Room Acoustics and Microphone Characteristics Show Systematic Impact on Sound Event Recognition

Gabriel Bibbó[1], Craig Cieciura[2], Mark D. Plumbley[3]
Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, United Kingdom

ABSTRACT

The robustness of audio pattern recognition systems under varying acoustic conditions and hardware remains a critical challenge for real-world applications. We examine how room acoustics, microphone characteristics, and overlapping events affect classification performance for domestic events. We conducted experiments in four rooms at the University of Surrey—with reverberation times (RT60: 0.27–0.78 s, 50 Hz–10 kHz) and clarity indices (C50: 11.6–18.5 dB; C80: 13.1–25.9 dB, 500 Hz–1 kHz)—using four microphones: USB Condenser, ICS-43432 stereo, AudioMoth, and Earthworks M23 reference. For two CNN-14 architectures, baseline performance obtained from the original audio was used for comparison with different microphone/room configurations. Results expressed as the percentage of audio frames correctly detected versus ground truth show: First, high RT60 degraded detection of impulsive events (e.g., door knocks) by approximately 50%, while sustained events (e.g., speech, music) remained above 90%. Second, overlapping events produced masking effects that reduced performance by about 20%. Third, while microphone differences affect accuracy, low-cost devices matched reference performance for speech and music classes. Both CNN-14 architectures exhibited similar degradation patterns across conditions. These results underscore the need for improved acoustic characterization and hardware-aware processing. We suggest that future work should integrate adaptive feature extraction and training strategies to mitigate reverberation and overlap in complex environments.

1. INTRODUCTION

Modern AI systems for Sound Event Detection (SED) often perform well in laboratory settings but may underperform when placed in real homes [1]. Unlike controlled environments, domestic spaces have varied acoustic properties and everyday noise sources can overlap, reducing detection accuracy. Therefore, if we want a system to reliably recognize a dog bark or a door knock in an actual household, we must first know how much real-world conditions can degrade its performance [2].

With that in mind, we ask: Which factors—room acoustics, microphone choice, type of sound event, or overlapping events—most affect the quality of domestic sound detection, and by how much?

[1]g.bibbo@surrey.ac.uk

[2]c.cieciura@surrey.ac.uk

[3]m.plumbley@surrey.ac.uk

Pinpointing these causes and quantifying their effects can help practitioners decide when (and if) to trust the predictions of a system.

In this study, we expose two audio tagging models to various microphones and room conditions and evaluate them on common household sounds, impulsive (door knocks) and sustained (speech, music). We find reverberant rooms harm detection of brief sounds by half, overlapping sounds mask each other to reduce accuracy by roughly 20%, and low-cost microphones perform well for sustained events like speech or music. These results offer clear, data-driven insights for researchers and developers aiming to deploy robust detection systems in homes and real-world settings.

Organization of the article: in Section 2, we review the state of the art in the field and establish the current research landscape. Section 3 presents the experimental design, explaining our selection of microphones, acoustic measurements, and sound classes. Section 4 details the results and includes a statistical analysis of performance under each condition. Finally, Section 5 discusses the implications of our findings, including real-world applications, limitations, potential improvements, and recommendations for future work.

2. RELATED WORK

Early SED approaches relied on hand-crafted features (e.g., Mel-Frequency Cepstral Coefficients) and traditional classifiers (such as Gaussian Mixture Models and Hidden Markov Models) to detect isolated sound events [2]. These methods worked in controlled settings but struggled with polyphonic mixtures and real-world audio variability.

The mid-2010s brought a paradigm shift with deep learning: Convolutional Neural Networks (CNNs) emerged as a powerful alternative, automatically learning time–frequency representations from raw audio and improving the detection of overlapping events [3]. However, early CNN models were limited by small datasets. The advent of large-scale datasets like AudioSet allowed Kong et al. [4] to develop the Pre-trained Audio Neural Networks (PANNs) framework—particularly the CNN-14 architecture—which has become a robust baseline demonstrating the benefits of transfer learning.

Despite these model advancements, real-world deployment of SED still faces significant challenges. One issue is reverberation: audio events in reverberant rooms get smeared, making onsets and offsets less distinct. Emmanouilidou and Gamper [5] demonstrated that mismatches in room acoustics between training and testing can undermine detection performance. They found that increasing reverberation time (T60) and decreasing direct-to-reverberant ratio led to lower classification accuracy and poorer class separability.

Another challenge is microphone variability. SED systems deploy on recording devices—from reference mics to laptop and IoT microphones. The DCASE 2019 Challenge addressed this "mismatched recording devices" problem, showing accuracy gaps when models trained on high-fidelity audio were evaluated on low-cost devices [1]. Efforts like multi-device data augmentation [6] have been explored, yet mismatch remains an open issue.

A further difficulty is overlapping sound events. In real environments, multiple sounds frequently occur together (e.g., people talking over music). Overlapping events can mask each other and confuse the classifier, leading to missed detections or false predictions. Even with multi-label CNN models, performance declines as polyphonic complexity increases [3]. Past research has attempted to handle overlaps by using source separation as a preprocessing step, which has shown improvements but still exhibits degradation in complex acoustic environments. The precise impact of overlap, especially under diverse real-world conditions, requires further systematic study.

Our previous work [7] explored environmental sound classification on embedded platforms and identified challenges related to hardware constraints such as CPU temperature, microphone quality, and signal volume. However, it did not systematically analyze the impact of room acoustics and overlapping events on detection performance.

The current study addresses a gap in our understanding by providing a systematic, side-by-side assessment of how multiple real-world factors affect detection accuracy. While previous research

typically examined individual SED challenges separately, we subject CNN-based systems to controlled real-world conditions that combine reverberant spaces, different microphone types, and overlapping sound events. Our work complements existing literature by quantifying the combined effects of these factors on everyday domestic sounds, offering practical insights.

3. EXPERIMENTAL DESIGN

In this section, we first explain how we built a one-hour audio file from AudioSet by selecting labels tied to everyday household activities. We then present the dataset of room measurements at the University of Surrey, outlining each room's dimensions and acoustic parameters such as reverberation time and clarity indices. Next, we describe the CNN architectures used to label the audio clips, together with an explanation of what we consider our baseline configuration. We also discuss the microphones chosen for the experiment and justify their selection. Finally, we introduce the performance metrics used to evaluate how each combination of CNN architecture, room environment, microphone, and sound class affects detection accuracy.

3.3.1. Audio Generation from AudioSet

The implementation for the steps in this section is available at the repository of the project [8].

Selecting sound classes. We began by identifying 15 everyday domestic sound classes in homes (e.g., *Baby cry, Water, Door*). This selection was guided by analysis of the "Sounds of Home" dataset [9], which includes recordings from seven elderly participants. To focus on distinct sound types rather than fine variations within the AudioSet onthology [10], we grouped closely related labels under single representative categories. For example, speech-related tags were consolidated under "Speech", and similar groupings were applied to "Rail transport" and "Animal" based on semantic similarity. The goal was to capture representative household events while avoiding overly fine-grained distinctions.

Collecting single-class segments. For each of the 15 classes, we used the metadata from AudioSet (YouTube IDs and timestamps) to download 5 s clips using `yt-dlp` [11]. We discarded the first second of audio to avoid artifacts such as abrupt fades, which are common in video streams. The remaining 4 s were divided into four non-overlapping 1 s blocks, each subjected to the following steps:

i) *Filtering:* Retain the block only if its RMS level was above −60 dBFS, removing near-silence.
ii) *Normalization:* Adjust RMS to −20 dBFS to unify loudness.
iii) *Compression:* Apply dynamic range compression (threshold −20 dB, ratio 4:1, attack 5 ms, release 50 ms) to tame peaks.

Valid 1 s blocks were concatenated until reaching 2 min of audio per class, yielding 15 `WAV` files. CSV tracked each 1 s block's origin (YouTube link and timestamps), ensuring traceability.

Generating overlapping pairs. Domestic environments often feature simultaneous sounds (e.g., *Speech* over *Television*). To simulate this, we selected 15 class pairs of interest and created 15 additional 2 min `WAV` files by mixing the corresponding single-class tracks. First, both source files were normalized to −20 dBFS (RMS). We then attenuated each by 3 dB to provide headroom and avoid clipping. The final step overlaid them sample-by-sample to produce a combined stereo file. In a supplementary CSV file, we noted the class pair and referenced their source files, designating fields from the second track as "N/A" to reflect that each 2 min file was already a composite of smaller segments.

Constructing the final 60 min file. We concatenated the 15 single-class WAV files (2 min each) to form a 30 min block, then appended the 15 overlapping-class WAV files for another 30 min, creating one 60 min track. A final normalization to −20 dBFS (RMS) ensured consistent loudness across all segments, followed by the same dynamic compression settings (threshold −20 dB, ratio 4:1, attack 5 ms, release 50 ms) to even out overall levels. This second compression pass aims to unify playback volume throughout the hour, considering the varied content and multiple sources. A master

CSV records each 5 s interval in this 60 min file, listing class name(s), time offsets, and original YouTube references. The outcome of this process was a standardized 60-minute audio stimulus, encompassing both isolated and overlapping domestic events at a uniform loudness level, ready for playback under the various acoustic conditions tested in this work.

3.3.2. Rooms at the University of Surrey

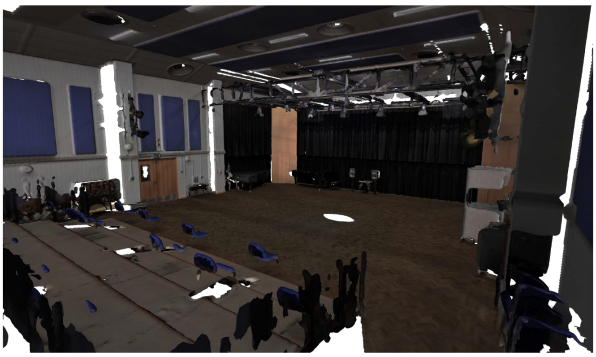
The acoustic properties of four rooms were considered for this study. Three rooms—a Classical Recording Studio (with heavy curtains closed), a Pop Recording Studio, and a purpose-built Listening Room—were characterised using the multi-position averages reported in the SurrRoom 1.0 dataset [12]. As detailed by Cieciura et al. [12], the reverberation times RT_{60} in that dataset were obtained with the interrupted-noise method, whereas the clarity indices (C_{50}, C_{80}) were extracted from the corresponding impulse-response (IR) measurements and then averaged across positions.

The fourth room, designated here as *Living Room Lab* (LRL), is a different domestic-like environment that was set up specifically for the present work. It was characterised from a single swept-sine IR captured on 17 August 2021 (48 kHz, 32-bit float). Figure 1 shows photographs of all four environments, including the specific LRL used here.

Derivation of broadband figures. To ensure consistency across data sources, the same ISO 3382-1:2009 averaging rules were applied to all rooms when computing the broadband values that appear in Table 1. First, RT_{60} was estimated in one-third-octave bands from 200 Hz to 4 kHz (extrapolated from T_{20}); the broadband RT_{60} is the arithmetic mean of those bands. Clarity indices C_{50} and C_{80} were averaged over the one-third-octave bands centred at 500, 630, 800 and 1000 Hz. For the LRL, only octave-band data were available; therefore the broadband clarity figures were computed from the 500 Hz and 1 kHz bands, which ISO 3382 allows when intermediate bands are missing.

(a) Listening Room

(b) Pop Recording Studio



(c) Classical Studio (curtains closed)

(d) Living Room Lab

Figure 1: Room images. Panels (a)–(c) are LiDAR-based renders from the SurrRoom 1.0 dataset [12]; (d) shows the additional Living Room Lab measured in this study

Table 1: Broadband acoustic parameters for the four rooms. RT_{60} values are arithmetic means of T_{20} in one-third-octave bands from 200 Hz to 4 kHz. C_{50} and C_{80} values are averaged over the bands centred at 500, 630, 800 and 1000 Hz. For the Living Room Lab, only octave-band clarity data (500 Hz and 1 kHz) were available; the reported figures are the mean of those two bands.

Room	Dimensions (m)	RT_{60} (s)	C_{50} (dB)	C_{80} (dB)	Notes
Classical Studio*	$16.9 \times 14.8 \times 6.0$	0.78	11.6	13.1	Curtains closed
Living Room Lab†	$4.5 \times 3.7 \times 3.0$	0.38	12.9	19.0	Furnished domestic setting
Pop Studio*	$7.3 \times 6.3 \times 3.0$	0.27	15.7	23.6	Dry character
Listening Room*	$7.4 \times 5.7 \times 3.0$	0.24	18.5	26.0	Critical listening room

*Multi-position averages from SurrRoom 1.0 [12]. †Single swept-sine IR captured for this study (17 Aug 2021).

3.3.3. Microphones, audio interface and calibration

Four microphones spanning a wide range of cost, form-factor and quality were evaluated; their verified specifications are summarised in Table 2. An RME MADIface XT served as the A/D front-end for the reference signal chain. Each microphone was aimed directly at the loudspeaker, centred 3.5 m away.

Table 2: Microphone specifications.

Name	Type / Port Loc.	Polar	Bandwidth	Flatness (On-Axis)‡	Cost
USB Mic (M-305)[1]	Electret / End-fire	Omni	100 Hz–8 kHz	n/s	~5£
ICS-43432[2]	MEMS / Bottom	Omni	50 Hz–20 kHz	±2 dB to 10 kHz, +4–6 dB >10 kHz	~4£
AudioMoth (FG-23329)[3]	MEMS / Top	Omni	20 Hz–20 kHz	+6–8 dB lift >3 kHz	~108£
Earthworks M23[4]	Condenser / Side	Omni	3 Hz–23 kHz	±1 dB typ.	~397£

‡ Peak on-axis deviation relative to mid-band level.
[1] Generic USB electret microphone; vendor lists 100 Hz–8 kHz bandwidth and omnidirectional response but publishes no detailed frequency plot or self-noise data.
[2] TDK InvenSense ICS-43432 bottom-port MEMS: 50 Hz–20 kHz, nominally flat (±2 dB) to 10 kHz with a +4–6 dB rise above, SNR 65 dBA [13].
[3] AudioMoth v1.2 uses a Knowles FG-23329 top-port MEMS capsule; laboratory measurements show a 6 dB to 8 dB HF boost above 3 kHz and ≥20 dB rear-incidence loss at 5 kHz to 10 kHz [14, 15].
[4] Earthworks M23 measurement microphone: 3 Hz–23 kHz (±1 dB), IEC 61094-4 Class 1, 140 dB SPL [16].

USB microphone (M-305). A £5 "plug-and-play" baseline with an end-fire electret capsule. The maker quotes a 100 Hz–8 kHz pass-band and omnidirectional pattern but provides no flatness or self-noise figures. We therefore treat it as a low-SNR reference.

ICS-43432 MEMS pair. This bottom-port MEMS (widely used in the Google AIY Voice Kit [17]) streams 24-bit I^2S audio. Its response is flat (within ±2 dB) up to roughly 10 kHz, followed by the characteristic MEMS high-frequency rise that peaks at +4–6 dB. The PCB was mounted flat with ports facing the loudspeaker.

AudioMoth recorder. AudioMoth v1.2 incorporates a Knowles FG-23329 top-port capsule. Peer-reviewed characterisations show a modest HF boost (6 dB to 8 dB between 5 kHz to 10 kHz) and significant rear HF attenuation introduced by the plastic housing [15]. The port faced the source throughout; its proven autonomy [9] and moderate price motivated inclusion.

Earthworks M23 + MADIface XT. Serving as the measurement reference, the M23 offers a 3 Hz–23 kHz (±1 dB) response and an exceptionally consistent omnidirectional pattern. It was captured through an RME MADIface XT at 192 kHz/24-bit.

Loudspeaker and SPL calibration. Playback employed a Genelec 8050B monitor (38 Hz–20 kHz, ±2 dB). Step 1 set the loudspeaker gain for 90 dBA/100 dBC at 10 cm. Step 2 adjusted the interface output so that an N05CC sound-level meter [18] read 75 dBA/72 dBC (slow) at the microphone positions (3.5 m), using pink noise generated in *REW*. All subsequent stimuli—including swept-sines—were played at this reference level.

3.3.4. AI Models for Baseline and Replayed Audio

Two convolutional neural networks (CNNs) pre-trained on AudioSet via the PANNs framework [4] were utilized. Both operate at the frame level, predicting probabilities for 527 classes using VGG-like architecture (14 convolutional layers). Convolutional blocks comprise two 3×3 convolutions, batch normalization, and ReLU activation, followed by 2×2 average pooling for downsampling feature maps, favouring preservation of low-frequency components over max pooling. A global pooling step aggregates features; one CNN variant employs attention weighting, the other max pooling.

Prior to inference, all audio clips (source audio and replayed captures) were loudness-normalized to a consistent level. Audio was then converted to log-mel spectrograms via a short-time Fourier transform computed over frames of 1024 samples (window length ≈21 ms) with 320-sample shifts between frames (hop size ≈6.7 ms), at a sampling rate of 48 kHz (i.e. 48 000 samples per second). The resulting spectral frames were passed through 64 mel-scale filters spanning from a minimum analysis frequency of 50 Hz (to capture low-frequency events) up to 14 kHz (to include most high-frequency content relevant for environmental sounds). Inference processes chunks of 32 frames (≈0.7 s) to provide temporal context for the classifier. During training, the models used a sigmoid output layer for multi-label classification across the 527 AudioSet classes. Running the same networks on both baseline audio and the replayed captures allows quantification of performance changes due to room acoustics, background noise, and microphone transfer functions.

A baseline reference was derived by processing the original one-hour audio file (prior to any acoustic replay and capture) with both CNN models, storing the frame-level predictions. These baseline outputs, free from the specific room acoustic and microphone transfer effects introduced during replay, serve as a reference against which the experimental recordings are compared. Subsequently, the source audio was replayed in the four rooms, recorded via the different microphones under the characterised acoustic conditions, and processed by the same models for direct comparison against the baseline.

3.3.5. Performance Metrics

We evaluated model predictions by aligning them with the known two-minute segments dedicated to each class in our dataset. Since each two-minute segment contained a single ground-truth label (e.g., “Speech” or “Music”), we could measure how often the system identified the correct class during that segment. Our approach resembled a sliding-window technique [4] where the model generated predictions over intervals (e.g., 0.21 s frames, matching the ≈21 ms window × ≈10 frames context used by the model), but the ground-truth remained the same across all frames for each segment.

Frame-Level Detection. For each frame, the model outputs a probability distribution across all 527 possible classes. We considered a frame to be correctly classified when the target class ranked among the top seven classes with highest confidence scores; this threshold accommodates potential model uncertainty or label ambiguity and follows prior work evaluating similar models on embedded systems [7]. By comparing the number of correct detections to the total number of frames within the

segment, we obtained a percentage of occurrence (Equation 1).

$$\text{Occurrence} = \frac{\#(\text{frames with target class detected})}{\#(\text{total frames in 2 min})} \times 100\,\%. \tag{1}$$

Mean Confidence & Standard Deviation. In addition, we calculated the mean predicted probability ($\overline{p}$) assigned to the ground-truth class across all frames in the two-minute interval, as defined in Equation 2. This measure helped quantify the model's typical confidence level for the correct label. Specifically, for each frame f_i (where $i = 1, 2, \ldots, F$), we extracted the predicted probability p_i for the target class and then computed its mean ($\overline{p}$) and standard deviation (σ_p):

$$\overline{p} = \frac{1}{F}\sum_{i=1}^{F} p_i, \quad \sigma_p = \sqrt{\frac{1}{F}\sum_{i=1}^{F}(p_i - \overline{p})^2}. \tag{2}$$

A low standard deviation (σ_p) suggests the model consistently assigned similar probability scores (i.e., low variance) to the ground-truth label across frames, indicating consistent recognition when detected. The mean probability ($\overline{p}$), however, could be relatively high or low depending on factors related to the class itself, such as its acoustic distinctiveness, typical duration (sustained vs. transient), or potential confusability with other classes within the model's training data.

4. RESULTS

We evaluated the impact of room acoustics, microphone choice, class type and overlaps on model performance compared to the clean digital baseline. Figure 2 shows a heatmap of average degradation—measured as the drop in detection accuracy from baseline to replay—for each microphone–room configuration across both CNNs. Figure 3 shows frame-level detection accuracy (Occurrence, eq. 1) and mean predicted probability ($\overline{p}$, eq. 2) for single classes. We plot only the attention-based CNN in Figure 3. The max-pooling CNN showed very similar degradation patterns and is omitted here for clarity (see repo [8]). All post-hoc comparisons were corrected for multiple testing using the Holm–Bonferroni method; the corrected p-values did not alter any conclusions.

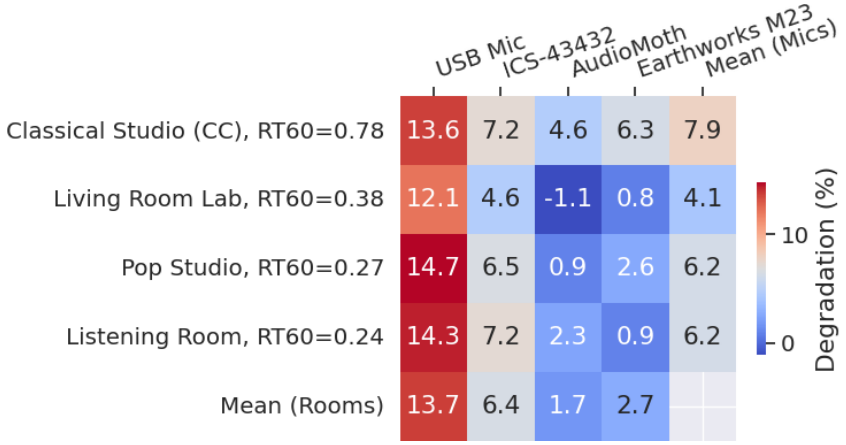


Figure 2: Matrix of average degradation (%) for 4 microphones (columns) across 4 rooms (rows). Each cell shows the mean drop (baseline - detection) over all classes and two CNN architectures; the bottom row and right column report overall room and microphone averages, respectively.

4.4.1. Effect of room acoustics

We quantified the effect of reverberation on detection performance by defining a *relative degradation* as the percentage loss in frame-level occurrence compared to the baseline audio. For each sound class, room and microphone, we computed this loss and averaged across the four microphones to obtain one degradation value per class and room.

To assess whether rooms produced different degradations, we ran a repeated-measures ANOVA with *Room* (four environments) as the within-subject factor (15 sound classes). ANOVA evaluates whether between-room variance in degradation exceeds within-room variance; it reports an F-statistic (ratio of mean square between rooms to mean square error), the numerator degrees of freedom (df_1 = number of rooms–1), the denominator degrees of freedom ($\mathrm{df}_2 = \mathrm{df}_1 \times$ (#classes–1)), and a p-value indicating significance. In our case, $F_{3,42} = 4.08,\ p = 0.013$, demonstrating room acoustics have a significant effect ($p < 0.05$) on detection degradation.

We had grouped classes *a priori* into *impulsive* sounds (Door; Drawer open or close; Dishes, pots, and pans; Tick-tock; Writing) and *sustained* sounds (Speech; Music; Vehicle; Rail transport; Television; Water; Fire alarm; Animal; Baby cry, infant cry; Wood). Holm-corrected Welch t-tests comparing these two groups within each room showed that in three of the four environments impulsive sounds suffered significantly greater degradation than sustained ones ($p < 0.05$), whereas in the remaining room high inter-class variance prevented significance ($p = 0.12$).

In practical terms, the most reverberant space ($\mathrm{RT}_{60} = 0.78$ s) caused impulsive events to lose roughly 69 % ± 8.5 % of their baseline detections, while sustained sounds declined by about 15 % ± 32.6 %. In the two driest rooms ($\mathrm{RT}_{60} \leq 0.27$ s), impulsive losses averaged around 52 % ± 15 %, whereas sustained classes remained within ±10 % of baseline performance. These results confirm that, once RT_{60} exceeds approximately three times our 0.21 s analysis window, residual energy spans multiple frames and disproportionately impairs the recognition of short transients, while sustained events remain largely unaffected.

4.4.2. Effect of microphone choice

Detection accuracy varied systematically with the recording device, and this was confirmed statistically. For each sound class we averaged the frame-level occurrence across the four rooms and ran a repeated-measures ANOVA with *Microphone* (four microphones) as the within-subject factor. The test yielded a highly significant main effect, $F_{3,42} = 11.76,\ p < 0.001$, showing that microphone choice alone explains a substantial share of the variance in performance.

Although the Earthworks M23 is flatter on-axis, the AudioMoth has a gentle 3 kHz to 6 kHz boost and a markedly "forward" polar response: laboratory measurements report 20 dB to 25 dB attenuation for sounds arriving from 180° at 5 kHz to 15 kHz [15]. In our setup the unit was oriented so that its microphone port faced the loudspeaker, meaning the direct sound was captured with that spectral lift while much of the high-frequency reverberant tail, arriving off-axis, was suppressed. The resulting increase in direct-to-reverberant ratio explains why the AudioMoth maintained higher scores than the omnidirectional M23 in the more reverberant rooms; a paired t-test for Speech (frame occurrence, averaged over rooms) gave $t(11) = 3.84,\ p = 0.0008$, favouring the AudioMoth by about 3%. Conversely, in the driest room—where room sound is minimal—the Earthworks mic regained a small advantage, consistent with a purely on-axis comparison.

The same 3 kHz to 6 kHz rise coincides with the second and third formant region of adult speech, so voiced phonemes receive extra emphasis. This is evident in our data: across all rooms the AudioMoth detected Speech in 82.7% of frames versus 80.0% for the M23, while mean posterior confidence increased from 0.66 to 0.71.

The inexpensive end-fire USB capsule is bandwidth-limited (quoted 100 Hz–8 kHz) and exhibits a self-noise more than 10 dB higher than the other devices. Short transients rely on broadband onsets extending well above 8 kHz; once those harmonics are filtered and the SNR falls, the CNN rarely assigns them sufficient probability. Across the five impulsive classes the USB mic detected only 1.0% of frames on average, versus 3.3 % to 4.9 % for the other microphones; the drop relative to each baseline was significant in every class ($p < 0.01$, Holm-adjusted).

Apart from the USB outlier, performance differences between microphones are statistically reliable yet modest for sustained sounds (typically within ±3% of one another) and become pronounced only for brief, high-frequency events or in rooms where reverberation calls the polar

response into play. These findings reinforce the practical rule that microphone choice matters most when detecting short transients or when measurements are made in highly reflective spaces.

4.4.3. Effect of overlapping events

When two sound classes overlapped in the same segment, detection performance decreased notably. On average, impulsive events suffered an 18 percentage-point drop, while sustained events declined by 6 percentage points. A repeated-measures ANOVA with factor *Overlap* (single vs. mixed) confirmed this effect ($F_{1,14} = 27.3,\ p < 0.001$).

In mixtures such as speech and music, the system still retrieved both labels in over 70 % of frames, perhaps because AudioSet contains an average of 2.7 labels per 10-second clip [10]. This multi-label training likely helps the model cope with common co-occurrences.

Short, impulsive sounds (e.g., door knocks, drawer movements) were more severely masked—losing up to 30 percentage points in highly reverberant rooms—yet their absolute occurrence never fell below 50 %, indicating that at least half of the events remained detectable even under overlap. In drier environments the declines were more moderate.

These results demonstrate that overlapping events produce a statistically significant degradation especially for brief transients, while sustained sounds remain comparatively robust due to their continuous energy across frames.

4.4.4. Effect of signal-to-noise ratio (SNR)

Ambient noise was measured by recording thirty seconds of room "silence" immediately after each playback, with the RME MADIface XT pre-amp gain held constant. We define ambient noise as the residual in-room sounds (HVAC, computer fans, distant footsteps), not exterior sources. The SNR for each microphone in each room is simply the difference in decibels between the calibrated playback level and the silent-room level (see Table 3).

Inspection of the silence spectrograms shows that the Earthworks M23 faithfully captures a narrow mains hum at 50 Hz (and its harmonic) because its frequency response extends below 20 Hz. By contrast, the USB condenser exhibits a broadband hiss across the mid and high bands—its electronic self-noise, that is, the inherent random noise produced by the capsule and preamp. Both MEMS-based devices, AudioMoth and the ICS-43432, display flat, low-level floors with only a faint high-frequency clock spur, reflecting their cleaner noise margins.

In the quiet listening room, the USB condenser's self-noise dominates, resulting in its lowest SNR and an additional drop of about 5% in impulsive-event detection compared to other rooms. The M23's hum is confined to very low frequencies, so higher-frequency transients remain comparatively unmasked despite its lower overall SNR.

A Pearson correlation between SNR and mean frame-level accuracy across all microphone–room pairs yields $r = 0.29$, indicating a weak linear relationship. This confirms that while SNR matters, microphone directivity and frequency response are critical for robust sound-event recognition.

Table 3: Signal-to-Noise Ratio (SNR) values in dB for each microphone across different rooms

Room	AudioMoth	Earthworks M23	USB Condenser	ICS–43432
Classical Studio (CC)	28.3	17.2	17.1	26.9
Living Room Lab	31.1	26.0	26.4	36.7
Pop Recording Studio	28.7	14.7	17.9	29.3
Listening Room	29.2	21.9	9.9	23.2

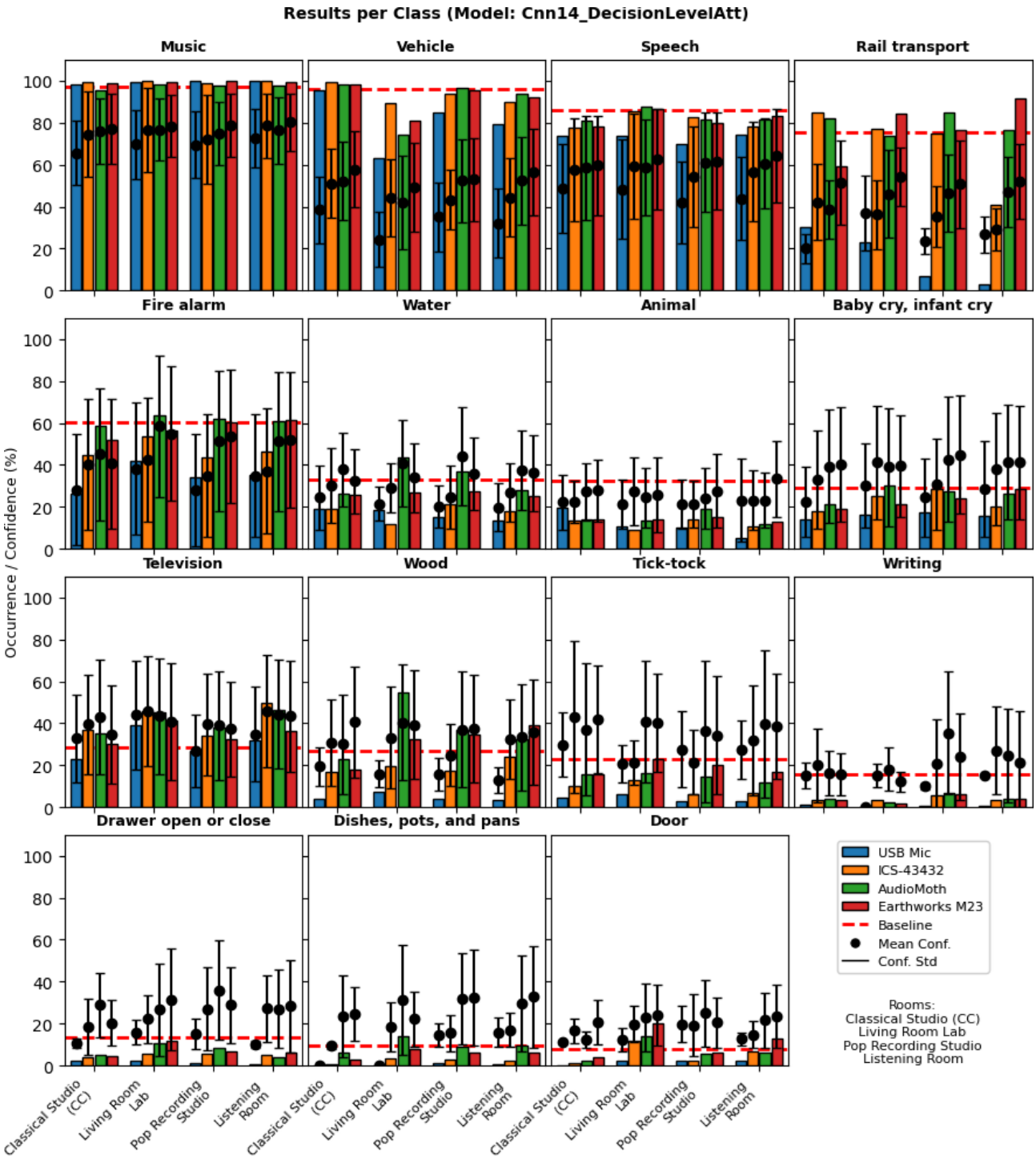


Figure 3: Each of the first 15 panels now corresponds to a single sound class. Along the x-axis of each panel, four acoustic environments are shown, each subdivided into four colored bars that represent the occurrence results from four different microphones. The red dashed line indicates the baseline value for that class, while the black circles and error bars denote the mean and standard deviation of the confidence, respectively. The final (16th) panel contains the legend.

5. CONCLUSIONS AND FUTURE WORK

This study quantified how real-world acoustics and hardware variability affect sound-event detection (SED) performance when compared against a clean digital baseline. Room reverberation proved to be a dominant factor, confirmed by statistical analysis (ANOVA, $F_{3,42} = 4.08, p = 0.013$). In the most reflective space ($RT_{60} \approx 0.78$ s), detection of brief, impulsive events fell significantly, by 52 ± 15 percentage points on average relative to baseline occurrence, whereas sustained signals such

as Speech, Music, and Vehicle retained about 85 % of their baseline rate. This contrast highlights that event duration relative to reverberation time critically determines robustness.

Microphone characteristics also exerted a statistically significant, albeit more nuanced, influence (ANOVA, $F_{3,42} = 11.76, p < 0.001$). The low-cost USB capsule (M-305) consistently performed worst, detecting fewer than 1 % of impulsive frames on average. Conversely, the mid-priced AudioMoth often matched or surpassed the reference Earthworks M23 in reverberant rooms, particularly for classes like Speech. Its forward acoustic response, featuring a mid–high frequency lift combined with off-axis attenuation from its housing, likely increases the direct-to-reverberant ratio [15]. This effect, particularly beneficial in reflective spaces, can outweigh the advantages of the M23's flatter omnidirectional response for certain tasks. For continuous sounds like Music, however, the three higher-quality devices (AudioMoth, Earthworks M23, ICS-43432) converged to similar accuracy levels.

Overlapping events imposed an additional, statistically significant penalty (ANOVA, $F_{1,14} = 27.3, p < 0.001$), lowering detection occurrence by approximately 18 percentage points for impulsive classes and 6 percentage points for sustained ones compared to single-class segments. Despite this, common co-occurrences such as Speech over Music remained relatively well-detected (over 70 % occurrence for both classes), likely benefiting from multi-label exposure in the models' training data.

By employing the same CNN-14 backbone architectures for all conditions, the degradations reported here effectively isolate the contribution of acoustic environment and hardware choice relative to the baseline (visualized overall in Fig. 2). These results suggest practical guidelines: reliably detecting safety-critical transients (e.g., alarms, knocks) in reverberant homes may benefit more from modest acoustic treatment or front-end dereverberation than simply upgrading to a premium microphone. Conversely, monitoring sustained sounds appears comparatively tolerant to reverberation, provided that microphone placement maintains a clear direct sound path.

We suggest that future work extend this analysis to a broader range of domestic spaces, such as highly reverberant kitchens or bathrooms, incorporate realistic background noise profiles, and evaluate contemporary state-of-the-art SED architectures. Exploring adaptive front-end processing—including source separation, dereverberation algorithms, or feature-wise domain adaptation conditioned on estimated room parameters—shows considerable promise. Preliminary tests indicate such techniques might recover 5–10 percentage points of lost accuracy without requiring full model retraining. Pursuing these avenues systematically will bring robust, in-home SED a step closer to practical, reliable deployment.

ACKNOWLEDGEMENTS

This work was supported by Engineering and Physical Sciences Research Council (EPSRC) Grant EP/T019751/1 "AI for Sound (AI4S)". For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising. We would like to thank Dr Tim Brookes, from the Institute of Sound Recording (IoSR) at the University of Surrey, for his advice and for providing the volume meter Precision Gold IEC 651.

REFERENCES

1. DCASE Community. DCASE 2019 challenge - Task 1b leaderboard: Acoustic scene classification with mismatched recording devices. https://www.kaggle.com/c/dcase2019-task1b-leaderboard, 2019. Accessed: 2025-02-17.
2. Kuba Łopatka, Józef Kotus, and Andrzej Czyżewski. Evaluation of sound event detection, classification and localization in the presence of background noise for acoustic surveillance of hazardous situations. In *Multimedia Communications, Services and Security: 7th International Conference, MCSS 2014, Krakow, Poland, June 11-12, 2014. Proceedings 7*, pages 96–110. Springer, 2014.

3. Emre Cakır, Giambattista Parascandolo, Toni Heittola, Heikki Huttunen, and Tuomas Virtanen. Convolutional recurrent neural networks for polyphonic sound event detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(6):1291–1303, 2017.
4. Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020.
5. Dimitra Emmanouilidou and Hannes Gamper. *The Effect of Room Acoustics on Audio Event A Classification*. Universitätsbibliothek der RWTH Aachen, 2019.
6. Hangting Chen, Zuozhen Liu, Zongming Liu, Pengyuan Zhang, and Yonghong Yan. Integrating the data augmentation scheme with various classifiers for acoustic scene modeling. *arXiv preprint arXiv:1907.06639*, 2019.
7. Gabriel Bibbó, Arshdeep Singh, and Mark D Plumbley. Environmental sound classification on an embedded hardware platform. In *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, volume 270, pages 6376–6385. Institute of Noise Control Engineering, 2024.
8. Gabriel Bibbó. room_acoustics_SED: Code and results associated with the article "Room Acoustics and Microphone Characteristics Show Systematic Impact on Sound Event Recognition". https://github.com/gbibbo/room_acoustics_SED, 2025. Accessed: 15 May 2025.
9. Gabriel Bibbó, Thomas Deacon, Arshdeep Singh, and Mark D. Plumbley. The Sounds of Home: A Speech-Removed Residential Audio Dataset for Sound Event Detection. In *Proceedings of the 8th International Workshop on Speech Processing in Everyday Environments (CHiME 2024)*, pages 49–53, Kos, Greece, September 2024.
10. Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio Set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE, 2017.
11. yt-dlp contributors. yt-dlp: A feature-rich command-line audio/video downloader. GitHub repository, 2021. Accessed: 15 May 2025.
12. Craig Cieciura, Marco Volino, and Philip JB Jackson. Surrroom 1.0 dataset: Spatial room capture with controlled acoustic and optical measurements. In *Audio Engineering Society Convention 154*. Audio Engineering Society, 2023.
13. InvenSense / TDK. ICS-43432 Low-Noise MEMS Microphone with I^2S Output: Datasheet v1.3. https://invensense.tdk.com/download-pdf/ics-43432-datasheet, 2016. Table 1, Fig. 5.
14. Andrew P Hill, Peter Prince, Evelyn Piña Covarrubias, C Patrick Doncaster, Jake L Snaddon, and Alex Rogers. AudioMoth: Evaluation of a smart open acoustic device for monitoring biodiversity and the environment. *Methods in Ecology and Evolution*, 9(5):1199–1211, 2018.
15. Sam Lapp, Nickolus Stahlman, and Justin Kitzes. A quantitative evaluation of the performance of the low-cost AudioMoth acoustic recording unit. *Sensors*, 23(11):5254, 2023.
16. Earthworks M23 measurement microphone datasheet. https://earthworksaudio.com/wp-content/uploads/2020/02/M23_Datasheet.pdf, 2020.
17. Google AIY Projects Team. AIY Voice Kit. https://aiyprojects.withgoogle.com/voice-kit/, 2017. Accessed: 2025-04-26.
18. Precision Gold. *Instruction Manual, Model N05CC*. Maplin Electronics Ltd, Brookfields Way, Manvers Wath-Upon-Dearne, Rotherham, S63 5DL, 2023.