

MPAI CAE-ASD Audio Scene Description: An international standard for analysis-based description of acoustic scenes

Mert Burak Çötel
TED University
Ankara, Türkiye
mert.coteli@tedu.edu.tr

Marina Bosi
CCRMA, Stanford University
Stanford, CA, USA
mbosi@stanford.edu

Leonardo Chiariglione
MPAI
Geneva, Switzerland
leonardo@chiariglione.org

Hüseyin Hacıhabiboğlu
METU
Ankara, Türkiye
hhuseyin@metu.edu.tr

Abstract—Object-based audio (OBA) is becoming extremely widespread due to the flexibility it offers in producing content that can be adapted for different user requirements or repurposed for different consumption scenarios. Specifically, tracked wireless audio capture systems require grounded, standardized approaches for registering representing and attributing spatial context via appropriately designed metadata structures. Existing standards such as AES69-2022 Spatially Oriented Format for Acoustics (SOFA) target static scenes and impulse responses for archival purposes. While ITU-R BS.2076 Audio Definition Model (ADM) targets broadcast scenarios and MPEG-H metadata mainly targets user-facing applications in broadcast media, a description model suitable for AI-based immersive media workflows had largely been missing. Such a model has recently been developed through the work carried out within Moving Picture, Audio and Data Coding with Artificial Intelligence (MPAI) Context-based Audio Enhancement (CAE) standard. This paper explains the components of the MPAI-CAE Audio Scene Description (ASD) model, discusses its capabilities and presents ongoing work towards conformance testing for MPAI-CAE ASD.

I. INTRODUCTION

Object-based audio (OBA) refers to the set of techniques for recording, storing, transmitting and rendering real or synthetic acoustic scenes as a combination of 1) isolated sounding entities, called audio objects, 2) the spatial attributes of thereof, and 3) the attributes of the acoustic environment, in order to allow presenting an audio representation of a scene that immerses its user in a real or imagined acoustic environments.

The advent of object-based audio (OBA), supported by the development of standards such as ISO/IEC MPEG-H 3D Audio [1], [2], the industry-wide adoption of object-based playback technologies (e.g. Dolby ATMOS [3], DTS:X [4], Auro-3D [5]) and its compatibility with volumetric audiovisual content creation workflows, for example, through game engines such as Unity [6] or Unreal Engine [7], made OBA a popular paradigm for immersive sound. More recently, ongoing work within ISO/IEC MPEG-I Immersive Audio development also has a strong object-based emphasis, based in part on MPEG-H 3D Audio [8], [9].

An important aspect of OBA is the capture and analysis of an acoustic scene. Existing OBA standards do not specify

how a scene should be captured or how the source signals and attributes pertaining to sources therein are extracted. Such flexibility is welcome and indeed expected for synthetic spatial audio scenes (e.g. when dry audio objects are synthetically positioned in a scene). However, when audio captured from real scenes is considered, both the source signals corresponding to the sources that comprise the scene and their spatial attributes must be obtained.

This need is not limited to audio production and coding: The Internet of Sounds (IoS) field, which formalizes networks of heterogeneous devices acquiring, processing and exchanging sonic information, requires common formats, protocols and representations fostering interoperability. The fragmentation of ad-hoc solutions is identified as one of principal open challenges for IoS [10]. MPAI-CAE ASD addresses this gap on the analysis side by standardising how object-based descriptions of real acoustic scenes are extracted from microphone array captures.

Moving Picture, Audio and Data Coding for Artificial Intelligence (MPAI) Context-Based Audio Enhancement (CAE) development committee started working on the Audio Scene Description (ASD) component of the standard in 2023 and it was included in version 2.1 of the MPAI-CAE technical specifications at the MPAI General Assembly in March 2024 [11]. MPAI-CAE ASD builds on the MPAI-CAE Enhanced Audioconference Experience (EAE) use case which is part of the MPAI-CAE standard technical specifications version 1.4 [12]. EAE use case aims to improve audioconference experiences under multiple-talker scenarios, possibly under adverse acoustical conditions (e.g., high reverberation and background noise) via an acoustic scene analysis approach which naturally gives way to object-based audio representation. This article presents the rationale behind MPAI-CAE ASD and discusses the technical specifications of its constituent AI modules.

II. MPAI AND AI-BASED CODING STANDARDS

MPAI is an international, unaffiliated, non-profit organisation developing standards for AI-based data coding with clear Intellectual Property Rights licensing frameworks founded in 2020. MPAI's rationale is that AI development would benefit

from operational reference standards comparable to those that have enabled the rapid development of digital media. [13], [14].

MPAI standards adopt a modular architecture in which AI applications are specified as compositions of interconnected AI Modules (AIMs), each with defined functions and interfaces, implemented using AI or traditional data-processing technologies and executed within a standard AI Framework (AIF).

The key components are [15]:

- **AIM (AI Module)** is the basic building block, with defined functions and interfaces. In MPAI, AI data coding denotes the transformation of data into an equivalent representation better suited to a specific application.
- **AIF (AI Framework)** provides the architecture, interfaces, protocols, and APIs required to execute AI-based implementations, including mixed AI and traditional data-processing workflows. It also includes a Security Abstraction Layer (SAL) for secure functionality.

This modular architecture enables interoperable AI components from different sources and vendors, supporting the development of advanced and explainable AI applications.

III. MPAI/IEEE CAE

The MPAI-CAE standard leverages automated AI tools, to enhance the audio experience, continuing a long tradition of international audio coding standardization [16], [17]. MPAI-CAE version 1.4 [12] was adopted by the IEEE Standards Association as IEEE 3302-2022 in December 2022 [18]. It addresses applications including entertainment, communication, teleconferencing, gaming, post-production, preservation, and restoration [19]. Version 1.4 defines four use cases targeting different listening contexts, including home, car, on-the-go, and studio environments [19]:

- 1) Emotion Enhanced Speech (EES)
- 2) Audio Recording Preservation (ARP) [20]
- 3) Speech Restoration System (SRS)
- 4) Enhanced Audioconference Experience (EAE)

The MPAI AI Framework (AIF), specified in the MPAI-AIF/IEEE 3301-2022 standard [15], [21], provides the infrastructure for implementing MPAI-CAE with interconnected AIMs.

MPAI-CAE normatively defines the semantics and syntax of input and output data, the functions of AIMs, and their interconnections. Interoperability is ensured by allowing an AIM implementation to be replaced by a functionally equivalent implementation while preserving the specified input and output formats.

The MPAI-CAE standard version 2.1 has introduced the Audio Scene Description (ASD) use case which leverages the capabilities of microphone arrays and multichannel audio recordings to provide detailed descriptions of audio objects and the spatial geometry of audio scenes. The goal is to enhance the parameters required for creating immersive audio experiences.

IV. CAE-ASD

The CAE-ASD model relies on the extraction of spatial and acoustic attributes of audio scenes from recordings which are subsequently converted to Audio Scene Descriptors. We will first discuss how analysis-based synthesis is performed and then describe the extraction of these descriptors within the technology-agnostic MPAI-CAE ASD framework.

A. Audio Scene Analysis

Analysis of an audio scene can be at different levels of abstraction depending on the requirements of the use case. For example, a fully parametric analysis might be useful in cases where a scene-based reconstruction is the final goal. Such an analysis would completely disregard the particular attributes of sound sources and consider the sound field as a temporally evolving spatial distribution of sound pressure in the region in which it is being captured. On the other extreme, the audio scene analysis may be concerned with extracting a higher-level, quasi-semantic representation that comprises a simplified geometric description of the scene.

The audio scene analysis approach taken in MPAI-CAE ASD is of the latter type and deconstructionist in the sense that different salient components of the sound field (i.e., audio objects) and their spatial attributes (e.g., direction, distance) are extracted from microphone array recordings to support an object-based audio (OBA) framework in use cases such as in mixed reality and immersive telepresence applications.

B. Requirements

The main functional requirement for audio scene analysis is processing of microphone array signals along with metadata representing microphone array attributes to generate a compact, object-based representation of the audio scene amenable to inform subsequent flexible immersive audio scene reconstruction methods. It is then clear that an audio scene analysis system should be capable of 1) separating sound sources from complex acoustic scenes containing not only sources themselves but also reflections and reverberation as well as background noise, 2) extracting the directions and distances of sound sources, and 3) extracting acoustic attributes of the environment in which the audio scene is being captured.

The audio scene analysis approach is based on representing the sound field as a sparse composition of plane waves. Such representation requires certain assumptions regarding the statistical stationarity of the sound sources as well as their incoherence. Short-time stationarity is satisfied after a suitable time-frequency transform. Incoherence is also satisfied if it is assumed that the sources are approximately W-disjoint orthogonal (e.g. Hadamard product of their windowed Fourier transforms is substantially zero) [22], [23]. The audio scene analysis, therefore, is carried out in the transform domain. This requires block-based processing of captured audio signal from microphone arrays. Each block of audio signals are transformed into frequency domain, and each *time-frequency bin* is processed individually for extracting spatial attributes

in shorter time intervals with further enhancement using long-term statistics.

The underlying assumption in the ASD model is that sound fields that require a description are in general reverberant, and the recorded signals constitute convolutive as opposed to instantaneous mixtures [24]. This means that the sound field comprises not only the direct path but also the reflections and reverberation.

The Audio Scene Description (ASD) composite AIM, depicted in Fig. 1 and summarised in Table I, introduces technology that captures and analyzes audio environments to create detailed, immersive soundscapes. ASD identifies and localizes audio objects such as, for example, speech and ambient sounds. This data is used to map the spatial geometry of the audio scene, accounting for dynamic changes over time. By standardizing its parameters, ASD ensures compatibility across different platforms and enhances realism in applications like virtual reality (VR), augmented reality (AR), and gaming. The advancements presented in ASD enable more engaging content creation and open new research opportunities in audio processing and immersive technologies.

As shown in Fig. 1, the inputs to the ASD composite AIM are the Microphone Array Geometry, which describes the spatial microphone arrangement, and Multichannel Audio which represents the audio signal captured by the Microphone Array. The output consists of Audio Scene Descriptors, a data type that includes detailed descriptions of the objects within an audio scene. The use of microphone arrays is essential in capturing sound from multiple directions. These arrays enable a detailed spatial analysis of the audio environment enhancing the accuracy of audio spatial resolution and the ability to isolate different audio sources within a scene. Multichannel audio offers a rich dataset utilised to produce a detailed audio scene description.

Recording techniques may include Ambisonics, binaural recording, as well as distributed wireless microphone arrays among others which can lead to different audio scene descriptions. The choice of technique influences how the audio data is processed and analysed. Different algorithms are employed to detect distinct audio objects within the scene, such as speech, music, quotidian sounds, sound effects and ambient noise. In order to precisely localize audio objects, i.e., to determine the position of audio objects within the audio environment, ASD accounts for both the direction and distance from the microphones. A detailed audio spatial mapping is described in the Audio Scene Geometry. This process involves creating a map of the audio scene, including the positions of detected audio objects, where it is essential to account for movement and changes within the audio scene over time to create realistic and immersive experiences. The gathered data can be used to simulate immersive audio environments for applications like virtual reality (VR) and augmented reality (AR), enhancing realism. ASD enhances the realism of audio experiences in VR, AR, and gaming by accurately describing audio objects and their spatial relationships. It provides content creators with detailed tools and data, enabling the production of more

engaging and lifelike audio content. Furthermore, standardized descriptions and parameters facilitate easier integration and interoperability between different audio systems and platforms.

The core of the ASD workflow (see Fig. 1) consists of five key AIMs: *Audio Analysis Transform*, *Audio Source localisation*, *Audio Separation and Enhancement*, *Audio Synthesis Transform*, and *Audio Description Multiplexing*. These components are described in detail in the following subsections.

C. Audio Analysis Transform AIM

Audio Analysis Transform is the AIM used for transforming the time-domain multichannel signals into time-frequency domain multichannel signals via extracting frequency bands. *Multichannel Audio* represents interleaved time-domain signals captured through the *Microphone Array* constituting several channels of which directional info are described in the *Microphone Array Audio*. Each microphone's directional characteristics is not independently used during this transformation and full channels are individually transformed into the time-frequency domain using an overlapped transform to obtain the *Transform Multichannel Audio*. The output is a data structure comprising complex valued audio samples in the frequency domain. While *Audio Analysis Transform* can include any invertible time-frequency transform such as modified DCT (MDCT), windowed Fourier transform is typically preferred since phase information is often essential for downstream processing by other AIMs.

D. Audio Source Localisation AIM

Audio Source localisation is the AIM used for detecting sound sources and their locations in the scene using the *Microphone Array Geometry* which is the metadata that describes the recording microphones and *Transform Multichannel Audio* obtained by the *Audio Analysis Transform* AIM. The output of this AIM is *Audio Spatial Attributes* which constitute the directional information of the sound sources (i.e. *Audio Objects*) detected in the scene. While this AIM typically uses sparse plane wave decomposition (SPWD) [25] other classical (e.g. EB-MUSIC [26], EB-ESPRIT [27]) or learning-based sound source localisation methods (e.g. [28]) can be used.

E. Audio Separation and Enhancement AIM

Audio Separation and Enhancement is the AIM used for separating audio objects from the *Transform Multichannel Audio* with the help of the *Audio Spatial Attributes*. The background noise is also suppressed from the separated sources for enhancement. This AIM outputs clean and separated audio sources. The structure of this data format is described in the *Transform Enhanced Audio*. In one implementation, this is achieved by a dictionary-based sparse plane wave decomposition [29]. The specifications of the model is open, however, to other approaches, AI-based or otherwise. For example, for sufficiently dry environments TF-GridNet [30] can be used for single-channel speech separation. Similarly, direction-of-arrival (DOA) deep learning algorithms such as DBnet [31] can potentially be used for multichannel source separation.

TABLE II
AUDIO SCENE DESCRIPTORS SEMANTICS

Label	Size	Description
HEADER	9 Bytes	
Standard	7 Bytes	The string CAE-ASD
Version	1 Byte	Major version
Subversion	1 Byte	Minor version
ASDID	16 Bytes	UUID Identifier of Audio Scene Descriptors set.
Time	17 Bytes	Collects various data expressed with bits
TimeType	0 bit	0 = Relative: time starts at 0000/00/00T00:00 1 = Absolute: time starts at 1970/01/01T00:00.
Reserved	1–7 bits	Reserved
StartTime	8 Bytes	Start of current Audio Scene Descriptors (in μ s).
EndTime	8 Bytes	End of current Audio Scene Descriptors (in μ s).
AudioObjectCount	1 Byte	Number of Audio Objects in the Audio Scene.
AudioObjectsData	N1 Bytes	Data associated to each Audio Object.
AudioObjectID	1 Byte	ID of a specific Audio Object in the Audio Scene.
SamplingRate	0–3 bits	0:8, 1:16, 2:24, 3:32, 4:44.1, 5:48, 6:64, 7:96, 8:192 (all kHz)
SampleType	4–6 bits	(aka sample precision) 0:8, 1:16, 2:24, 3:32, 4:64 (bits/sample)
Reserved	7 bit	
Spatial Attitude	N2 Bytes	Spatial Attitude of Audio Object.
AudioObject	N3 Bytes	Set of Audio Object Data.
FormatID	1 Byte	Format Identifier of Audio Object.
ObjectLength	4 Bytes	Number of Bytes in Audio Object.
DataInObject	N4 Bytes	Data of Audio Object.

to provide clear speech substantially free from background noise and acoustics-related artefacts, reducing audioconference fatigue. In addition to separating and denoising the speakers, EAE extracts the Spatial Attitudes of the speakers with respect to the microphone array, represented in the Audio Scene Geometry format, so that the receiver side can spatially re-render each participant’s speech.

- **Avatar-Based Videoconference and Mixed-Reality Collaborative Spaces (MPAI-MMC V2 / MPAI-MCS).** Virtual spaces where humans and avatars collaborate (e.g., Conversation About a Scene, Avatar-Based Videoconference). The audio scene description separates and spatially locates each speaker in the room so that their speech can be attached to the correct avatar and positioned consistently in the virtual meeting environment.
- **Human-CAV Interaction (MPAI-CAV, HCI Subsystem).** The HCI creates AV Scene Descriptors providing access to the individual audio and visual objects – in particular the speech and faces of the humans – and, inside the cabin, separates and locates the speech of the passengers from the rest of the environment, enabling the vehicle to identify which occupant issued a command or query and from which position.
- **Audio-Visual Scene Description (MPAI-OSD).** The Object and Scene Description standard specifies technologies for object description and localisation in space, shared across MPAI use cases. Its Audio-Visual Basic Scene Description (OSD-AVB) Composite AIM combines the Audio Scene with the Visual Scene, assigning identifiers to audio-visual objects such as a human and

their speech.

- **Audio Object Rendering (CAE-AOR).** Supports rendering of Audio Scenes comprising digital representations of Audio Objects that can be rendered in a perceptually veridical fashion in a renderer-agnostic way, targeting six degrees-of-freedom immersive VR/XR applications with free listener movement.

VII. EXPERIMENTS

We evaluated the Audio Scene Description (CAE-ASD) composite AIM on a two-speaker scene rendered with measured room impulse responses. The impulse responses come from the METU SPARG Eigenmike em32 Acoustic Impulse Response dataset [32], recorded in a classroom with a reverberation time T60 of 1.12 s using a 32-channel rigid spherical microphone array. We selected two source positions on the dataset’s measurement grid: position 310, at (0.0, 1.0, −0.6) m relative to the array centre, and position 350, at (0.0, −1.0, −0.6) m (Fig. 2). Both lie 1.17 m from the array, at azimuths of 90° and 270° and an elevation of −31°, so the two sources face each other across the array from below its equator as in Fig. 2. The classroom measures 8.3 × 6.5 × 2.9 m; the array centre was 1.5 m above the floor and the two source positions 0.9 m, i.e. on the lowest of the dataset’s five measurement layers.

Dry speech was taken from the TSP Speech Database [33]. Speaker FA (female) was placed at position 310 and speaker MJ (male) at position 350. For each talker, consecutive sentences were concatenated to 3.0 s and scaled to equalise RMS. Each dry signal was convolved with the 32 impulse responses of its position, the two 32-channel audio signals were summed, and the mixture was normalised to a peak of

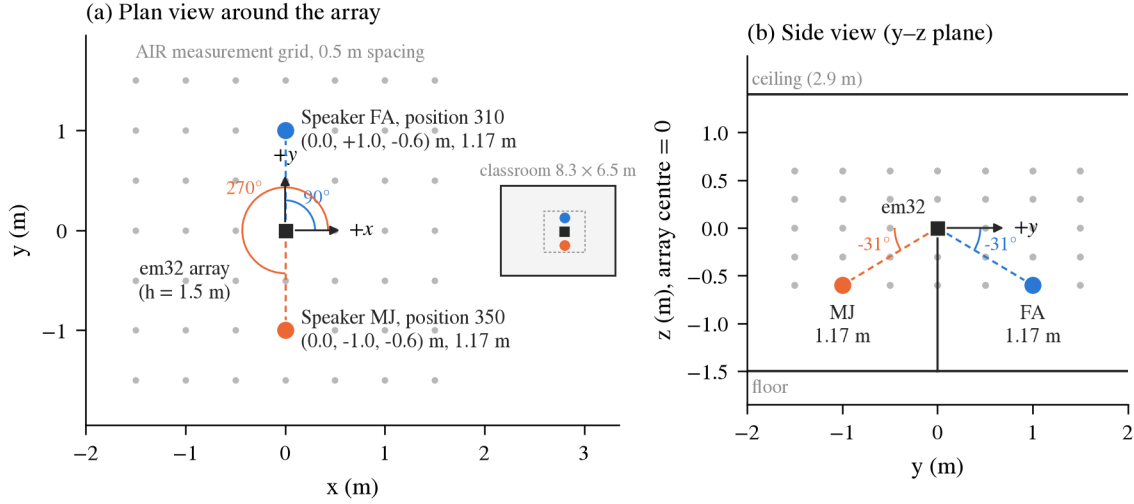


Fig. 2. Scene layout. (a) Plan view of the classroom with the AIR measurement grid, the em32 array and the two talker positions; the array’s position inside the room is approximate, taken from the dataset’s floor plan. (b) Side view in the y - z plane through the talkers, showing the -31° elevation and the 1.17 m source distance.

0.5 full scale and stored as a 16-bit WAV file, which is the input format of the reference application. Both speakers are active throughout the 3 s, so the scene contains no single-talker segments.

The recording was processed block by block by the CAE-ASD chain in its reference implementation, with the sub-AIMs mapped as follows:

- Audio Analysis Transform: 2048-sample (42.7 ms) non-overlapping blocks, rectangular window, 1024 frequency bins, 48 kHz.
- Audio Source Localisation: spherical harmonic decomposition of the array signals with a rigid-sphere modal filter, projection onto a HEALPix grid of order 2 (192 directions, mean angular spacing 14.7°) with a von Mises kernel ($\kappa = 6$), and the residual-energy test (RENT) over bins 169–512 (3.9–12 kHz) to count sources [29]. Detected directions are tracked across blocks; a candidate within 57° (1.0 rad) of an existing source is merged into it, and the source direction is updated when the new estimate lies within 23° (0.4 rad).
- Audio Separation and Enhancement: orthogonal-matching-pursuit separation on the GPU for the tracked directions, followed by a Wiener time-frequency mask built from a spherical-harmonic (order 4) pseudo-inverse of the source kernels [29].
- Audio Synthesis Transform: inverse FFT of each separated spectrum.
- Audio Description Multiplexing: one Audio Scene Descriptors document per block in the MPAI-CAE V2.1 JSON layout.

Each block therefore yields one JSON blob with a header, the block time, the number of audio objects, and for each object its identifier (a UUID is kept stable across blocks by the tracker), the sampling parameters, an OSD Spatial Attitude with spherical position [azimuth, elevation, distance]

in radians and metres, and the 2048 time-domain samples of the separated object.

Ground-truth directions follow from the measured source positions. Each reported audio object was matched to the speaker nearest to the object’s mean direction over all blocks. The angular error of every block estimate was measured as the great-circle angle to that speaker’s true direction. No temporal smoothing was applied to the output.

The 3 s recording produced 71 ASD JSON blobs. The first five blocks (0–0.21 s) report no audio object while the tracker accumulates evidence. From block 5 onwards, every block reports exactly two objects, and the two object identifiers persist unchanged to the end of the recording, so tracking was not lost, split to multiple sources beyond two or duplicated. Each block’s spatial attributes can be followed through Fig. 3.

It may be observed that both estimates are within the tolerance of HEALPix resolution (azimuth 90° and 270° , elevation -30°) for the large majority of blocks. The excursions visible in the figure are single-block jumps to the closest possible grid position, which is why they cluster at 15 – 20° , roughly the cell spacing of the order-2 grid, and return immediately. The resulting ASD JSON blobs for two processing blocks are shown in Fig. 4.

VIII. DISCUSSION AND CONCLUSIONS

This paper presented the Audio Scene Description (ASD) component of the MPAI-CAE standard, an analysis-based model for describing acoustic scenes captured with microphone arrays in terms of audio objects, their spatial attributes, and the geometry of the scene.

A. Comparison with Other Existing Standards

Existing metadata models for spatial audio each address a specific portion of the production and delivery chain. AES69 (spatially oriented format for acoustics, SOFA) [34] targets

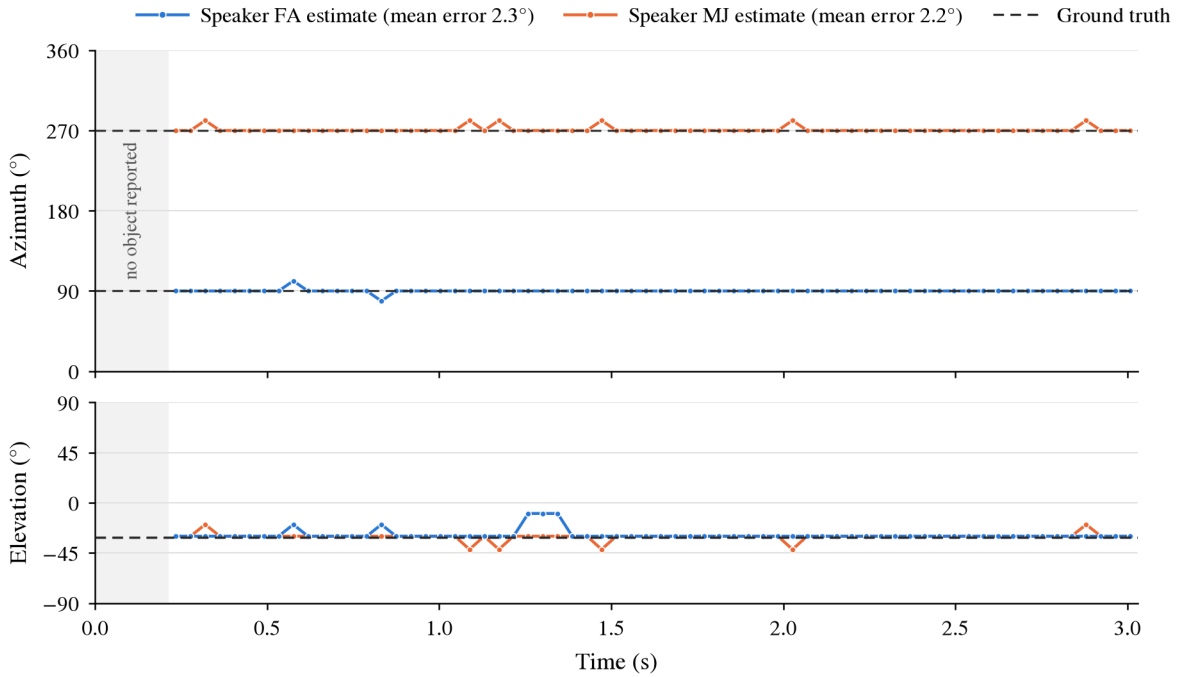


Fig. 3. Azimuth (top) and elevation (bottom) of the two reported audio objects per block, with the nominal speaker directions as dashed lines. The grey band marks the start-up interval in which no Audio Object is produced.

the archival exchange of spatially oriented acoustic data such as impulse responses measured under static conditions. The ITU-R BS.2076 Audio Definition Model [35] describes the composition of broadcast content, and MPEG-H 3D Audio metadata [1] primarily serves user-facing rendering and personalization in broadcast delivery. None of these models, however, specifies how an object-based representation is to be *obtained* from real captured scenes. MPAI-CAE ASD addresses an adjacent, upstream problem these standards do not: it standardizes the interfaces, data types, and workflow of the analysis stage itself, so that the extraction of audio objects and their spatial attributes from multichannel recordings becomes an interoperable, substitutable process rather than a proprietary preprocessing step. The resulting Audio Scene Descriptors can then feed downstream representations and renderers, including those based on the aforementioned standards. Table III presents a comparison of the mentioned standards in terms of their typical use cases, similarities and differences.

B. Technology Agnosticism and Interoperability

A distinguishing property of the ASD model is that it is technology-agnostic. Each of the five constituent AIMs i.e. Audio Analysis Transform, Audio Source localisation, Audio Separation and Enhancement, Audio Synthesis Transform, and Audio Description Multiplexing, is normatively specified only in terms of its function and its input and output data. An implementer is free to realize, for example, the Audio Source localisation AIM with classical array processing techniques or with a learned neural model, provided that the interfaces

conform to the specification. Because MPAI-CAE inherits the AIW/AIM/AIF architecture of MPAI-AIF [15], [21], functionally equivalent AIMs from different vendors can be substituted for one another within the same workflow. This modularity supports explainability, since intermediate data such as the Audio Spatial Attributes are exposed at standardized interfaces and can be inspected independently of the implementation.

C. Conformance Testing and Reference Implementation

Work is ongoing within the MPAI-CAE Development Committee towards a complete conformance testing framework for CAE-ASD. Conformance of the composite AIM is assessed at the data level: Audio Scene Descriptors produced by an implementation shall validate against the normative JSON schema, and the embedded Audio Objects, Spatial Attitudes, and Audio Scene Geometry shall validate against their respective schemas. Conformance of the individual AIMs is specified by the corresponding AIM specifications, allowing an implementation to be certified at the level of a single module, of the composite AIM, or of a complete workflow such as CAE-EAE. The conformance testing methods for the CAE-EAE AIW and its individual AIMs are specified in the MPAI-CAE Use Cases specification [36]. A reference software implementation is available through the MPAI Git repositories [37] and serves both as a baseline for interoperability testing and as a starting point for implementers.

D. Future Work

Several directions are being pursued. First, the definition of profiles will allow lightweight deployments (e.g., the Audio

```

{
  "$schema": "https://schemas.mpai.community/CAE/V2.1/data/AudioSceneDescriptors.json",
  "Header": { "Standard": "MPAI-CAE", "Version": 2, "Subversion": 1 },
  "ASDID": "c9550240-0392-4a34-a06e-89ee0d873cb7",
  "Time": { "TimeType": false, "StartTime": 1, "EndTime": 1 },
  "AudioObjectCount": 2,
  "AudioObjectsData": [
    {
      "AudioObjectID": "701b3286-3267-4075-ba57-fb4095b4f634",
      "SamplingRate": 4, "SamplingType": 1,
      "SpatialAttitude": {
        "General": { "CoordType": 1, "ObjectType": 1, "Precision": 0, "MediaType": 0 },
        "SpherPosition": [4.7124, -0.5236, 0.0]
      },
      "AudioObject": {
        "FormatID": 0, "ObjectLength": 2048,
        "DataInObject": [0.0050304, 0.0031826, 0.001388, ... 2045 further samples]
      }
    },
    {
      "AudioObjectID": "b1931e1a-9db2-4d81-b21d-abedb4b9ad4b",
      "SamplingRate": 4, "SamplingType": 1,
      "SpatialAttitude": {
        "General": { "CoordType": 1, "ObjectType": 1, "Precision": 0, "MediaType": 0 },
        "SpherPosition": [1.5708, -0.5236, 0.0]
      },
      "AudioObject": {
        "FormatID": 0, "ObjectLength": 2048,
        "DataInObject": [-0.0052767, -0.0021161, -0.0014908, ... 2045 further samples]
      }
    }
  ]
}

```

MPAI-CAE V2.1 schema
new UUID per block
scene-geometry BlockStart/End
sources found by RENT
tracker id, kept across blocks; talker MJ
48 kHz, 24 bit
OSD V1.0: spherical coordinates, speech
[azimuth, elevation, distance] = 270°, -30°, unknown
samples carried in the JSON, 42.7 ms
separated object, time domain
tracker id, kept across blocks; talker FA
48 kHz, 24 bit
OSD V1.0: spherical coordinates, speech
[azimuth, elevation, distance] = 90°, -30°, unknown
samples carried in the JSON, 42.7 ms
separated object, time domain

(a) Block 6 (0.256–0.299 s)

```

{
  "$schema": {...}, "Header": { "Standard": "MPAI-CAE", "Version": 2, "Subversion": 1 },
  "ASDID": "ce0c5849-33ea-431d-8ed7-65c5fd2d1a99",
  "Time": { "TimeType": false, "StartTime": 1, "EndTime": 1 }, "AudioObjectCount": 2,
  "AudioObjectsData": [
    {
      "AudioObjectID": "701b3286-3267-4075-ba57-fb4095b4f634",
      "SamplingRate": 4, "SamplingType": 1, "SpatialAttitude": {...},
      "AudioObject": {
        "FormatID": 0, "ObjectLength": 2048,
        "DataInObject": [-0.027523, -0.027649, -0.027245, ... 2045 further samples]
      }
    },
    {
      "AudioObjectID": "b1931e1a-9db2-4d81-b21d-abedb4b9ad4b",
      "SamplingRate": 4, "SamplingType": 1, "SpatialAttitude": {...},
      "AudioObject": {
        "FormatID": 0, "ObjectLength": 2048,
        "DataInObject": [0.067802, 0.062321, 0.06156, ... 2045 further samples]
      }
    }
  ]
}

```

as in (a)
new UUID per block
as in (a)
same tracker id as in (a), talker MJ
as in (a): 270°, -30°
samples carried in the JSON, 42.7 ms
separated object, time domain
same tracker id as in (a), talker FA
as in (a): 90°, -30°
samples carried in the JSON, 42.7 ms
separated object, time domain

(b) Block 40 (1.707–1.749 s)

Fig. 4. Audio Scene Descriptor (ASD) documents written by the CAE-ASD chain for two blocks of the two-speaker scene. Of the 2048 samples in each DataInObject array only the first 3 are shown, long values are rounded, and in (b) the field groups that are identical to (a) are folded to {...}. The grey comment text indicate the meaning of the coded fields. The highlighted AudioObjectID values are the tracker identifiers, which persist from block to block; their colours are those of the speakers in Fig. 2. Positions are [azimuth, elevation, distance] in radians and metres.

TABLE III
COMPARISON OF MPAI-CAE ASD, MPEG-H AUDIO OBJECT, SOFA, AND ADM.

	MPAI-CAE ASD	MPEG-H Audio Object	SOFA	ADM
Standard	MPAI-CAE 2.1	ISO/IEC MPEG 23008-3	AES69-2022	ITU-R BS.2076
What is it used for	Separation and localisation of sources from a raw multichannel capture as an AI task	Perceptual coding and transport of object-based audio to a decoder/renderer	Storage and exchange of measured multichannel acoustic transfer functions (HRTFs, BRIRs, array responses)	Codec-agnostic metadata describing file content for correct playback/rendering
Input	Raw microphone array signal and array geometry	(Typically) authored object waveform and metadata	Not applicable. A data container, not a process	Produced audio content (channel, object, or scene-based)
Core operation	Sound source localisation and acoustic source separation for downstream AI tasks	Compression and multiplexing of audio and spatial metadata into a bitstream	Storage of transfer functions with defined measurement geometry and coordinates	Declaration of audio type and structure, authored before encoding
Output	Structured scene description (source count, direction, distance)	Renderer-ready encoded bitstream	Data file with transfer functions and geometry	Metadata carried in a broadcast-wave container
Relevance to OBA	Produces the object specification from raw microphone array data prior to encoding/compression	Encoding and delivery of an (typically) manually specified audio object	Not object-based audio per se, but supplies impulse response/transfer-function data for subsequent rendering	One of several content types (with channel and scene-based) for authoring objects before encoding
Time-varying position	Yes, if the scene changes over time	Yes, via interpolated metadata updates	No. Fixed measurement geometry per entry	Yes, via time-segmented position data

Basic Scene Description variant, which describes a scene only in terms of its composing audio objects) alongside full-featured ones. Second, the description of environmental acoustic attributes, such as reverberation characteristics, is being extended to better support perceptually veridical re-rendering in six-degrees-of-freedom scenarios addressed by CAE-6DF. Third, methodologies for the performance assessment of ASD implementations, including objective separation and localisation metrics as well as subjective evaluation protocols, are under study. Finally, the alignment of ASD with the audio-visual scene description work in MPAI-OSD is expected to yield a unified treatment of audio-visual objects across MPAI standards.

In conclusion, MPAI-CAE ASD provides the missing standardized link between the capture of real acoustic scenes and their object-based representation, enabling interoperable, AI-based immersive media workflows spanning audioconferencing, mixed reality, automotive, and six-degrees-of-freedom audio applications.

IX. ACKNOWLEDGMENT

The authors thank the members of the MPAI Context-based Audio Enhancement Development Committee (CAE-DC) for their contributions to the development of the MPAI-CAE standard. This work was supported in part by the Council of Higher Education of Türkiye (YÖK), Research Universities Support Programme Research Grant ADEP-704-2024-11499 “Development of a deep-learning based MPAI standard for scene-based 6DoF immersive audio coding and compression” (2024-2026).

REFERENCES

- [1] ISO/IEC 23008-3:2022, “Information technology—high efficiency coding and media delivery in heterogeneous environments—Part 3: 3D

- audio,” International Organization for Standardization, Geneva, Switzerland, 2022.
- [2] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, “MPEG-H audio—the new standard for universal spatial/3D audio coding,” *J. Audio Eng. Soc.*, vol. 62, no. 12, pp. 821–830, Jan. 2015.
- [3] Dolby Laboratories, “Dolby Atmos specifications,” Dolby Laboratories, Inc., San Francisco, CA, USA, accessed 2024 Nov. [Online]. Available: <https://www.dolby.com/technologies/dolby-atmos/>
- [4] DTS, Inc., “DTS:X,” Xperi Inc., Calabasas, CA, USA, accessed 2024 Nov. [Online]. Available: <https://dts.com/dtsx/>
- [5] Auro Technologies, “Auro-3D,” Auro Technologies NV, Mol, Belgium, accessed 2024 Nov. [Online]. Available: <https://www.auro-3d.com/>
- [6] Unity Technologies, “Unity real-time development platform,” accessed 2024 Nov. [Online]. Available: <https://unity.com/>
- [7] Epic Games, “Unreal Engine,” accessed 2024 Nov. [Online]. Available: <https://www.unrealengine.com/>
- [8] ISO/IEC 23090-4, “Information technology—coded representation of immersive media—Part 4: Immersive audio,” International Organization for Standardization, Geneva, Switzerland.
- [9] S. R. Quackenbush and J. Herre, “MPEG standards for compressed representation of immersive audio,” *Proc. IEEE*, vol. 109, no. 9, pp. 1578–1589, Sep. 2021.
- [10] L. Turchet, G. Fazekas, M. Lagrange, H. S. Ghadikolaei, and C. Fischione, “The internet of audio things: State of the art, vision, and challenges,” *IEEE Internet of Things Journal*, vol. 7, no. 10, pp. 10 233–10 249, 2020.
- [11] MPAI, “Technical specification: Context-based audio enhancement (MPAI-CAE) V2.1,” Moving Picture, Audio and Data Coding by Artificial Intelligence, Geneva, Switzerland, Mar. 2024. [Online]. Available: <https://mpai.community/standards/mpai-cae/>
- [12] —, “Technical specification: Context-based audio enhancement (MPAI-CAE) V1.4,” Moving Picture, Audio and Data Coding by Artificial Intelligence, Geneva, Switzerland, 2022. [Online]. Available: <https://mpai.community/standards/mpai-cae/>
- [13] A. Artusi, A. Basso, M. Bosi, S. Canazza, L. Chiariglione, M. Choi, F. Columbano, M. B. Çöteli, N. Dalla Pozza, R. Dini, M. Guarise, H. Hacıhabiboğlu, R. Iacoviello, C. Jia, J. Kang, P. Kudumakis, V. Lazzaroli, M. Mazzaglia, G. Perboli, N. Pretto, P. Ribeca, M. Rosano, and M. Seligman, *Towards Pervasive and Trustworthy Artificial Intelligence: How Standards Can Put a Great Technology at the Service of Humankind*. Independently published, 2021.
- [14] A. Basso, P. Ribeca, M. Bosi, N. Pretto, G. Chollet, M. Guarise, M. Choi, F. Yassa, and R. Iacoviello, “AI-based media coding standards,” *SMPTE Motion Imaging J.*, vol. 131, no. 4, pp. 10–20, 2022.

- [15] MPAAI, “Technical specification: AI framework (MPAAI-AIF) V2.0,” Moving Picture, Audio and Data Coding by Artificial Intelligence, Geneva, Switzerland, 2023. [Online]. Available: <https://mpai.community/standards/mpai-aif/>
- [16] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, and M. Dietz, “ISO/IEC MPEG-2 advanced audio coding,” *J. Audio Eng. Soc.*, vol. 45, no. 10, pp. 789–814, Oct. 1997.
- [17] M. Bosi and R. E. Goldberg, *Introduction to Digital Audio Coding and Standards*. Boston, MA, USA: Kluwer Academic Publishers, 2002.
- [18] IEEE 3302-2022, “IEEE standard for adoption of moving picture, audio and data coding by artificial intelligence (MPAI) technical specification context-based audio enhancement (CAE) v1.4,” IEEE Standards Association, Piscataway, NJ, USA, Dec. 2022.
- [19] M. Bosi, N. Pretto, M. Guarise, and S. Canazza, “Sound and music computing using AI: Designing a standard,” in *Proc. 18th Sound and Music Computing Conf. (SMC’21)*, Virtual Conference, 2021, pp. 215–218.
- [20] M. Bosi, S. Canazza, A. Russo, N. Pretto, and L. Chiariglione, “An MPAAI/IEEE international standard for audio: Overview of CAE audio recording preservation (ARP) technology,” in *AES International Conference on Audio Archiving, Preservation & Restoration*, Jun. 2023, pp. 21–28.
- [21] IEEE 3301-2022, “IEEE standard for adoption of moving picture, audio and data coding by artificial intelligence (MPAI) technical specification artificial intelligence framework (AIF) v1.1,” IEEE Standards Association, Piscataway, NJ, USA, 2022.
- [22] Ö. Yilmaz and S. Rickard, “Blind separation of speech mixtures via time-frequency masking,” *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.
- [23] O. Thiergart and E. A. P. Habets, “Sound field model violations in parametric spatial sound processing,” in *Proc. Int. Workshop on Acoustic Signal Enhancement (IWAENC)*, Aachen, Germany, 2012, pp. 1–4.
- [24] S. Makino, T.-W. Lee, and H. Sawada, Eds., *Blind Speech Separation*. Dordrecht, The Netherlands: Springer, 2007.
- [25] M. B. Çöteli, O. Olgun, and H. Hacıhabiboğlu, “Multiple sound source localization with steered response power density and hierarchical grid refinement,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 11, pp. 2215–2229, 2018.
- [26] H. Sun, E. Mabande, K. Kowalczyk, and W. Kellermann, “Localization of distinct reflections in rooms using spherical microphone array eigenbeam processing,” *J. Acoust. Soc. Am.*, vol. 131, no. 4, pp. 2828–2840, 2012.
- [27] A. Herzog and E. A. P. Habets, “Eigenbeam-ESPRIT for DOA-vector estimation,” *IEEE Signal Process. Lett.*, vol. 26, no. 4, pp. 572–576, 2019.
- [28] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, “A survey of sound source localization with deep learning methods,” *J. Acoust. Soc. Am.*, vol. 152, no. 1, pp. 107–151, 2022.
- [29] M. B. Çöteli and H. Hacıhabiboğlu, “Sparse representations with Legendre kernels for DOA estimation and acoustic source separation,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 2296–2309, 2021.
- [30] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, “TF-GridNet: Integrating full- and sub-band modeling for speech separation,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 3221–3236, 2023.
- [31] A. Aroudi and S. Braun, “DBnet: DOA-driven beamforming network for end-to-end reverberant sound source separation,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*. IEEE, 2021, pp. 211–215.
- [32] O. Olgun and H. Hacıhabiboğlu, “METU SPARG Eigenmike em32 acoustic impulse response dataset v0.1.0,” Zenodo, 2019.
- [33] P. Kabal, “TSP speech database,” Dept. Elect. Comput. Eng., McGill Univ., Montreal, QC, Canada, Tech. Rep., Sep. 2002.
- [34] AES69-2022, “AES standard for file exchange—spatial acoustic data file format,” Audio Engineering Society, New York, NY, USA, 2022.
- [35] ITU-R BS.2076-2, “Audio definition model,” International Telecommunication Union, Geneva, Switzerland, Oct. 2019.
- [36] MPAAI, “Technical specification: Context-based audio enhancement (MPAAI-CAE) use cases (CAE-USC) V2.4,” Moving Picture, Audio and Data Coding by Artificial Intelligence, Geneva, Switzerland, 2026. [Online]. Available: <https://mpai.community/standards/mpai-cae/>
- [37] —, “MPAAI-CAE reference software: Enhanced audioconference experience (CAE-EAE),” Moving Picture, Audio and Data Coding

by Artificial Intelligence, Geneva, Switzerland, mPAI Git; registered users. Reference software specified in Technical Specification MPAAI-CAE Use Cases (CAE-USC) V2.4. [Online]. Available: <https://experts.mpai.community/software/>