Audio Analytic

Why Real Sounds Matter for Machine Learning

Exploring the shortcomings of publicly available sources of audio data, such as YouTube and FreeSound, for sound recognition system training and evaluation.

By Adrian Stępień with contributions from Arnoldas Jasonas, Simon Worgan, Sacha Krstulović and Neil Cooper

Contents

1. Executive summary

All sounds are not created equal. When it comes to training and evaluating sound recognition systems which deliver top performance in consumer applications, you cannot rely on recordings downloaded from the internet, due to a range of legal and technical limitations.

Despite the fact that it can be accessed at a click of the mouse, audio and video content shared online is made available for specific purposes only – typically for social networking, entertainment and media production sound effects. In many cases its usage for commercial purposes is prohibited and the copyright ownership is held by the person or company who created, filmed or recorded it. From a legal and ethical perspective, you need appropriate licence and copyright permissions to use audio data for training a system that will be commercialised. Launching a consumer product into the global market without such permission – or the ability to prove it – opens you up to significant risk.

Technical limitations fall into two broad but significant categories. First, internet-downloaded recordings don't usually match the target consumer usage scenarios. For example, they suffer from certain biases such as being recorded in the wrong environment (indoors instead of outdoors with additional environmental reverberations, etc) or at the wrong distance from the microphone.

Second, when you play back internet-downloaded recordings to evaluate a sound recognition system, a range of environmental conditions such as echoes and other room responses – as well as the limitations of the playback device – mean that the sounds do not always match reality and are, more often than not, unsuitable for emulating real world scenarios.

In this whitepaper, we've explored the legal and ethical issues, and various technical limitations of using internet-downloaded recordings, explaining why a detailed and considered approach is required when training and evaluating sound recognition technology that is fit for the real world.

On two occasions I have been asked, "Pray, Mr Babbage, if you put into the machine wrong figures, will the right answers come out?" ... I am not able rightly to apprehend the kind of confusion of ideas that could provoke such a question.

Charles Babbage, Passages from the Life of a Philosopher

2. Introduction

Many of us take our sense of hearing for granted. However, when giving machines their own accurate sense of hearing, it is important to understand that humans and machines both listen in very different ways.

A human uses additional sensory inputs and past experiences to identify a sound to effectively 'fill in the blanks'; we hear the meaning of the sounds from their raw vibrations. By contrast, a machine is trained to hear using information that is often irrelevant or imperceptible to our human ears. Every fragment of audio information – whether that's perceptible to the human ear or not – is required by a sound recognition system to effectively do its job.

For example, in a movie, the hero may get thrown through a window. To you and I, the glass-smashing sound effects sound like a window being broken. This is because the sound effect has been designed that way using Foley Art and we are likely to have heard the sound of a window being broken in movies more frequently than in real life. However, a real window being smashed sounds nothing like this. So, to a machine with a sense of hearing, this sound effect used for the movie isn't a **real** glass window being smashed.

Correctly, the machine doesn't recognise this imposter sound as it has been trained using real windows being broken in real homes using acoustic and temporal features that we, as humans, do not fully appreciate.

Appreciating the need for machines to extract features that are irrelevant to us is critical when training and evaluating products featuring sound recognition. In this whitepaper, we provide you with an understanding of the legal and technical limitations of internet-downloaded recordings, giving you a glimpse into why sound recognition is a burgeoning and specialist field of machine learning (ML). This knowledge will also help you to better assess the performance of sound recognition systems.

A sound recognition system running on diverse products, such as smart speakers, smartphones, true-wireless earbuds and smart home devices, should be evaluated in real-world conditions and using real sounds, which include everyday, naturally-occurring issues such as acoustic room effects and microphone distortions. Sound recognition systems should perform well under such field conditions. After all, the commercial success of sound recognition depends on its ability to work in the real world, not the lab or the silver screen.

3. Legal and ethical issues

The legal and ethical implications of training sound recognition systems with recordings downloaded from the internet, that you do not have the clear permission to use, are significant. These include being sued for breach of licence conditions and copyright, fines from regulators, and reputational damage through negative media coverage. The regulations around AI and data processing are also becoming more and more restrictive, so it is important to adopt approaches that can adapt to protect your organisation now and in the future. The same is true when licensing software and technology from third parties.

Those working in machine learning have a legal and ethical responsibility to make sure that they have clear permission to use the data for training a system for commercial purposes. For example, when an individual like you and me uploads a video or an audio file to a site such as YouTube or Freesound, they retain full ownership rights of that content. In order to use that data for commercial purposes you would need the permission of each 'owner' before that data could be used for training purposes. Even if you are using datasets like AudioSet, which includes a list of 200,000 non-speech, human-labelled 10-second clips from YouTube, an organisation may still need permission to use the raw data files that the list provides. As an example, audio clips listed in the AudioSet dataset labelled as 'baby cry' include the cartoon character George from Peppa Pig.

Aside from sounding very little like a real baby, the George character is the intellectual property of its creator and is subject to copyright.

As IBM found out when they used Creative Commons images on Flickr to train a facial recognition system, ethical issues exist around the use of public sources of data for machine learning applications, even if the content owner was happy to allow commercial usage. In the case of IBM's usage of Flickr, some users were unhappy as the commercial application they had in mind was the usage of photos on posters or in adverts where people could marvel at their photographic skills, not the training of a facial recognition system.

The legal and ethical landscape is also likely to change as national and international regulators catch up with industry and data traceability implications, as we have seen with the recent California Consumer Privacy Act (CCPA) and GDPR legislation in the US and Europe respectively. For example, in February 2020, the EU published its 'On Artificial Intelligence - A European approach to excellence and trust' whitepaper, which sets out a potential regulatory framework around AI. This includes a suggestion that the regulatory framework could prescribe that accurate records should be kept regarding the dataset used to train and test the AI systems, including a description of the main characteristics of the dataset and how it was selected.

Such legal and ethical issues can only be resolved by acquiring full data ownership and achieving full GDPR compliance. This is easier to achieve with primary-sourced data than with internet-downloaded audio.

4. Technical issues

Depending on the evaluation circumstances, an internet-downloaded sound event may be correctly classified, misclassified or missed altogether by a sound recognition system. This is highly dependent on the distance between playback speaker and microphone, the type of sound, its frequency content and the effects applied to it. This is because internet-downloaded audio files are unknown quantities – where information about the recording environment and processes are unclear. These inconsistencies make the files unsuitable to train, and undesirable to evaluate, a sound recognition system fit for consumer applications.

The inconsistencies and problems introduced by internet-downloaded audio recordings span a vast range of factors, as highlighted in figure 1, which compares the issues facing real sounds in real environments with files found on the internet. The combination – or mixture – of these factors makes it difficult to pinpoint the root cause (or causes) of any resulting issues in the source file, leading to mistakes in the evaluation and training of sound recognition systems.

We will now look at each of these issues in turn, explaining the impact that each can have.

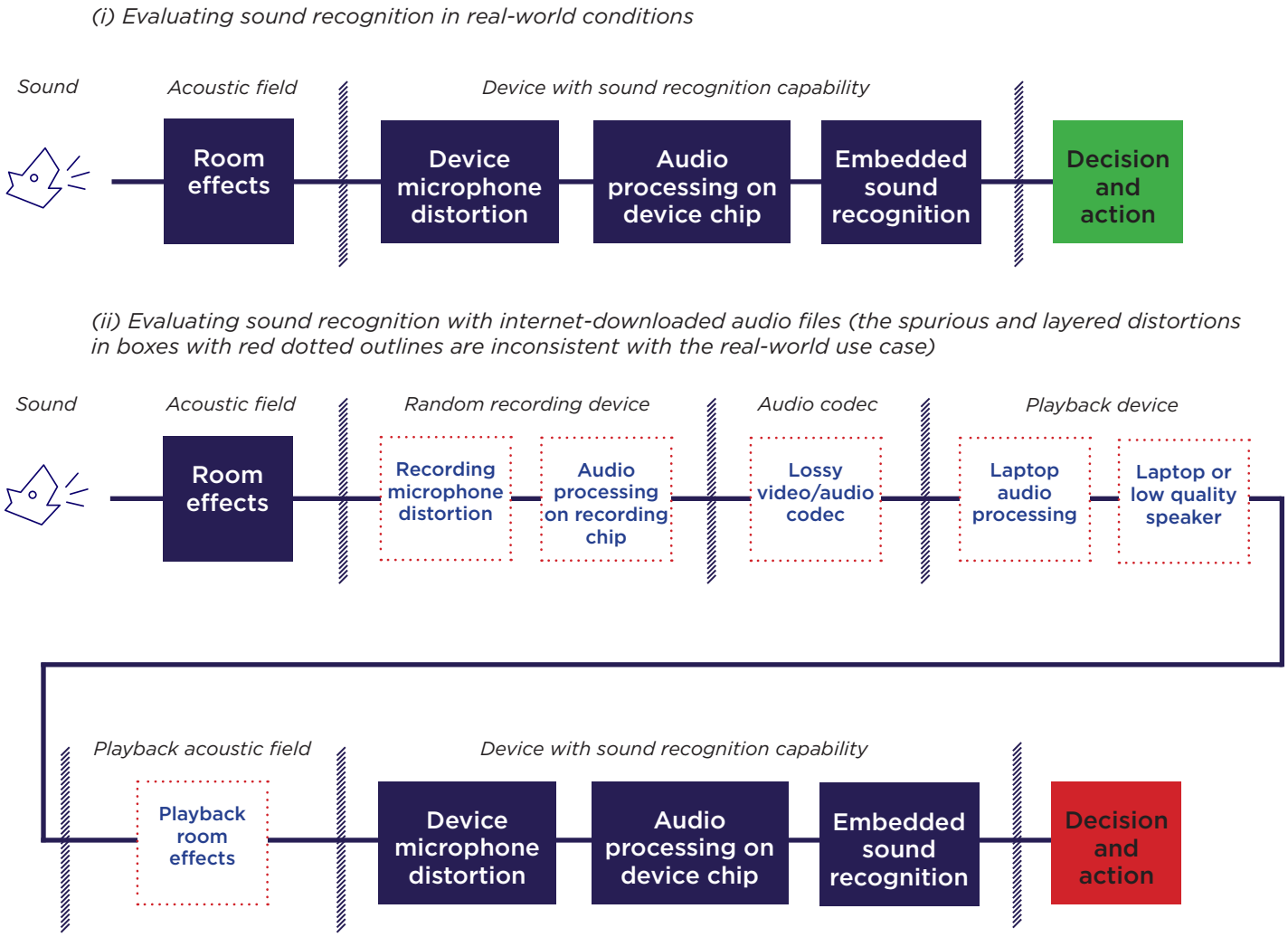


Figure 1: Using downloaded and unknown audio recordings for training or evaluation introduces a range of distortions. This 'pipeline of distortions' is clearly illustrated when comparing evaluations (i) in real-world conditions and (ii) using an audio file from the internet

4.1. Subject matter diversity and variability of sounds

A major training limitation when using internet-downloaded audio files is that they may not represent the diverse range of real sounds. Sometimes this may be down to the fact that the person training the system doesn't have the available subject information from the recording, or that the file was just tagged according to the uploader's judgement rather than according to a consistent taxonomy. As we mentioned at the start, these files are often unknown quantities and this lack of metadata may mean that systems are trained using what is available, including a large proportion of noise, rather than what is representative of the real world.

Not only do ML and data engineers need to understand the quality and variability of their training data, they also need to fully understand the target application of the end products in order to carefully plan the data collection stage.

To highlight the impact of using narrow or unknown data, consider this example: if a system built to detect a glass window being broken is only trained using recordings featuring one type of glass – because audio recordings featuring only one type of glass are all that were available online – then that system may not cope with the multitude of other types of glass windows which are present in consumers' homes. It becomes even worse if the category tagged as 'glass' includes sounds of kitchen glasses being broken or being merely 'chinked' for a toast.

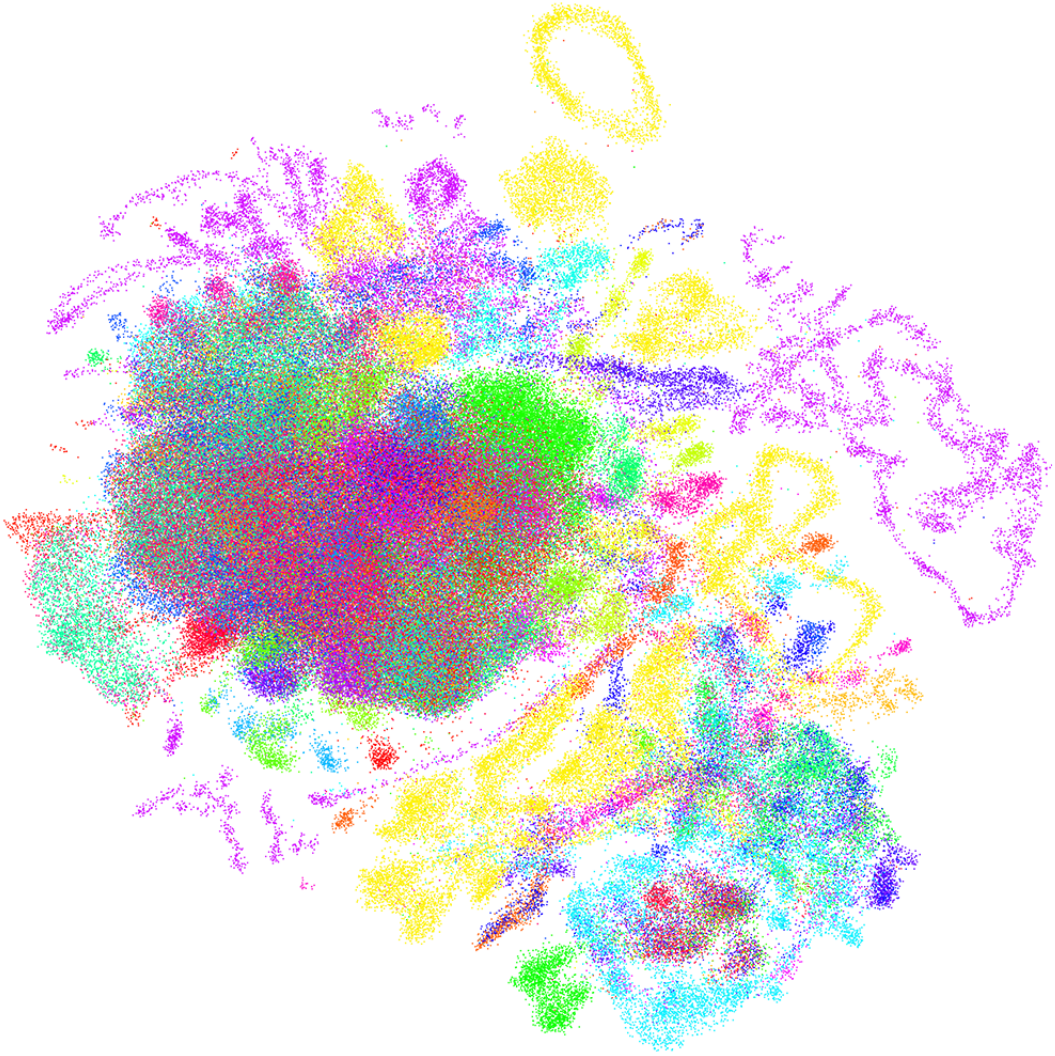


Figure 2: Sound Map generated from the Alexandria™ dataset (each colour represents a different class of audio event)

Fortunately, this does not happen with Audio Analytic's data, because we know precisely what has been recorded, and all the sounds are labelled according to a coherent taxonomy. Also, we make sure that the recorded sound variations are consistent with what happens in real life: 'glass' really means windows in your home. Using our Alexandria™ dataset, figure 2 shows a visualisation of 15 million sounds in two dimensions, where each colour represents a different class of audio event, such as smoke alarms, emergency vehicles, glass breaks, people shouting, etc. Similar sounds, such as speech or laughter, are closer to each other and, conversely, sounds that differ are distant.

This data visualisation highlights the variations of real-world sounds within their own classes, as well as the overlap between sounds, thus highlighting the complexity of dealing with real-world sounds.

4.2. Random recording devices

The majority of internet-downloaded audio files are recorded on the simple microphones found in smartphones, laptops, tablets and other personal devices.

Sound recognition systems must tolerate the channel distortions from consumer microphones - the frequency responses of those microphones have significant variations, with a lack of low and high frequencies, and additional audio processing applied to them. When unrealistically or inappropriately using internet-downloaded audio files to evaluate or train a sound recognition system, such channel effects are overlaid, and the reproduced audio is not true to life. Further problems arise during playback, when system distortions, codec distortions and playback environments are combined with these.

We will look at these other issues in more detail in the following sections.

4.3. Codec compression

The poor application of codecs for the compression of downloaded audio is a major factor affecting the quality of the audio files from the internet. Many such files are encoded using lossy codecs, such as MP3, AAC or Vorbis. Once encoded, a portion of audio information is lost and the quality is lower compared to lossless formats such as WAV or FLAC. This may be fine if decoded audio is fed directly from a device's audio channel to a sound recognition module, for example when sound recognition is required to deal with a security camera's AAC audio stream. However, encoding/decoding effects can cumulate or amplify in undesired ways if propagating decoded files further down the chain for system testing or for data augmentation.

These codecs are specifically designed for encoding music and speech, often removing information that humans cannot perceive. They may, for example, utilise frequency masking, which can make it difficult for a machine to detect specific sound events, especially those with a specific time-frequency structure.

Figures 3 and 4 clearly demonstrate this point, where spectrograms of original high-quality recordings and encoded versions are shown. The black zones represent frequency content holes and band-limiting effects caused by the encoder, which are redundant for human listeners but are likely to affect the performance of the sound recognition system if used without care.

In addition to the encoding and normalisation of audio files, other audio effects may be applied, which are difficult to tell by just listening to the recording. These effects could include:

- Dynamic Range Compression: where the volume of the loud parts of the audio is reduced and quiet parts are amplified, compressing an audio signal's dynamic range
- Equalisation: further alters the frequency response of the recording
- Synthesisers: adds artificial sounds on top of the original recording or synthesises the sound
- Noise reduction: which can corrupt target sounds.

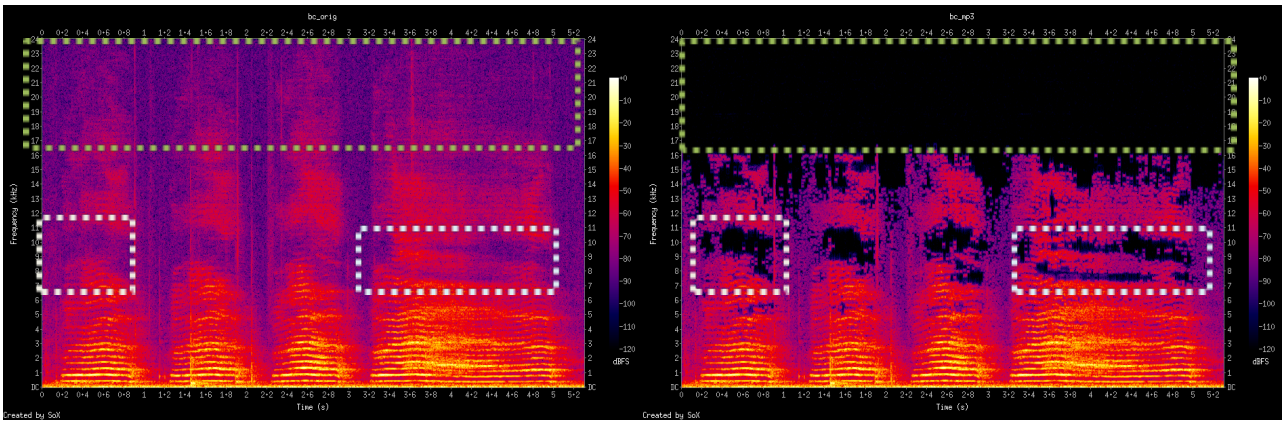


Figure 3: Spectrograms of high-quality baby cry recording (left) and MP3 encoded version (right)

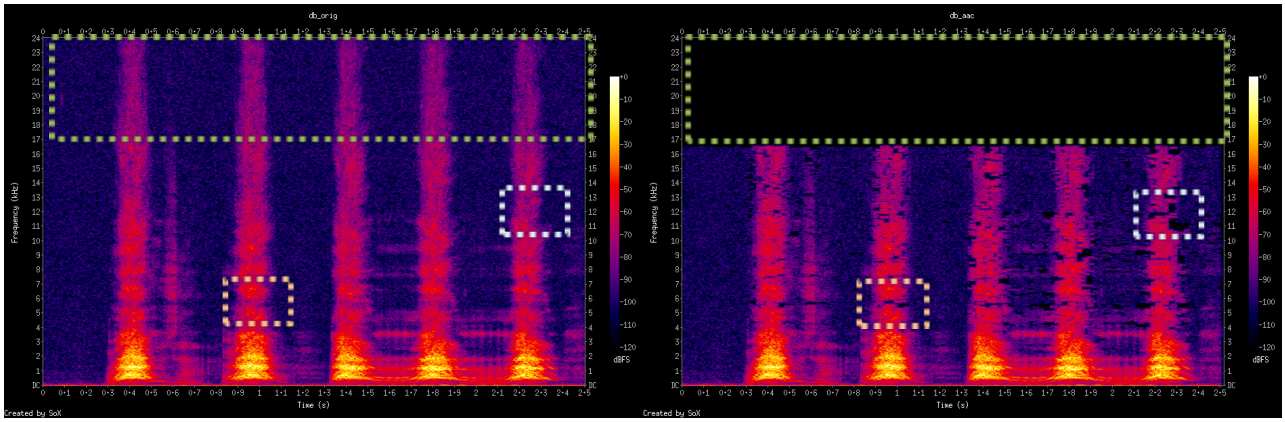


Figure 4: Spectrograms of high-quality dog bark recording (left) and AAC encoded version (right)

4.4. Lack of Sound Pressure Level (SPL) information

Most internet-based recordings lack the Sound Pressure Level (SPL) information for an acoustic event. SPL is an indication of the natural loudness at which sounds physically occur in the real world. As a result, the underlying question is whether, for example, the recording of a baby crying will be as loud as a real baby in a real room. If you are not training a model based on a correct representation of the target sound then the performance of the model will be affected, often missing occurrences of that sound. If evaluating the performance of a high-quality system using this same recording you may be playing it at a volume that doesn't represent its natural loudness and, as a result, the system may not confidently recognise it because a key acoustic feature is not present. It may help at this point to briefly explain what SPL information is. When a sound is recorded, the analogue signal is converted into a digital waveform. This digital waveform is directly related to the SPL of an acoustic event.

To play back a recording with the exact same SPL, we must know the sensitivity of the microphone as well as any other gains applied at each stage of the audio path. Generally, this requires a calibration signal with a known SPL to make that mapping.

Furthermore, the amplitude of the microphone signal is often boosted before the recordings are uploaded, to increase the volume on consumer devices. This makes it infeasible to playback the recording at the correct natural sound level.

Due to the fact that all of our audio data is primary-sourced, we embed the SPL information in the recording alongside other

important metadata. This is very useful for system evaluation, because we can reproduce the evaluation sounds at their original and natural loudness, just like they would appear in the natural environment around a particular consumer device.

4.5 Limitations of playback devices

When evaluating a sound recognition system in a simulated environment, accurate playback is required to ensure the accuracy of that evaluation. To achieve this, you must use loudspeakers with a flat frequency response to minimise distortions. This is usually not the case for most consumer-grade speakers found in our laptops, smartphones and many other consumer devices.

In figure 5, you can see the frequency responses of a high-quality loudspeaker and two consumer-grade laptop speakers, where the latter struggle to reach the high sound pressure levels (SPLs) of sounds in the real world. The low-frequency content (below 500 Hz) is significantly reduced.

In addition, Laptop 1 struggles to reproduce high frequencies above 4,000 Hz. In comparison, the high-quality loudspeaker can reach realistically high SPLs with a flat frequency response, meaning that it can play back the sounds at their natural volume and without distortions across the whole frequency range.

Such frequency response deviations are tolerated by listeners in consumer-grade speakers because these are optimised for speech and music playback, whereas more general sound playback requires different measurements and calibration.

For example, the speaker must reproduce sufficiently high SPL values across the full frequency range, which is off-limits for laptop speakers, where the maximum measured SPL levels are 84 dBA for Laptop 1 and 86 dBA for Laptop 2.

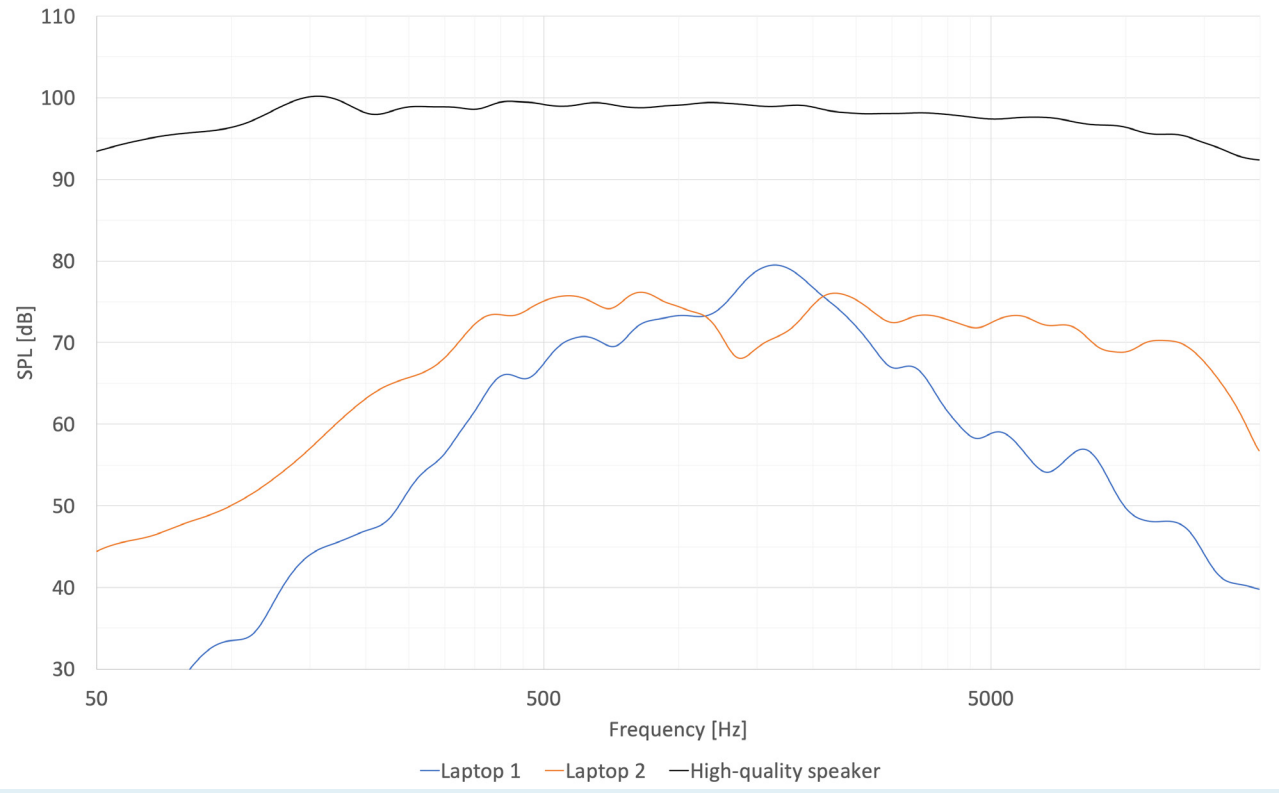


Figure 5: Examples of measured speaker frequency responses of two laptops and a high-quality speaker

In figure 6, the distribution of SPLs for four different types of audio events is shown to illustrate its importance. For accurate sound reproduction, a speaker must be capable of reaching these levels. However, these are generally outside of laptop speakers' range. Reaching levels of 100 dB SPL is possible in principle if the volume is cranked all the way up, but this results in significant sound distortions.

In addition to poor quality speakers, some devices, especially laptops using the Windows operating system, use additional audio processing. Advanced users can disable this type of processing. However, the only way to guarantee that no unwanted processing is applied through omitted or invisible processing modules is to conduct a test measurement over the whole audio chain, as highlighted in figure 1(ii).

There again, the point is to make sure that the sounds haven't deviated too much from the application domain when testing or training; what you want is recognition of real sounds from real devices, not a playback recogniser.

4.6. Playback room effects

Think about shouting in a cave, compared to a wide-open space. Your voice is going to sound different, depending on where you are. The same applies to any recorded sound. The room response will affect the recording.

Many audio files are recorded in a reverberant environment, as opposed to an anechoic chamber where the sound is not impacted by the room. When training a sound recognition system these room responses need to be fully understood so that the model is able to adapt to different environments when used by consumers.

When playing back such files in a regular room to evaluate a sound system, you inadvertently add **another** room response on top of the existing one within the recording. If you are using an internet-downloaded file where the room response is not understood in the first place, this causes further problems when testing a sound recognition system. You now have to contend with the room effects of the room where you are playing back the

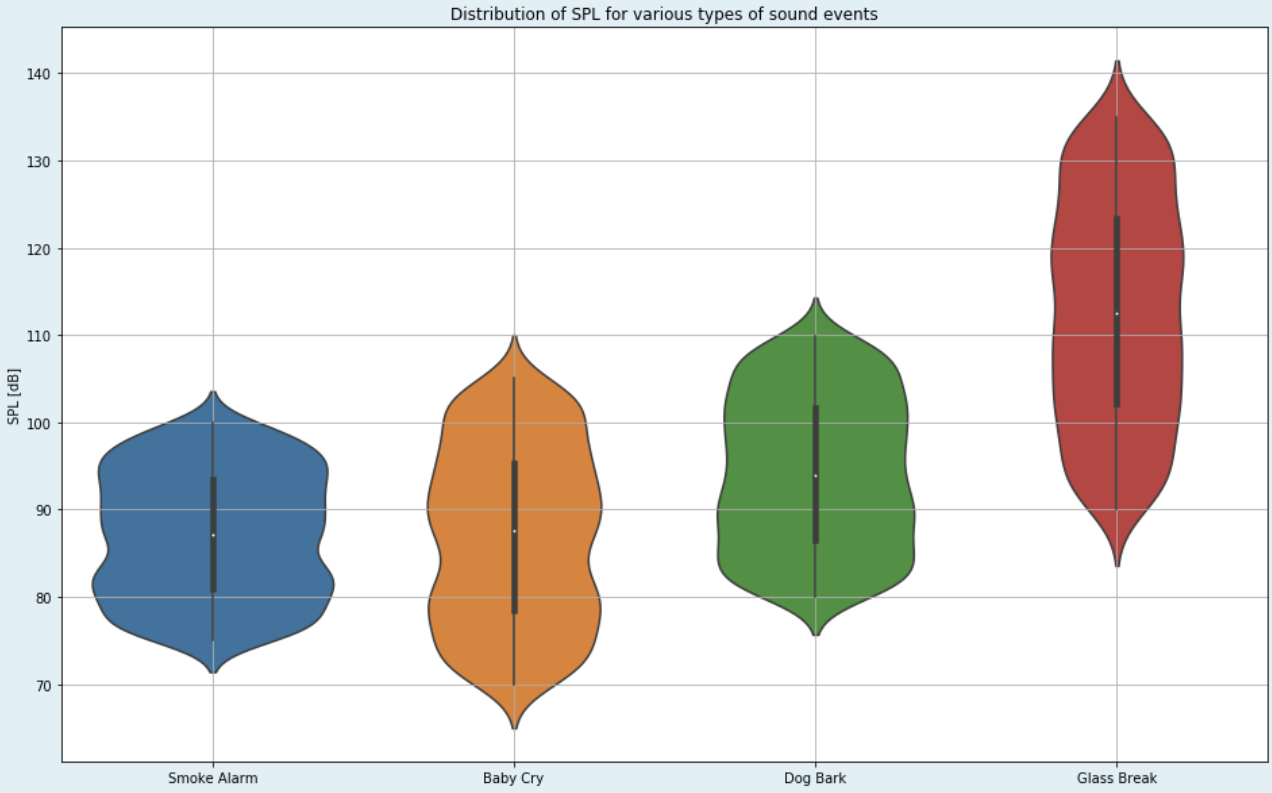


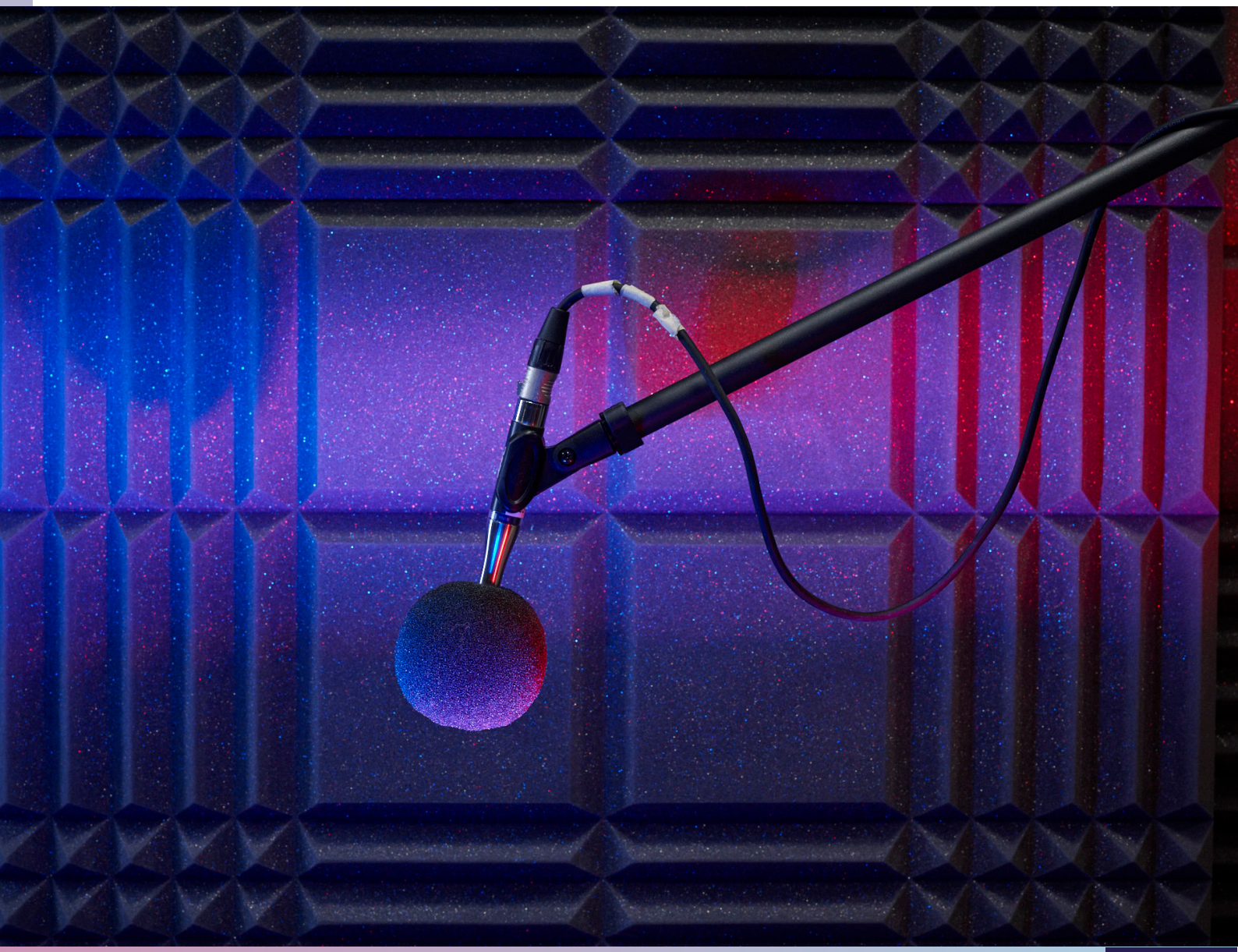
Figure 6: Distributions of SPL for various types of sound events with estimated maximum SPL of consumer-grade speakers

sound **and** the room where it was originally recorded.

As a result, it is difficult to determine whether the performance deterioration comes from the room effects in the original space in which the audio was recorded, or in the room where the sound is being played back.

To understand the detrimental effect of these multiple layers of room responses, this video from Alvin Lucier on YouTube enables you to hear and appreciate the impact: https://youtu.be/fAxHlLK3Oyk.

So, one layer of room effects is fine, and indeed a high-performance sound recognition system should be robust across the huge range of environmental responses observed in consumer applications, both indoors and outdoors. On the contrary, overlaying many room effects might end up sounding as crazy and as unreal as Alvin Lucier's experiment.



5. How do you give machines a sense of hearing?

Thanks to the information contained in this whitepaper you will have a better understanding of the impact of internet-downloaded audio recordings on machine learning. As a result, you will be better informed when it comes to judging whether a sound recognition system is fit for running on smart consumer products in the real world. While it may be possible to trigger a correct response using random internet audio, you are now in a position to understand why system evaluation results obtained using audio that is acceptable for social media will be totally inconsistent with the results experienced by your users once the device is deployed in the field.

In our experience, there are four key considerations when it comes to collecting the right audio data: data management, facilities, equipment and purity.

5.1. Invest considerable time and effort in data, labelling and management

Machine learning relies heavily on diverse high-quality data, and the sound recognition space is no exception. To represent the variability of real-world sounds you need a well-labelled dataset that supports the task.

In addition, the dataset must contain both target and non-target sounds collected in the field across thousands of acoustic environments, allowing for accurate and representative training of a reliable, gold-standard sound recognition system.

To address this challenge, we built Alexandria™, the world's largest, commercially-exploitable audio dataset for machine learning with over 15 million labelled sound events, over 700 label types, and over 200 million metadata points. The data it contains is expertly labelled, featuring three data labelling levels (fine, episodic and weak), which are essential for model training.

5.2. State-of-the-art facilities are essential

With access to anechoic facilities and very precise recording equipment, it is possible to achieve full control over sound variability, and in particular to simulate an infinity of application environments with a very high degree of realism. The idea is to record pristine sounds which are untouched by the environment, in order to be able to control the environmental variations ourselves later on in ways which are consistent with the desired application. Indeed, sounds recorded through such equipment are virgin of any unwanted noise contributions: their background noise levels fall below the threshold of human hearing or the 'self-noise' of professional measurement microphones, and they don't suffer from any room effects or other environmental effects. For example, figure 7 shows how a recording in a meeting room alters the sound signal, compared to an anechoic environment. As you can see, the frequency changes substantially due to reflections in the meeting room.

Furthermore, anechoic facilities (figure 8) support the controlled and precise usage of augmentation techniques as they can be used as 'green screens' for audio. The sound recognition system can thus learn from a dataset that has been constructed to cover precisely specified use-cases that are consistent with consumer applications, for example recognising door knocks but ignoring other banging sounds, or simulating an infinity of realistic living room responses. We can do this because we have detailed knowledge around acoustics, to be able to both record virgin sounds and blend them with a dataset of room effects and common environmental sounds in very realistic ways.

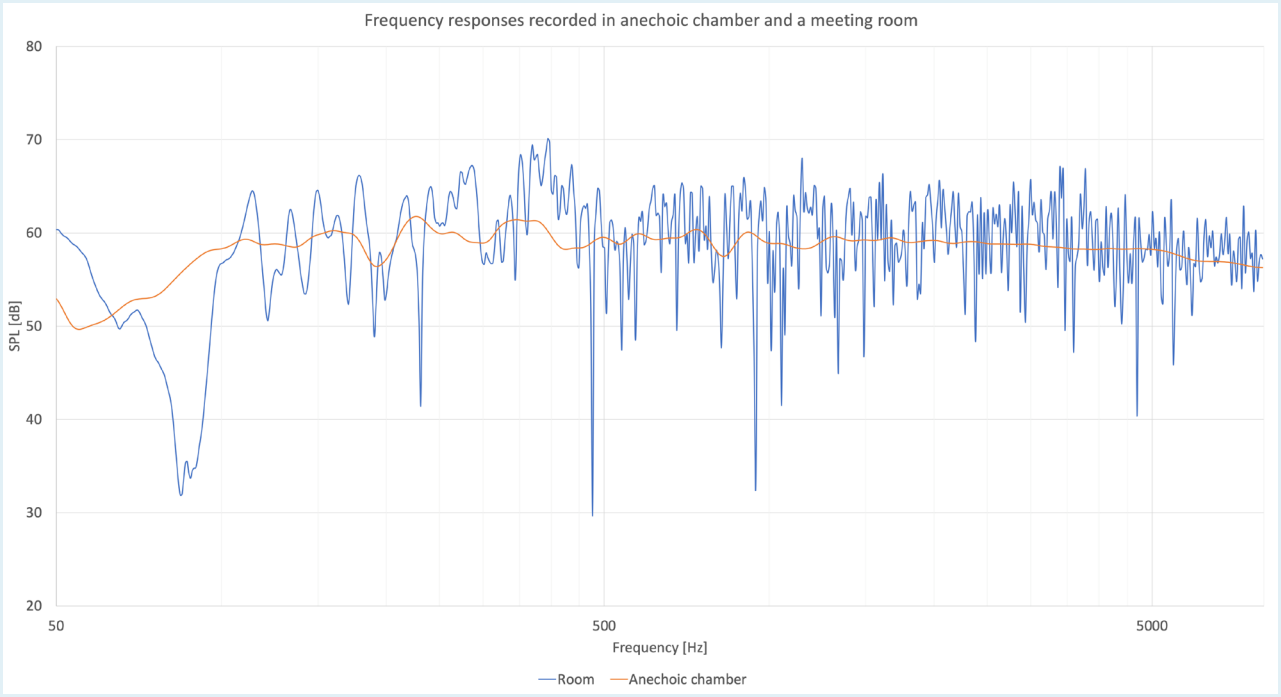


Figure 7: Frequency responses measured in anechoic chamber and a meeting room at 0.5m distance

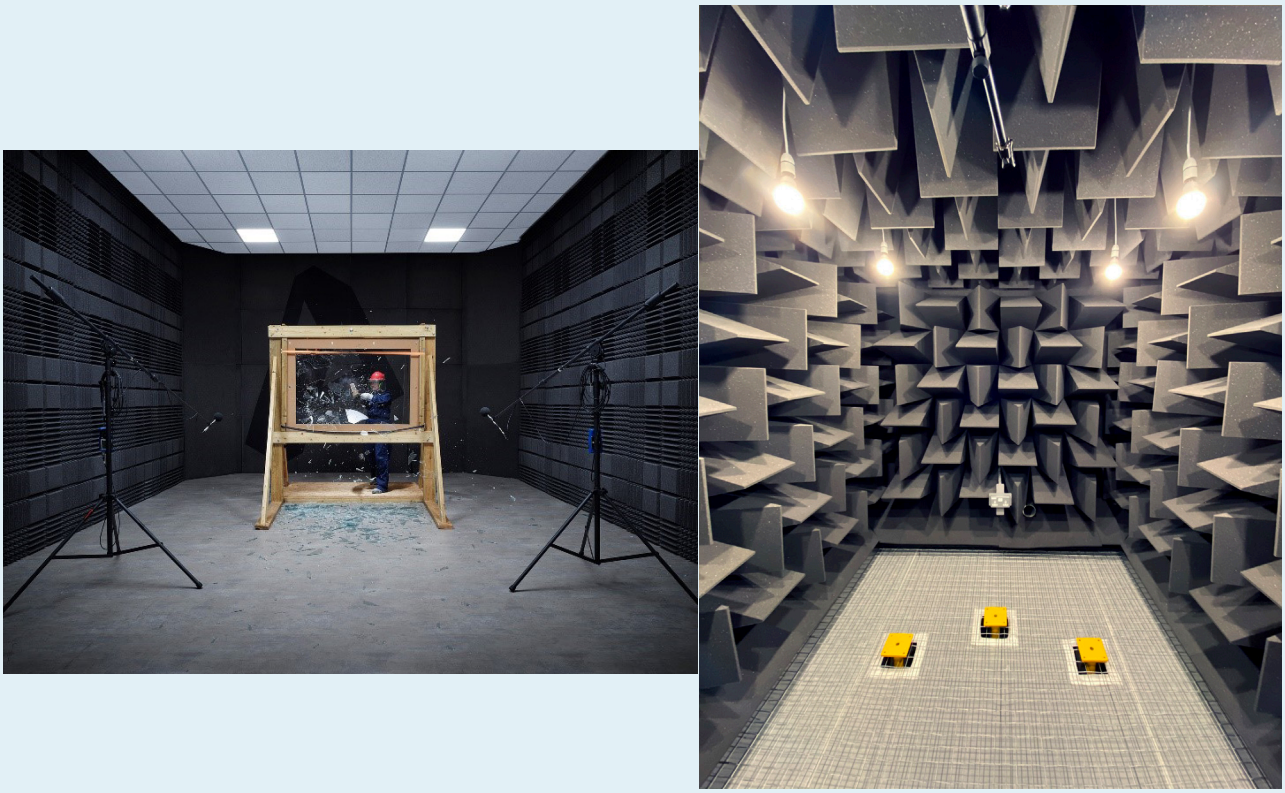


Figure 8: Some of our anechoic facilities which are used to create an audio 'green screen'

5.3. Use the right equipment

So, how do we record the very useful virgin sounds described in the previous section? High-quality audio equipment is essential to collect and utilise the full range of frequencies present in sound. In figure 11, the frequency responses of a typical Class 1 reference microphone and a consumer device are compared.

At Audio Analytic we collect audio data using high-quality, certified Class 1 reference microphones, which have passed our own audio integrity checks. This ensures the accuracy of our recordings across a range of frequencies, where these microphones have a negligible frequency variation of +/- 1 dB, and a self-noise lower than any consumer device. We use high-quality microphones alongside the standard ones that are found in a wide array of consumer devices, so that we can model the differences between perfect and real audio channels in our simulations. This allows the application of advanced augmentation techniques and ensures that the system being trained will work with the components and audio paths that OEMs typically use.

5.4. Use unfiltered, pure sounds and data augmentation techniques

Using the right equipment and facilities is so important because they provide a more precise understanding and control of sound variability, which influences training quality. Virgin audio is crucial to be able to apply high-quality data augmentation techniques, which are used to upscale our training data way beyond what would be achievable with in-situ recordings or internet-downloaded audio.

Indeed, our recordings avoid the application of codecs, compression, equalisation, synthesiser or noise reduction effects that would negatively impact on the realism of the augmented data, and in turn negatively impact on the fitness of sound recognition for real world application. As a result, our methods enable a system that works in real life across a much wider range of imperfect microphones and variety of locations than what could ever be achieved with internet-downloaded audio.

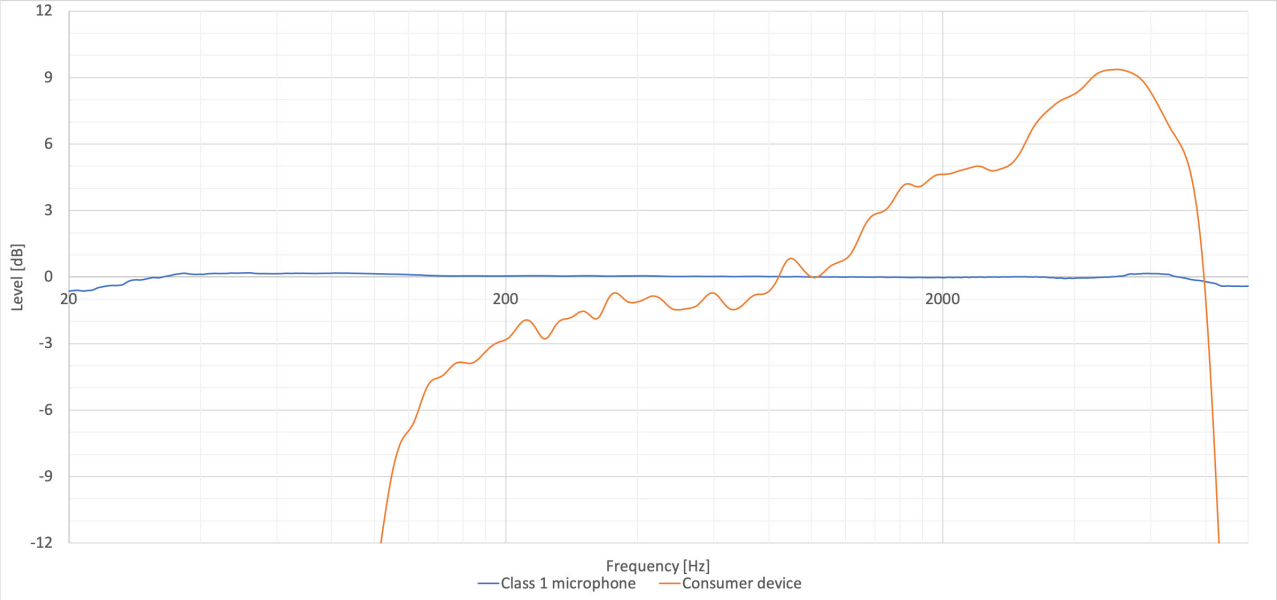


Figure 9: Examples of frequency responses of Class 1 microphone used by Audio Analytic and a consumer device

6. In summary

Sound recognition is a field that is radically different from other types of machine learning, such as image, facial, speech or music recognition.

But it faces the same issue that all machine learning systems encounter: if you put garbage in, you get garbage out. The particular challenge for sound recognition is that our ability as humans to understand what is garbage is limited by our own sense of hearing.

This whitepaper has highlighted the key technical and legal limitations of recordings downloaded from the internet. This will help you fully understand why diverse, high-quality data – as well as the ability to capture, specify and manage it – is so important to machine learning for real world sound recognition systems running on consumer devices.

The whitepaper has also explained why many of those same issues (and some additional test-specific challenges) make it difficult to assess the true performance of sound recognition technology using random files from the internet.

The impact of getting this wrong is poor performance, a lack of user trust in the technology, limited commercial success and significant legal risks. If you get it right, you deliver consumer delight, reliability and value.

About Audio Analytic

Audio Analytic is the pioneer of AI sound recognition technology. The company is on a mission to map the world of sounds and give machines a compact sense of hearing. By transferring our sense of hearing to consumer products and digital personal assistants we give them the ability to react to the world around us, helping satisfy our entertainment, safety, security, health, wellbeing, convenience, and communication needs.

Learn more: audioanalytic.com