MPEG Audio Coding for Machines

Thomas Sporer[1], Sang-Wook Kim[2]

[1] *Fraunhofer IDMT, Ilmenau, Germany, E-Mail:Thomas.Sporer@idmt.fraunhofer.de*
[2] *Chung-Ang University, Seoul, Korea, E-Mail: swkim4@cau.ac.kr*

Introduction

In the acoustic community MPEG is recognized as the place where coding standards have been created. These standards exploit properties of the human auditory system to reduce data rate for transmission and storage. Newer standards, like MPEG-H and MPEG-I, improve the perceived quality to 3D and 6DoF.

MPEG is now addressing data formats for AI algorithms and, in respect to acoustics, machine listening. A key challenge for all AI algorithm is the quality of training data: Bad data or data of insufficient size used for training can lead to unexpected behaviour of systems (wrong classification, hallucinations).

The data used for training must be described as precise as possible with metadata, including description of content ("what"), recording environment ("where") and recording configuration ("how").

To enable exchange of annotated audio files between laboratories, applications, and projects a standardized format for audio and metadata is needed.

MPEG started with collecting use cases and requirements in 2022 [1]. In the beginning the focus was on industrial applications, only but with more stakeholders becoming involved applications now include ambient assisted living, medical monitoring, and user generated media classification. The scope was extended from training-only to the use of the format for the storage and/or transport of audio data from sensor (arrays) to the classification. MPEG plans to issue a Call for Proposals in April 2025.

This document gives some insights in the status. Contributions and collaboration in the work towards the standard is encouraged.

Motivation

Humans are very good in recognition of different acoustic signals. Examples are speech recognition and speaker recognition. We decompose complex sound scenes and can recognize important sound events in background noise. Experienced workers recognize malfunction of machines before they are totally broken. To mimic human listening with machine listening is a challenging task which need proper training data. In AI different approaches are used for training: supervised learning, where the algorithm receives content and annotations ("ground truth"), and unsupervised learning, where the training is based on huge data sets and the algorithm learns statistics of "normal" data just by observing. In Machine Listening (ML) pure unsupervised learning was not successful in the past: In general, at least part of the training data must be annotated (semi-supervised learning).

Applications and Use Cases

Experts in MPEG collected applications and use cases for ML. Another important source of inspiration about use cases was the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE) [2]: The challenges organized by this community consist of different tasks, and each task addresses different application areas, and in general, defines every time a new metadata format.

MPEG clustered the use cases [3,4]:

- **Industrial Applications**
  - **Predictive maintenance** (UC1)
    Detect early that a machine might be broken soon.
  - **Production Process Control** (UC2)
    Control production parameters for aging components to preserve best manufacturing quality.
  - **Inline testing in manufacturing** (UC3)
    Check quality of semi-finished products in (noisy) manufacturing.
  - **End-of-line testing** (UC4)
    Check quality of end products in measurement environment.
- **Surveillance of outdoor environments**
  - **Traffic and Public spaces** (UC5)
    Detect and analyse crowds concerning movements and mood (e.g. concerts, sport events, tourists). Acoustically track vehicles and control traffic lights.
  - **Construction sites** (UC6)
    Track vehicles and detect intruders in construction sites to reduce risk of accidents and theft.
- **Surveillance and control in personal (resp. indoor) environments**
  - **Assistance and health care**
    Example: Detect unusual behaviour like noises beside bed in stead of inside bed.
  - **Safety and security**
    Example: Recognize visitor whether known person or unknown person by voice/sound.
  - **Comfort services**
    Example: Monitor and control household devices by sound detection.

This includes voice/sound detection and acoustic Scene Analysis (UC7)

- **Biomedical applications**
  - **For medical doctors**: recording and analysis of time series for documentation, diagnosis and research (UC8)
    Examples: (longterm) EEG and ECG
  - **For consumers**: monitoring live with mobile phones (UC9)
    Examples: Heartbeat and snorring

  These applications together are called coding of Timed Medical Data.

  Not all biomedical signals are acoustic signals, but they share similar properties as being several related channels of 1D time series.

  UC8 is also addressed in a joint group of ITU-T SC21 and MPEG [5].
- **Entertainment and Content analysis**
  - **Analyses of User Generated Content** (UC10)

    Analysis and annotation of user generated content (UGC) e.g. during upload to simplify later retrieval of content with search engines.
  - **Live Stream Content Analysis** (UC11)
  - **Artistic Creation** (UC12)

Requirements

The data format must be capable to store the audio signals and metadata describing the recording. Depending on the application many audio channels might be necessary: If a complete manufacturing plant should be analysed where many sound sources are simultaneously active the recordings of many sensors can be used to separate these sound sources and to analyse them separately. Most audio codecs standardized in MPEG are based on perception: With the help of a psychoacoustical model the quantization error is shaped to become non-perceivable for humans (irrelevancy reduction). If such perceptual coding would be uses machine listening would be limited to the performance of humans. Especially when many channels and unusual frequency ranges are used the relevant information might be in the tiny differences between channels. Currently standardized algorithms for lossless audio coding are restricted in the number of channels they can analyse for redundancy reduction.

Therefore, the codec to be used in the new format must be multichannel and lossless.
Depending on the application the metadata might be very different but there are at least three categories which must be present:

- Description of the recording: Pose and type of each sensor, parameters of room acoustic, climate and time of recording. Note: time might be used for data fusion if more than one sensor array is used to capture the same acoustical scene.
- Description of the object: What was recorded, what was the object doing and the pose of the object.
- Ownership and usage rights of the recording.

Depending on the application the precision of coordinates in the metadata is limited: for tiny scenes the position of sensors must be in the range of 1 mm, while for large scenes a precision of several cm might be sufficient. The coding of geometric data must be in the limits of this application specific precision.

Scope

Based on these requirements the scope of the MPEG project on Audio Coding for Machines was defined:
A format to loss-less encode multichannel audio and near-lossless encode metadata to be used for storage of data for training of algorithms for machine listening, and for transport between sensor (arrays) and a central data processing and analysis unit. The coding of the audio essence should be application agnostic. The metadata format must be scalable: depending on the application only (available) subsets of metadata are stored resp. transmitted.

MPEG Call for Proposals (CfP)

For the CfP a package is currently in preparation. The following list describes the status. There might be changes until the CfP is issued.

- A definition of the use cases used in the CfP.
- Test data to be used for the CfP: The audio essence, the metadata file, and a usage rights file for each item.
- A specification of the metadata format. (see below)
- A definition of an anchor resp. baseline codec: Candidates for the baseline codec are MPEG-4 SLS (for audio essence) and zip for the metadata.
- Rules for submission and test evaluation: Every proposal must consist of the bitrate reduced file for every test item and a decoder executable. A proposal which is not lossless for the audio essence, or where the precision for the metadata is not within the limits specified, is excluded. Every proponent published performance numbers for each item. Other proponents check this numbers based on the submitted decoder.
- A specification of the figure of merit: For each test item the performance of each proposal is calculated relative to the performance of the baseline. Both the average performance over all test items and the worst-case performance are considered for selecting a winner of the CfP. Based on the winner the so-called Reference Model 0 (RM0) is implemented and available for further improvements of the standard.
- The rules for the Core Experiment procedure. In the CE phase everybody can propose improvements on the "winner": If either the performance is significantly better, or additional requirements not tested in CfP are met without decreasing the

performance, these improvements are added to or replacing parts of the RM.

Metadata format

The metadata file for each item can contain five groups of values:

- **General** describe the recording situation
- For each channel **Sensor** specifies the file and the microphone resp. loudspeaker parameters
- **Annotation** describes information that a machine is expected to infer while performing a task. Annotations might be time-stamped.
- **Object** describes the object recorded
- **Rights** describes the usage rights of the data

Depending on the use case (application) only a subset of the fields is present. Non optional fields are the number of sensors and the file names of the files containing the audio essence.

The formal specification of metadata will be in XML or json.

Group/Field	Description
Group: General	
samplingRate	Sampling rate in Hz
audioFormat	Specifies int16/int24/int32/float32 or wav
Humidity	Relative humidity
Temperature	Temperature in C
airPressure	Air pressure in hPascal
numberOfSensors	Number of AudioStreams
numberOf-Annotations	Number of Annotations
dateTime	Date and time of recording in UTC
Group: Sensor	
Id	Unique identifier of sensor channel
File	File name of recording channel relative to metadata file
SensorType	Brand name and type of sensor
SensorDirectivity	Either omni, cardioid, super-cardioid, hypercardiod, or filename of balloon data specifying sensor
SensorPosition	Comma separated of cartesian coordinates of sensor in meters (1) or string defining medical sensor label (2) or position label of loudspeaker (3)
SensorPosition-Tolerance	Comma separated of tolerance of cartesian coordinates of sensor in meters
SensorDirection	Cartesian vector of sensor direction
SensorDirection-Tolerance	Tolerance of sensor direction
SensorGain	Scaling of sensor signal in dB (positive: data in file is amplified compared to raw output of sensor)
Group: Annotation	
Id	Unique identifier of annotation
File	File name of annotation
AnnotationType	Type of annotation
Group: Object	
Name	Verbal description of object measured
OperationMode	OperationMode like “heavy load”, “broken”, “gear defect”, “motor blocked”, …. This might also include a specification of material processed and tool used (if machine has different tools)
ObjPosPrecision	Comma separated of tolerance of cartesian coordinates of object in meters
ObjOrientation	Cartesian vector of object direction
ObjOrientDescr	Text describing semantics of ObjOrientation
ObjOrient-Tolerance	Tolerance of ObjOrientation
Transcript	For content containing speech recordings this might contain a transcript of spoken words (ground truth)
Group: Rights	
RightsType	Define the usage rights of the data. Valid values are “public_domain”, “real-time_processing”, “MPEG_internal”
RightsFile	link to the rights file relative to directory of audio data
ContentOwner	Ownership of data (name, Institute, department)
Internal Use Only	Flag indicating that all data (incl. the metadata file should not be used outside the company of content owner.
Group: Further	
Category	Kind of content (e.g. industry, domestic, tool)
Usage	All further relevant information

Expected Timeline

The following timeline depends on the finalization of all CfP documents latest April 4th, 2025. If one of components for the package is not ready in-time there will be a delay of at least three months.

Date	Action
April 2025	Issue Call for Proposals package
April 2025	Verified version of test material is available
Aug 1, 2025	Registration of proposals
Sep 1, 2025	Upload bitstream files, results and decoders
Sep 8, 2025	Begin cross-checking others' results.

Date	Action
Sep 22, 2025	Complete cross-checking others results
Sep 26, 2025	Submit results of cross-checking as a contribution to the 21st WG 6 meeting
Sep 26, 2025	Submit proponent documentation as a contribution to the 21st WG 6 meeting
Oct. 2025 (10/6-10/10)	Evaluate submissions received and select technology. Respondents must be present at the meeting and present the proposals for information sharing. The source code of candidate technologies for the reference model will be published.
Oct 24, 2025	Source code of the selected technologies published on Git
Jan. 2026	Submit reference model (RM) for selected technologies as input contribution.
Jan. 2026	Submit Working Draft (WD) for selected technologies as input contribution.
	Improvements of WD and RM in Core Experiments
Oct. 2026	Committee draft (CD)
Jan. 2027	Draft International Standard (DIS)
Jul. 2027	Final Draft International Standard (FDIS)
Jan. 2028	International Standard published

Invitation to collaborate

Standardization is a joint activity of many companies, research institutes and universities. To participate in meetings, it is necessary to be sent by a national committee (in Germany DIN, in Denmark DS). Convenors are allowed to invite experts which are not members for a limited time.

There are different roles to participate [6]. For ACoM possibilities might be:

- **Helping to define use cases and requirements**
  If you need a file format for an application which is not covered in the list of use cases you can influence the standardization to create what you need.
- **Donating content for test**
  For the CfP annotated content is available, but for the CE phase, verification tests, and for a future version (non-lossless, feature coding) additional content is highly appreciated.
- **Proposing and improving algorithms**
  The call for proposal will be public (see press release of MPEG) and everybody (including non-members) can propose algorithms. An important requirement is that you have to comply to ISO rules [7].

If you are interested, you can contact the authors who are Co-Chairs of the AhG on ACoM.

Summary and Outlook

This document focuses on the CfP for first version of ACoM. It is expected that a second version will follow, which will be based on feature coding and provide a lower total bitrate. This second version probably will be application specific.

References

[1] MPEG140 (2022-10-28): Use Cases and Requirements for Audio Coding for Machines, (WG6 N0172/WG2 N0252),
(MPEG output document, not public).

[2] IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE),
URL: `https://dcase.community`

[3] MPEG146 (2024-04-29):
Use Cases and Requirements for Audio Coding for Machines, (WG6 N0252),
URL: `https://www.mpeg.org/wp-content/uploads/mpeg_meetings/146_Rennes/w23972.zip`

[4] MPEG149 (Geneva, 2025-01-24): Use Cases and Requirements for Audio Coding for Machines, (WG6 N0308),
(MPEG output document, not public).

[5] ITU-T: Call for Proposals for a new ITU-T Recommendation H.BWC on the coding of biomedical waveform data (Rennes, France, 26 April 2024).
URL: `https://www.itu.int/en/ITU-T/studygroups/2022-2024/16/Documents/docs/CfP-H.BWC-TD-PLEN-0286-R1-Clean.pdf`

[6] ISO: Developing standards,
URL: `https://www.iso.org/developing-standards.html`

[7] ISO: Directives and polices,
URL: `https://www.iso.org/directives-and-policies.html`