WHITE PAPER

The potential of “audio in”

Capturing and processing sounds for scene awareness and evidence

June 2021

AXIS COMMUNICATIONS

Table of Contents

1 Summary

Audio capturing capability, either integrated and ready to use in a video camera or provided by an external microphone, enables various important use cases. Responsible, professional use of "audio in" can add critical value and benefits to a security installation. It could, for example, provide the missing piece of evidence in a forensic investigation or enable realtime detection of events that require immediate attention of security guards or hospital staff. The mere fact that audio surveillance is taking place could also have a deterring effect and prevent crime.

Audio capture (often combined with instant analytics action) can be deployed as a standalone technology, enabling several use cases in crime prevention, protection, and forensics.

But combined with video, audio capture also has the potential to reinforce the great majority of existing surveillance use cases. For example, security operators can get a significantly better overview of scene events if their video stream is complemented with an audio stream.

Just as you may employ several types of *video* analytics for automatic event detection and alarming based on visual detection, *audio* analytics can monitor the audio streams and react when something stands out.

Audio analytics software can be set to trigger automatic alarms and other actions when a microphone picks up sounds associated with people shouting, glass breaking, or gunshots. This provides early warning that enables quick responses and intervention.

Audio analytics can also be about detecting whether an unexpected sound came from the left or the right and automatically redirect a PTZ camera towards the sound source. In a hospital or care facility, audio analytics can be used to detect high sound levels implying that a patient is in distress and send an automatic notification to a nurse. This use case can also benefit from sound visualization analytics which makes it easier to simultaneously monitor sound from many locations.

There is a difference between capturing sounds and recording them. For many types of use there is no need to record audio to achieve the goal, and this may help managing privacy concerns and complying with regulations regarding personal data. Audio analytics applications in general do not record sound continuously. They typically just process the incoming audio to search for specific patterns, levels, or frequencies. When analytics run on the edge (in the camera), no digital audio data needs to leave the camera - only the results from the performed analytics, that is, metadata or triggers, do.

Axis does not provide any legal advice. Laws that regulate surveillance vary by region, state, and country, and it is the user of the products (typically the end customer) who is responsible for making sure that any surveillance is conducted in a compliant manner. Just like the case of video surveillance, the installation of audio surveillance must be preceded by an investigation, and understanding, of the legal aspects of such an installation.

Once the necessary measures have been taken to meet legal conditions, the installation should be carefully considered regarding placement and configuration of the equipment. This may require some planning but is generally not difficult and simple measures go a long way to achieve audio usability.

2 Introduction

Audio information can be a valuable asset for crime prevention, protection, or forensic usage. Captured audio can also be processed in real time by analytics software which enables very efficient audio monitoring for detection of activities, behavior, or events.

This white paper describes the potential of audio in security with examples of typical use cases. Various types of audio analytics are presented along with brief overviews of how they work.

This paper does not provide any legal advice, but present different technical solutions that may be helpful for setting up an installation. Depending on how you choose to implement audio analytics, it may be possible to navigate regional laws and recommendations and employ this powerful tool where needed.

The scope of this paper is limited to the *capturing and possible recording* of audio, i.e., audio *input*. Another common use of audio in security solutions concerns the *broadcast* of audio, i.e., audio *output*, typically for playing voice messages or alarms to deter trespassers or shoplifters. More information about audio broadcast in security is available on www.axis.com/products/audio.

2.1 Capturing without recording

It is possible to capture and use audio without recording it. Capturing audio basically means digitizing it and making it available for use in software. This is done through registering the sound vibrations in the air using a microphone, converting these analog signals to digital signals (using A/D conversion equipment) and passing them on to a processing unit.

If the captured audio is not placed on any permanent medium such as a flash memory or a hard drive, it is not recorded. Recordings may be unnecessary for some use cases, such as when a human operator is listening in real time to the captured audio. In some situations, there are even specific reasons to *not* record the audio. There may be differences in legal restrictions depending on whether audio is recorded or just captured.

In general, audio analytics do not record sound continuously. To function properly, they temporarily buffer sound. Many systems could be set up to record what was buffered just before and after a detection to allow security to verify the detection and, possibly, preserve the sound for forensic evidence.

3 Navigating through the obstacles

Many people have concerns regarding the use of microphones in video surveillance situations. These concerns are typically linked to the recording of plain speech along with the video material.

We can move past this initial obstacle if we understand that there are many more possibilities with "audio in" than just recording it. There are many use cases where there is no need to record any sound information.

Laws that regulate surveillance vary by region and by country, so be sure to know what is permitted before adding audio to your surveillance system.

3.1 What do laws and regulations say?

Just like the case of video surveillance, the installation of audio surveillance must be preceded by an investigation, and understanding, of the legal aspects of such an installation. If applicable, the appropriate application documents must be submitted, and permits acquired. Signs or public statements must be used where required.

Audio usage and/or recording can be prohibited or require special consideration for several reasons, by national legislature or various types of local rules and regulations. While one region or environment may allow audio capture, it may still prohibit audio recordings. Companies can also prohibit audio surveillance usage within their premises.

3.1.1 US examples

The laws and regulations in the US vary between different states.

Some states require one-party consent for recording audio. This means that only one party in a conversation needs to consent for the surveillance to be legal.

Other states require all-party (or two-party) consent, meaning that all parties must consent to being recorded before a recording can take place. Exceptions to the all-party consent may apply in public places where a person cannot expect to be private.

Your legal assessment may also lead to another outcome in some regions when using an audio analytics application which does not record audio. You therefore need to investigate what laws and regulations apply in your specific state.

3.1.2 European examples

Audio surveillance is regulated through national laws in the European countries. You therefore need to investigate what laws and regulations apply in your specific country.

Audio recordings may contain personal data which is subject to the GDPR (General Data Protection Regulation). The GDPR does not necessarily prohibit audio recordings but the capture or recording of audio require special considerations. When adding audio to your existing video surveillance you need to consider if your legal ground for processing the personal data according to the GDPR still applies.

3.2 Investigate the possibility

There is a general misconception that audio is never allowed in surveillance. This misconception is so widespread that in many cases, the possibility to reinforce a surveillance installation with audio is never even looked into.

But many types of installations could be allowed, for example, if people are informed, if you have consent, and so on. You need to investigate what laws and regulations apply in your region and to your use case. Even if the *record and store* use case would not be allowed in your security installation, many use cases may be adapted to not infringe on privacy rights, such as *listen and interact*, *listen and witness*, and *detect incidents using audio analytics*.

4 Installation matters

The positioning of the microphone in a scene defines the potential applications. Before installing audio equipment, its placement and configuration should be carefully considered. This may require some planning but is generally not difficult and simple measures go a long way to increase audio usability.

Reflecting on the proper placement of a microphone and choosing an acoustically good spot will increase the probability of achieving your surveillance goals. A microphone must, of course, be placed so that it can easily capture the sounds that are relevant. Typical placements are in the middle of a room, in conjunction with a camera, or close to where specific events of interest may take place. A microphone

should not be placed close to a noise source, such as ventilation or machinery, which could overshadow sounds that are weaker or come from further away.

Typical microphone placement

1 *Where actions of interest take place*
2 *In a camera*
3 *In the middle of the room*

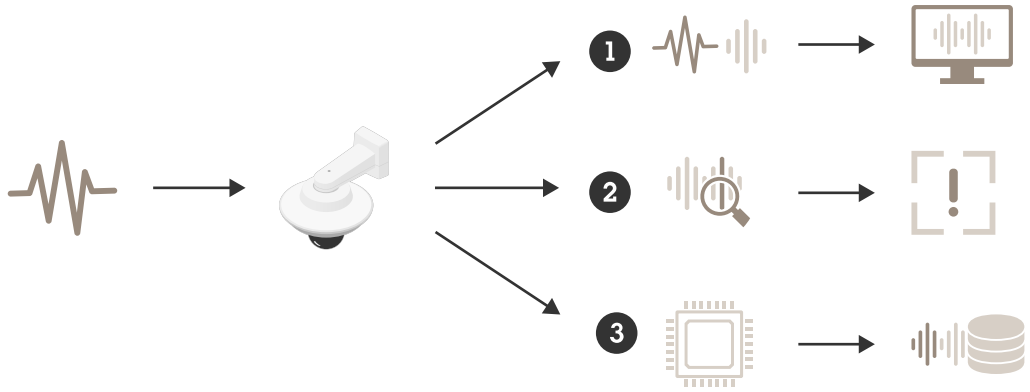
The acoustic environment, such as sound absorbing properties of walls or ceiling/floor and dimensional complexities (such as very long corridors), will yield different reverberation and echoes that can severely affect the sound field in certain locations. As an example, a voice will sound very different in a heavily attenuated area (such as an acoustically treated conference room), compared to in a church or in a fully tiled bathroom. In acoustically challenging situations, microphone placement can become critical.

Both installation and configuration (for example, the audio gain setting) of the equipment are important, as well as the integration of the audio equipment with the surveillance system. System installers and integrators can provide recommendations for specific situations.

For audio analytics, specific recommendations sometimes apply, which may differ from the recommendations for general audio recordings. Always study the user documentation to be aware of the applicable prerequisites.

5 Audio preparation

After the initial audio capturing, the captured information is prepared for the next processing steps. The different preparations can be made in parallel or exclusive.



1 *Transformation*
2 *Realtime edge analytics*
3 *Processing and encoding*

- **Transformation.** The sound is made abstract and converted into, for example, visual information as a graph showing the sound spectrum. This process cannot be reversed: you cannot retrieve the original sound from the spectrum graph.
- **Realtime edge analytics.**
  A **sound classifier** can be used if the sound is processed on the edge. The outcome will be metadata describing the sound's characteristics. The original sound cannot be recreated from its metadata.
  A **sound detector** can be used to recognize patterns, levels, or frequency and provide status information. Again, the original sound is not restorable.
- **Processing and encoding.** For cases where the original audio will be used (not transformed or analyzed), some processing and encoding is normally performed to prepare the audio data for the intended use cases. These use cases can involve storing audio data on the edge, streaming to external clients for additional processing (on server or cloud), or external storage.

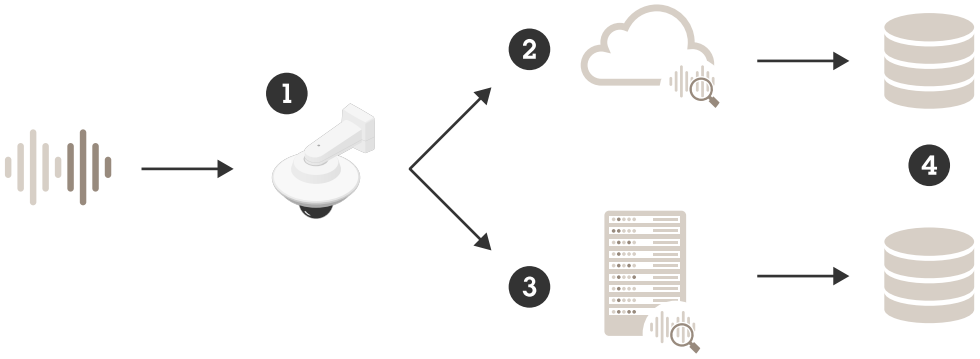
6 Analytics topology

The location of the analytics engine in the system is important for many reasons. Especially for managing privacy concerns and complying with regulations regarding personal data, it matters *where* the software algorithm analyzes the audio data. There are situations where audio data cannot be sent over the network and it is critical that captured (but not stored) audio data can be analyzed locally. If very computation-intensive algorithms are needed, such that cannot run on the edge, it might be required to send digital audio data to the cloud or a server.

- **Edge analytics.** When analytics run on the edge, no digital audio data needs to leave the camera. In the case of audio capture without recording, only the result from the performed analytics, that is, metadata or triggers, will be sent.
- **Server analytics.** When executed on a server, digital audio needs to leave the camera. If preprocessed on the camera (edge), this data could be abstracted or depersonalized metadata. A server is normally a

part of a closed system (a system owner is in control), so privacy concerns of transported audio can be managed. Nonetheless, it must be ensured that applicable rules and regulations are followed.

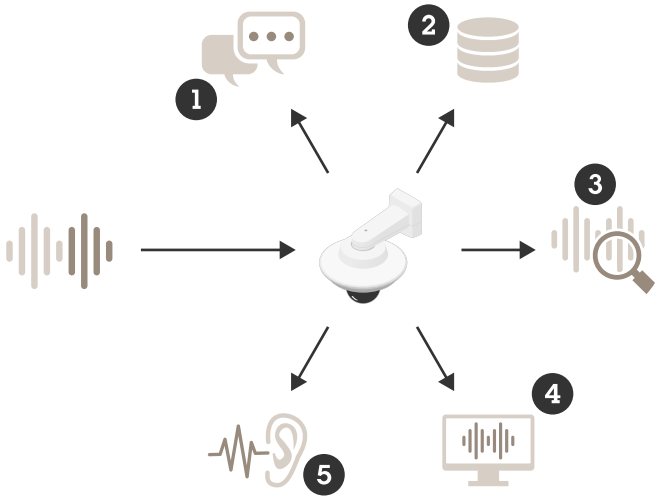
- **Cloud analytics.** Digital audio can also be transported to a server in a cloud context. As in the server analytics case, the audio information can be preprocessed into metadata. Cloud usage is often decentralized, so it is even more important to address privacy issues and ensure that regulations are complied with.



1 Edge
2 Cloud
3 Server
4 Storage

7 Use cases and examples

Audio capability is often integrated and ready to use in video cameras. There are various use cases where responsible and professional use of "audio in" can provide critical value and several potential benefits. It can, for example, be used to present the missing piece of evidence in a forensic investigation or enable realtime detection of events that require the immediate attention of security guards or hospital staff. The fact that audio surveillance is taking place could also have a deterring effect and prevent crime.



Typical purposes of audio capture:
1 Communicate
2 Record

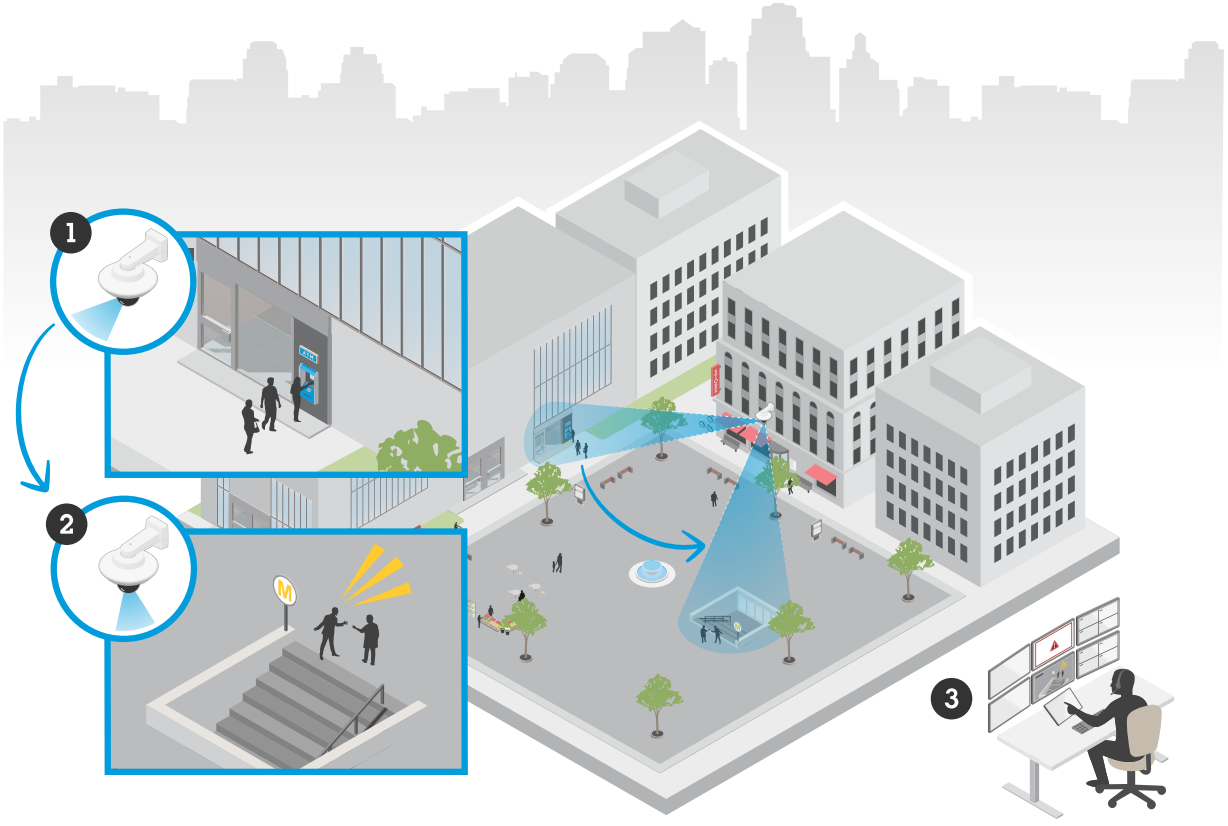
3 *Analyze*
4 *Visualize*
5 *Listen*

7.1 Detect incidents using audio analytics

Audio analytics applications are software programs that process captured audio in order to find and extract specific information. These are used to detect events such as gunshots, glass breakage, or aggression. They could, for example, process input audio to provide a yes-or-no answer to the question "did a window break?" Upon detection, the system typically sends an automatic notification to staff through a visual alert or by triggering an alarm. This provides early warning that enables quick responses and intervention.

7.1.1 Redirect a camera

Another example of audio analytics is a PTZ camera redirection application. This combines the audio and video functionalities by detecting where audio is coming from and automatically redirecting the camera towards the audio source.



1 *A PTZ camera is monitoring an ATM machine.*
2 *The camera microphone picks up a loud, sudden noise and the camera instantly redirects to the incident.*
3 *The operator receives an alarm and verifies the incident.*

7.2 Visualize sound in video

The sound captured in a video can be visualized and displayed as a sound spectrum diagram on a monitor. If a set threshold is exceeded, the diagram will start to indicate an alarm.

Such sound visualization can be valuable in situations where you need to monitor sounds from multiple sources at once, for example, several patient rooms in a hospital. While it would be too difficult to listen to many audio sources (sound from many rooms) simultaneously, it would be much easier to view many visualizations on a monitor in the nurse's station. If video feeds from the rooms are available, the visualizations can be added as overlay to the video image.

Sound visualization added as overlays to video feeds in a hospital.

7.3 Listen and interact

Perhaps the most basic and intuitive use case is audio surveillance with direct operator interaction to increase scene awareness. Examples are typically perceiving a suspicious conversation and be able to send a security guard to investigate it further. Or, in a hospital or care facility, to hear if a patient is in distress and call for a nurse. It could also be about detecting whether a 'strange' sound came from the left or the right and point a PTZ camera towards the sound source.

These use cases involve one or several operators having access to the audio environment from a control room or via a security application on a mobile device. The human ear captures sounds and the brain extracts what is relevant in the scene or the situation. If used in conjunction with video surveillance, audio adds another dimension of information for decision making. In some cases, audio will actually be the only dimension, if the audio source is outside the camera's field of view or if the light conditions are challenging.

7.3.1 Listen and witness

Audio surveillance can also be used for the purpose of direct testimony based on witnessed (heard) events. This use case differs from the *listen and interact* usage because the purpose is not decision making, but the use cases often coexist. For example, upon hearing an escalating argument with incriminating speech, an operator can not only send guards but later also bear witness about what was heard.

7.4 Record and store

If appropriate, the use case of capturing and recording audio data can provide great additional evidence. This could concern incriminating speech or gun fire. Recorded audio can provide proof of who said what, how many gunshots were fired, or similar events of forensic interest.

When audio is recorded in a forensic context, care should be taken to conserve the original data and avoid processing (which, in other contexts, may be required or beneficial). For forensic recordings, all types of processing could be considered evidence tampering. Voice-enhancing algorithms can be used to increase the audibility of recorded speech; it may improve the forensic value. But such algorithms should be applied afterwards, on a copy of the recorded material. By keeping the recording as unprocessed as possible, options are kept open as to how the material can be used later.

7.5 Get more out of your surveillance system

Surveillance systems often incorporate several types of sensors. The camera's image sensor is one, of course, registering the visual aspect of a scene. Non-visual sensors are also commonly used, such as motion detectors based on radar technology or infrared radiation emissions. Sometimes, video surveillance is not appropriate and non-visual sensors are therefore used as standalone devices. But in many cases, non-visual sensors are used to complete the camera installation by adding other types of information.

By also employing audio sensors (microphones) in a surveillance installation, the great majority of all possible use cases are reinforced. Adding audio capability to a non-audio system enables multisensor interaction, either via analytics or via operator interaction.

The *listen and interact* use case is a simple example, where the operator gets a significantly better overview of scene events when also receiving an audio stream. It may be difficult to detect aggressive behavior by only looking at people, but much easier if you can also hear them.

Another typical example is using video analytics, such as video motion detection. If the analytics application is challenged by, for example, low-light conditions, the presence of audio analytics can increase the detection confidence.

8 Monitoring and detection

Audio contains several kinds of information that can be used both for monitoring and audio analytics. Various types of processing and characterization assist in extracting and refining this information for easier usage and interaction with the surrounding system.

8.1 Sound characteristics

Characteristics such as loudness and pitch may comprise important information in a surveillance context. For how long it is audible, whether it is moving, or whether it is coming from near or from far, are all examples of pieces that add to the puzzle when we draw conclusions about a sound we hear. Hardware

and software for audio monitoring and detection are designed to work with the same types of information, "listening" for complex combinations of characteristics from decibel level to the energy in different frequencies over time.

- **Spatial information.** This concerns the physical world around us, including concepts such as location, direction, and distance. Spatial information can be used to focus or zoom the audio capture in different directions to enable better recordings. It can also be used by analytics to determine which direction a sound is coming from, or how far away its source is located.
- **Temporal information.** Temporal (time) information is important both in the dynamic sense (change over time) and the absolute sense (when did something happen?), often seen in relation with information from other sensors, such as video. Temporal information plays an important role in behavioral analysis – to know what happened when, and for how long.
- **Spectral information.** This concerns frequencies, such as how high pitch a sound has, or the combination of pitches in more complex sounds. Microphones used in audio surveillance are designed to have a flat frequency response, that is, they try to capture all the frequencies within the audible range (20 Hz – 20 kHz) equally. This differs from how the human hearing system works, because we can more easily detect those frequencies that are typically occurring in human speech, than the other frequencies.
- **Amplitude information.** This is about how intense or loud a sound is. Amplitude information can complement spectral information and be used together to paint an image of how incoming audio is structured.

8.2 Signal processing

Within audio surveillance, signal processing is typically about improving transmission, storage efficiency, or subjective quality, or to emphasize or detect components of interest. This is done through software algorithms which modify or analyze audio in various ways.

8.2.1 Modifying signals

Algorithms can be used to change the signal for a specific purpose, typically to:

- improve the signal, for example, increase audibility through automatic gain control.
- alter the signal, for example, by changing relative frequency content with an equalizer.
- limit the signal by removing specific frequencies or amplitudes. This could be about keeping the data volume down through compression or about ensuring privacy through voice scrambling.

8.2.2 Analyzing signals

Audio analytics use captured (but normally not recorded) audio data and analyze the relevant sound characteristics to generate non-audio results. The applications essentially convert the audio data to a more actionable asset in another format. There are analytic applications especially developed to detect, for example, aggression, gunshots, breaking glass, or car alarms.

If machine learning algorithms are used, they can be trained from large amounts of data to learn to make predictions without being explicitly programmed to do so. One example in an audio context could be an algorithm that can reliably detect the sound of a door closing after having been trained with thousands of such sounds.

8.3 The human hearing

The human ear is one of the best tools available to detect and analyze audio. In very noisy environments the human ear and brain can still detect and interpret speech where most algorithms might not succeed.

Using our ears, we can derive spatial information from a scene such as where a sound is coming from and whether the audio source is moving. Because we have two ears, we can hear if a sound is coming from the left or the right or somewhere in between. The ears and the head are also designed so that we hear whether a sound is coming from above or below, and from the front or the back. Several "filter steps" in the brain work with temporal differences between the ears, instantly detecting deviations as small as microseconds to make us aware of specific types of events. We have a well-developed capability for audio signal analysis, especially concerning human voices but also sounds associated with historical dangers.

Under the right circumstances (such as good sound quality, stereophonic sound, not too much delay) a human operator can be a powerful "analysis tool" and complement detection hardware or software. Using an audio surveillance product with only two microphones, an operator can derive spatial information from a scene such as from where a sound is coming and the movement of that sound.

9 Disclaimer

This document and its content is provided courtesy of Axis and all rights to the document or any intellectual property rights relating thereto (including but not limited to trademarks, trade names, logotypes and similar marks therein) are protected by law and all rights, title and/or interest in and to the document or any intellectual property rights related thereto are and shall remain vested in Axis Communications AB.

Please be advised that this document is provided "as is" without warranty of any kind for information purposes only. The information provided in this document does not, and is not intended to, constitute legal advice. This document is not intended to, and shall not, create any legal obligation for Axis Communications AB and/or any of its affiliates. Axis Communications AB's and/or any of its affiliates' obligations in relation to any Axis products are subject exclusively to terms and conditions of agreement between Axis and the entity that purchased such products directly from Axis.

FOR THE AVOIDANCE OF DOUBT, THE ENTIRE RISK AS TO THE USE, RESULTS AND PERFORMANCE OF THIS DOCUMENT IS ASSUMED BY THE USER OF THE DOCUMENT AND AXIS DISCLAIMS AND EXCLUDES, TO THE MAXIMUM EXTENT PERMITTED BY LAW, ALL WARRANTIES, WHETHER STATUTORY, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, TITLE AND NON-INFRINGEMENT AND PRODUCT LIABILITY, OR ANY WARRANTY ARISING OUT OF ANY PROPOSAL, SPECIFICATION OR SAMPLE WITH RESPECT TO THIS DOCUMENT.

Appendix 1 Audio quality terminology

Digital audio:

Digital audio is a representation of analog audio (often an acoustic signal captured with a microphone) recorded in digital form. In digital audio, the sound wave of the audio signal is typically encoded as a continuous sequence of numerical samples. The accuracy is dependent on the number of significant digits that the encoder records. For example, in CD audio, samples are taken 44,100 times per second, each with a 16-bit sample depth.

Noise:

Noise is unwanted (and sometimes unavoidable) sound that will define or limit the silent end of the loudness range. It is generated by all parts of an audio chain, from the recorded source (for example, a fan in the room), via the microphone (*e.g.,* self-noise, vibrations, wind) and cabling (*e.g.,* interference, crosstalk), to the capture device (*e.g.,* self-noise, digital sampling noise), all combined together creating what is normally called the noise floor.

Noise is normally defined by SNR (signal-to-noise ratio), the entire range from a defined level (sometimes the loudest sound the system can handle) to the noise floor.

The video equivalent is video noise, seen as random (normally) static pixel pattern, "snow"; limits what you can see in dark images (just as it limits what you hear for silent signals).

Distortion:

All unwanted alterations of a signal subtract from the original "truth" and this is called distortion (noise, as explained above, is normally excluded from the distortion specification). Distortion reduces the subjective quality (normally, there is distortion that sounds "nice") and obscures the objective information content, making the signal harder to listen to, especially for content analysis, and reduces analytics functionality.

THD (total harmonic distortion) and IMD (inter-modulation distortion) are two properties normally used to quantify distortion.

Distortion correlates to video as artefacts, such as chromatic aberration, vignetting, blur, etc.; makes an image look "bad" and limits how much details you can see.

Sample rate and frequency response:

In a digital system, audio is sampled a set number of times per second. This is the sample rate (normally from 8000 to 48,000 times per second, or Hz). To adequately capture a sound, signal theory (specifically the Nyquist Shannon sampling theorem) tells us that the sample rate needs to be at least twice that of the highest desired or required frequency in the analog signal.

A normal human ear hears frequencies from 20 Hz to approximately 15-20 kHz depending on age and other factors. Roughly speaking, the low frequency range, upwards of hundreds of Hz, often defines the foundation of specific sounds (like fundamentals in voices), while the upper frequency range, above a few thousand Hz, contains more 'details'.

The frequency range in audio correlates to resolution and frame rate in video; the lower you set it, the less details you get.

Bit depth:

Every time audio is sampled, an analog value is captured and translated to a digital representation. In the digital domain there are no infinites, so the amount of detail is limited to a defined bit depth. Every bit represents a factor of two (0 or 1, low or high, etc.) which, combined with a defined amplitude range (*e.g.,*

a chosen voltage or sound pressure level), creates fractions of this range. Two bits yield four fractions, three bits yield eight, and so on. Simplified, a one-volt signal, sampled with three bits, would be split and represented in 1/8-volt steps.

For sufficient audio quality, 16 bits are normally enough (representing 65 536 steps), at least for the human ear. This is what CD-audio is using. For analytics or more demanding use, 24 bits are more relevant.

Bit depth correlates to contrast in video, the range of luminance or chrominance each pixel can reproduce.

T10167735/EN/M3.2/2106

About Axis Communications

Axis enables a smarter and safer world by creating solutions for improving security and business performance. As a network technology company and industry leader, Axis offers solutions in video surveillance, access control, intercom, and audio systems. They are enhanced by intelligent analytics applications and supported by high-quality training.

Axis has around 4,000 dedicated employees in over 50 countries and collaborates with technology and system integration partners worldwide to deliver customer solutions. Axis was founded in 1984, and the headquarters are in Lund, Sweden