 ISO/IEC JTC 1/SC 29/WG 2 N0343

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

**Document type:** Output Document

**Title:** Scope and Roadmap for Audio Coding for Machines (ACoM)

**Status:** Approved

**Date of document:** 2024-01-26

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

# Expected action: None

# Action due date: None

**No. of pages:** 25 (with cover page)

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

**Committee URL:** <https://isotc.iso.org/livelink/livelink/open/jtc1sc29wg2>

**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 N** **0343**

**Online - January 2024**

|  |  |
| --- | --- |
| **Title** | Scope and Roadmap for Audio Coding for Machines (ACoM) |
| **Source** | **WG 2, MPEG Technical requirements** |
| **Status** | **Approved** |
| **Serial Number** | **23689** |

# Summary

WG2 is currently exploring market needs, use cases, and requirements for an audio coding format for machine analysis of audio content. Discussions in MPEG has shown that the topic needs a clear definition what ACoM is about.  
In this document the scope is defined, and a draft roadmap shown.   
Attached are the most recent working documents concerning “use cases and requirements”, and about “market needs” for some application areas.

# Scope

Audio Coding for Machines (ACoM) is a format for coding

* of audio data and metadata
* used for training in machine listening applications
* used for interoperability and exchange of training data

The Format

* is use case agnostic
* might use different subsets of metadata for different use cases
* considers privacy issues if need in use cases
* includes metadata describing microphone configurations (e.g. 3D microphones and distributed microphones)

Audio data

* in general, consists of several independent 1D time series of samples (“multi-channel”)
* biomedical data has similar structure and might be considered special case of audio

The format might be used in later applications also to specify microphone signals, too.

# Roadmap

* Collect information on use cases.
* Collect list of companies to contribute either on use cases, requirements, or with providing technologies.
* Define list of metadata necessary for each use case.  
  Necessary metadata might be related to the use case. Protection of privacy important for use cases which are not related to speech processing.
* Define requirements on audio essence.
* Result of these preparation phase is a requirements document.
* Collect data necessary for specification and development of the format.
* Work on **first phase** of creating a standard:
  + Audio coding use **loss-less** coding (including raw audio)
  + Specification and efficient representation of **metadata**
  + Carriage of flexible subsets of metadata
* Work on **second phase** of standard:
  + Find use case specific **features** replacing the loss-less coding.

# Attachments

The attached documents show the current status of work.

|  |  |
| --- | --- |
| Annex A | Use Cases and Requirements for Audio Coding for Machines |
| Annex B.1 | Market and practical considerations - ACoM Industrial |
| Annex B.2 | Market and practical considerations - ACoM Site Monitoring |
| Annex B.3 | Market and practical considerations - ACoM Lifelogging |
| Annex B.4 | Market and practical considerations - ACoM Medical |

**Annex A**

**Use Cases and Requirements for Audio Coding for Machines (Source: WG2 N252, Serial number: 22180)**

# Abstract

This Document explains the scope, the use cases and requirements of the proposed new work item for WG6 “Audio Coding for machines”. **More input on use cases and requirements is needed.**

# Introduction

A topic not yet covered by WG 6 is **audio coding for machines (ACoM)**: If the receiver of coded audio content is not a human being but a computer program which has to analyze the audio scene, the use of perceptual models, like being used in previous MPEG audio standards, might not be adequate. Many Use Cases will be based on machine learning, that is the learn about different typical situation. Machine learning in general relies on big data. Therefore, data bases of many audio recordings are necessary. To store large data bases bitrate and coding becomes an issue.

Depending on the application the sounds to be analyzed might be very different. For the general case it is not possible which features of the sound are relevant for analyzing. There for in a first phase it is expected that the compressed format will not include feature extraction.

This document lists Use Cases for ACoM and Requirements on such schemes. Currently both lists are incomplete and should be amended as necessary.

Figure 1 gives a simplified block diagram where in a value chain the activities of WG 6 on the topic will start: A machine producing noise is measured (or simulated) from different directions.

Two formats will be worked on in WG 6: A raw format containing the PCM data and metadata and a bitrate reduced version (compressed) format. These formats can be understood as complete acoustical representations of the machine: an acoustical twin of the machine.

The data base using the acoustical twin is used to train the evaluation algorithm.

The work in WG 6 will be on the formats and the encoding algorithms.

# 

# Figure 1: Simplified block diagram of value chain

Figure 2 gives some additional insights: ACoM is used to store the data many machines in a standardized format. That way many different evaluation systems can use this data. Creation of the acoustical twin might be done either by the manufacturer of the machine or external service providers. Creation and implementation of evaluation systems in general will not be the manufacturers of the machines but other companies. Many production sites will use machines from different vendors and the format enables interoperability concerning audio monitoring.

# 

# Figure 2: Interoperability between machines and evaluation systems based on ACoM

# Use Cases

The following list gives examples of use cases:

* Industrial applications  
  In industrial applications the focus is on improving quality of products and reducing production costs. With the fourth industrial revolution a small number of machine operators are in the control room and no more near the machines. In the past operators used their ears to know what is going on but in the control room they loose the ability to do so.
  + **In-line testing**:

During production the sound of the machines are monitored to detect misfunction of the machines and also the sound of semi-ready products (for example when they fall out of a machine).   
In general in-line production is happening in a very noisy environment.

There might be the measurement of many machines and/or products in parallel (i.e. based on microphone array). This application therefor might merge and analyze the data from data bases of several acoustic twins.

*Examples*:

- Crack detection on vents  
- Leakage detection in pneumatic systems

- Leakage detection of vacuum needed for filling beer bottles

- Detection of bad welding connection

…

* + **End of line testing**

In general end of line will happen in a controlled environment.

*Examples:*

- Using build in loudspeakers for end of line testing of car interior

- Testing electrical motors  
- Testing roof tiles   
- Testing of microphone and loudspeakers

….

* + **Predictive monitoring  
    *Examples***
  + **Process control**

*Examples*:   
- Arc welding: Acoustic data is used to control parameters like current and distance of electrode.

- Acoustic monitoring of laser ablation processes: For recycling electronic components a laser is used to separate coating from expensive materials. Based on acoustical monitoring power and duration of the laser is controlled.

* **Prediction of noise exposure**
  + The acoustic twins are used to predict sound distribution to fulfil legal requirements concerning noise exposures to workers and environmental.
* **Traffic measurement**
  + Counting number and type of vehicles in a town. A special case of vehicle type are ambulance, police, or firefighters. The data can be used to control traffic lights.
  + Detecting and tracking crowds of people in towns, for instance to reduce risk of congestions and panic (i.e. sport events, musical events).
* **Construction site monitoring**
  + Counting and tracking vehicles in (large) construction sites to avoid unloading goods at wrong place or that vehicles are approaching dangerous regions.

# Requirements

* Storage of spatial acoustic data
  + Raw format as interface to measurement and acoustic simulation
  + Compressed format of the raw format   
    (parameters: bitrate and computational load)

The acoustic data to be stored might be based on measurement from many positions around the machine using an intensity probe. Encoding should be virtual lossless but might use redundancy reduction based on similarity of measurements done in parallel.

Note: The relation between such measurements to HOA is like the relation of “cameras pointing” inside to omni-cameras (360 degrees pointing outside).

* Storage of (measurement related) metadata
  + Number, position(s) and type(s) of sensors
  + Measurement conditions (temperature, humidity, air pressure, …)
  + Room acoustics
  + Operation mode of machine

**Annex B.1   
Market and practical considerations - ACoM Industrial  
(Source: N64340)**

# Abstract

This contribution highlights some market & practical considerations regarding potential needs for audio coding for machines (ACoM) with the focus on **industrial applications** using the template developed in the market Needs AhG.

# Introduction

In WG6/N0203, an overview of use cases and requirements on audio coding for machines (ACoM) is given. At the 11th meeting, WG6 decided to ask for assistance from WG2 to create a report on market and practical considerations. In AG6/N023, eight uses cases are listed. This document focuses on the scenarios in industrial context:

* Predictive Maintenance
* Process Control
* In-line Testing
* End-of-line Testing

**Market and practical considerations questionnaire**

# Topic presentation

|  |  |
| --- | --- |
| A – Topic presentation | |
| A.1 | Please introduce the activity/domain that requires the development of new technology |
| *Algorithms to analyze acoustic data for predictive maintenance, process control, in-line and end-of-line testing need training data. This activity should specify a format to store the spatial acoustical data in an interoperable format. The format must 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.*  *Note: This activity is about the* ***data format*** *used for training but not the detectors.* |
| A.2 | Please provide a brief description of high-level scenarios expected to be impacted by the technology.  For each of them, are there any associated high-level requirements? |
| *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.*  ***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 often the trained algorithm might be limited to an abnormally detection. 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 material processed. Acoustic sensors can be outside a machine and therefore not influenced by dirt or other disturbance. Neural networks are used to train 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 be capable of 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, 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 be capable of 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 process 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, there is no environmental noise enabling very precise measurement. Usually, 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 type of failure) of processed good. In the application fast reaction to failure is important.* |

# Timing aspects

|  |  |  |
| --- | --- | --- |
| B - Timing aspects | | |
| B.1 | Necessary date for completion of specification, if known. | |
| *Not known* | |
| B.2 | When are compliant products/services required to be available? | |
| *Recordings of typical acoustical data are necessary. Due to the reason that there is no standard for the exchange of such data, such recordings have to be part of the work within MPEG.* | |
| B.3 | Is the technology already implemented? | |
| Please select | * Other: some incomplete and proprietary file formats exist (non-interoperable prototypes) |

# MPEG Technologies

|  |  |
| --- | --- |
| C - MPEG technologies | |
| C.1 | Are there existing MPEG or other standards/technologies addressing the proposal? |
| *Some standards are related but do not address the proposal completely.*  ***MPEG****:*  *MPEG-4 SLS (ISO/IEC 14496-3) defines ways for lossless and near-lossless coding of a small number of channels.*  *MPEG-H 3D Audio (ISO/IEC 23008-3) specifies ways to encode spatial audio as individual channels, objects and in HOA format.*  ***AES:***  *AES 69 Spatially Oriented Format for Acoustics" (SOFA) specifies ways to encode spatial audio impulse responses.*  ***OPC****:*  *The open platform communications foundation (OPC) creates and maintains standards for open connectivity of industrial automation devices and systems, such as industrial control systems and process control generally. The specifications enable to create a digital twin of machines.*  ***DICONDE****:*  *Digital Imaging and Communication in Non-Destructive Evaluation (DICONDE) is an open standard format for the display, transfer and storage of digital non-destructive evaluation data.* |
| C.2 | If yes, why are they deficient or don’t meet the need? |
| *MPEG-4 SLS is not able to store the necessary metadata and does not provide tools to exploit spatial redundancies between many channels.*  *MPEG-H 3D Audio is not able to store the necessary metadata. MPEG-H 3D audio is based on perceptual audio coding which is neither lossless nor near-lossless. ACoM is not intended for human consumption and the perceptual coding of MPEG-H 3D Audio might delete the components needed for acoustic processing by machines.*  *AES 69 is not able to store the necessary metadata and can only store (short) impulse responses. The format is lossless but does not provide efficient storage (no means to exploit redundancies to reduce bitrate).*  *The set of standards in OPC currently does not include acoustical data.*  *The set of standards in DICONDE currently does not include acoustical data.* |
| C.3 | If not, will a new specification be required? Any dependency to existing MPEG or other standards |
| *Yes, a new specification will need to be made as no existing format addresses the problem of big data storage and transfer.* |
| C.4 | Which MPEG WG(s) is(are) expected to be impacted by the development of the technology? (if known) |
| *WG6 Audio.* |

# Market reach

|  |  |
| --- | --- |
| D - Market reach | |
| D.1 | Please indicate the target ecosystem (horizontal market, retail, vertical, professional…) and any geographical limitation (worldwide, region based…) |
| *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 of 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 machine that listen the necessary memory to learn about acoustical situations.* * *With aging societies and missing expert workers the need to support workers is increasing. Automatic manufacturing based on all senses which a skilled worker would use are necessary.* |
| D.2 | Is there any market research available associated to the proposed requirements? |
| *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. Besides the wireless links (5G), this number includes also the data formats (https://www.marketsandmarkets.com/Market-Reports/communications-test-measurement-market-1309.html). Due to the fact that there is no standardized format for acoustic data in industrial context, no statistics for acoustic formats is available.* |

# MPEG Member support

|  |  |  |  |
| --- | --- | --- | --- |
| E - Member support | | | |
| E.1 | Please list the members supporting the activity, including (for each of them) their level of support and constituency.   1. Please indicate the level of support for each member by selecting one of the following options:  * **interested** (only) * **active** (participation to the development of a technical solution)  1. Please indicate the most representative constituency of each member by selecting one of the applicable following options:  * **Academic** * **Research lab** * **Service provider** * **device manufacturer** * **equipment vendor** * **network operator** * **technology provider** * other (specify) | | |
|  | **Member** | **Level of support** | **Constituency** |
|  | *Fraunhofer* | *Active* | *Technology provider* |
|  |  |  |  |

# Additional information

|  |  |
| --- | --- |
| F – Additional information | |
| F.1 | Please provide any additional information that you think relevant for assessing the market needs |
| *The* ***DCASE*** *Community (Detection and Classification of Acoustic Scenes and Events, https://dcase.community/) is a scientific community where experts on algorithms work together. DCASE organizes workshops and challenges to compare algorithms based on real world data.*  *The 2022 challenges included the following tasks:*  *1. Low-Complexity Acoustic Scene Classification*  *2. Unsupervised Anomalous Sound Detection for Machine Condition Monitoring Applying Domain Generalization Techniques*  *3. Sound Event Localization and Detection Evaluated in Real Spatial Sound Scenes*  *4. Sound Event Detection in Domestic Environments*  *5. Few-shot Bioacoustic Event Detection*  *6. Automated Audio Captioning and Language-Based Audio Retrieval*  *Tasks 2 and 3 of DCASE2022 are related to the scenarios listed here.*  *In the context of DCASE the following comment from DCASE was received: “blind quality assessment is still an open issue and needs further research before standardization on algorithms should start.”  This is very much in line with the intention of the ACoM which focuses on a lossless format. Currently different formats for the audio data are used in this research community making exchange of data difficult. The DCASE community would benefit from a common format.*  *However, in future there might be a phase 2 of ACoM with extensions based on audio features adequate for algorithmic detection tasks providing lower data rates.*  *Several members in WG6 are also active in DCASE.*  ***Additional Information***   * *Automotive applications are not explicitly named, but are somehow monitoring the status of a machine (car). This might be a large market, too.* * *Business model: There might be a split between manufacturers of machines and of monitoring systems. Manufacturing sites usually will combine machines of different vendors but want to monitor the whole process. ACoM might be the format these two groups can interoperate.* |

**Annex B.2   
Market and practical considerations - ACoM Site Monitoring (Source: N64341)**

# Abstract

This contribution highlights some market & practical considerations regarding potential needs for an audio coding for machines (ACoM) technology with a focus on **site monitoring** using the template developed in the market Needs AhG.

# Introduction

In WG6/N0203 an overview of use cases and requirements on audio coding for machines (ACoM) is given. At the 11th meeting, WG6 decided to ask for assistance from WG2 to create a report on market and practical considerations. In AG6/N023, eight uses cases are listed. This document focuses on the scenarios in the context of site monitoring:

* Traffic Monitoring and Control
* Construction Site Monitoring

**Market and practical considerations questionnaire**

# Topic presentation

|  |  |
| --- | --- |
| A – Topic presentation | |
| A.1 | Please introduce the activity/domain that requires the development of new technology |
| *Algorithms for acoustic surveillance, for instance to monitor and control traffic flows of vehicles and pedestrians in public spaces, or to monitoring and tracking vehicles entering and passing through construction sites, need training data. This activity should specify a format to store the spatial acoustic data in an interoperable format. The format must contain measured (recorded) audio data either in raw format or as acoustic features together with metadata describing the data. Part of the metadata is used as ground truth in the training of algorithms.*  *Note: This activity is about the* ***data format*** *used for training but not the monitoring.* |
| A.2 | Please provide a brief description of high-level scenarios expected to be impacted by the technology.  For each of them, are there any associated high-level requirements? |
| *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 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 (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 must 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, like position (and trajectory) of audio objects, classification of the objects. Metadata must include properties of the sensor nodes like position, orientation and type.* |

# Timing aspects

|  |  |  |
| --- | --- | --- |
| B - Timing aspects | | |
| B.1 | Necessary date for completion of specification, if known. | |
| *Not known* | |
| B.2 | When are compliant products/services required to be available? | |
| *Recordings of typical acoustical data are necessary. Due to the reason that there is no standard for the exchange of such data such recordings have to be part of the work within MPEG. Scene description might be hand-annotated data.* | |
| B.3 | Is the technology already implemented? | |
| Please select | * already products/services deployed * **Proof of concepts/prototype:** exist for the scene analysis and classification algorithms * Planned roadmaps * Nothing yet * **Other:** some incomplete and proprietary file formats exist (non-interoperable) |

# MPEG Technologies

|  |  |
| --- | --- |
| C - MPEG technologies | |
| C.1 | Are there existing MPEG or other standards/technologies addressing the proposal? |
| *Some standards are related but do not address the proposal completely.*  ***MPEG****:*  *MPEG-4 SLS (ISO/IEC 14496-3) defines ways for lossless and near-lossless coding of a small number of channels.*  *MPEG-H 3D Audio (ISO/IEC 23008-3) specifies ways to encode spatial audio as individual channels, objects and in HOA format.*  ***AES:***  *AES 69 Spatially Oriented Format for Acoustics" (SOFA) specifies ways to encode spatial audio impulse responses.* |
| C.2 | If yes, why are they deficient or don’t meet the need? |
| *MPEG-4 SLS is not able to store the necessary metadata and does not provide tools to exploit spatial redundancies between many channels.*  *MPEG-H 3D Audio is not able to store the necessary metadata. MPEG-H 3D audio is based on perceptual audio coding which is neither lossless nor near-lossless. ACoM is not intended for human consumption and the perceptual coding of MPEG-H 3D Audio might delete the components needed for acoustic processing by machines.*  *AES 69 is not able to store the necessary metadata and can only store (short) impulse responses. The format is lossless but does not provide efficient storage (no means to exploit redundancies to reduce bitrate).*  *All formats are missing tools to respect privacy issues.* |
| C.3 | If not, will a new specification be required? Any dependency to existing MPEG or other standards |
| *Yes, a new specification will need to be made as no existing format addresses the problem of big data storage and transfer.* |
| C.4 | Which MPEG WG(s) is(are) expected to be impacted by the development of the technology? (if known) |
| *WG6 Audio* |

# Market reach

|  |  |
| --- | --- |
| D - Market reach | |
| D.1 | Please indicate the target ecosystem (horizontal market, retail, vertical, professional…) and any geographical limitation (worldwide, region based…) |
| ***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 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 improving acceptance.* * *Audio sensors can “listen around occluders” and are less prone to failure caused by accident or intentionally.*   ***Construction Site Monitoring:***   * *Construction sites are among the sites where still many different contractors have to work together. Coordination of 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.* |
| D.2 | Is there any market research available associated to the proposed requirements? |
| ***Traffic Monitoring and Control***  *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*](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 are missing.*  ***Construction Site Monitoring:***  *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*](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.* |

# MPEG Member support

|  |  |  |  |
| --- | --- | --- | --- |
| E - Member support | | | |
| E.1 | Please list the members supporting the activity, including (for each of them) their level of support and constituency.   1. Please indicate the level of support for each member by selecting one of the following options:  * **interested** (only) * **active** (participation to the development of a technical solution)  1. Please indicate the most representative constituency of each member by selecting one of the applicable following options:  * **Academic** * **Research lab** * **Service provider** * **device manufacturer** * **equipment vendor** * **network operator** * **technology provider** * other (specify) | | |
|  | **Member** | **Level of support** | **Constituency** |
|  | *Fraunhofer* | *Active* | *Technology provider* |
|  |  |  |  |

# Additional information

|  |  |
| --- | --- |
| F – Additional information | |
| F.1 | Please provide any additional information that you think relevant for assessing the market needs |
| *Several European projects on smart cities addressed the issue of algorithms, but did not look on interoperable data formats to train the algorithms.* |

**Annex B.3**

**Market and practical considerations - ACoM Lifelogging (Source: N66360)**

# Abstract

This contribution highlights some market & practical considerations regarding potential needs for an audio coding for machines (ACoM) technology with a focus on home and office usage scenarios of **lifelogging** using the template developed in the market Needs AhG.

# Introduction

In WG6/N0203, Use Cases and Requirements for Audio Coding for Machines, an overview of use cases and requirements on audio coding for machines (ACoM) is given. In this document, it is focused on the scenarios in the context of lifelogging:

* Speech Recognition and Acoustic Scene Analysis

**Market and practical considerations questionnaire**

# Topic presentation

|  |  |
| --- | --- |
| A – Topic presentation | |
| A.1 | Please introduce the activity/domain that requires the development of new technology |
| *Algorithms for lifelogging, for example, detect event, record and summarize dialogs during communication at home and work, or listen and identify people of interest in group speeches, need training data. This activity should specify a format to store the spatial acoustic data in an interoperable format. The format must contain measured (recorded) audio data either in raw format or as acoustic features together with metadata describing the data. Part of the metadata is used as ground truth in the training of algorithms.*  *Note: This activity is about the* ***data format*** *used for training but not the monitoring.* |
| A.2 | Please provide a brief description of high-level scenarios expected to be impacted by the technology.  For each of them, are there any associated high-level requirements? |
| *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.*  ***Speech Recognition and Acoustic Scene Analysis***  *Acoustic sensors are used to capture sound for speech recognition and acoustic scene analysis, which can be applied in everyday life, whether at home or at work. In the field of voice recognition, technology to convert voice into text has been researched and developed. These developed speech-to-text technologies typically provided the user with an intended recording starting point in the intended environment. Conversely, when a trigger signal to start recording is not provided, lifelogging requires determining the beginning and end of a sound of interest from a complex mixture sounds and detecting the location and identification of the sound source itself. When several types of sounds exist together in everyday life, it is not easy to separate the signals from each sound source. By analyzing scenes experienced in daily life, it detects events and determines the direction of the sound source, who is speaking, and what home device is special sound generating, helping to determine whether recording is necessary and what response is needed. This allows people to lead more comfortable and safer daily lives. Neural networks are used to train the detecting, 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 must include properties of the sensor nodes like position, orientation and type.* |

# Timing aspects

|  |  |  |
| --- | --- | --- |
| B - Timing aspects | | |
| B.1 | Necessary date for completion of specification, if known. | |
| *Not known* | |
| B.2 | When are compliant products/services required to be available? | |
| *Recordings of typical acoustical data are necessary. Due to the reason that there is no standard for the exchange of such data such recordings have to be part of the work within MPEG. Scene description might be hand-annotated data.* | |
| B.3 | Is the technology already implemented? | |
| Please select | * already products/services deployed * **Proof of concepts/prototype:** exist for the speech to text conversion algorithms, scene analysis and classification algorithms * Planned roadmaps * Nothing yet * **Other:** some incomplete and proprietary file formats exist (non-interoperable) |

# MPEG Technologies

|  |  |
| --- | --- |
| C - MPEG technologies | |
| C.1 | Are there existing MPEG or other standards/technologies addressing the proposal? |
| *Some standards are related but do not address the proposal completely.*  ***MPEG****:*  *MPEG-4 SLS (ISO/IEC 14496-3) defines ways for lossless and near-lossless coding of a small number of channels.*  *MPEG-H 3D Audio (ISO/IEC 23008-3) specifies ways to encode spatial audio as individual channels, objects and in HOA format.*  ***AES:***  *AES 69 Spatially Oriented Format for Acoustics" (SOFA) specifies ways to encode spatial audio impulse responses.* |
| C.2 | If yes, why are they deficient or don’t meet the need? |
| *MPEG-4 SLS is not able to store the necessary metadata and does not provide tools to exploit spatial redundancies between many channels.*  *MPEG-H 3D Audio is not able to store the necessary metadata. MPEG-H 3D audio is based on perceptual audio coding which is neither lossless nor near-lossless. ACoM is not intended for human consumption and the perceptual coding of MPEG-H 3D Audio might delete the components needed for acoustic processing by machines.*  *AES 69 is not able to store the necessary metadata and can only store (short) impulse responses. The format is lossless but does not provide efficient storage (no means to exploit redundancies to reduce bitrate).*  *All formats are missing tools to respect privacy issues.* |
| C.3 | If not, will a new specification be required? Any dependency to existing MPEG or other standards |
| *Yes, a new specification will need to be made as no existing format addresses the problem of big data storage and transfer.* |
| C.4 | Which MPEG WG(s) is(are) expected to be impacted by the development of the technology? (if known) |
| *WG6 Audio* |

# Market reach

|  |  |
| --- | --- |
| D - Market reach | |
| D.1 | Please indicate the target ecosystem (horizontal market, retail, vertical, professional…) and any geographical limitation (worldwide, region based…) |
| ***Speech Recognition and Acoustic Scene Analysis***   * *Speech Recognition and Acoustic Scene Analysis can be used as core technologies for human-machine interfacing tools and devices,* * *Speech Recognition and Acoustic Scene analysis can be applied on electronic documentation market, it can be a key technology in next generation word processor, voice recorder and electronic diary.* * *At home, machines can help older people recognize the presence of guests at the front door, remembrance of previous discussion and conversations, and monitor the operation of home appliances such as shutting down the washing machine, boiling, and cooking voice recognition and acoustic scene analysis.* * *In offices, devices can adopt speech recognition and acoustic scene analysis to provide conversation summaries, meeting summaries.* |
| D.2 | Is there any market research available associated to the proposed requirements? |
| ***Speech Recognition and Acoustic Scene Analysis***  *The global speech and voice recognition market is expected to grow from $12.62 billion USD in 2023 to $59.62 billion USD by 2030, at a CAGR of 24.8%.*  (*https://www.fortunebusinessinsights.com/industry-reports/speech-and-voice-recognition-market-101382*)  *The global smart audio devices market is expected to grow from $92.92 billion USD in 2022 to $352.89 billion USD by 2029, at a CAGR of 21%.*  *(*[*https://www.maximizemarketresearch.com/market-report/smart-audio-devices-market/145977/*](https://www.maximizemarketresearch.com/market-report/smart-audio-devices-market/145977/)*)* |

# MPEG Member support

|  |  |  |  |
| --- | --- | --- | --- |
| E - Member support | | | |
| E.1 | Please list the members supporting the activity, including (for each of them) their level of support and constituency.   1. Please indicate the level of support for each member by selecting one of the following options:  * **interested** (only) * **active** (participation to the development of a technical solution)  1. Please indicate the most representative constituency of each member by selecting one of the applicable following options:  * **Academic** * **Research lab** * **Service provider** * **device manufacturer** * **equipment vendor** * **network operator** * **technology provider** * other (specify) | | |
|  | **Member** | **Level of support** | **Constituency** |
|  | *Chung-Ang Univ.* | *Active* | *Academic* |
|  |  |  |  |

# Additional information

|  |  |
| --- | --- |
| F – Additional information | |
| F.1 | Please provide any additional information that you think relevant for assessing the market needs |
|  |

**Annex B.4**

**Market and practical considerations - ACoM Medical (Source: N64342)**

# Abstract

This contribution highlights some market & practical considerations regarding potential needs for an audio coding for machines (ACoM) with the focus on **medical applications** using the template developed in the market Needs AhG.

# Introduction

In WG6/N0203, an overview of use cases and requirements on audio coding for machines (ACoM) is given. At the 11th meeting, WG6 decided to ask for assistance from WG2 to create a report on market and practical considerations. In AG6/N023 eight uses cases are listed. This document focuses on the scenarios in medical context:

* Timed Medical Data Process Control

**Market and practical considerations questionnaire**

# Topic presentation

|  |  |
| --- | --- |
| A – Topic presentation | |
| A.1 | Please introduce the activity/domain that requires the development of new technology |
| *Algorithms to analyze timed medical data like EEG, EMG and ECG for research and diagnosis. This activity should specify a format to store the multi-dimensional data in an interoperable format. The format must contain measured data in raw format together with metadata describing the data. In the context of the application, raw data includes lossless compression by exploiting redundancies in the data.  Data about a patient and additional diagnosis stored in the format might be necessary, but it is still unclear whether this causes privacy issues.*  *Timed medical data has a very similar structure as spatial audio data: the frequency range of each individual channel is similar to audio, and the channels often are correlated to each other like the audio channels of a multi-channel audio recording.*    *Note: This activity is about the* ***data format*** *used for training but not the algorithms. In medical applications, the same data format might be useful for long time storage of data with offline analysis for diagnosis.* |
| A.2 | Please provide a brief description of high-level scenarios expected to be impacted by the technology.  For each of them, are there any associated high-level requirements? |
| *Timed Medical Data like EEG and ECG consist of several one-dimensional streams, and these are very similar to the data captured by microphones. Long-time monitoring and multi-electrode measurements create huge data sets. Currently, exploitation of measurements of many patients in research is limited due to storage constraints and privacy issues. Using large data sets, both diagnosis and therapy methods can benefit from big data algorithms. In addition, the format enables to record and store data of a single patient for longer time series for offline processing by a medical doctor either manually or via machine learning algorithms.*  *Key Requirements: The format to be developed has to store large data for training of algorithms. Therefore “efficient storage” which means low bitrate is essential. In addition, privacy by design is essential, too.* |

# Timing aspects

|  |  |  |
| --- | --- | --- |
| B - Timing aspects | | |
| B.1 | Necessary date for completion of specification, if known. | |
| *Not known* | |
| B.2 | When are compliant products/services required to be available? | |
| *During the development of the storage format, access to medical data is necessary. It is expected that anonymized data sets will be made available by university hospitals.* | |
| B.3 | Is the technology already implemented? | |
| Please select | * already products/services deployed * Proof of concepts/prototype * Planned roadmaps * **Nothing yet** * Other: (please specify) |

# MPEG Technologies

|  |  |
| --- | --- |
| C - MPEG technologies | |
| C.1 | Are there existing MPEG or other standards/technologies addressing the proposal? |
| ***MPEG****:*  *MPEG-4 SLS (ISO/IEC 14496-3) defines ways for lossless and near-lossless coding of a small number of channels.*  *MPEG-H 3D Audio (ISO/IEC 23008-3) specifies ways to encode spatial audio as individual channels, objects and in HOA format.*  ***AES:***  *AES 69 Spatially Oriented Format for Acoustics" (SOFA) specifies ways to encode spatial audio impulse responses.*  ***EDF:*** *The European Data Format (EDF) is used for EEG / ECG / PSG data. This is a freely available open standard, with implementations in the common scientific programming languages.*  ***DICOM****:*  *The Digital Imaging and Communications in Medicine (DICOM) is the standard for exchange of imaging data in medicine applications. Currently the DICOM consortium plans to extend their work to timed medical data (including electroencephalography (EEG), video-electroencephalography (VEEG), electromyography (EMG), evoked potentials (EP), polysomnograms (PSGs), electrocardiograms (ECGs), and other types of neurophysiology signals). First candidates to store such data are some MPEG audio formats.* |
| C.2 | If yes, why are they deficient or don’t meet the need? |
| *MPEG-4 SLS is not able to store the necessary metadata and does not provide tools to exploit spatial redundancies between many channels.*  *MPEG-H 3D Audio is not able to store the necessary metadata. MPEG-H 3D audio is based on perceptual audio coding which is neither lossless nor near-lossless. ACoM is not intended for human consumption and the perceptual coding of MPEG-H 3D Audio might delete the components needed for processing by machines.*  *AES 69 is not able to store the necessary metadata and can only store (short) impulse responses. The format is lossless but does not provide efficient storage (no means to exploit redundancies to reduce bitrate).*  *EDF uses just PCM and does not include any efficient coding. It is therefore not useful for long time recording and big data applications.*  DICOM currently has no standardized format for storing timed medical data. Using existing MPEG standard will either limit the compression rate achieved or lead to a loss of information. |
| C.3 | If not, will a new specification be required? Any dependency to existing MPEG or other standards |
| *Yes, a new specification will need to be made as no existing format addresses the problem of big data storage and transfer.*  *Collaboration with DICOM would be beneficial.* |
| C.4 | Which MPEG WG(s) is(are) expected to be impacted by the development of the technology? (if known) |
| *WG6 Audio.* |

# Market reach

|  |  |
| --- | --- |
| D - Market reach | |
| D.1 | Please indicate the target ecosystem (horizontal market, retail, vertical, professional…) and any geographical limitation (worldwide, region based…) |
| *The health and medical sector is huge:*   * With an aging society, the gap between the number of older (and less healthy) patients and fewer medical doctors makes it necessary to use algorithms for diagnosis and recommendation of therapy. * Automatic data collection and analysis becomes more and more important. * Research in medicine is currently limited by the availability of medical data in an interoperable format: often data of EEG and ECG is only available as paper or image file but without metadata and computerized processing capabilities. |
| D.2 | Is there any market research available associated to the proposed requirements? |
| *The global medical devices market size was valued at $512.29 billion in 2022 & is projected to grow from $536.12 billion in 2023 to $799.67 billion by 2030 (*[*https://www.fortunebusinessinsights.com/industry-reports/medical-devices-market-100085*](https://www.fortunebusinessinsights.com/industry-reports/medical-devices-market-100085)*). This market includes a large variety of devices for diagnosis and treatment. Detailed numbers for timed medical data are not available* |

# MPEG Member support

|  |  |  |  |
| --- | --- | --- | --- |
| E - Member support | | | |
| E.1 | Please list the members supporting the activity, including (for each of them) their level of support and constituency.   1. Please indicate the level of support for each member by selecting one of the following options:  * **interested** (only) * **active** (participation to the development of a technical solution)  1. Please indicate the most representative constituency of each member by selecting one of the applicable following options:  * **Academic** * **Research lab** * **Service provider** * **device manufacturer** * **equipment vendor** * **network operator** * **technology provider** * other (specify) | | |
|  | **Member** | **Level of support** | **Constituency** |
|  | *Fraunhofer* | *Active* | *Technology provider* |
|  |  |  |  |

# Additional information

|  |  |
| --- | --- |
| F – Additional information | |
| F.1 | Please provide any additional information that you think relevant for assessing the market needs |
| *Free response* |