 ISO/IEC JTC 1/SC 29/WG 2 N272

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

**Document type:** Output Document

**Title:** Draft Use Cases and Draft Requirements on Audio Coding for Machines (ACoM)

**Status:** Approved

**Date of document:** 2023-01-20

**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 ORGANISATION FOR STANDARDISATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC1/SC29/WG2**

**MPEG Technical Requirement**

**ISO/IEC JTC1/SC29/WG2 N272**

**January 2023, Virtual**

|  |  |
| --- | --- |
| **Title** | **Draft Use Cases and Draft Requirements on Audio Coding for Machines (ACoM)** |
| **Source** | **WG 2, MPEG Technical Requirements** |
| **Status** | **Approved** |
| **Serial Number** | **55071** |

# 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 they learn about different typical situations. In many applications recordings from different directions around the sound source are necessary. Machine learning in general relies on big data. Therefore, huge 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 to decide which features of the sound are relevant for analyzing.   
Therefore in a **first phase** it is expected that the compressed format will not include feature extraction.   
A **second phase** of ACoM might achieve higher compression rate by storing features describing the audio data (AI technologies).

This document lists Use Cases for ACoM and Requirements on the ACoM data format. Currently both lists are incomplete and should be amended as necessary. At the end of the document other groups, which might be interested in the outcome and/or which might contribute in defining use cases and requirements, are listed.

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.

In the first phase WG 6 will work 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 from 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 by 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 lose 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 therefore 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 connections after the welding process

…

* + **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 – detecting any loose parts creating noise when exposed to sound stimuli

- Testing electrical motors  
- Testing roof tiles by clapping on tiles  
- Testing of microphones and loudspeakers

….

* + **Predictive monitoring  
    *Examples***

- Detecting the change of sound of components before failure (tools in drilling machines, motors in vehicles, gears, ...)

* + **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 environment.
* **Traffic measurement**
  + Counting number and type of vehicles in a town. Interesting categories of vehicles are ambulances, police vehicles, or firefighters. The data can be used to control traffic lights.
  + Acoustic sensors build into vehicles to get information about the acoustic scene around a vehicle even when out of sight
  + 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 and/or prevent vehicles from approaching dangerous regions.
* **Speech Recognition and acoustic scene analysis**
  + Training data for speech recognition
  + Training data for speaker recognition
  + Training data for adaptive, real-time enhancement of speech in noisy situations  
    (digital cocktail-party processor)
  + Training data for sentiment analysis (i.e. emotion recognition)
* **Storage of timed medical data** (e.g. EEG, ECG)

# Phase 1 Requirements

As visible in Figure 1 two formats are foreseen:

* The raw acoustic twin (RAT)
* The compressed acoustic twin (CAT)

From RAT and CAT share the same requirements with one exception: RAT is not considering data rate. The acoustic data of individual channels are represented as PCM data. Metadata is stored as XML. CAT exploit redundancies between channels and uses a bitrate efficient storage of metadata.

In the following list RAT resp. CAT is mentioned in brackets when the requirement is targeted to only one of the formats. If neither RAT or CAT is attached to a requirement this requirement is for both.

* **Storage of spatial acoustic data** 
  + Raw format as interface to measurement and acoustic simulation (RAT)
  + Compressed format of the raw format (CAT)  
    (parameters: bitrate and computational load)
  + For applications where edge computing is not possible (i.e. monitoring of blades of wind turbines) the **data rate** of measured data (not only the training data base!) is an additional important parameter. (CAT)

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
* **Privacy**

The storage format must consider issues of privacy: For any non-speech application it must be done in a way that prevents unintended analyzation of speech content. For some use cases, especially the monitoring of crowds which might be part of traffic monitoring, this implies that the format is not lossless. For other use cases, like industrial applications, it might be sufficient to reject encoding of content which contains speech.

* **Data protection/copyright**

Collecting and storing huge data sets might include the fusion of data sets from different owners. Such compiled data sets in general make training of algorithms more successful and more stable. However, the **copyright** of the content owners must be preserved. Some content might even contain trade secrets of content owners.

* **Annotations for speech**

For use cases around speech recognition and scene analysis the metadata in the data format must be capable to carry also semantic knowledge about speech content, language, speakers and recording condition.

* **Metadata to overcome domain mismatch**

It was highlighted that it is essential that in general the training data and measured data might differ in several aspects like type of microphones, position of microphone, orientation of microphones and room acoustics. The format must carry all metadata to enable the adaption between domains.

# Other relevant groups

Several other groups which are active in the application area have been identified.

## DCASE

The DCASE Community (Detection and Classification of Acoustic Scenes and Events) 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 very close to the already identified use cases. Tasks 1, 4, 5 and 6 are use cases hadn’t been listed in document WG6/N0172.

An important comment from DCASE is that the topic “blind quality assessment” is still an open issue and needs further research before standardization on algorithms should start. This is in line with document WG6/N0172 which focuses on a lossless 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.

As mentioned above 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.   
Several members in WG6 are also active in DCASE.

## OPC

The OPC Foundation (Open Platform Communications) is an industry consortium which creates and maintains standards for open connectivity of industrial automation devices and systems, such as industrial control systems and process control generally. The OPC standards specify the communication of industrial process data, alarms and events, historical data and batch process data between sensors, instruments, controllers, software systems, and notification devices.

Some experts reported that OPC enables them to optimize the whole manufacturing plant especially concerning fast reconfiguration, which is essential for cost efficient manufacturing of small batch sizes. To enable this each machine and each tool has a digital twin. That way production planning can model the complete production chain before the machines are reconfigured.

The set of standards in OPC currently does not include acoustical data.

## 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. DICONDE belongs to ASTM International (formerly known as “American Society for Testing and Materials”) which is part of ISO. The list of method specific standards in DICONDE includes “Computed Radiography”, “Digital Radiography”, “Computed Tomography”, “Ultrasonic Test” and “Eddy Current Test”. DICONDE originally started as an extension of the well-established data format for medical image data DICOM.

The set of standards in DICONDE currently does not include acoustical data.

## DICOM

The Digital Imaging and Communications in Medicine (DICOM) currently 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.