Towards Textless Multilingual Audio Question Answering (TM-AQA) System Using Audio-MAMBA

Anonymous ACL submission

Abstract

Audio Question Answering (AQA) is a complex task in Multi-Modal Learning, where a system interprets audio inputs and associated questions to produce appropriate answers. Previous AQA research has primarily focused on text-based queries, exploration into spoken questions in languages like English has been limited. Since speech is a primary mode of communication, integrating spoken queries could significantly enhance AQA system capabilities. To bridge this gap, this paper introduces a Spoken AQA system utilizing the Textless Multilingual Audio Question Answering (TM-AQA) dataset. This dataset comprises 107,514 question-answer pairs in English, Hindi, and Bengali, derived from 1991 environmental audio recordings corresponding to various environmental scenes. The study establishes baseline performance metrics by evaluating several multimodal (MML) AQA frameworks that employ diverse acoustic features and architectures. The experimental results demonstrate that the proposed Audio-MAMBA (A-MAMBA) based MML framework, incorporating a Continuous Scanning Mechanism (CSM), surpasses Transformer-based MML frameworks in performance and computational efficiency.

1 Introduction

AQA systems (Lipping et al., 2022; Behera et al., 2023; Anderson et al., 2018; Sun and Fu, 2019) are designed to respond to queries related to environmental sounds, functioning as multi-modal systems. These systems analyze audio signals containing sounds like footsteps, bird songs, rain, and wind, among others, alongside associated queries to generate suitable responses. However, current AQA systems primarily operate on text-based queries (Behera et al., 2023; Li et al., 2023; Sudarsanam and Virtanen, 2023; Fayek and Johnson, 2020), which can limit usability and prove cumbersome due to the manual input required for typing questions. To overcome these limitations and enhance the user experience by offering a more natural, hands-free interaction mode (Alasmary and Al-Ahmadi, 2023; Patil et al., 2019; Alasmary and Al-Ahmadi, 2023; Chowdhury et al., 2017) that supports multitasking, this work proposes the development of a speech-based AQA system. Responses generated by the system can still be presented in textual form for better comprehension, as humans often find it easier to comprehend information through reading rather than writing. This initiative is motivated by the potential to pioneer advancements in speech-based AQA technology.

Audio Question Answering (AQA) is a rapidly emerging research field facilitated by the introduction of datasets such as CLEAR (Abdelnour et al., 2019), DAQA (Fayek and Johnson, 2020), and ClothoAQA (Lipping et al., 2022) in the English language. Implementing a speech-based AQA system necessitates training on datasets containing triplets of environmental sounds, corresponding spoken queries, and their responses. Without such a dataset, the Clotho-AQA (Lipping et al., 2022) dataset was expanded. Text-based questions were converted into spoken form using Text-to-Speech Synthesis (TTS) systems (Indurthi et al., 2019; Duquenne et al., 2022; Zhang et al., 2022; Xue et al., 2024; Chen et al., 2021). To create a multilingual system, these text-based questions were translated into Hindi and Bengali, popular Indian languages, using Machine Translation (MT) systems. The translated texts were then synthesized to generate spoken questions. This effort led to developing the first-ever Textless Multilingual Audio Question Answering (TM-AQA) dataset, featuring spoken questions in English, Hindi, and Bengali. Significantly, this work also implements and evaluates the first multilingual speech-based AQA systems using a Transformer-based (Vaswani

et al., 2017) Multi-Modal Learning (MML) framework on two different types of questions. The first type involves binary "Yes/No" questions, while the second type consists of open-ended questions with multi-class answers. To the best of our knowledge, this dataset represents a pioneering effort in facilitating research into speech-based AQA.

The TM-AQA dataset undergoes evaluation through a detailed comparative analysis of systems utilizing MML frameworks based on Transformer architectures as described in Transformer (Vaswani et al., 2017). These systems employ a variety of features extracted from different Large Acoustic Models (LAMs), including Wav2Vec2.0 (Baevski et al., 2020), Whisper (Radford et al., 2023), and Hu-BERT (Hsu et al., 2021), alongside traditional mel filterbanks and raw audio. To address the computational complexities associated with Transformer-based architectures (Pau and Aymone, 2024; Katharopoulos et al., 2020), this study proposes a new architecture based on Structures State Space Sequence using Selective Sequence (S6) models (Smith et al., 2022; Gu et al., 2021). S6 models have demonstrated effective solutions for Transformer challenges and have exhibited strong capabilities in modeling long sequences across various tasks (Liu et al., 2024). This research introduces a novel MML framework, A-MAMBA, based on the S6 model, tailored for AQA implementation. A-MAMBA accepts raw audio inputs from speech-based queries and environmental sound clips, generating appropriate responses. Moreover, A-MAMBA integrates a novel continuous scanning mechanism to efficiently capture contextual relationships among adjacent audio sequences divided into patches.

In summary, this paper presents three primary contributions: (i) a Speech-based TM-AQA dataset, (ii) comprehensive baseline evaluations of the TM-AQA dataset using Transformer and A-MAMBA-based MML frameworks, and (iii) the introduction of a novel Continuous Scan Mechanism (CSM). The structure of this paper is organized as follows: Section 2 provides a detailed description of the TM-AQA dataset. Section 3 outlines the proposed methodologies. Experimental settings are discussed in Section 4, followed by results and discussions in Section 5. Finally, conclusions are drawn in Section 6, with limitations discussed in Section 7.

2 TM-AQA Dataset

2.1 Overview

TM-AQA includes spoken questions in English, Hindi, and Bengali, extending from the ClothoAQA dataset chosen for its diverse content and real-world challenges. Figure 1 illustrates spectrogram representations generated for raw environmental signals and corresponding questions in these languages. Each audio clip is associated with 18 question-answer (QA) pairs across all three languages, resulting in 107,514 ($1991 \times 18 \times 3$) QA pairs based on 1991 audio clips containing numerous environmental sounds. Figure 2 presents word-cloud representations of the answers in their respective languages. The distribution of the first five words across all questions in the training set of TM-AQA is visualized in Figure 3, where the innermost ring represents the first word and subsequent rings represent subsequent words. The arc lengths are proportional to the frequency of each word in the questions, with words occurring less than 30 times omitted for clarity.

2.2 Spoken Question Generation

The SeamlessM4T (Barrault et al., 2023) is an AI model designed for translation and transcription tasks, capable of performing speech-to-text (S2T), speech-to-speech (S2S), text-to-speech (T2S), and text-to-text (T2T) translations across 100 languages. This model has demonstrated more realistic translations than similar ones, assessed manually on a subset of translated texts. In this work, SeamlessM4T is employed to translate textual questions from English to Hindi and Bengali. Native speakers and linguists proficient in Hindi and Bengali evaluate these translations to ensure higher quality. Subsequently, the final translations are converted from T2S using the SeamlessM4T model, generating spoken questions in Hindi, Bengali, and English.

2.3 Statistical Overview of TM-AQA

Table 1 provides an overview of the specifications of the TM-AQA dataset. It details the distribution of the dataset across training, validation, and test sets, along with the duration of audio and speech files categorized by the three languages in hours.

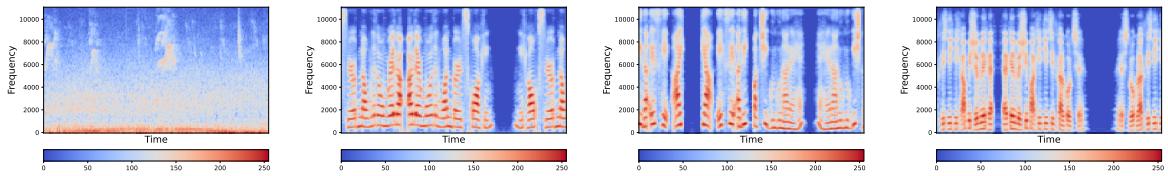

Figure 1: Spectrogram visualization of a raw environmental signal(leftmost) and the corresponding speech-based question in English, Hindi, and Bengali (*from right to left*)

Figure 2: Word clouds representation of answers for all question types in the TM-AQA training set in English, Hindi & Bengali (*from left to right*).

Table 1: Overview of TM-AQA dataset including the number and duration of sound files (in hours) containing environmental sounds, Question and Answer pairs related to the sound files along with duration of spoken questions (in hours) generated in four languages viz. English, German, French and Spanish for train, validation and test sets.

Sl. No.	Set	# of Sound files	Sound duration	# QA pairs	Speech duration (in hours)		
					English	Hindi	Bengali
1	Train	1174	7.35	21132	10.28	11.92	24.55
2	Validation	344	2.13	6192	3.05	3.53	6.99
3	Test	473	2.98	8515	4.15	4.81	8.78
4	**Total**	**1991**	**12.46**	**35839**	**17.48**	**20.26**	**40.32**

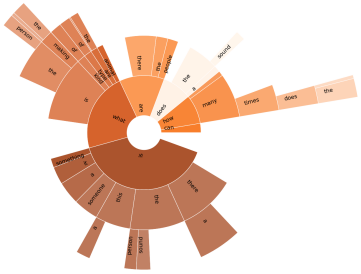

Figure 3: Chart representing the distribution of the first four words for all questions in the training set of the TM-AQA dataset

3 Proposed Methodology

3.1 A-MAMBA: A Structured State Space Sequence Using Selective Scan (S6) Model

S6 models (Gu et al., 2021) are neural network architectures designed to handle long data sequences efficiently. Unlike Transformers, which use self-attention for capturing long-range dependencies but suffer from quadratic complexity, S6 models like MAMBA (Gu and Dao, 2023) offer effective management of sequential long-range dependencies without significantly increasing computational cost. MAMBA, specifically, is a linear-time S6 model (Gu and Dao, 2023) known for faster inference and state-of-the-art performance across various modalities. It maps input sequences through an implicit latent state and applies the Zero-Order Hold rule for output discretization. To this end, this work introduces A-MAMBA, an audio-specific adaptation of MAMBA tailored for MML tasks. A-MAMBA integrates a continuous scan mechanism and processes raw signals by grouping similar features. This approach capitalizes on MAMBA's efficient management of sequential data, enhancing its suitability for audio-related applications.

MAMBA maps the input sequences $x(t) \in \mathbb{R}$ through implicit latent state $h(t) \in \mathbb{R}^N$ to $y(t) \in \mathbb{R}$. It uses evolution parameter **A** and projection parameters **B** & **C**. Mathematically, they are represented as follows:

$$h^{'}(t) = \mathbf{A}h(t) + \mathbf{B}x(t) \quad (1)$$

$$y(t) = \mathbf{C}h(t) \quad (2)$$

The Equations 1 & 2 represents continuous systems and the continuous paremeters ($\mathbf{A}$ & $\mathbf{B}$) were transformed to discrete parameters ($\overline{\mathbf{A}}$ & $\overline{\mathbf{B}}$) using a Zero-Order Hold discretization rule utilizing an additional timescale parameter Δ. Mathematically, it is given as:

$$\overline{\mathbf{A}} = exp(\Delta\mathbf{A}) \tag{3}$$

$$\overline{\mathbf{B}} = (\Delta\mathbf{A})^{-1}(exp(\Delta\mathbf{A}) - I).\Delta\mathbf{B} \tag{4}$$

Finally, the discretized form of Equations 1 & 2 are given as:

$$h_t = \overline{\mathbf{A}}h_{t-1} + \overline{\mathbf{B}}h_t \tag{5}$$

$$y_t = \mathbf{C}h_t \tag{6}$$

The resultant State Space Model (SSM) is the global convolution between the input sequences x and kernel $\overline{\mathbf{K}} \in \mathbb{R}^k$ and is represented as:

$$\overline{\mathbf{K}} = (\mathbf{C}\overline{\mathbf{B}}, \mathbf{C}\overline{\mathbf{A}\mathbf{B}}, ..., \mathbf{C}\overline{\mathbf{A}}^k\overline{\mathbf{B}}) \tag{7}$$

$$y = x * \overline{\mathbf{K}} \tag{8}$$

3.2 Continuous Scan Mechanism

The original Mamba architecture (Gu and Dao, 2023) utilized a selective scan mechanism (SSM), which may not effectively capture contextual relationships between adjacent tokens. To this end, this work employs a continuous scan mechanism (CSM) to overcome the limitation above while ensuring spatial continuity within each region. For instance, consider a zone with four regions: $(1,1), (1,2), (2,1), (2,2)$, corresponding to *top-left, top-right, bottom-left, and bottom-right*, respectively. In the SSM (Gu and Dao, 2023), the scanning order is $(1,1) \rightarrow (1,2) \rightarrow (2,1) \rightarrow (2,2)$, whereas with the CSM, the scanning order is $(1,1) \rightarrow (1,2) \rightarrow (2,2) \rightarrow (2,1)$. The CSM approach efficiently organizes input tokens based on proximity and aligns them accordingly, enabling the model to capture intricate temporal relationships within audio sequences. This work proposes an A-MAMBA model that utilizes the CSM to traverse through patches of audio sequences, enhancing its ability to model complex temporal dependencies.

3.3 A-MAMBA-Based MML Frawework

Figure 4 illustrates the entire workflow of the proposed MML framework. The proposed MML framework utilizing A-MAMBA can be segmented into four components: (i) a shallow feature extractor, (ii) a deep feature extractor, (iii) a fusion-MAMBA module, and (iv) a classification module. The proposed MML architecture utilizes two distinct modalities related to audio: (i) speech-based questions (A_1) and (ii) environmental sounds (A_2). These modalities are combined using a MAMBA-based fusion module to produce responses.

Shallow Feature Extractor: This module comprises a sequence of one-dimensional Convolution (Conv-1D) layers stacked one after another. Convolutional networks excel in capturing local features and mapping them to a higher-dimensional feature space. The input audio sequences (A_1) and (A_2) were processed through multiple Conv-1D layers with LeakyReLU activation functions. Each layer employed a 3×3 kernel size and a stride of 1. The resulting output feature from the Conv-1D block has dimensions $B \times M \times D$, where B represents the batch size, M signifies the sequence length of the audio, and D denotes the embedding dimension.

Deep Feature Extractor: Considering CNNs' potential limitations in effectively capturing temporal audio patterns and the quadratic complexity issues encountered by Transformers, this module opted for Mamba blocks. These blocks have been recognized for efficiently extracting advanced specific features, as evidenced in previous research (Xie et al., 2024; Peng et al., 2024; He et al., 2024). Consequently, the outputs from the CNN blocks $(A_1{}')$ and $(A_2{}')$ underwent processing through multiple Bi-Mamba blocks (Zhu et al., 2024), resulting in $(A_1{}'')$ and $(A_2{}'')$. This process can be expressed mathematically as follows:

$$A_1{}'' = \textit{Bi-Mamba}(A_1{}') \tag{9}$$

$$A_2{}'' = \textit{Bi-Mamba}(A_2{}') \tag{10}$$

The *Bi-Mamba* block (Zhu et al., 2024) introduces a 2-way feature extraction and scanning (forward & backward) of the input sequences, which enables multi-directional spatial-aware processing. The input feature sequence with dimension $B \times M \times D$ is normalized by passing it through a *LayerNorm* block. The output is then projected into two branches, x and z, using two MLP layers. In

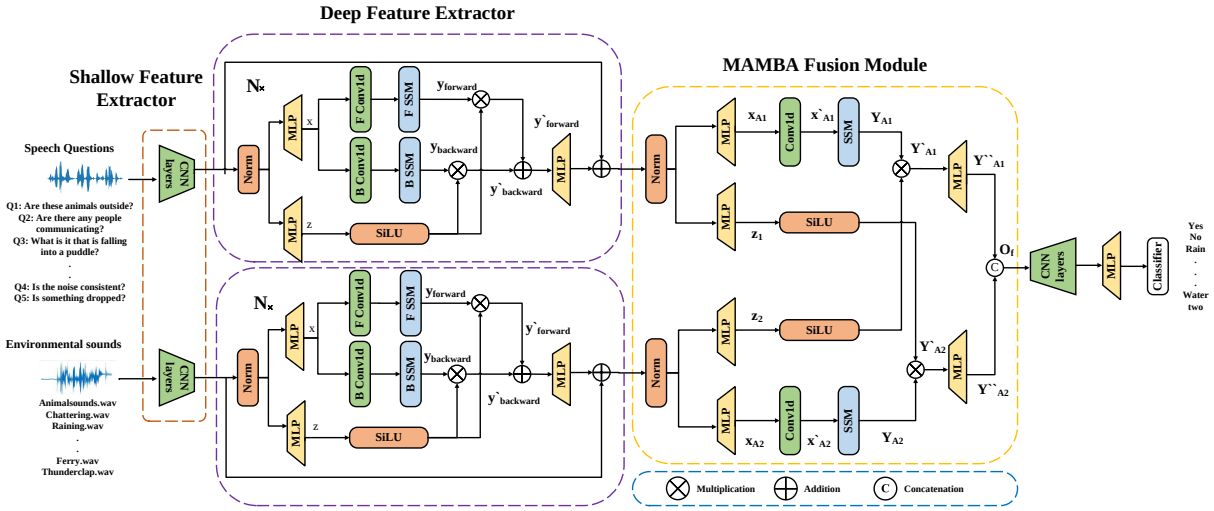

Figure 4: Block diagram of the proposed A-MAMBA-based MML framework.

the first branch, x is again fed into two separate branches consisting of *Conv-1D* layers and the State Space Models to compute $x'_{backward}$ and $x'_{forward}$. The z is initially passed through a *SiLU* activation function on the other branch. Further, z passed through *SiLU* is multiplied with $x'_{forward}$ and $x'_{backward}$ to get the resulting features $y'_{forward}$ and $y'_{backward}$. The final input to the next MLP layer is the summation of $y'_{forward}$ and $y'_{backward}$ $(y'_{forward} \oplus y'_{backward})$. After completing the above steps, the final output of the *Bi-Mamba* block is processed through an MLP layer and residual connection to obtain the ultimate features. These final features are denoted as A''_1 and A''_2, corresponding to the input features from modalities A_1 and A_2, respectively. Algorithm 1 comprehensively illustrates the operational principles of the *Bi-Mamba* block.

Algorithm 1 Bi-Mamba Block

Input: token sequence $T_{l-1} : (B, M, D)$
Output: token sequence $T_l : (B, M, D)$
/* normalize the input sequence T_{l-1} */
$T'_{l-1} : (B, M, D) \leftarrow \text{Norm}(T_{l-1})$
$x : (B, M, E) \leftarrow \text{MLP}^x(T'_{l-1})$
$z : (B, M, E) \leftarrow \text{MLP}^z(T'_{l-1})$
for o in {forward, backward} **do**
  $x'_o : (B, M, E) \leftarrow \text{SiLU}(\text{Conv1d}_o(x))$
  $B_o : (B, M, N) \leftarrow \text{MLP}^B_o(x'_o)$
  $C_o : (B, M, N) \leftarrow \text{MLP}^C_o(x'_o)$
  /* softplus ensures positive Δ_o */
  $\Delta_o : (B, M, E) \leftarrow \log(1 + \exp(\text{Linear}^{\Delta}_o(x'_o) + \text{Parameter}^{\Delta}_o))$
  /* shape of Parameter is (E, N) */
  $\overline{A_o} : (B, M, E, N) \leftarrow \Delta_o \otimes \text{Parameter}^A_o$
  $\overline{B_o} : (B, M, E, N) \leftarrow \Delta_o \otimes B_o$
  /* SSM denoted by Eq (7) & Eq (8) */
  $y_o : (B, M, E) \leftarrow \text{SSM}(\overline{A_o}, \overline{B_o}, C_o)(x'_o)$
end for
/* get gated y_o */
$y'_{\text{forward}} : (B, M, E) \leftarrow y_{\text{forward}} \odot \text{SiLU}(z)$
$y'_{\text{backward}} : (B, M, E) \leftarrow y_{\text{backward}} \odot \text{SiLU}(z)$
/* residual connection */
$T_l : (B, M, D) \leftarrow \text{MLP}^T(y'_{\text{forward}} + y'_{\text{backward}}) + T_{l-1}$
return T_l

Fusion-MAMBA Module: To facilitate the cross-modal interaction and fusion, a Mamba-based fusion module is introduced. This fusion module takes the output from both the *Bi-Mamba* blocks and uses a gating mechanism to learn features from each other. Through this block, the fused features O_f from features A_2'' and A_1'' were obtained. Mathematically, it can be represented as follows:

$$O_f = FM([A_1'', A_2'']) \tag{11}$$

The algorithm for the MAMBA fusion module is depicted in Algorithm 2:

Classification: The final fused output O_f undergoes processing through a sequence of linear layers before being fed into the MLP classification head. In the case of multiclass classification, the classification layer accommodates 829 classes, whereas, for binary classification, it reduces to two classes: "Yes" or "No." The predicted answer is ultimately represented as:

$$y_{pred} = Softmax(Conv\text{-}1d(MLP([O_f]))) \tag{12}$$

4 Experimental Settings

4.1 Representation of Different Modalities

Speech Representation: To expand the comparative analysis, audio features were extracted from both speech-based questions and raw environmental sounds (both resampled at 16 kHz) using advanced Large Acoustic Models (LAMs) including

Algorithm 2 The Fusion-Mamba Module

Input: $T_{l-1}^{A1} : (B, M, D), T_{l-1}^{A2} : (B, M, D)$
Output: $O_f : (B, W, D)$
for o in {A1, A2} **do**
  /* Normalize both the input sequence */
  $T_{l-1}^{o} : (B, M, D) \leftarrow \text{Norm}(T_{l-1}^{o})$
  $x'_o : (B, M, E) \leftarrow \text{SiLU}(\text{Conv1d}_o(x))$
  $B_o : (B, M, N) \leftarrow \text{MLP}(x'_o)$
  $C_o : (B, M, N) \leftarrow \text{MLP}(x'_o)$
  /* softplus ensures positive Δ_o */
  $\Delta_o : (B, M, E) \leftarrow \log(1 + \exp(\text{Linear}(x'_o) + \text{Parameter}_o^{\Delta}))$
  /* shape of Parameter is (E, N) */
  $\overline{A_o} : (B, M, E, N) \leftarrow \Delta_o \otimes \text{Parameter}_o^{A}$
  $\overline{B_o} : (B, M, E, N) \leftarrow \Delta_o \otimes B_o$
  /* SSM denoted by Eq (7) & Eq (8) */
  $Y^o : (B, M, E) \leftarrow \text{SSM}(\overline{A_o}, \overline{B_o}, C_o)(x'_o)$
end for
$z_1 : (B, M, E) \leftarrow \text{MLP}(T_{l-1}^{A_1})$
$z_2 : (B, M, E) \leftarrow \text{MLP}(T_{l-1}^{A_2})$
/* get output Y^o */
$Y^{A1\prime} : (B, M, E) \leftarrow Y^{A1} \odot \text{SiLU}(z_2)$
$Y^{A2\prime} : (B, M, E) \leftarrow Y^{A2} \odot \text{SiLU}(z_1)$
$Y^{A1\prime\prime} : (B, M, O) \leftarrow \text{MLP}(Y^{A1})$
$Y^{A2\prime\prime} : (B, M, O) \leftarrow \text{MLP}(Y^{A2})$
/* Concatenate the final output */
$O_f : (B, W, D) \leftarrow \text{Cat}(Y^{A_1\prime\prime}, Y^{A_2\prime\prime})$
return O_f

Wav2Vec2 (Baevski et al., 2020), Hu-BERT (Hsu et al., 2021), and Whisper (Radford et al., 2023). Here's how each LAM was utilized:

- **Wav2Vec2**: Features were extracted using the XLS-R 128 pre-trained model (Babu et al., 2021), renowned for its training across 128 languages. Features were derived from the L2-normalization layer following the encoder.
- **Hu-BERT**: Features were obtained from the 11^{th} layer of the encoder of the multilingual Hu-BERT model. Before input into the TM-AQA system, these features underwent normalization.
- **Whisper**: Utilizing its large-V3 pre-trained model with 1.5 billion parameters, Whisper extracted audio features from the final encoder layer.

Across all the models, 80-dimensional Mel filterbanks were computed with a frame length of 400 and a hop length of 160. Feature extraction was conducted from the frozen encoders of the LAMs, ensuring consistency in utilizing the specified layers or endpoints for a fair comparison within the TM-AQA system.

Sound Representation: Due to the absence of pre-trained models specific to environmental sounds, the LAMs mentioned above (Learnable Audio Models) underwent fine-tuning to perform Environmental Scene Classification (ESC) on a combined dataset. This combined dataset includes signals from widely-used ESC or acoustic scene classification (ASC) datasets: ESC-50 (Piczak, 2015), DCASE-2019-task-1(A) (Mesaros et al., 2019), and FSC22 (Bandara et al., 2023). After fine-tuning using this combined dataset, these models were then utilized to extract features from raw environmental sounds.

4.2 Implementation Details

The TM-AQA systems were trained and evaluated according to the dataset split specified in Table 1 across all the languages. For transformers-based models, various combinations of audio features were employed for evaluation, whereas A-MAMBA utilized raw audio inputs. Both Transformer and A-MAMBA-based MML frameworks were optimized using the Adam optimizer. A learning rate of 1×10^{-4} was used for all the baseline models, while A-MAMBA used a fixed learning rate of 2×10^{-4} without weight decay. The β values for the Adam optimizer were set to $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Training procedures included early stopping criteria, using cross-entropy loss for multi-class classification and binary cross-entropy loss for binary classification tasks. Typically, convergence was achieved well before completing 100 epochs for Transformer-based models, whereas A-MAMBA models were trained for 500 epochs. Both models employed a batch size of 128 during training. The AST-MAMBA network (A-MAMBA using spectrogram inputs) was trained using spectrogram features extracted from the audio signal. These spectrograms have dimensions of $128 \times 100 \times t$, where t and $100 \times t$ represent the length of the signal in seconds and the total number of frames, respectively. The total number of Mel filters used was 128, and the signal length t was set to 5 seconds.

4.3 Evaluation Metrics

The TM-AQA systems were evaluated based on Top-1, Top-5, and Top-10 accuracy metrics. Top-k (where k=1, 5, 10) accuracy measures indicate the percentage of instances where the correct answer is included among the top "k" predicted answers. These metrics offer a flexible evaluation by considering a range of potentially correct answers, not just the top-ranked ones. In the context of AQA, where questions may have multiple plausible answers, Top-5 and Top-10 accuracy metrics are particularly relevant. They acknowledge the ambiguity

by recognizing acceptable answers within the top 5 and 10 predictions. This approach aligns with real-world scenarios where users expect systems to present a set of potential answers rather than a single definitive answer. Therefore, Top-5 and Top-10 accuracy metrics reflect the complexity of understanding and responding to questions involving visual contexts more comprehensively. While Top-1 accuracy is suitable for binary classification tasks, Top-1, Top-5, and Top-10 accuracy metrics were employed for multi-class classification evaluations to provide a broader assessment of model performance across varying degrees of correctness in prediction.

5 Results & Discussions

Table 2 provides a performance comparison across the Transformer-based MML framework using different audio features for the three languages in the TM-AQA dataset. The table highlights that the optimal performance across all languages was consistently achieved when utilizing features extracted from the Whisper pre-trained model. This model benefits from extensive training on diverse audio data across multiple languages, allowing for nuanced feature capture. Table 3 compares the performance of architectures employing Transformer and A-MAMBA with MLP classifiers. All models described in this table utilize raw audio inputs to minimize architectural complexity. Notably, the small variant of A-MAMBA utilizing CSM across all languages consistently achieved the best performance across all classification tasks. Table 4 provides detailed information regarding the different variants of A-MAMBA employed in this work.

Table 2: Performance comparison between baseline benchmark systems utilizing Transformer architecture and distinct features extracted from speech-based questions and environmental sounds. The performances are presented in percentages. In the given table *RA: Raw Audio*, *MLB: Mel filter bank*, *W2V: Wav2Vec*, *HuB: HuBERT*

ENGLISH					
Features (Questions)	**Features (Sounds)**	**Top-1**		**Top-5**	**Top-10**
		Binary	**Multiclass**	**Multiclass**	**Multiclass**
RA	RA	50.08	**38.8**	60.47	64.76
MLB	MLB	50.23	36.09	60.22	65.45
W2V	W2V	50.46	37.81	61.64	66.15
HuB	HuB	50.7	38.12	61.87	66.45
Whisper	Whisper	**51.82**	38.66	**62.15**	**66.7**
HINDI					
Features (Questions)	**Features (Sounds)**	**Top-1**		**Top-5**	**Top-10**
		Binary	**Multiclass**	**Multiclass**	**Multiclass**
RA	RA	49.67	37.1	60.21	64.54
MLB	MLB	49.1	37.81	59.81	63.87
W2V	W2V	50.32	38.16	62.67	66.21
HuB	HuB	50.32	38.43	62.32	66.45
Whisper	Whisper	**50.78**	**38.81**	**63.65**	**67.32**
BENGALI					
Features (Questions)	**Features (Sounds)**	**Top-1**		**Top-5**	**Top-10**
		Binary	**Multiclass**	**Multiclass**	**Multiclass**
RA	RA	49.22	37.36	59.38	64.56
MLB	MLB	49.45	38.09	60.5	62.78
W2V	W2V	49.89	38.23	62.64	64.64
HuB	HuB	50.21	38.2	62.78	**66.95**
Whisper	Whisper	**51.34**	**38.93**	**63.62**	66.1

Transformers vs. MAMBA: From Table 3, it is clear that AQA systems employing A-MAMBA outperformed Transformer-based MML frameworks. Effective modeling of contextual information from questions, sounds, or other modalities is crucial for generating accurate answers in question-answering systems. The Continuous Scan Mechanism (CSM) inherent in A-MAMBA enhances its ability to capture this contextual information more effectively than Transformers.

Spectrogram vs. Raw Audio Features: The A-MAMBA-based AQA systems incorporating raw audio demonstrated superior performance compared to spectrograms, as illustrated by Table 3. Raw audio signals retain amplitude and phase information, preserving temporal dynamics and fine-grained details of speech signals often lost in spectrograms. This temporal information is crucial for the AQA system to enhance its contextual understanding of spoken questions paired with environmental sounds. The Raw audio also maintains full resolution without distortions introduced by methods like Short-Time Fourier Transform (STFT), Mel Frequency Cepstrum Coefficients (MFCCs), Mel filterbanks, or spectrograms. This capability likely contributes to better performance using raw audio instead of spectrograms in MAMBA-based architectures.

SSM vs CSM: As observed in Table 3, systems utilizing CSM performed better than SSM for all variants of MAMBA across the three languages. This may be attributed to better contextual understanding between adjacent tokens by maintaining spatial continuity throughout the scanning phase. This improved the decision-making ability and is reflected in the performance.

6 Conclusion

This study aims to develop a multilingual speech-based Audio Question Answering (AQA) system using a state-space model-based Meta-Learning (MML) framework. To achieve this objective, the study introduces the TM-AQA dataset, which in-

Table 3: Performance comparison (expressed in %) between Transformer, AST-MAMBA (Spectrogram), and A-MAMBA(Raw Audio) MML frameworks using MLP classifier for binary and multi-class classification task. (Performance differences with the best model are represented in blue.)

ENGLISH					
MODEL	**VARIANT**	**Top-1**		**Top-5**	**Top-10**
		Binary	**Multiclass**	**Multiclass**	**Multiclass**
Transformer (Vaswani et al., 2017)	Base	50.08 ↓ (5.48)	38.8 ↓ (4.46)	60.47↓ (8.47)	64.76 ↓ (10.01)
AST-MAMBA + SSM	Small	50.67↓ (4.89)	38.79↓ (4.47)	66.45↓ (2.66)	67.89↓ (6.88)
	Medium	48.67↓ (6.89)	36.54↓ (6.72)	64.31↓ (4.8)	66.13 ↓ (8.64)
	Large	42.35 ↓ (13.21)	32.31↓ (9.95)	60.48↓ (8.63)	61.53 ↓ (13.24)
AST-MAMBA + CSM	Small	50.70 ↓ (4.86)	38.56↓ (4.7)	66.87↓ (2.24)	68.97 ↓ (5.8)
	Medium	48.6↓ (6.96)	37.37↓ (5.89)	65.23↓ (3.88)	66.77 ↓ (8)
	Large	42.22 ↓ (13.34)	33.11↓ (10.15)	60.85↓ (8.26)	61.89 ↓ (12.88)
A-MAMBA + SSM	Small	52↓ (3.56)	40.59↓ (2.67)	68.34↓ (0.77)	74.4↓ (0.37)
	Medium	51.62↓ (3.94)	37.26↓ (6)	65.47↓ (3.64)	68.54 ↓ (6.23)
	Large	44.62↓ (10.94)	33.48↓ (9.78)	62.3↓ (6.81)	65.31↓ (9.46)
A-MAMBA + CSM	Small	**55.56**	**43.26**	**69.11**	**74.77**
	Medium	51.98↓ (3.58)	41.42↓ (1.84)	68.44↓ (0.67)	71.21↓ (3.56)
	Large	45.86↓ (9.7)	35.94↓ (7.32)	63.8↓ (5.31)	66.48↓ (8.29)
HINDI					
MODEL	**VARIANT**	**Top-1**		**Top-5**	**Top-10**
		Binary	**Multiclass**	**Multiclass**	**Multiclass**
Transformer (Vaswani et al., 2017)	Base	49.67 ↓ (3.97)	37.1 ↓ (5.46)	60.21 ↓ (12.46)	64.54 ↓ (12.06)
AST-MAMBA + SSM	Small	50.35 ↓ (3.29)	38.53 ↓ (4.03)	66.67 ↓ (6.00)	68.12 ↓ (8.48)
	Medium	48.74 ↓ (4.90)	35.14 ↓ (7.42)	63.1 ↓ (9.57)	65.12 ↓ (11.48)
	Large	41.35 ↓ (12.29)	31.42 ↓ (11.14)	60.85 ↓ (11.82)	62.95 ↓ (13.65)
AST-MAMBA + CSM	Small	50.44 ↓ (3.20)	38.56 ↓ (4.00)	66.68 ↓ (5.99)	68.65 ↓ (7.95)
	Medium	49.08 ↓ (4.56)	35.43 ↓ (7.13)	63.66 ↓ (9.01)	65.43 ↓ (11.17)
	Large	41.23 ↓ (12.41)	31.52 ↓ (11.04)	61.33 ↓ (11.34)	63.17 ↓ (13.43)
A-MAMBA + SSM	Small	51.98 ↓ (1.66)	40.44 ↓ (2.12)	69.18 ↓ (3.49)	73.61 ↓ (2.99)
	Medium	49.42 ↓ (4.22)	37.59 ↓ (4.97)	65.72 ↓ (6.95)	67.09 ↓ (9.51)
	Large	44.98 ↓ (8.66)	33.77 ↓ (8.79)	61.18 ↓ (11.49)	65.19 ↓ (11.41)
A-MAMBA + CSM	Small	**53.64**	**42.56**	**72.67**	**76.6**
	Medium	51.8 ↓ (1.84)	40.55 ↓ (2.01)	66.35 ↓ (6.32)	68.65 ↓ (7.95)
	Large	45.64 ↓ (8.00)	35.71 ↓ (6.85)	62.69 ↓ (9.98)	66.86 ↓ (9.74)
BENGALI					
MODEL	**VARIANT**	**Top-1**		**Top-5**	**Top-10**
		Binary	**Multiclass**	**Multiclass**	**Multiclass**
Transformer (Vaswani et al., 2017)	Base	49.22 ↓ (3.56)	37.36 ↓ (2.08)	59.38↓ (10.73)	64.56↓ (11.2)
AST-MAMBA + SSM	Small	49.84↓ (2.94)	37.31↓ (2.13)	66.31↓ (3.8)	69.83↓ (5.93)
	Medium	48.84↓ (3.94)	36.45↓ (2.99	64.11↓ (6)	67.41↓ (8.35)
	Large	41.8↓ (10.98)	32.34↓ (7.1)	59.31↓ (10.8)	62.23↓ (13.53
AST-MAMBA + CSM	Small	49.88↓ (10.9)	37.36↓ (2.08)	66.67↓ (3.44)	68.88↓ (6.88)
	Medium	48.8↓ (3.98)	36.57↓ (2.87)	64.17↓ (5.94)	67.56↓ (8.2)
	Large	41.89↓ (10.89)	33.11↓ (6.33)	60.61↓ (9.5)	62.34↓ (13.42)
A-MAMBA + SSM	Small	51.21↓ (1.57)	38.42↓ (1.02)	68.75↓ (1.36)	74.32↓ (1.44)
	Medium	50.34 ↓ (2.44)	36.43 ↓ (3.01)	65.87 ↓ (4.24)	68.76↓ (7)
	Large	44.89↓ (7.89)	33.37 ↓ (6.07)	62.75 ↓ (7.36)	66.76↓ (9)
A-MAMBA + CSM	Small	**52.78**	**39.44**	**70.11**	**75.76**
	Medium	50.79 ↓ (1.99)	38.2 ↓ (1.24)	65.78 ↓ (4.33)	70.87 ↓ (4.89)
	Large	45.78 ↓ (7.00)	34.61 ↓ (4.83)	63.11 ↓ (7.00)	66.88 ↓ (8.88)

Table 4: Computational Complexity of A-MAMBA and its variants (*M=Millions, G=Giga*)

Model Type	Layers	Hidden Size	Expand	d_state	Params (M)	Flops	MACS
A-MAMBA-small	1	192	1	8	11.0065 M	1.6548 G	820.2 M
A-MAMBA-medium	2	256	1	8	18.4939 M	2.8176 G	1.3991 G
A-MAMBA-large	4	384	2	16	45.2617 M	6.3927 G	3.1815 G

cludes spoken questions in English, Hindi, and Bengali. Additionally, the study proposes a novel state-space model-based MML framework called A-MAMBA, which incorporates the Continuous Scan Mechanism. The performance of A-MAMBA is compared with transformer-based MML frameworks to establish rigorous baseline benchmarks for the TM-AQA task. Experimental results demonstrate that the proposed A-MAMBA-based MML framework outperforms transformer-based MML frameworks in the context of TM-AQA.

7 Limitations

The TM-AQA dataset includes synthesized spoken questions generated by the T2S system. However, it is crucial to incorporate real voice recordings for spoken questions to develop a robust system. This work focuses solely on using raw audio as input features for the proposed A-MAMBA model, aiming to minimize computational demands since features derived from pre-trained models necessitate additional computations. These constraints are anticipated to guide future research directions in the speech-based AQA domain.

8 Ethical Statement

Native speakers of Hindi and Bengali, proficient in both languages and English, were employed to check the manual quality of translated texts. They were compensated on a pre-sentence basis. Additionally, the authors ensured proper documentation of their employment.

References

Jerome Abdelnour, Giampiero Salvi, and Jean Rouat. 2019. Clear: A dataset for compositional language and elementary acoustic reasoning.

Faris Alasmary and Saad Al-Ahmadi. 2023. Sbvqa 2.0: Robust end-to-end speech-based visual question answering for open-ended questions. IEEE Access, 11:140967–140980.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6077–6086.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. arXiv preprint arXiv:2111.09296.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems, 33:12449–12460.

Meelan Bandara, Roshinie Jayasundara, Isuru Ariyarathne, Dulani Meedeniya, and Charith Perera. 2023. Forest sound classification dataset: Fsc22. Sensors, 23(4):2032.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. 2023. Seamlessm4t-massively multilingual & multimodal machine translation. arXiv preprint arXiv:2308.11596.

Swarup Ranjan Behera, Achyut Mani Tripathi, Krishna Mohan Injeti, Jaya Sai Kiran Patibandla, Praveen Kumar Pokala, and Balakrishna Reddy Pailla. 2023. Aquallm: Audio question answering data generation using large language models. arXiv preprint arXiv:2312.17343.

Junkun Chen, Mingbo Ma, Renjie Zheng, and Liang Huang. 2021. Direct simultaneous speech-to-text translation assisted by synchronized streaming ASR. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4618–4624, Online. Association for Computational Linguistics.

Iqbal Chowdhury, Kien Nguyen, Clinton Fookes, and Sridha Sridharan. 2017. A cascaded long short-term memory (lstm) driven generic visual question answering (vqa). In 2017 IEEE International Conference on Image Processing (ICIP), pages 1842–1846.

Paul-Ambroise Duquenne, Hongyu Gong, Benoît Sagot, and Holger Schwenk. 2022. T-modules: Translation modules for zero-shot cross-modal machine translation. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 5794–5806, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Haytham M Fayek and Justin Johnson. 2020. Temporal reasoning via audio question answering. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28:2283–2294.

Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752.

Albert Gu, Karan Goel, and Christopher Ré. 2021. Efficiently modeling long sequences with structured state spaces. arXiv preprint arXiv:2111.00396.

Xuanhua He, Ke Cao, Keyu Yan, Rui Li, Chengjun Xie, Jie Zhang, and Man Zhou. 2024. Pan-mamba: Effective pan-sharpening with state space model. arXiv preprint arXiv:2402.12192.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29:3451–3460.

Sathish Indurthi, Houjeung Han, Nikhil Kumar Lakumarapu, Beomseok Lee, Insoo Chung, Sangha Kim, and Chanwoo Kim. 2019. Data efficient direct speech-to-text translation with modality agnostic meta-learning. arXiv preprint arXiv:1911.04283.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. 2020. Transformers are RNNs: Fast autoregressive transformers with linear attention. In Proceedings of the 37th International Conference on Machine Learning, volume 119 of Proceedings of Machine Learning Research, pages 5156–5165. PMLR.

Guangyao Li, Yixin Xu, and Di Hu. 2023. Multi-scale attention for audio question answering. arXiv preprint arXiv:2305.17993.

Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen. 2022. Clotho-aqa: A crowdsourced dataset for audio question answering. In 2022 30th European Signal Processing Conference (EUSIPCO), pages 1140–1144. IEEE.

Zicheng Liu, Li Wang, Siyuan Li, Zedong Wang, Haitao Lin, and Stan Z Li. 2024. Longvq: Long sequence modeling with vector quantization on structured memory. arXiv preprint arXiv:2404.11163.

Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. 2019. A multi-device dataset for urban acoustic scene classification. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019), pages 9–13.

Annapurna P. Patil, Amrita Behera, P Anusha, Mitali Seth, and Prabhuling. 2019. Speech enabled visual question answering using lstm and cnn with real time image capturing for assisting the visually impaired. In TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), pages 2475–2480.

Danilo Pietro Pau and Fabrizio Maria Aymone. 2024. Mathematical formulation of learning and its computational complexity for transformers' layers. Eng, 5(1):34–50.

Siran Peng, Xiangyu Zhu, Haoyu Deng, Zhen Lei, and Liang-Jian Deng. 2024. Fusionmamba: Efficient image fusion with state space model. arXiv preprint arXiv:2404.07932.

Karol J. Piczak. 2015. ESC: Dataset for Environmental Sound Classification.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In International Conference on Machine Learning, pages 28492–28518. PMLR.

Jimmy TH Smith, Andrew Warrington, and Scott W Linderman. 2022. Simplified state space layers for sequence modeling. arXiv preprint arXiv:2208.04933.

Parthasaarathy Sudarsanam and Tuomas Virtanen. 2023. Attention-based methods for audio question answering. arXiv preprint arXiv:2305.19769.

Qiang Sun and Yanwei Fu. 2019. Stacked self-attention networks for visual question answering. In Proceedings of the 2019 on International Conference on Multimedia Retrieval, ICMR '19, page 207–211, New York, NY, USA. Association for Computing Machinery.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc.

Xinyu Xie, Yawen Cui, Chio-In Ieong, Tao Tan, Xiaozhi Zhang, Xubin Zheng, and Zitong Yu. 2024. Fusionmamba: Dynamic feature enhancement for multimodal image fusion with mamba. arXiv preprint arXiv:2404.09498.

Zhengshan Xue, Tingxun Shi, Xiaolei Zhang, and Deyi Xiong. 2024. Speaker voice normalization for end-to-end speech translation. Expert Systems with Applications, 248:123317.

Biao Zhang, Barry Haddow, and Rico Sennrich. 2022. Revisiting end-to-end speech-to-text translation from scratch. In International Conference on Machine Learning.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. 2024. Vision mamba: Efficient visual representation learning with bidirectional state space model. arXiv preprint arXiv:2401.09417.