白皮书

“音频”的潜能

捕捉并处理声音，深入场景分析，保障证据调查

六月 2021

AXIS COMMUNICATIONS

目录

1 概述

音频捕捉能力——无论是集成在视频摄像机中的现成能力，亦或是由外置麦克风提供的能力，均能够实现多样化的重要应用。以负责任且专业的方式使用“音频输入”能够为安防系统带来重要的价值和诸多潜在有益效果。例如，它可在司法调查中提供证据补充，或者实时侦测需要立即调动安保人员或医务人员的事件。音频监控的存在还具有威慑作用，能够预防犯罪。

音频捕捉（通常与分析工具的实时操作相结合）可以作为单独的技术来部署，广泛应用于预防罪犯、提供保护和司法证据等场合。

而在结合视频的情况下，音频捕捉还能够增强大多数应用场合的监控效果。例如，通过以音频流补充视频流，安防操作人员能够更好地理解监控场景的事态。

正如可以采用多种类型的*视频*分析工具来自动侦测事件并基于视觉侦测发出报警，*音频*分析工具也能够监视音频流，并在发现异常时做出应对。

音频分析软件可被设置为：在麦克风拾取到与呼喊、玻璃破碎或枪击相关的声音时，触发自动报警和其他操作。这能够提供预警，确保快速响应和干预。

音频分析工具还可以用来侦测非预期声音是来自左侧还是右侧，并自动调转PTZ摄像机的方向，使其指向声源。在医院或护理中心，音频分析工具能够用来侦测表明患者病发的高声级，并自动通知护士。这种应用情形也可受益于声音可视化分析工具，它能够更轻松地同时监测来自多个声源的声音。

声音捕捉与声音记录是有区别的。在许多类型的应用场合中，不需要记录音频，亦可实现监控目标，这有助于保护隐私，遵守有关个人数据的法律法规。音频分析应用通常不会持续录音。它们通常仅处理传入的音频，以搜索特定声音形式、声级或频率。在前端（摄像机中）运行分析工具时，数字音频数据不需要离开摄像机，只有分析结果（即，元数据或触发信号）才会离开。

安讯士不提供任何法律建议。各国家、地区和州的监控法律各不相同，产品的用户（通常是最终客户）应负责确保以合法方式实施一切监控。跟视频监控一样，在安装音频监控之前，也必须调查并理解这种安装所涉及的法律方面。

在为满足法律法规要求而采取了必要的措施之后，在安装时应仔细考虑设备的安放位置和配置。这可能需要一定的规划，但通常不会太难，简单的措施就足以长效保障音频的可用性。

2 引言

音频信息可能是用于预防罪犯、提供保护或司法证据的宝贵资产。所捕捉的资产也可以被分析软件实时处理，这就能够实现非常高效的音频监测，从而侦测不良活动、行为或事件。

本白皮书以典型应用情形为例，介绍了音频在安防领域的潜力。其中介绍了多种音频分析工具，并简要概述了它们的工作方式。

本白皮书未提供任何法律建议，但介绍了有助于构建监控系统的不同技术解决方案。根据音频分析工具的具体实现方式，可以相应地参照相关地区法律和建议，并在需要时应用这个强大的工具。

本文的范围仅限于音频的*捕捉和可能记录*，即音频*输入*。音频在安防解决方案中的另一个常见用途涉及音频*广播*，即，音频*输出*，这通常是播放语音消息或报警，以吓退闯入者或小偷。有关安防领域音频广播的更多详情，请访问www.axis.com/products/audio。

2.1 捕捉但不记录

可以在不录音的情况下捕捉和使用音频。捕捉音频基本上意味着，对音频进行数字化处理，并使其可在软件中使用。具体实现方式是，使用麦克风寄存空气中的声音振动，（使用模数转换设备）将这些模拟信号转换为数字信号，然后将这些信号传送到处理单元。

如果所捕捉的音频未置于闪存、硬盘等永久存储介质上，那么便不会记录这些音频。在某些应用情形下，比如，在人类操作员正实时监听所捕捉的音频时，可能不必录音。在某些情况下，出于特定原因，甚至*不得*录音。在音频的记录或仅捕捉方面，可能有不同的法律限制。

一般情况下，音频分析工具不会持续录音。为确保正确工作，它们会临时缓存声音。许多系统可以设置为仅在侦测前后记录缓存内容，以便安保人员能够验证侦测，并在需要时保存声音以用作司法证据。

3 障碍概述

许多人对视频监控场合中的麦克风使用存在担忧。这些担忧通常涉及语音和视频材料的记录。

如果我们明白相比单纯的音频记录，“音频输入”的可能性要多得多，那么就可以越过这种初始障碍。在许多情况下，其实不需要记录全部声音信息。

各国家和地区的监控法律各不相同，因此，在视频监控系统中纳入音频功能之前，务必要知道哪些是允许的。

3.1 法律法规有何要求?

跟视频监控一样，在安装音频监控之前，也必须调查并理解这种安装所涉及的法律方面。根据相关要求，必须提交相应的申请文件，且必须获得许可证。在需要的地方，必须提供指示牌或公共说明。

出于某些原因，国家法律法规或各类地方法规条例可能会禁止使用和/或记录音频，或者要求满足特殊要求。尽管某个区域或环境可能允许音频捕捉，但它也可能禁止录音。企业也可能禁止在其场所范围内使用音频监控。

3.1.1 美国应用示例

在美国，不同的州有着不同的法律法规。

某些州要求，在录音之前，应获得一方当事人的许可。这就意味着，只需要获得谈话中一方当事人的许可，监控即为合法。

而其他某些州要求获得所有当事人（或双方当事人）的许可，这就意味着，在录音之前，必须获得所有谈话方的许可。在人们无法拥有私密性的公共场所中，可以豁免各方当事人许可。

即使在使用不支持录音功能的音频分析应用时，您的法律评估也可能表明，在某些地区会有不同的法律后果。因此，您需要调查具体州所施行的法律法规。

3.1.2 欧洲应用示例

音频监控受到欧洲国家的相关法律监管。因此，您需要调查具体国家所施行的法律法规。

录音可能涉及个人数据，需要遵守GDPR（一般数据保护法案）。GDPR不一定会禁止录音，但音频捕捉或记录需要满足特殊要求。在现有视频监控中纳入音频功能时，需要考虑个人数据的处理是否符合GDPR的要求。

3.2 可行性调研

一个常见的误解是，在监控中绝不允许使用音频 。这种误解广泛存在，以致于在许多情况下，人们甚至根本都不考虑是否可以使用音频来增强监控系统。

但实际上，许多安装都是允许的，例如，在向人们做了相应告知的情况下、在获得了相关方的许可的情况下，等等。您需要研究当地以及具体的应用场合适用哪些法律法规。即使安防系统中不允许*记录与存储*，也可以调整许多其他应用情形，以免侵犯隐私权，诸如调整*监听与交互*、*监听与目击*以及 *利用音频分析工具侦测事件*。

4 安装说明

麦克风在场景中的位置决定了潜在的应用能力。在安装音频设备之前，应仔细考虑其安放位置和配置。这可能需要一定的规划，但通常不会太难，简单的措施就足以长效保障音频的可用性。

通过仔细考量麦克风的安放位置，精心选择收音点，将能够更好地实现监控目标。当然，麦克风的安放必须使得它能够轻松捕捉相关声音。通常是与摄像机一起安放在房间中央，或者

靠近特定目标事件可能发生的地方。麦克风不应靠近噪声源，如通风设备或机械装置，否则可能掩盖本就较弱的或者来自远处的声音。

典型的麦克风安放位置

1 目标行为发生点
2 摄像机中
3 房间中央

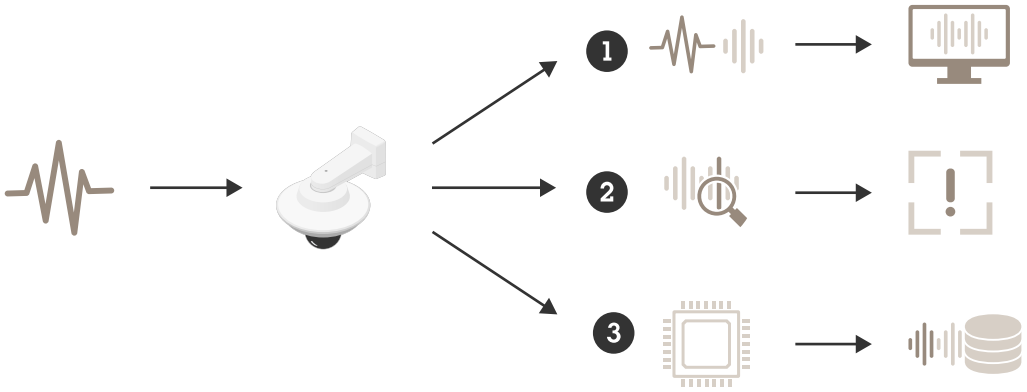
声音环境，比如墙壁或天花板/地板的吸音特性以及结构尺寸上的复杂性（比如非常长的走廊），会产生不同的混响和回声，在某些地方，这可能严重影响到声场。例如，相比教堂或贴满瓷砖的浴室，在经过了高强度消音处理的区域（比如，做过声音处理的会议室）中，声音听起来将非常不同。在难以听清的环境中，麦克风的安放位置可能就变得非常重要。

设备的安装和配置（比如，音频增益设置）、以及音频设备与监控系统的集成都是非常重要的。系统安装商和集成商可以就具体的环境提供建议。

对于音频分析工具，有时需要遵守特定建议，这可能不同于一般录音方面的建议。请务必研读用户文档，知悉相应的前提条件。

5 音频预处理

在初始音频捕捉之后，需要对捕捉的信息进行预处理，以便再进行后续处理。不同的预处理可以同时或单独进行。



1 转换
2 实时前端分析
3 处理和编码

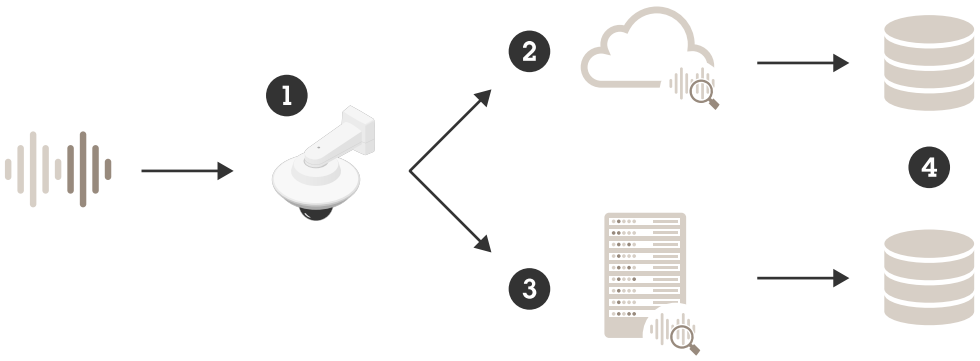
- **转换。** 声音被提取并转换成（比如）视觉信息，以图形形式显示声谱。这个过程不可逆：无法再通过声谱图获取原始声音。
- **实时前端分析。**
  如果在前端处理声音，可以使用**声音分类器**。这将得到描述声音特征的元数据。无法通过元数据再造原始声音。
  可以使用**声音检测器**来识别声音形式、声级或频率，并提供状态信息。它同样也无法还原原始声音。
- **处理和编码。** 如要使用原始（未经转换或分析的）音频，通常需要执行某些处理和编码，由此对音频数据进行预处理，以供预期应用场合之用。这些应用场合涉及在前端存储音频数据、将数据流传送到外部客户端以供（在服务器或云端）进一步处理、或者外部存储。

6 分析架构拓扑

出于多方面的原因，分析引擎在系统中的位置非常重要。尤其是在保护隐私和遵守有关个人数据的法律法规方面，它涉及到软件算法*在哪里*分析音频数据。有时，可能无法通过网络发送音频数据，那么在本地分析所捕捉（但未存储）的音频数据的就变得非常重要。如果所用的算法需占用大量计算资源，进而导致其无法在前端运行，则可能需要将数字音频数据发送到云或服务器。

- **前端分析。** 当分析工具在前端运行时，数字音频数据不需要离开摄像机。如果仅捕捉音频而不录音，那么将仅发送分析结果，即，元数据或触发信号。
- **服务器分析。** 在服务器上执行分析时，数字音频需要离开摄像机。如果在摄像机（前端）上进行了预处理，这些数据就可以成为经提取或去个性化处理的元数据。服务器通常是封闭式系统（由系统所有者控制）的组成部分，因此就能够确保所传输的音频的隐私保护。但也必须遵守相应的法律法规。

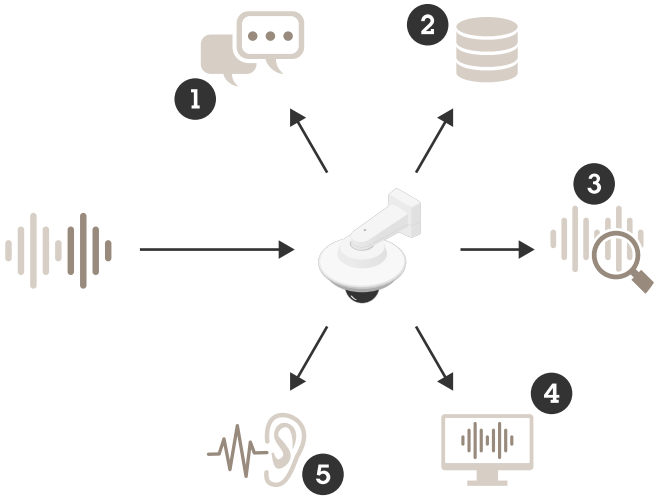
- **云端分析。**也可以将数字音频传输到云端服务器。跟服务器分析的情况一样，可以将音频信息预处理成元数据。云应用通常是分布式应用，因此需要更谨慎地对待隐私问题，确保符合相关法律法规。



1 Edge
2 云
3 服务器
4 存储容量

7 应用情形与示例

音频能力通常是集成在视频摄像机中的现成能力。在许多应用情形下，通过以负责任且专业的方式使用“音频输入”，能够带来重要的价值和诸多潜在有益效果。例如，它可用于在司法调查中提供证据补充，或者实时侦测需要立即调动安保人员或医务人员的事件。音频监控的存在还具有威慑作用，能够预防犯罪。



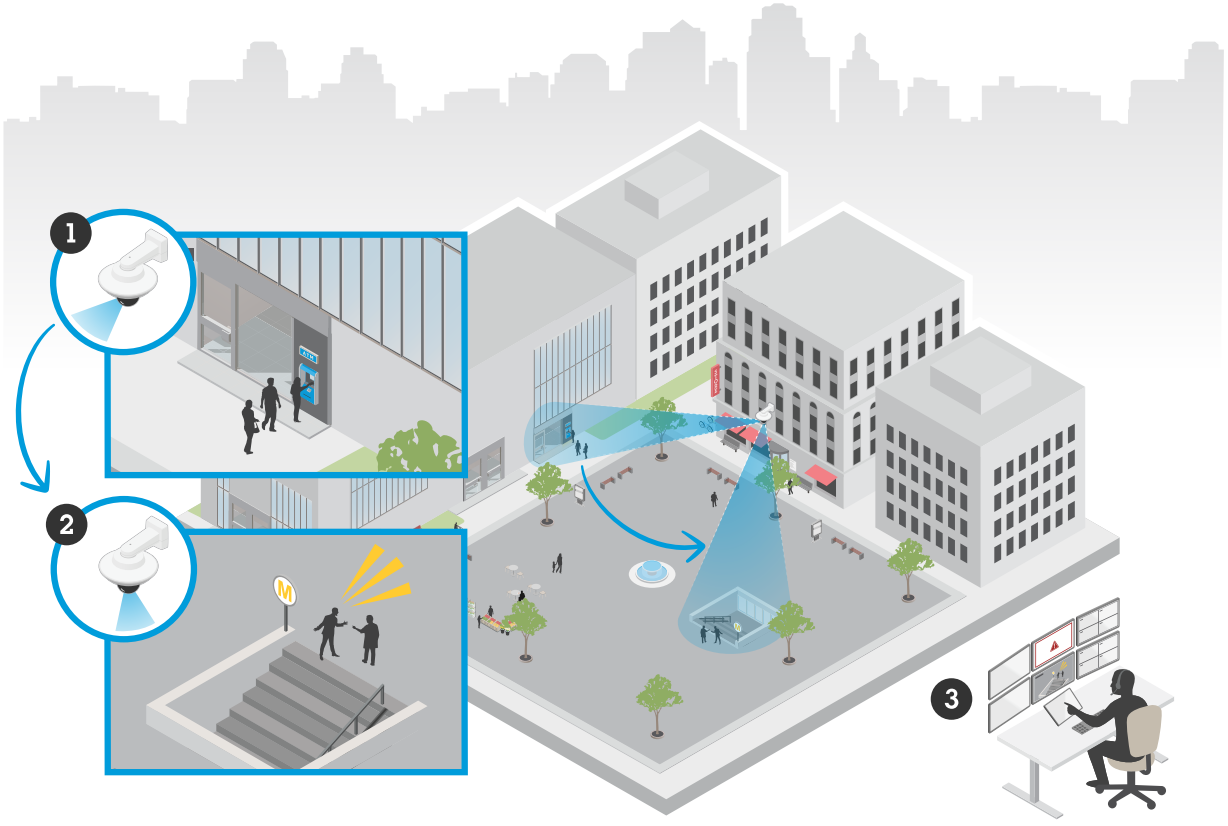
音频捕捉的典型用途：
1 对话
2 记录
3 分析
4 显示
5 监听

7.1 利用音频分析工具侦测事件

音频分析应用是软件程序，能够处理所捕捉的音频，从而帮助找到并提取特定信息。它们用于侦测诸如 枪击、玻璃破碎或暴力等的事件。它们能够，例如，处理输入音频，从而就“是否发生了破窗事件？”这一问题提供肯定或否定答案。一旦侦测到此事件，系统通常会通过视觉警报或通过触发报警的方式，向工作人员自动发送通知。这能够提供预警，确保快速响应和干预。

7.1.1 调转摄像机方向

音频分析工具的另一个示例是PTZ摄像机调转应用软件。它将音视频功能组合到一起，侦测音频来自哪里，并自动调转PTZ摄像机的方向，使其指向音频源。



1 PTZ摄像机正在监视ATM机。
2 摄像机麦克风拾取响亮且突发的噪声，摄像机立即调转方向，指向事件发生位置。
3 操作人员收到报警，并验证事件。

7.2 视频声音的可视化

视频中捕捉的声音可以被可视化，并作为声谱图显示在监视器上。如果超过设定阈值，声谱图将显示报警。

在需要一次性监测来自多个声源（例如，医院中的多个病房）的声音时，这样的声音可视化就可能非常宝贵。虽然要同时监听众多音频源（来自多个房间的声音）非常困难，但在护士

站的监视器上查看多个声音显示，则容易得多。如果同时还有来自房间的视频画面，则可以将这些声音可视化叠加到视频图像上。

叠加到医院视频画面上的声音可视化。

7.3 监听与交互

可能更基本、更直观的应用情形是支持操作人员直接交互的音频监控，它能够更深入地理解相关场景。其中的典型示例是，觉察可疑对话，并能够派遣安保人员进一步调查。或者，在医院或护理中心，它能够用来监听患者是否发病并呼叫护士。它还可以用来侦测“异常”声音是来自左侧还是右侧，并将PTZ摄像机调转为指向声源。

这些应用情形可能涉及一名或多名操作人员，他们可以在控制室中访问音频环境，或者也可以通过移动设备上的安防应用软件来进行这种访问。人的耳朵捕捉声音，大脑提取与场景或情境相关的信息。在结合视频监控的情况下，音频能够为决策制定另外增添了一个信息维度。在某些情况下，如果音频源不在摄像机视野范围内，或者如果光照条件欠佳，音频实际上就可能是仅剩的信息维度。

7.3.1 监听与目击

音频监控也可以用作基于目击（所听到的）事件的直接证据。这种应用情形不同于*监听与交互*用途，因为其目的不是决策制定，但这两种应用情形通常是共存的。例如，在听到激烈争吵且其中包含有犯罪指向的语言时，操作人员不仅可以派遣安保人员，而且还可以作为目击证人就所听到的内容提供证词。

7.4 记录与存储

在适当情况下，通过捕捉和记录音频数据，能够提供更详细的证据。这可能涉及有犯罪指向的语言或枪声。所记录的音频能够证明谁说了什么、开过多少枪、或类似的涉案事件。

在出于司法证据目的而记录音频时，应注意保留原始数据，避免信息处理（而这在别的情形下可能是必要或有利的）。对于司法级别的记录，不同形式的处理都可能被视为篡改证据。可以使用语音增强算法来提高录音的辨识度；它可以提高证据价值。但这样的算法应在后期应用于录像材料的副本。通过尽可能保持原始的录像内容，后期在材料的使用方式上便有更多选择。

7.5 进一步发掘监控系统的潜力

监控系统通常采用多种类型的传感器。当然，摄像机的图像传感器就是其中之一，它用于寄存场景视觉画面。此外，通常也会使用非视觉传感器，比如，基于雷达技术或红外辐射的移动侦测器。有时，视频监控并不适用，因此仅使用非视觉传感器。但在许多情况下，非视觉传感器用于补充其他类型的信息，对摄像机系统做进一步完善。

通过在监控系统中纳入音频传感器（麦克风），能够增强大多数应用场合的监控效果。为非音频系统赋予音频能力，能够通过分析工具或操作员交互，实现多传感器交互。

*监听与交互*的应用情形就是一个简单的示例，其中操作人员还能够同时接收到音频流，从而更好地理解监控场景的事态。仅通过观看画面，可能难以判断暴力行为，但如果同时还能够听到声音，这种判断就容易得多。

另一个典型的示例是使用视频分析工具，如视频移动侦测。如果在例如低照度条件下，视频分析应用难以发挥效用，那么通过 音频 分析工具，就能够提高 侦测可靠性。

8 监控与侦测

音频包含多种类型的信息，这些信息既适用于监控，也适用于音频分析。不同类型的处理和特征描绘有助于提取并优化这种信息，从而实现更轻松的使用以及与周围系统的更轻松交互。

8.1 声音特征

在监控环境中，诸如响度和音高等的特征可能包含重要信息。而声音的可听持续时间、声音是否在移动、或者声音是来自近处还是远处，则都是其中一些例子，为我们的声音判断增添了多方面的维度。用于音频监测和侦测的软硬件被设计成能够处理相同类型的信息，即“监听” 多种特征的复杂组合，这些特征包括分贝等级，甚至随时间推移在不同频率下的能量，等等。

- **空间信息。** 这涉及我们周围的真实世界，包括位置、方向和距离等概念。空间信息可用于在不同方向上突出或缩放声音捕捉，以便更好地记录。它也可以被分析工具用来确定声音来自何方、或者与声源相距的距离。
- **时间信息。** 时间信息在动态意义（随时间推移而改变）和绝对意义（事件发生在何时? ）上都具有较大的重要性，这种信息的处理通常需结合来自其他传感器的信息（如视频）。时间信息在行为分析方面有着重要意义，它让您能够知道何时发生了什么以及持续了多长时间。
- **频谱信息。** 这涉及频率，比如声音的音高，或者在较复杂的声音中，还涉及多种音高的组合。音频监控中使用的麦克风被设计成具有平坦的频率响应，即，它们试图同等地捕捉可

听频率范围内的所有频率 (20 Hz – 20 kHz)。这不同于人类听觉系统的工作方式，因为相比其他频率，人类能够更容易地侦测到通常出现在人类语音中的频率。

- **振幅信息。**这涉及声音的强度或响度。振幅信息可以补充频谱信息，并共同用于绘制所传入的音频的结构图。

8.2 信号处理

在音频监控中，信号处理 通常涉及改善传输、存储效率或主观音频质量，或者突出或侦测目标部分。这通过软件算法来实现，这些算法能够以多种方式修改或分析音频。

8.2.1 修改信号

可以利用算法来更改信号，使其能够用于特定用途，这种用途通常是：

- 改善信号，例如，通过自动增益控制，提高声音辨识度。
- 更改信号，例如，这通过使用均衡器更改相关频率内容来实现。
- 通过移除特定频率或振幅，来限制信号。这可能涉及通过压缩的方式减小数据量，或者涉及通过语音加密来确保隐私。

8.2.2 分析信号

音频分析工具使用所捕捉（但通常未记录）的音频数据，并分析相关的声音特征，从而得到非音频结果。从本质上讲，这些应用软件将音频数据转换为其他格式的更具执行性的资产。有些分析应用软件专门设计用于侦测（比如）暴力事件、枪击、玻璃破碎或汽车报警。

如果使用了机器学习算法，那么就可以用大量数据来训练这些算法，让它们在未经专门编程的情况下，学会做出预测。音频应用环境的一个示例可能是，在使用数以千计的关门声来训练某种算法之后，此算法能够可靠侦测这种声音。

8.3 人类听觉

人的耳朵是侦测和分析音频的理想工具之一。在非常嘈杂的环境中，人耳和大脑仍能够侦测并理解话语 ，而大多数算法可能却无法做到。

利用耳朵，我们能够推断场景的空间信息，比如，声音来自哪里以及音频源是否正在移动。由于拥有两只耳朵，我们能够判断声音是来自左侧还是右侧，亦或是这之间的某个地方。耳朵和头部在构造上也使得我们能够判断声音是来自上方还是下方，是来自前方还是后方。大脑中的若干“过滤步骤”与耳朵之间的时间差相结合，能够立即分辨小至微秒的偏差，从而让我们注意到特定类型的事件。我们拥有发达的音频信号分析能力，这尤其体现在人类声音以及与过往危险相关的声音方面。

在适当的环境条件（如良好的声音质量、立体声、延迟较小）下，人类操作员可以是强大的“分析工具”，并能够补充侦测软硬件。利用仅配备两个麦克风的音频监控产品，操作人员能够推断场景的空间信息，比如，声音来自哪里以及该声音的移动状态。

9 免责声明

本文档及其内容经由安讯士提供，与本文档或其中所涉及的任何知识产权（包括但不限于其中的商标、商业名称、徽标以及类似标志）有关的所有权利均受到法律保护，本文档或其中所涉及的任何知识产权中的或关联的所有权限、权利和/或权益都并且应都归属于Axis Communications AB。

请注意，本文档系“按原本”提供，不包含任何类型的保证，仅供参考之用。本文档中提供的信息不构成且不意在构成法律建议。本文档不意在构成且不应构成Axis Communications AB和/或其任何附属公司的任何法律义务。Axis Communications AB和/或其任何附属公司的与任何安讯士产品相关的义务仅遵从在安讯士与直接从安讯士购买此产品的实体之间的协议中所规定的条款和条件。

为避免疑问，与本文档的使用、结果和效用有关的风险均由本文档的使用者承担，安讯士在法律允许的范围内否认并排除保证，无论是法定的、明确的还是隐含的保证，这其中包括但不限于适销性、对特定用途的适用性、权益和非侵权保证以及产品责任、或者因与本文档相关的任何提议、规格指定或样本所致的任何保证。

附录 1　音频质量术语

数字音频：

数字音频是模拟音频（通常是使用麦克风捕捉的声音信号）的一种表示方式，它以数字形式记录。在数字音频中，音频信号的声波通常被编码为一系列连续的数字样本。它的准确度取决于编码器所记录的有效数位的数量。例如，在CD音频中，采样频率为每秒44,100次，每次的采样深度为16位。

噪声：

噪声是不期望（但有时又无法避免）的声音，这种声音会限定或限制响度范围的静音限值。它可由音频链上的众多组成部分产生，从所记录的声源（比如，房间内的风扇），经由麦克风（*比如，*自有噪声、振动、风）和线缆（*比如，*干扰、串扰），直至捕捉设备（*比如，*自有噪声、数字采样噪声），所有这些组成部分相互影响，即形成通常所说的“本底噪声”。

噪声通常由SNR（信噪比）来定义，涉及从定义级别（有时称为系统能够处理的最大响声）到本地噪声的整个范围。

在视频中，与之相当的术语是视频噪声，它的表现形式是随机（通常）静态像素图案，即“雪花”，能够限制阴暗图像中的内容呈现（与音频噪声对静音信号的听觉限制如出一辙）。

失真：

以不期望的方式从原始“真”信号中减去某个信号，即被称为失真（上文所述的噪声通常不属于失真）。失真会降低主观音频质量（听起来“好听”的失真也是常有的）并掩盖客观信息内容，从而使得信号难以被听到，尤其是在内容分析中，这会降低分析工具的性能。

THD（总谐波失真）和IMD（互调失真）是常用来量化失真的两个特性。

与视频相关的失真表现为伪影，比如色差、光晕、模糊等；它使得图像的呈现效果“差”，并限制了细节呈现的丰富程度。

采样率和频率响应

在数字系统中，以每秒若干次的频率对音频采样。这便是采样率（通常为每秒8000至48,000次，或者以Hz计）。为了适当地捕捉声音，信号原理（尤其是Nyquist–Shannon（奈奎斯特–香农）采样定理）告诉我们，采样率需要为模拟信号中最高所需或必需频率的至少两倍。

正常人耳能够听到的频率范围为20 Hz至大约15–20 kHz，具体取决于年龄和其他因素。总的来说，低频范围（不超过数百Hz）通常定义特定声音的基本特性（比如，语音的基频），而高频率范围（数千Hz以上）则包含更多“细节”。

音频的频率范围与视频的分辨率和帧速相似；设置值越低，细节呈现越少。

位深：

每次对音频采样时，都会捕捉模拟值，并将其转换为数字表示。在数字域中，没有无穷大的值，因此细节量受限于所定义的位深。每个位代表二元中的一元（0或1，低或高，等等），再结合定义的振幅范围（*比如，*所选择的电压或声压级），能够形成这个范围的分段范围。2个位可以产生4个分段，3个位可以产生8个分段，以此类推。以3个位采样的简化（1伏）信号可被拆分并以1/8伏步长表示。

为获得足够好的音频质量，至少对于人耳来说，16位采样通常就已足够（代表65 536个步长）。这也是CD音频所使用的位深。对于分析或更严苛的应用，建议使用24位。

位深在视频方面表现为对比度，即，每个像素能够再现的亮度或色度范围。

T10167735/zh/M2.2/2106

关于 Axis Communications

Axis 通过打造解决方案，不断提供改善 以提高安全性和业务绩效。 作为网络技术 公司和行业领导者，Axis 提供视频监控解决方案， 访问控制、对讲以及音频系统的相关产品和服务。 并通过 智能分析应用实现增强，通过高品质培训提供支持。

Axis 在 50 多个国家/地区拥有约 4,000 名敬业的员工 并与全球的技术和系统集成合作伙伴合作 为客户带来解决方案。 Axis 成立于 1984 年，总部 在瑞典隆德