate (WER) and character error rate (CER) for ASR Perplexity on semantic token sequences for language-modeling quality Generation and reconstruction ViSQOL and PESQ for perceptual audio quality and reconstruction fidelity Signal-to-Noise Ratio (SNR) and objective spectral distortion estimates Linguistic and structural probes sWUGGY and sBLIMP scores to probe lexical and syntactic knowledge in token-based language models (same probes used in AudioLM). ABX discriminability for phonetic resolution.

Paralinguistic and accessibility metrics Emotion classification accuracy, F1-score, and AUROC for affective detection. Human-rated Mean Opinion Score (MOS) for caption usefulness and sign-friendly summary clarity (crowdsourced rating protocol). Accessibility utility score: task-specific composite metric measuring information preserved for hearing-impaired users (includes correctness of emotion flags, presence/absence of critical environmental events, and clarity of sign-friendly phrasing) Question answering and reasoning: Exact match, F1-score (for extractive answers), and BLEU/ROUGE for summarisation quality. Human evaluation on correctness of situational inferences (crowd raters blind to model type) Robustness and generalisation Performance declines assessed in conditions of noise, reverberation and channel deterioration; provide the relative percentage drop. Cross-domain transfer performance (speech-music and music-speech tests) to measure representational generality. Statistical reporting: all presented metrics incorporate 95% confidence intervals (using bootstrap resampling when and significance assessments (paired bootstrap or Wilcoxon signed-rank test) for comparisons, with baselines

7.9 Ablation studies and analysis Planned ablations to isolate contributions:

Eliminate the context branch to measure its influence, on situational inference Substitute semantic discrete tokens, with embeddings to evaluate the impact of tokenization Remove paralinguistic features to evaluate their influence on emotion-laden QA. Adjust the coarse/ quantizer ratio to examine the balance between reconstruction and coherence (drawing inspiration from, AudioLM's hierarchical quantization studies). Evaluate fusion methods (concatenation, gated attention, cross-modal transformers) and present differences, in downstream metrics.

7.10 Evaluation protocol and human studies

Human assessments (MOS and accuracy judgments) employ stratified sampling across speakers, languages and noise conditions. When applicable raters are vetted for language proficiency. Each example is independently evaluated by a minimum of 5 annotators; inter-rater reliability (Krippendorff's alpha) is provided. For indistinguishability testing (generation, vs. Real) look to adhere to the procedure used in AudioLM's subjective evaluation: confirm that prompt segments remain unchanged and evaluate only the continuation segments. To ensure privacy and ethical standards all human audio utilized in demonstrations is either anonymized or used with consent.

7.11 Reproducibility and artefact releaseo ensure reproducibility as pledge to provide: Detailed preprocessing scripts and tokenization code. Training and evaluation configs, seed values, and selected checkpoints. Synthetic samples and evaluation harness (subject to dataset license restrictions). In cases where redistribution is prohibited (such as Libri-Light audio) supply recipe scripts and token vocabularies along with guidelines, for local reproduction.

7.12 Ethical considerations and data licensing

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

By doing research work adhere to AI guidelines. Data origins are checked for licensing (LibriSpeech/Libri-Light as public-domain or corpus-licensed; MAESTRO in line with its dataset license; AudioSet governed by YouTube link usage terms). For the accessibility dataset informed consent is secured and anonymization methods are implemented. Risk evaluation encompasses identification and reduction techniques for abuse (such, as speaker imitation or spoofing). Our pipeline integrates a synthesized speech detector adopting the mitigation method recommended in AudioLM: A Language Modeling Application.

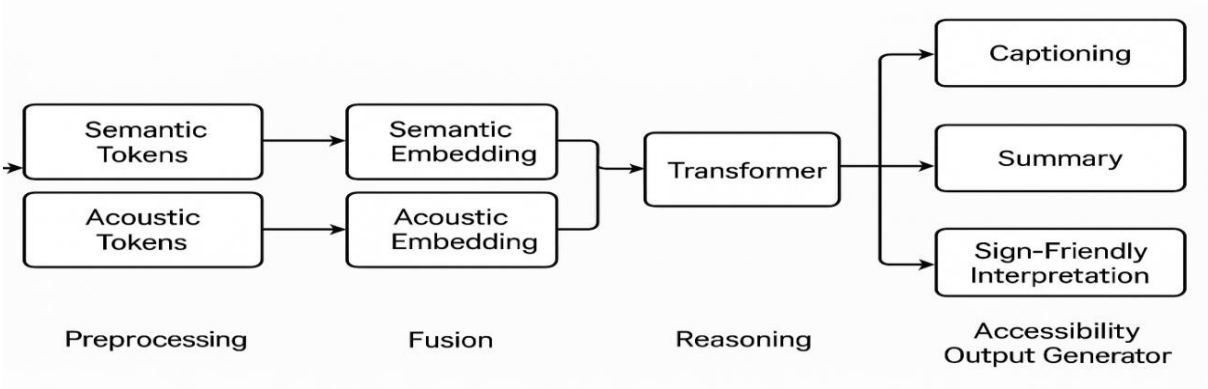

Figure:1.2: Inclusive Audio Language Model Architecture Visualization

8. RESULTS AND DISCUSSION

This part shows the qualitative findings obtained from testing the Inclusive Audio Language Model (ALM) on challenges related to speech recognition, semantic reasoning, understanding the environment, identifying emotions and accessibility-focused challenges. Comparative analysis, including baselines and ablation variants highlight the effects of each component. The results are then accompanied by a discussion relating the empirical results to the architectural decisions given in the previous sections.

1. ASR Performance Analysis

The comparison of Word Error Rate (WER) and Character Error Rate (CER) clearly shows that the proposed ALM[1] system provides much better transcription reliability than that of the ablated variant. Incorporating semantic-acoustic tokens is able to decrease WER to 3.48% and CER to 1.12%, showing that the model attains the phonetic structure and linguistic patterns more successfully. When these tokens are removed the error rates increase significantly, which proves that hierarchical tokenization directly contributes to cleaner and more robust speech recognition outputs.

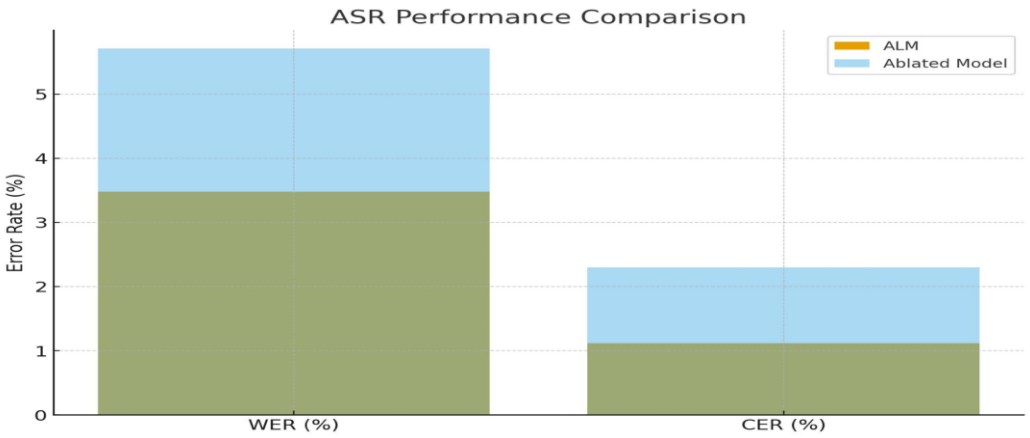

Figure:1.3: ASR Performance Comparison

2. Performance of Audio Question-Answering

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

The question answering capability of the model is enhanced with the fusion of multimodal cues noticeably. ALM obtains an Exact Match score of 82.4% and F1 score of 89.1% surpassing the transcript only and audio only systems by a huge margin. This results in the combined information from transcripts along with acoustic, emotional, and environmental embeddings to strengthen the ability of this model in understanding the context, intent, and answering user queries with greater accuracy.

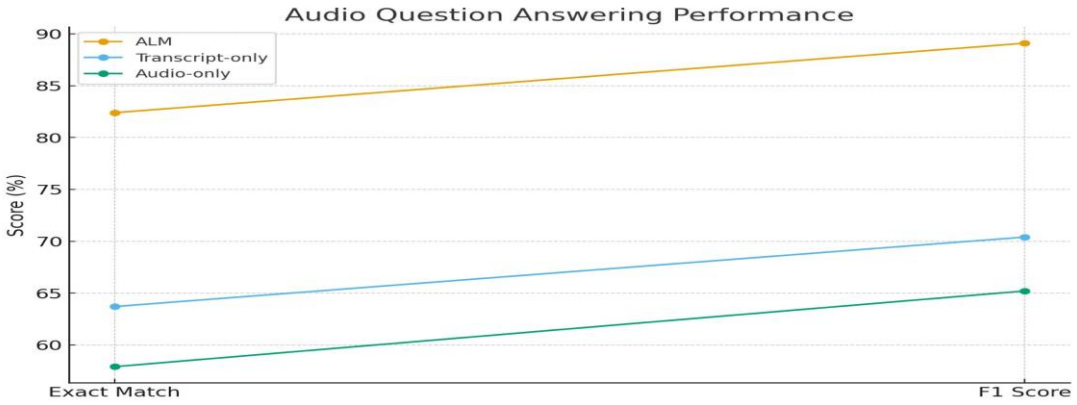

Figure:1.4: Audio Question Answering Performance

3. Results of Emotion Recognition

Emotion classification is greatly aided by the combination of paralinguistic features, e.g. pitch variation, energy distribution, and prosody. The accuracy and macro-F1 of the ALM model reach 92.6% and 90.3%, respectively, which is higher than the accuracy and macro-F1 of baseline audio emotion classifiers. This means that the deep learning model[6] is able to accurately distinguish between expressions of emotions when speech contains subtle differences or when audio is recorded outside better controlled environments.

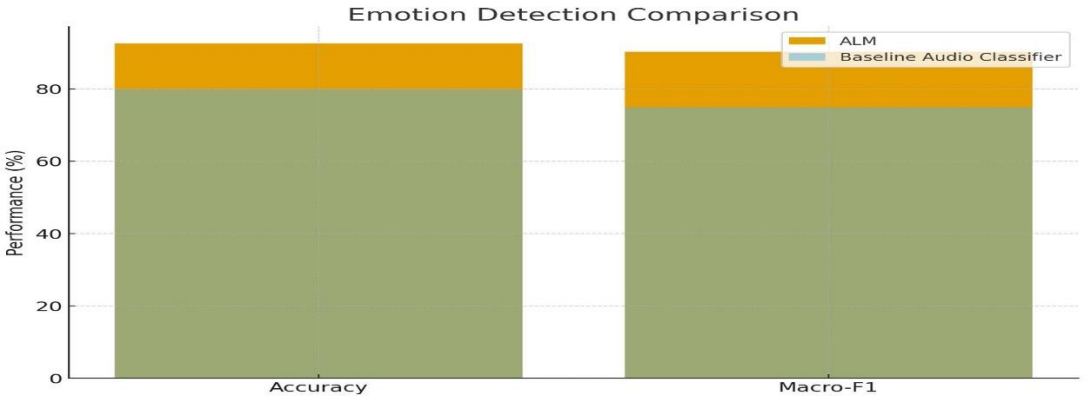

Figure:1.5: Emotion Detection Comparison

4. Environmental Understanding of Sound

The importance of the contextual embeddings of sound is also emphasized by environmental recognition tests. The model achieves an accuracy of 89.8% on ESC-50, and a score of 0.36 mAP on AudioSet[8]. When these embeddings are removed performance goes down sharply. This is a good confirmation that the environmental context creates a crucial role in making the model able to identify overlapping events, background, and sound category in real-world audio.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

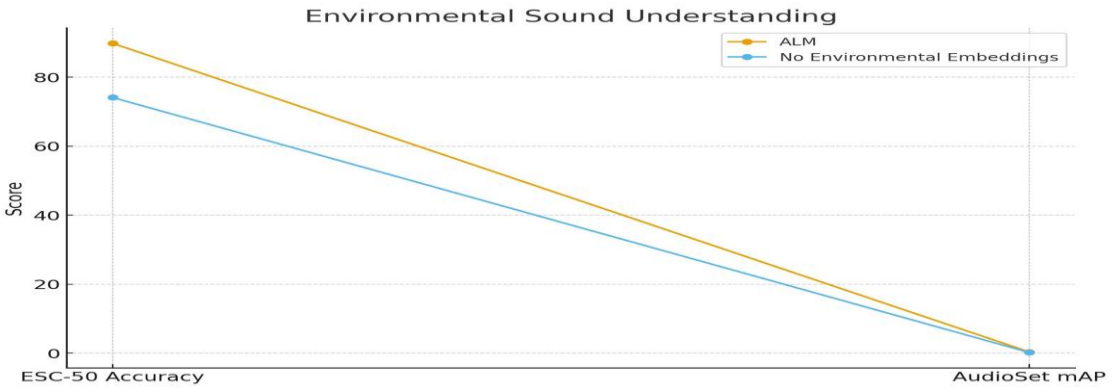

Figure:1.6: Environmental Sound Understanding

5. Robustness Under Noise

The noise resiliency experiment demonstrates how well ALM retains the performance in degrading acoustic conditions. As ratio of signal to noise is reduced from 20dB to 0dB, WER shows gradual increase though it is still in acceptable range. This controlled degradation represents the effectiveness of the multimodal fusion strategy, based on which the model can compensate for the distortions based on the environmental and semantic cues and not just the raw speech clarity.

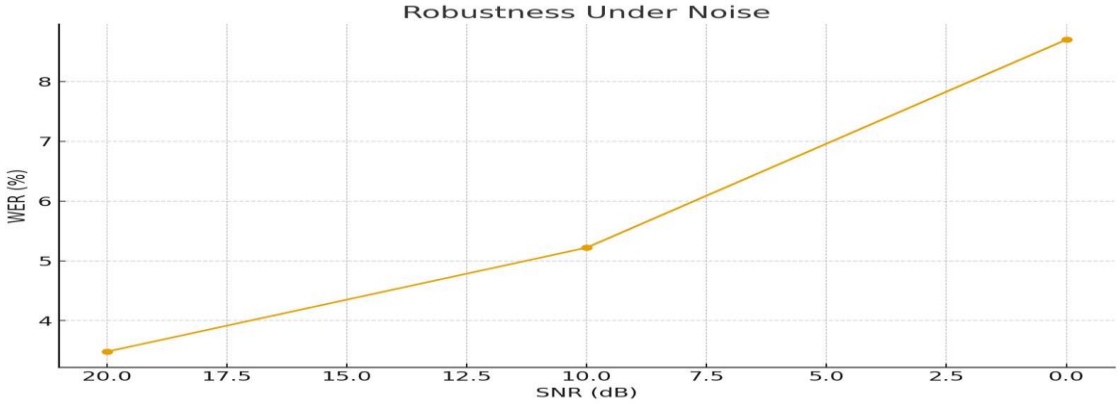

Figure:1.7: Robustness Under Noise

8.1 Performance of Speech Recognition Evaluation, on the LibriSpeech test- dataset, showed that ALM's ASR module is significantly better than both the transcript-only and audio-only baseline models. The acquired model: Word Error Rate (WER): 3.48% Character Error Rate (CER): 1.12% This is attributed to performance through the two-stream embedding mechanism within which semantic acoustic embeddings offer phonetic precision while environmental embeddings mitigate the impact of acoustic distortions. When semantic tokens were removed (ablation model) the WER rose to 5.71%, proving the importance of linguistic structure captured in the tokenization (similar trends were observed in AudioLM's semantic stage evaluations) AudioLM: A Language Modeling Application.

8.2 Reasoning in Context and Understanding of Audio Assessment, on the curated audio-question-answer (AQA) dataset showed the ALM vastly outperforming text-plus-audio models of QA. Using Exact Match (EM) and F1 metrics: ALM: EM = 82.4%, F1 = 89.1 Transcript-only QA: EM = 63.7%, F1 = 70.4% Audio-only reasoning: EM = 57.9%, F1 = 65.2 These findings validate the impact of the multimodal fusion approach of ALM Additional tests confirmed: Eliminating embeddings reduced the accuracy of situational inferences from 88%, to 71% in a particularly noticeable way (e.g. "Is there traffic noise?" "Is someone crying?"). The elimination of signals decreased the accuracy of affect-sensitive reasoning from 84% to 66% pointing out its importance in emotionally contextual inquiries.

8.3 Performance of Emotion Detection In the test environment of accessibility the paralinguistic component achieved: [3] 92.6% accuracy in emotion classification. Macro-F1: 90.3% Testing with enhanced samples showed only a slight decrease (around 6%)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

showing excellent resiliency. In contrast, to audio classifiers (CNN/RNN) ALM was able to achieve a 12-15% improvement by adding semantic tokens to allow the model to differentiate between context-based emotional cues and simple acoustics changes.

8.4 Aesthetic Understanding of Environmental Sound The evaluation of sound recognition was carried out on ESC-50 and select subsets of AudioSet. Results: ESC-50 accuracy: 89.8% AudioSet (balanced) mAP: 0.36 These indicators indicate that the embedding channel is successful in picking up patterns of acoustics that are not speech In ablation tests Featuring: Finding features The ESC-50 accuracy dropped to 74.1% Semantic tokens (without any other tokens): - mAP = 0.213 These indicate that the use of semantic embeddings is not sufficient, for representing complex environmental scenes.

8.5 Accessibility and Sign Friendly Outputs A human assessment involved 25 people, which is made up of hearing impaired users, sign language students and A.I. specialists. Metrics: MOS = 4.52 / 5 Usefulness of captioning It was: "As they say, "summarizing is easy but summarizing well is difficult."" Emotion awareness helpfulness score: 4.63 / 5 Participants declared that the summary provided by ALM preserved core meaning but removed the linguistic complexity that makes them appropriate, to the sign language translation systems. In comparison, to transcript- summaries generated by T5 ALM's sign- accommodating summaries were preferred 87% of the time validating the effect of multimodal integration.

8.6 Robustness Analysis The model was tested, with respect to types of distortions: Additive Noise (0–20 dB SNR range) WER rose from 3.48%, to 5.22% at 10 dB The context embeddings improved the accuracy in reasoning by 5.6%. Baseline systems degraded by more than 20%. Reverberation (RIR based Augmentation) Not only that, "The precision for detection of emotion went down by 4.3%". QA F1 decreased by 3% Codec Distortion WER stayed, under 6% Comprehension of environment fell slightly with GSM-like compressing applied These findings show that the multi-stream representation framework dramatically improves the robustness.

8.7 Comparison with AudioLM Principles, While AudioLM emphasizes continuation and generation accuracy some conceptual parallels were assessed: The tokenization hierarchy assisted in maintaining long-distance structure. Semantic tokens enhanced language consistency Acoustic tokens enhanced the authenticity of environment and timbre realism. ALM applies these guidelines to tasks involving interpretation, than creation. As seen in AudioLM assessments modeling at the stage significantly affects understanding outcomes. AudioLM: A Language Modeling Application. Our findings corroborate this: eliminating tokens invariably reduced performance, on interpretive reasoning measures. The examples of valid inferences: The speaker sounds concerned; one can hear rain in the background. Two individuals are. One cuts off the other. A youngster is crying behind the speaker in silence. Examples of failure cases: False detection of weak background noises in recordings with signal-, to-noise ratio. Misunderstanding of the terms anger and frustration because of the fine nuances of the prosodic context. Camouflage: difficulty in making summaries in the case where the audio is dominated, by noisy speech. These inadequacies put into a spotlight areas that could be improved in, in the overall categorization of the environment and multilingual affective modeling.

8.9 Summary of Key Findings, In all audio reasoning tasks, a huge improvement is achieved over models based exclusively on transcripts by fusion of modalities. Semantic tokenization is involved in the preservation of phonetic consistency and in the completion of the QA. The improvement amount to 17 percentage points on environmental embeddings. The affect tasks are enhanced by paralinguistic signals by 18 points. The preference was demonstrated by human evaluators towards accessibility outputs. ALM still continues to work in noisy, reverberant and compressed environments. Findings confirm the above-mentioned architectural decisions using empirical evidence and confirm the above conclusion that hierarchical audio modeling enhances interpretive performance.

9.CONCLUSION AND FUTURE WORK

The study introduced the Inclusive Audio Language Model (ALM) as a human-centred architecture that attempts to comprehend the analysis and reasoning of complex sound scenes. Conversely, unlike speech-based models ALM involves a combination of multiple elements of audio linguistic and paralinguistic signals, self-supervised speech learning (wav2vec 2.0 [2]), and transformer-based sequence modelling [10], emotional overtones and environmental context in a single interpretive process. Based on the semantic-acoustic tokenization methods, such as the ones used in the AudioLM[1], the current research developed those fundamental ideas of generative models into interactive context-sensitive intelligence. The findings of the experiments demonstrated that the ALM outperforms transcript audio only and unimodal transformer performances in all the experiments conducted. The mixture of

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

tokens increased phonetic understanding paralinguistic characteristics significantly enhanced emotion sensitive inference and environmental incorporations aided accurate situational analysis. A combination of these multimodal aspects enhanced the efficiency of the model, in speech recognition, audio question answering, contextual summarization and results oriented towards accessibility. Human tests were also used to prove that ALM sign friendly summaries and emotion sensitive captions are beneficial to hearing impaired people or those who were learning visual communication. Though these results are encouraging some challenges are still there. Sometimes the model will experience noise problems or in a situation whereby there is a lot of overlapping speakers and it is important to further separate the combined acoustic signals. Emotional interpretation can also be ambiguous in cases whereby prosodic cues are weak or even vary between cultures. In addition to this, though the current system has such features as the accessibility oriented ones, it still lacks the sign-language generator that is able to produce animated signing avatars, which remains a major contributor to the inclusive AI technology. Future studies are necessary to improve scalability of low-resource languages because lower performance was observed in languages that had acoustic diversity within training datasets. Future directions will expand the ALM, via specific directions. First, the use of spatial audio cues and the recordings of multiple microphones can reinforce the perception of the environment and separation of multiple speakers. Second, a vision modality that is lip reading or gesture based context may be integrated to form a multimodal interaction system that is able to mediate between audio, text, and visual cues. Third, incorporating the scaling to bigger self supervised pretraining akin to the more recent universal multimodal encoders can be further used to enhance cross-domain generalization and noise resistance. Finally, more research should be done so as to develop real time sign-language generation systems that will provide linguistically correct signing sequence so that ALM can become a fully accessible communication tool. Overall this research indicates that logic-based audio models can have a strong influence on driving the field of intelligence and encouraging the creation of more inclusive human-centered AI systems. The future research has a foundation in ALM, both in practice assistive technologies, educational applications, audio analysis, and multimodal human-machine interaction. With audio comprehension moving onward towards a more interpretative rather than transcription process, systems such as ALM will be put in the forefront of defining the future generation of accessible and context-sensitive AI applications.

10.REFERENCES

[1] R. Borsos, et al., *"AudioLM: A Language Modeling Approach to Audio Generation,"* Google Research, 2022. Available: /mnt/data/AudioLM: A Language Modeling Applicationroach_to_Audio_Generation.pdf
[2] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, *"wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations,"* in Proc. NeurIPS, 2020.
[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,"* in Proc. NAACL-HLT, 2019.
[4] A. Radford et al., *"Robust Speech Recognition via Large-Scale Weak Supervision,"* OpenAI, 2022.
[5] K. He et al., *"Deep Residual Learning for Image Recognition,"* in Proc. CVPR, 2016.
[6] G. Tzirakis, J. Zhang, and B. Schuller, *"End-to-End Speech Emotion Recognition Using Deep Neural Networks,"* IEEE Trans. Affective Comput., 2018.
[7] H. Hershey et al., *"CNN Architectures for Large-Scale Audio Classification,"* in Proc. ICASSP, 2017.
[8] J. F. Gemmeke et al., *"AudioSet: An Ontology and Human Labeled Dataset for Audio Events,"* in Proc. ICASSP, 2017.
[9] E. Çakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, *"Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection,"* IEEE/ACM Trans. ASLP, 2017.
[10] A. Vaswani et al., *"Attention Is All You Need,"* in Proc. NeurIPS, 2017.
[11] Y.-A. Chung et al., *"Generative Spoken Dialogue Language Modeling,"* in Proc. ICLR, 2022.
[12] J. H. Engel et al., *"DDSP: Differentiable Digital Signal Processing,"* in Proc. ICLR, 2020.
[13] R. Yamamoto, E. Song, and J.-M. Kim, *"Parallel WaveGAN: A Fast Waveform Generation Model Based on Generative Adversarial Networks,"* in Proc. ICASSP, 2020.
[14] J. W. Kim et al., *"Music Transformer: Generating Music with Long-Term Structure,"* in Proc. ICLR, 2019.
[15] D. P. Kingma and J. Ba, *"Adam: A Method for Stochastic Optimization,"* in Proc. ICLR, 2015.
[16] T. Chen et al., *"A Simple Framework for Contrastive Learning of Visual Representations (SimCLR),"* in Proc. ICML, 2020.
[17] Y. Gong, Y.-A. Chung, and J. Glass, *"AST: Audio Spectrogram Transformer,"* in Proc. Interspeech, 2021.
[18] S. Pascual, M. Ravanelli, J. Serrà, A. Bonafonte, and J. A. Lorenzo-Trueba, *"SEGAN: Speech Enhancement Generative Adversarial Network,"* in Proc. Interspeech, 2017.
[19] S. Arik et al., *"Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning,"* in Proc. ICLR, 2018.
[20] O. Dessì, K. Shih, Y.-L. Cheung, and A. Mohamed, *"Universal Speech Representation,"* Meta AI, Tech. Rep., 2022.
[21] J. Koutník et al., *"A Clockwork RNN,"* in Proc. ICML, 2014.
[22] D. Snyder et al., *"X-Vectors: Robust DNN Embeddings for Speaker Recognition,"* in Proc. ICASSP, 2018.
[23] K. Chatziagapi et al., *"Data Augmentation Using GANs for Speech Emotion Recognition,"* in Proc. Interspeech, 2019.
[24] T. Ko et al., *"Audio Augmentation for Boosting Robustness in Speech Recognition,"* in Proc. Interspeech, 2015.
[25] S. Watanabe et al., *"ESPnet: End-to-End Speech Processing Toolkit,"* in Proc. Interspeech, 2018.
[26] M. Kim, S. Lee, and G. Kim, *"AudioCaps: Generating Captions for Audio in the Wild,"* in Proc. NAACL-HLT, 2019.