 ISO/IEC JTC 1/SC 29/WG 6 N0203

**ISO/IEC JTC 1/SC 29/WG 6**

**MPEG Audio Coding Convenorship: US**

**Document type:** General

**Title:** Use cases and requirements for Audio Coding for Machines

**Status:** Approved

**Date of document:** 2023-04-28

**Source:** ISO/IEC JTC 1/SC 29/WG 6

**Expected action:** None

**Action due date:** None

**No. of pages: 12**

**Email of Convenor:** srq @ audioresearchlabs . com

**Committee URL:** https://isotc.iso.org/livelink/livelink/open/jtc1sc29wg6

**INTERNATIONAL ORGANISATION FOR STANDARDISATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC1/SC29/WG6, MPEG AUDIO CODING**

**ISO/IEC JTC1/SC29/WG6 N0203**

**April 2023, Virtual**

|  |  |
| --- | --- |
| **Title** | **Use cases and requirements for Audio Coding for Machines** |
| **Source** | **WG 6, MPEG Audio Coding** |
| **SN** | **22691** |

# Editorial Remark

# In section 6 at the end of this document a number of open issues and the status of discussions in WG6 are listed. This document should help to stimulate input for the analysis to be conducted by WG2 AdHoc Group on “Market Needs”.

# Introduction

## Motivation

Traditional coding methods aim for the best audio under certain bit-rate constraint for human consumption. However, with the rise of machine learning applications, along with the abundance of sensors, many intelligent platforms have been implemented with massive data requirements including scenarios such as connected vehicles, audio surveillance, machine diagnostics and smart city. In many of these applications the spatial distribution of audio contains important information. Medical data, like those coming from EEG and EKG measurements, while not inherently audio, often have a similar structure and similar requirements as audio data. Spatial audio and medical data will be referenced as multi-dimensional streams.

The sheer quantity of data being produced constantly leads previous methods with a human in the pipeline to be inefficient, and unrealistic in terms of latency and scale. There are additional concerns in transmission and archive systems which require a more compact data representation and low latency solution. This led to the introduction of Audio Coding for Machines.

In some cases, machines will communicate amongst themselves to perform tasks without a human in the mix, while in others there will be a need for additional human consumption of the specific decompressed stream. This specific scenario is possible in surveillance use cases, where a human “supervisor” may occasionally search for a specific sound, or scene in the audio. In other cases, the corresponding bitstream may be used for both human and machine consumption. In the case of cars, the features may be used to overcome the good sound insulation of modern cars (“detection of ambulance sirens”) and for monitoring failure of components (including predicted maintenance).

Any use cases in which audio features need to be transmitted for additional processing which may potentially be used for machine or human end users could benefit from a standard in the coded features (shared backbone). Interoperability is crucial where different manufacturers and platforms need communication to achieve a common goal.

Additionally, the feature stream must be efficient for both transmission and archive concerns for both latency and space. A standard for the compressed coding of this feature stream will establish an efficient protocol for machines to communicate.

## Scope

MPEG-ACoM aims to define a bitstream from compressing audio, multi-dimensional streams, or features extracted from such signal that is efficient in terms of bitrate/size and can be used by a network of machines after decompression to perform multiple tasks without significantly degrading task performance. The decoded audio, multi-dimensional streams or features can be used for machine consumption or hybrid machine and human consumption.   
In addition to the essence the format must also contain metadata describing how the data audio or multi-dimensional stream was captured.

The first phase should be application agnostic: Data is encoded near-lossless enabling the training of feature extraction schemes. The result from this phase is already useful for industry simplifying the exchange of data using this standardized format.

In a second phase feature extraction schemes are added. These features will be optimized for different applications.

## System Overview

The generic system architecture contains a pair of ACoM encoder and decoder. The input of the ACoM system could be metadata describing the input and either of

* audio signals (one or multi-dimensional)
* multi-dimensional streams (e.g. medical data)
* extracted features (phase 2).

In case of a feature stream, the type and format of feature should be specified, features may take different forms depending on the application.   
Feature extraction and coding is not in phase 1 but in phase 2.

The decompressed bitstream of audio and/or multi-dimensional streams and/or features may then be used for post-processing tasks, which may include machine consumption tasks or hybrid machine and human consumption tasks. The encoder can be optimized for either a single task or multiple, and the size of the compressed stream should compare favorably to traditional coding techniques on the unprocessed audio or medical data.

The MPEG activity on Audio Coding for Machines (ACoM) aims to standardize a bitstream format generated by compressing a previously extracted feature stream or data stream.

Metadata

Metadata

ACoM

Encoder

ACoM

Decoder

Compressed

bitstream

Task Analysis for

Machine Listening

Human Listening

(optional)

（optional）

Decompressed  
Audio/  
Feature/  
multi-dim.  
stream

Audio/   
Feature/   
multi-dim.   
stream

Fig1. Pipeline for ACoM (note: feature is not there in phase 1)

Transmission of encoded bitstream

ACoM decoder

ACoM encoder

Feature Extraction

Feature En-coding

Feature Decoding

Interface for NN

Interface for NN

Human Listening

Machine Listening

Sensor output

Audio Encoding

Bitstream Mux

Bitstream Demux

Audio Decoding

Feature Con-version

Fig 2. An example of potential ACoM architecture (phase 1 and phase 2).   
Coding of metadata not shown in figure.

Fig 2 shows an example of potential ACoM architecture. The ACoM codec could be an audio codec or a feature codec, or both. In case of a feature codec, the ACoM feature encoding is consisting of feature extraction, feature conversion and feature coding.   
There may be an interface to an external neural network (NN) for the feature extraction and the task specific networks. Not shown in the figure is the path for metadata encoding and decoding.

# Use Cases

|  |
| --- |
| * 1. UC1 Predictive Maintenance |
| Acoustic sensors are used in industrial environments to monitor the function of a single machine or all machines in a hall. Predictive maintenance is replacing components before failure but a late as possible. Acoustic sensors can be outside a machine. Therefore, retrofit of old machines is possible. Neural networks are used for to train detectors. Usually, the difference 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. |
| **Overlap with other use cases** |
| UC2 Process control, UC3 in-line testing and UC4 end-of-line testing |
| **Required properties of the algorithm** |
| * Capable to encode many microphone channels around a machine * Capable to encode metadata (see sub-tasks) * The compression must be near-lossless to adapt to many different types of machines and failures * GDPR compliance: Automatic detection and suppression of (unwanted) speech recordings |
| **Optional properties of the algorithm** |
| * In some cases, it might be feasible to auralize the reconstructed audio for human inspection. |
| **What are the different sub-tasks expected in this use case?** |
| * Capturing the whole acoustic scene exploiting redundancies of measurement signals * Coding of metadata describing type and position of sensors * Coding of metadata describing measurement conditions (e.g. temperature, humidity, air pressure) * Coding of process steps (material, tools) stored in the data file * Coding of operation condition (at least OK/NOK, in case of NOK optionally different type of predicted failure) |
| **What is the expected bandwidth of the distribution channel?**  **What is the maximum latency that is acceptable?**  **What are the power requirements?** |
| * Compared to other use-cases, the required bandwidth might be larger since the systems are stationary. The limitation is the amount of storage. * Near real-time processing is necessary. * Power is not a large concern. |

|  |
| --- |
| * 1. UC2 Process Control |
| Acoustic sensors are used in industrial environments 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 for to train process parameters. |
| **Overlap with other use cases** |
| UC1 Predictive Maintenance, UC3 In-Line Testing and UC4 End-of-Line Testing |
| **Required properties of the algorithm** |
| * Capable to encode a few microphone channels close to the machine. * Capable to encode metadata (see sub-tasks). * The compression must be near-lossless to adapt to many different types of machines. |
| **Optional properties of the algorithm** |
| * In some cases, it might be useful to code operation condition (OK/NOK). * Auralisation not necessary. |
| **What are the different sub-tasks expected in this use case?** |
| * Capturing the whole acoustic scene exploiting redundancies of measurement signals. * Coding of metadata describing type and position of sensors, machines and of the room acoustics. * Coding of metadata describing measurement conditions (e.g. temperature, humidity, air pressure) * Coding of process steps (material, tools) stored in the data file |
| **What is the expected bandwidth of the distribution channel?**  **What is the maximum latency that is acceptable?**  **What are the power requirements?** |
| * Compared to other use-cases, the required bandwidth might be larger since the systems are stationary. * Real-time processing is necessary. * Power is not a large concern. |

|  |
| --- |
| * 1. UC3 In-line Testing |
| Acoustic sensors are used in industrial environments 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 for to train detectors. Usually, the difference to be detected are small. Usually, there is a lot of environmental noise. Usually, there is much more training data available for correct operation. |
| **Overlap with other use cases** |
| UC1 Predictive Maintenance, UC2 Process Control, and UC4 End-of-Line Testing |
| **Required properties of the algorithm** |
| * Capable to encode a few microphone channels specially to discard noise * Capable to encode metadata (see sub-tasks) * The compression must be near-lossless to adapt to many different types of machines and failures |
| **Optional properties of the algorithm** |
| * None |
| **What are the different sub-tasks expected in this use case?** |
| * Capturing the whole acoustic scene exploiting redundancies of measurement signals * Coding of metadata describing type and position of sensors * Coding of metadata describing measurement conditions (e.g. temperature, humidity, air pressure) * Coding of process steps (material, tools) stored in the data file * Coding of operation condition (at least OK/NOK, in case of NOK optionally different type of predicted failure) |
| **What is the expected bandwidth of the distribution channel?**  **What is the maximum latency that is acceptable?**  **What are the power requirements?** |
| * Compared to other use-cases, the required bandwidth might be larger since the systems are stationary. The limitation is the amount of storage. * Real-time processing is necessary. * Power is not a large concern. |

|  |
| --- |
| * 1. UC4 End-of-line Testing |
| Acoustic sensors are used in industrial environments 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 for to train detectors. Usually, the difference 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. |
| **Overlap with other use cases** |
| UC1 Predictive Maintenance, UC2 Process Control and UC3 In-line Testing. |
| **Required properties of the algorithm** |
| * Capable to encode a few microphone channels * Capable to encode metadata (see sub-tasks) * The compression must be near-lossless to adapt to many different types of products and failures |
| **Optional properties of the algorithm** |
| * None |
| **What are the different sub-tasks expected in this use case?** |
| * Capturing the whole acoustic scene exploiting redundancies of measurement signals * Coding of metadata describing type and position of sensors * Coding of metadata describing measurement conditions (e.g. temperature, humidity, air pressure) * Coding of name of product stored in the data file * Coding of condition (at least OK/NOK, in case of NOK optionally different type of predicted failure) |
| **What is the expected bandwidth of the distribution channel?**  **What is the maximum latency that is acceptable?**  **What are the power requirements?** |
| * Compared to other use-cases, the required bandwidth might be larger since the systems are stationary. The limitation is the amount of storage. * Real-time processing is necessary. * Power is not a large concern. |

|  |
| --- |
| * 1. UC5 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 advancement of emergency vehicles. The acoustic sensor network can also be used to track crowds of people. |
| **Overlap with other use cases** |
| UC6 Construction Site Monitoring. |
| **Required properties of the algorithm** |
| * Capable to encode a few microphone channels in each sensor node * Capable to encode metadata (see sub-tasks) * The compression must be lossy to avoid speech decoding, but at the same time time-stamped to track individual vehicles or crowds. |
| **Optional properties of the algorithm** |
| * None |
| **What are the different sub-tasks expected in this use case?** |
| * Capturing the whole acoustic scene * Edge computing in each node * Coding of metadata describing type and position of sensors and exact time * Coding of metadata describing measurement conditions (e.g. temperature, humidity, air pressure) * GDPR compliance: Automatic detection and suppression of (unwanted) speech recordings and/or lossy coding to avoid possibility of auralisation. |
| **What is the expected bandwidth of the distribution channel?**  **What is the maximum latency that is acceptable?**  **What are the power requirements?** |
| * Compared to other use-cases, the required bandwidth should be low because sensor node connection might be wireless. * Near-Real-time processing is necessary. * Power should be low if the nodes are battery powered. |

|  |
| --- |
| * 1. UC6 Construction Site Monitoring |
| Acoustic monitoring is used for automatic detection and tracking of vehicles entering in and driving thru large construction sites. This enables monitoring the flow of goods, preventing accidents by predicting conflicts, and to avoid stealing 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. |
| **Overlap with other use cases** |
| UC5 Traffic Monitoring and Control. |
| **Required properties of the algorithm** |
| * Capable to encode a few microphone channels in each sensor node * Capable to encode metadata (see sub-tasks) * The compression must be either lossy to avoid speech decoding or mute all times speech is detected. * All captured data must be time-stamped to track individual vehicles. * The algorithm must be capable to detect position of vehicles and intruders to enable alarms for dangerous positions. |
| **Optional properties of the algorithm** |
| * None |
| **What are the different sub-tasks expected in this use case?** |
| * Capturing the whole acoustic scene * Edge computing in each node * Coding of metadata describing type and position of sensors and exact time * Coding of metadata describing measurement conditions (e.g. temperature, humidity, air pressure) * GDPR compliance: Automatic detection and suppression of (unwanted) speech recordings and/or lossy coding to avoid possibility of auralisation. |
| **What is the expected bandwidth of the distribution channel?**  **What is the maximum latency that is acceptable?**  **What are the power requirements?** |
| * Compared to other use-cases, the required bandwidth should be low because sensor node connection might be wireless. * Near-Real-time processing is necessary. * Power should be low if the nodes are battery powered. |

|  |
| --- |
| * 1. UC7 Speech Recognition and Acoustic Scene Analysis |
| Speech Recognition and Acoustical Scene Analysis is used as user interfaces and/or for improving speech in noise environments. Neural networks for such applications need training data:  o Training data for speech recognition  o Training data for speaker recognition  o Training data for adaptive, real-time enhancement of speech in noisy situations  (digital cocktail-party processor)  o Training data for sentiment analysis (i.e. emotion recognition). |
| **Overlap with other use cases** |
| None |
| **Required properties of the algorithm** |
| * Capable to encode a few microphone channels * Capable to encode metadata (see sub-tasks) |
| **Optional properties of the algorithm** |
| * None |
| **What are the different sub-tasks expected in this use case?** |
| * Capturing the whole acoustic scene * Edge computing * Coding of metadata describing type and position of sensors and exact time * Coding of metadata describing measurement conditions (e.g. temperature, humidity, air pressure) * Metadata for annotation of content, e.g. language, speaker, spoken word |
| **What is the expected bandwidth of the distribution channel?**  **What is the maximum latency that is acceptable?**  **What are the power requirements?** |
| * Bandwidth is not an issue because there is no transmission of content * Real-time processing is necessary. * Power should be low because devices are battery powered. |

|  |
| --- |
| * 1. UC8 Timed Medical Data |
| 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 measuremenst create huge data sets. Currently, exploitation of measurements of many patients is limited due to storage constraints and privacy issues. |
| **Overlap with other use cases** |
| None |
| **Required properties of the algorithm** |
| * Capable to encode a few channels * Capable to encode metadata (see sub-tasks) |
| **Optional properties of the algorithm** |
| * None |
| **What are the different sub-tasks expected in this use case?** |
| * Capturing and near-lossless coding of all streams exploiting redundancies * Coding of metadata describing type and position of sensors and exact time * Coding of metadata describing measurement conditions * Metadata for annotation of content, e.g. reason for recording (diagnosis, background). |
| **What is the expected bandwidth of the distribution channel?**  **What is the maximum latency that is acceptable?**  **What are the power requirements?** |
| * Bandwidth is not an issue because there is no transmission of content, but storage capacity is important * Real-time processing necessary for recording. * Power should be low because recording devices are battery powered. |

# Summary of Proposed Sub-tasks (phase 1)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Description | UC1 | UC2 | UC3 | UC4 | UC5 | UC6 | UC7 | UC8 |
| Spatial Scene Capturing | Capturing the whole (acoustic) scene exploiting redundancy of recorded signals | X | X | X | X | X | X | X | X |
| Scene Analysis | Localization of sound objects in the scene | (X) | (X) |  |  | X | X | X |  |
| Metadata  Sensors | Coding of metadata describing type and position of sensors | X | X | X | X | X | X | X | X |
| Metadata environ-ment | Coding of metadata describing measurement conditions (temperature, humidity, air pressure) | X | X | X | X | X | X | X | X |
| Metadata process | Coding of processing steps in recorded data (“what was measured?”) | X | X | X | X | X | X | X | X |
| Metadata condition | Coding of operation conditions (OK/NOK) | X | X | X | X |  |  |  | X |
| Speech activity | Detection of speech activity to suppress storage of speech | X |  | X |  | X | X | X |  |
| Speech content | Coding of spoken text in recorded data |  |  |  |  |  |  | X |  |
| Coding of timed data | (lossy) of scene with time stamp |  |  |  |  | X | X |  | X |

# Metrics for Key Tasks

To be defined.

# Requirements

The term “machine” refers to a process or algorithm that gets as input audio or feature (eventually after a decoding stage) in order to analyse it or process it. For example, a machine is a neural network with the task to detect people in the input audio.

The potential benefits of ACoM include compression efficiency, computational offloading, and privacy protection, etc. The following requirements reflect different use-cases. Requirements 1 to 4 and 6 are required for all use cases. Requirement 5 is required for most use cases. Requirements 7 and 8 are required for only some use cases.

|  |  |  |
| --- | --- | --- |
| **Number** | **Requirement** | **Description** |
| 1 | Efficiency of compression of bitstream | The size needed to represent the essence shall be less than the data stream under traditional lossless audio coding |
| 2 | Coding technologies can generate one bitstream to support single task or multiple tasks | The resulting coded features shall be usable and optimized for different scenarios   * One bitstream for single task. Example: scene-level classification. * One common bitstream for multiple tasks. Examples: object-level classification and action recognition.   The coding technology should support multiple tasks. There could be an advantage for using the bitstream for single tasks in comparison to using the bitstream for additional tasks. The bitstream for single tasks may either be smaller, or the encoder complexity may be lower than the encoder for additional tasks. |
| 3 | Varying degrees of performance for multiple tasks | Some machines may be required to perform more accurately than others (i.e., the tasks that some machines perform may have higher priority or importance than the tasks performed by other machines).  Priority may be a function of latency, bandwidth, or other application-specific requirements which may result in the varying encoding of the data stream.  The coding shall support varying levels of quality as measured by performance for different sub tasks. |
| 4 | Near-lossless coding | The coding should exploit redundancy in multi-channel data. Coding error must be lower than measurement precision. |
| 5 | Privacy protection | The features should be:  Not usable for reconstruction of spoken text and recognition of talker Note: this is no requirement for UC7! |
| 6 | Coding of metadata | The format should be capable to encode metadata about recording configuration and conditions, the process recorded and status. For UC7 this includes annotations of text spoken. |
| 7 | Hybrid machine and human consumption | A common bitstream should be used for machine and human consumption. (Only for UC1, UC5, UC6, UC7) |
| 8 | Edge Computing and time stamps | For UC5 and UC6 the transmission of data from sensor node is an issue. For these use cases edge computing in decentral sensor nodes plus a central instance to combine time aligned signals from sensor nodes is required. |

# Open Questions / Market Needs

Doc. WG06 M63138 raised some questions which need further attendance before deciding whether to proceed with the work on ACoM. Some of these questions would be best considered by WG02 AdHoc Group on Market Needs. The following lines are from M63138. Answers from the WG06 meeting are in ***italics*** and start with ***Addx***:

## Use cases

Generally, while the use cases listed in N187 seem reasonable, there is a question of whether there is sufficient market needs for such a standard. In many cases industries already have in place de-facto standards, or individual companies will have working systems tailored to their specific needs, for example:

* In-/End- of Line testing

In this situation a system makes real-time measurements and performs analysis of a device under test, either production machinery or final product. It is unclear what the use case for ACoM is. Is the intention that ACoM will be used to make baseline measurements that a model can be trained on for the actual In/End of line testing?  
***Add1****: In WG06 it was identified that Automotive applications are not explicitly named, but are somehow monitoring the status of a machine (car). This might be a large market.*

***Add2****: In WG06 it was added that 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.*

***Add3****: In phase 1 the major purpose of ACoM would be on storage of annotated data for training of neural networks or other models.*

* Speech Recognition training

Many datasets already exist for this, with typically file or folder names being used to provide the context of the content, such as CHiME-5 or TRIDIGITS. Many companies are highly secretive of the training datasets they use, so may prefer not to use an open standard.

* Storage of Medical data

There is already the widely used European Data Format (EDF) standard used for EEG / ECG / PSG data. This is a freely available open standard, with implementations in the common scientific programming languages. It is unclear why MPEG Audio should be involved in the storage of these signals?  
***Add4****: In WG06 it was added that the standardized formats are PCM (lossless) and huge. Therefore, in general it is not possible to store the data for long time and to make it available for research. DICOM, the most applied standard for storing medical images, is currently considering to use perceptual audio codecs to store EEG and ECG data. EEG and ECG signals are not audio, and therefore the coding will cause unknown effects. On the other side EEF and ECG signals are multichannel data, and a (near-)lossless multichannel scheme from WG06 would provide coding efficiency and quality.*

The primary question raised are:

* What does MPEG Audio bring above the current working solutions?
* How large is the perceived market?
* How do we propose to drive adoption of any new standard?

## Requirements

Some thoughts on the requirements that any future standard would need.

* Built-in privacy. As a manufacturer of medical devices many of the in-use measurements would be with patients and medical staff in the same space, which makes the recording of audio challenging from a legal and regulatory perspective so privacy, and possibly encryption, is needed.   
  ***Add5****: Some text about privacy is already in document M63244.*
* Segmentation of long data sets with short repetitive sections, and a need to link individually stored segments to each other.   
  ***Add6****: for process monitoring long sequences are foreseen. Segmentation would make it impossible to recognize changes in production rhythm.*
* Precise phase information and modulation frequencies for real time analysis of sensor data.  
  ***Add7****: yes!*
* Metadata about system usage, protocols (CT, MRI,…), environmental condition, calibration status
* Sensors generating data beyond human-perceivable frequencies (acceleration sensor, MEMS sensor, HF coils…). What is the upper sample rate that ACoM must support? For some cases >1 GHz sample rates may be needed.
* Clear definition of what is in scope, and what is out of scope. The given use cases include signals and sensors that are clearly not audio related (i.e. EEG), so why would they be in an MPEG Audio standard?