INTRODUCTION
• Sound event classification (SEC) is a task that automatically categorizes audio clips into labels that match their acoustic content.
• It can be used in abnormal event detection like a surveillance system, and sound recognition on edge devices like AI speakers or mobile devices, which usually have real-world scenarios.
• A well-trained SEC model breaks easily in the real-world scenario.
• Reverberation is one of the major reasons for performance degradation in the real-world.
• In this research, we experimentally verify performance degradation of the SEC for reverberant environments, through various reverberation conditions.
• Then we propose a performance enhancement technique, which utilizes room adaptive information, which is room impulse response (RIR).

BACKGROUND
Room Impulse Response (RIR)
• Shows the complete acoustic path of source sound with room reverberation under the LTI condition.
• Source audio is distorted more as reverberation time ($T_{60}$) increases and direct-to-reverberation ratio (DRR) decreases.
• RIR can be easily acquired on the edge devices like AI speakers by simple clapping or testing with sine sweep.

PROPOSED METHOD
Fig 1. The proposed method architecture
Room Impulse Response (RIR) Embedding block Input audio Embedding block Feature-wise transformation Classification Output
• The proposed method is a conditioning method that uses room adaptive information of the target room, which is RIR.
• We use two embedding blocks for scaling and biasing conditioning. Then the output is fed into the classification network.
• The proposed method can be attached to the conventional deep learning-based SEC models.

DATA SET
• The classification dataset : Real World Computing Partnership (RWCP) – 50 classes with 80 clips (total 4,000 clips).
• Clean test set : 20 clips of each classes of RWCP (total 1,000 clips).
• Simulated test set : Made by convolving real-world impulse response (IR) with Clean test set.
• Recorded test set : Re-recording the clean test set in 2 real-world reverberant environments (corridor and boardroom).

NETWORK ARCHITECTURE
• Multi-spectrogram layer: Convolutional layers are followed by max pooling layer.
• Embedding block: The RIR embedding block is applied to the input audio.

TRAINING STRATEGIES
• Base: The baseline model without the RIR embedding blocks. (Trained using original train set)
• Deconv: Same with Base, but at the inference time, deconvolve the test audio with the RIR of random points in the same room.
• Aug: Trained using an augmented train set and applied the proposed RIR conditioning method. The exact virtual RIR that convolved with the input audio is given as a RIR input pair

RESULTS
Fig 3. The results of each model in the original clean test set (Clean) and the simulated test set. (a) shows performance related to the T60 and (b) shows performance related to the DRR in the chosen six rooms.
• Reverberation significantly degrades the SEC performance and degradation is intensified as $T_{60}$ increases.
• Aug model works to some extent, but the proposed model shows a statistically significant additional performance improvement, especially in the room that has long $T_{60}$.
• The proposed method works for not only in the various $T_{60}$ environment but also various DRR environment.

CONCLUSION
• We experimentally verified the SEC’s performance degradation in reverberant environments through various reverb conditions ($T_{60}$ and DRR).
• We proposed the room adaptive conditioning methods which uses room impulse response (RIR) of the target room.
• We showed the proposed method tends to enhance performance with reverberation time-related information, which implies that only with the approximate RIR of the target room, our method still has benefits.