 ISO/IEC JTC 1/SC 29/WG 2 N305

**ISO/IEC JTC 1/SC 29/WG 2  
MPEG Technical requirements   
Convenorship: SFS (Finland)**

**Document type:** Output Document

**Title:** Call for Interest – Audio Coding for Machines

**Status:** Approved

**Date of document:** 2023-07-21

**Source:** ISO/IEC JTC 1/SC 29/WG 2

**Expected action:** None

**Action due date:** None

**No. of pages:** 6 (with cover page)

**Email of Convenor:** igor.curcio@nokia.com

**Committee URL:** <https://sd.iso.org/documents/ui/#!/browse/iso/iso-iec-jtc-1/iso-iec-jtc-1-sc-29/iso-iec-jtc-1-sc-29-wg-2>

**INTERNATIONAL ORGANIZATION FOR STANDARDIZATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC 1/SC 29/WG 2**

**MPEG TECHNICAL REQUIREMENTS**

**ISO/IEC JTC 1/SC 29/WG 2 N305**

**Geneva, Switzerland - April 2023**

|  |  |
| --- | --- |
| **Title** | **Call for Interest – Audio Coding for Machines** |
| **Source** | **WG 2, MPEG Technical requirements** |
| **Status** | **Approved** |
| **Serial Number** | **22908** |

# Abstract

WG2 is assessing the pros and cons of working on Audio Coding for Machines (ACoM) within MPEG. This document summarizes the current status of the discussions within WG2 “Requirements” and WG6 “Audio Coding” and invites contributions from within and outside MPEG on this topic. Contributions might include expression on interest on using such a format in applications, or in working on the creation and standardization of the technology.

Contributions might also amend the list of application areas listed in this document.

Interested parties might be, but are not limited to, from academic institutions, research labs, service providers, device manufacturers, equipment vendors, network operators, and technology providers.

# Introduction

All audio coding schemes and formats standardized by MPEG have so far targeted human consumption of audio content. High compression ratios have been achieved by incorporating models of the human auditory system. Such data formats might be not adequate for computerized analysis of sound and sound scenes. Audio Coding for Machines is targeting several new applications where computers, rather than humans, are listening and analyzing audio content.

In principle, two different operations modes are foreseen:

* The new data format is used to store and exchange data to be used in the development of audio analysis algorithms. In general, this will involve training artificial intelligence (AI) models using huge data sets.
* The new data format is used for the communication between acoustic sensors and some analysis and control algorithms. In general, such algorithms will be AI-based and trained on huge data sets.

For both operation modes, the data set will contain the audio essence together with rich metadata describing the content.

# Application areas

This section lists applications and key requirements already identified.

# Industrial Applications

Algorithms to analyze acoustic data for predictive maintenance, process control, in-line and end-of-line testing need training data. This activity will specify a format to store the spatial acoustical data in an interoperable format. The format will contain measured audio data either in raw format or as acoustical features together with metadata describing the data.

Data about a machine and its different operation states stored in the format can be considered as an acoustic twin of the machine.

Business model: There might be a split between manufacturers of machines and providers of monitoring systems. Manufacturing sites usually combine machines from different vendors, but there is a need to monitor the whole process. ACoM might be the format these two groups need to be able to interoperate.

**Predictive Maintenance**

Acoustic sensors are used to monitor the function of a single machine or all machines in a hall. Predictive maintenance aims at replacing components before failure but as late as possible. Acoustic sensors can be outside a machine. Therefore, retrofit of old machines is possible. Neural networks are used to train detectors. Usually, the differences to be detected are small compared to differences within the data in normal operation. Usually, there is much more training data available for correct operation. Therefore, the trained algorithm might often be limited to detecting abnormalities.  
Key Requirements: Efficient near-lossless storage of spatial audio data together with status of machine (ok versus (known) error-classes).

**Process Control**

Acoustic sensors are used to control the function of a machine. Process parameters are modified to compensate aging of components or to adapt to differences in the material being processed. Acoustic sensors can be outside a machine and therefore not influenced by dirt or other disturbance. Neural networks are used to train the process parameters.

Key Requirements: Efficient near-lossless storage of spatial audio data together with status of machine (ok versus (known) error-classes). In the application, fast reaction to failure is important. Therefore, the format should allow for fast access to data.

**In-line Testing**

Acoustic sensors are used during production to detect whether individually produced components are out of spec. Such components can be discarded before further processing. Acoustic sensors are close to the process in the machine. Neural networks are used to train detectors. Usually, the differences to be detected are small. Usually, there is a lot of environmental noise. Usually too, there is much more training data available for correct operation.

Key requirements: Efficient near-lossless storage of audio data (small number of channels) together with acoustical fingerprints (ok or nok) of processed good. In the application, fast reaction to failure is important. Therefore, the format should allow for fast access to data.

**End-of-line Testing**

Acoustic sensors are used to detect whether a final product is out of spec. Acoustic sensors are after the processing in the machines and might be in a sound insulated measurement cabin. Neural networks are used to train detectors. Usually, the differences to be detected are small. Usually too, there is no environmental noise, which enables very precise measurement. Usually again, there is much more training data available for correct products.

Key requirements: Efficient near-lossless storage of audio data (small number of channels) together with acoustical fingerprints (ok or type of failure) of processed good. In the application, fast reaction to failure is important.

**Automotive and Trains**

Acoustic sensors built into a car are used to detect any abnormalities in the sound of the vehicle. Examples of such abnormalities might be concerning engine noise, gears, wheels, or street conditions. Neural networks are used to train detectors. Especially in autonomous driving, such sensors might be a necessary replacement to the knowledge of an experienced driver and might be necessary for predictive maintenance and control of the vehicle.

# Site Monitoring Application

Algorithms for acoustic surveillance, for instance to monitor and control traffic flows of vehicles and pedestrians in public spaces, or to monitor and track vehicles entering and passing through construction sites, need training data. This activity will specify a format to store the spatial acoustic data in an interoperable format. The format will contain measured (recorded) audio data either in raw format or as acoustic features together with metadata describing the data. Part of the metadata can be used as ground truth in the training of algorithms.

The format to be developed has to store large data for training of algorithms. In all scenarios listed below, an important high-level requirement is “efficient storage” which means low bitrate. Another high-level requirement is related to the issue of privacy protection: the format should never store speech in a way that it can be understood by machines or humans.

**Traffic Monitoring and Control**

Acoustic sensors are used to monitor traffic flow in cities. For this purpose, a network of acoustic sensors is installed in a city. The sensors could not only count cars but also classify them into groups and control traffic signs by detecting siren, enabling faster progression of emergency vehicles. The acoustic sensor network can also be used to track and control crowds of people. The tracking of crowds might include prediction of the general mood of a crowd (for example: cheerful, aggressive, or happy).  
Neural networks are used to train the tracking and classification algorithm. In general, the number of audio objects to be tracked is larger than the number of sensors.

Key Requirements: Efficient near-lossless storage of audio data of spatially-separated sensors together with a time code and ground truth of the scene recorded, like position (and trajectory) of audio objects, classification of the object. Metadata will include properties of the sensor nodes like position, orientation, and type.

**Construction Site Monitoring**

Acoustic monitoring is used for automatic detection and tracking of vehicles entering in and driving through large construction sites. This enables monitoring the flow of goods, preventing accidents by predicting conflicts, and avoiding theft of goods. In addition, such systems can be used to identify the guilty party in case of noise pollution above legal restrictions.

Construction sites are usually dirty. Acoustic sensors are more robust and can even work when partially occluded by objects.

Key Requirements: Efficient near-lossless storage of audio data of spatially separated sensors together with a time code and ground truth of the scene recorded, such as position (and trajectory) of audio objects, classification of the objects. Metadata will include properties of the sensor nodes like position, orientation, and type.

# Market Relevance

The **industrial** market for a generic metadata-aware spatial audio data format is broad:

• While information and communication technology becomes more and more important, in the end, there are mechanical devices to be produced.

• Mechanical production is more and more controlled by computers.

• In such automated production plants, only a small number of experts are present. They are usually in the control room, but not anymore in dangerous, noisy and dirty proximity of machines. In the past, the worker was trained to detect the sound that characterizes miss-function of machines, to control manufacturing processes (examples: welder, cutting machine operator), and to notice miss-produced goods. This training is not happening anymore, and the machines have to learn how to listen. The ACoM format gives machines that listen the necessary memory to learn about acoustical situations.

• With aging societies and the growing shortage of expert workers, the need to support remaining workers is increasing. Automatic manufacturing based on all human senses on which a skilled worker relies becomes a necessity.

The market for mechanical engineering, as part of industrial production, is around 2.8 trillion USD worldwide, with about 33% in China, 32% in Europe, 18% in Asia (without China) and 14% in North-America (source: AUMA, 2019). The market for communication in test and measurement, that means just the link between sensors, is expected to grow to 10 billion USD in 2025. In addition to the wireless links (5G), this number also includes the data formats (https://www.marketsandmarkets.com/Market-Reports/communications-test-measurement-market-1309.html). Due to the fact that there currently is no existing standardized format for acoustic data in industrial context, no statistics for acoustic formats is available.

**Traffic Monitoring and Control**

• Traffic monitoring and control is necessary to protect people, to reduce the risk of traffic jams, and to provide priority for emergency vehicles.

• In some countries, video surveillance of public spaces is an important market. In other countries, there is strong opposition and such systems are even stopped. However, audio algorithms in general are more accepted, especially if privacy is guaranteed by design criteria.

• Systems based on audio can be completely invisible, which facilitates acceptance.

• Audio sensors can “listen around occluders” and are less prone to failure caused by accident or intentionally.

The global traffic management market size is expected to grow from 38.2 billion USD in 2022 to 68.8 billion USD in 2027 (https://www.marketsandmarkets.com/Market-Reports/traffic-management-market-1036.html). These numbers include not only the video surveillance but also components for predictive traffic monitoring and incident detection. Numbers for audio-only applications are missing.

**Construction Site Monitoring:**

• Construction sites are among the sites where many different contractors still have to work together. Coordination between vehicles and flow of goods on larger construction sites is still a major problem. But even with coordination on such sites, vehicles are controlled by humans, and humans make mistakes.

• Video surveillance at construction sites is mainly used to avoid unwanted access during night and weekend but not used during worktime.

• Systems based on video are often rendered useless because of objects (e.g., trucks, cranes, material) standing in the line of sight. Audio sensors listen around such occluders.

• Systems based on video are often rendered useless because of dirt or water. Audio sensors can work in such unfriendly environments.

The global Construction Site Monitoring System market size is projected to reach multimillion USD by 2030 (https://www.marketwatch.com/press-release/2023-2030-construction-site-monitoring-system-market-research-2023-06-12). Currently the market is dominated by video. Numbers for audio are missing.

# Further Information:

The following contributions concerning the topic of Audio Coding for Machines have already been received by WG06. Note that these documents are only available to MPEG members.

M61162: Proposal for New Work Item: Audio Coding for Machine

M61860: Thoughts on Audio Coding for Machines (ACoM)

M63138: Thoughts on Audio Coding for Machines

M63225: Use Cases and Requirements on Audio Coding for Machines

M64285: Market and practical considerations - ACoM Industrial

M64288: Market and practical considerations - ACoM SiteMonitoring

M64290: Market and practical considerations - ACoM Medical

# How to contribute

If you are an MPEG Member, you can use the Template of the MPEG WG2 (Technical Requirements) on "Market & practical considerations (Doc. W20949)" and submit an individual input to the next MPEG meeting via https://dms.mpeg.expert/.   
Please select WG2 and AdHoc Group "Market Needs".

If you are not an MPEG Member, or if you want to support the activity as a user but do not want to be visible, you can send an email to the Convenor of WG2 <igor.curcio@nokia.com>.