Room Adaptive Conditioning Method for Sound Event Classification in Reverberant Environments

PAPER NO.2573

Jaejun Lee[1], Donmoon Lee[1,2], Hyeong-Seok Choi[1], and Kyogu Lee[1]

Music and Audio Research Group, Seoul National University[1], Cochlear.ai[2]

jjlee0721@snu.ac.kr

SEOUL NATIONAL UNIVERSITY

MARG MUSIC & AUDIO RESEARCH GROUP

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, it utilizes room adaptive information, which is room impulse response (RIR). It is done by the feature-wise transformation conditioning method.

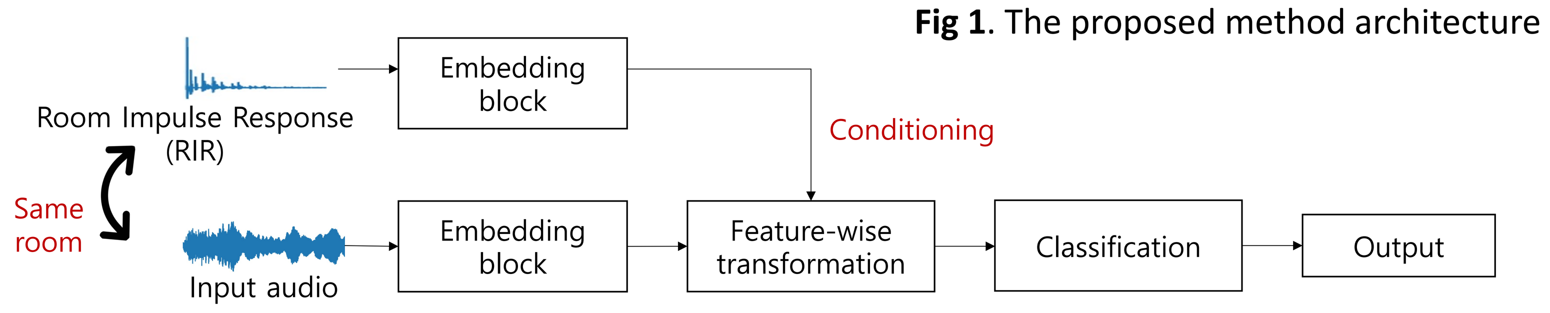
BACKGROUND

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



- The proposed method is a conditioning method that uses room adaptive information of the target room, which is RIR.

x : the input source audio
$H(x)$: the embedding block of x

r : the RIR of the room
$R_s(r)$: the (scaling) embedding block of r
$R_b(r)$: the (biasing) embedding block of r

$$y = \sigma\big(H(x) * R_s(r)\big) + R_b(r)$$

σ : non-linear activation
$*$: element-wise multiplication

- We use two embedding blocks for scaling and biasing conditioning. Then the output y is fed into the classification network.
- The proposed method can be attached to the conventional deep learning-based SEC models.

EXPERIMENT

Dataset

- 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)

Test set	RIR dataset	Room type	# of RIR	T_{60}(s)	Notation
Simulated Test set	AIR	Booth	12	0.27	R027
	WDR	CR7	360	0.29	R029
	AIR	Office	12	0.39	R039
	MARDY	-	73	0.55	R055
	AIR	Lecture	24	0.68	R068
	AIR	Stairway	78	0.77	R077
	QMUL	Classroom	130	0.134	R134
Recorded Test set	-	Corridor	1	0.20	Record1
	-	Boardroom	1	0.22	Record2

Tab 1. The specification of the simulated and recorded test sets.

Network Architecture

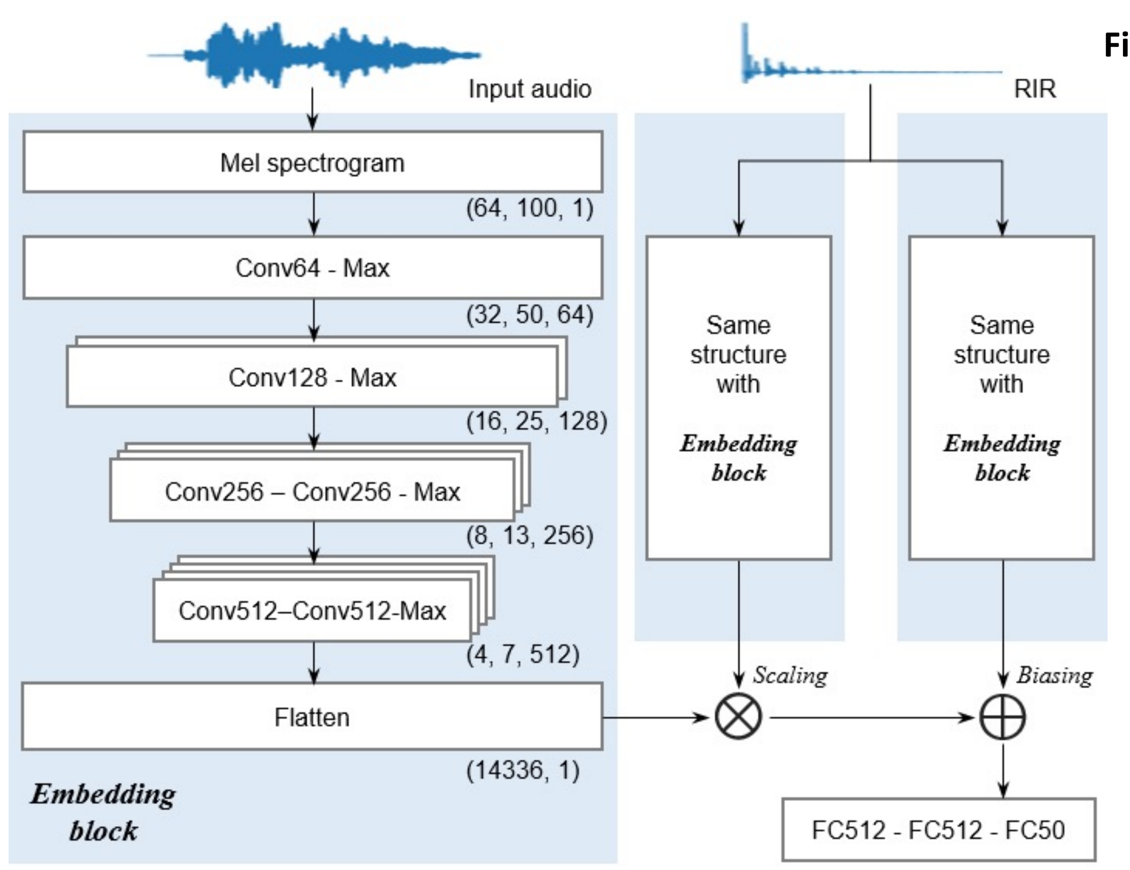


Fig 2. The networ architecture

Conv : convolution (3x3) filter size, (1,1) strides

Max : Max-pooling (2x2) pooling size, (2,2) strides

FC : Fully connected layer each with the filter dimension

Mel-spectrogram : 32 ms window length, 10 ms hop length, (64, 100) input size

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 that convolved random virtual RIR generated by image method[1,2].
- *Cndt* : 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

[1]Allen, Jont B., and David A. Berkley, 2017, [2]https://github.com/ehabets/RIR-Generator

RESULTS

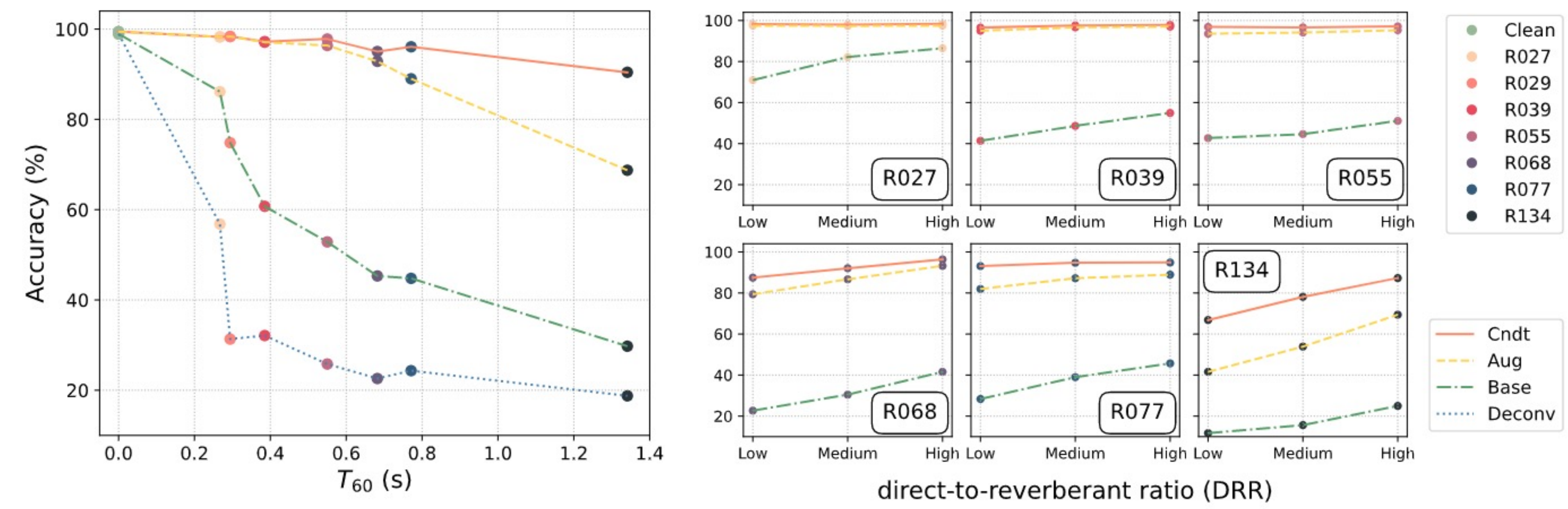


Fig 3. The results of each model in the original clean test set (Clean) and the simulated test sets. (a) shows performance related to the T60 and (b) shows performance related to the DRR in the chosen six rooms.

- 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.

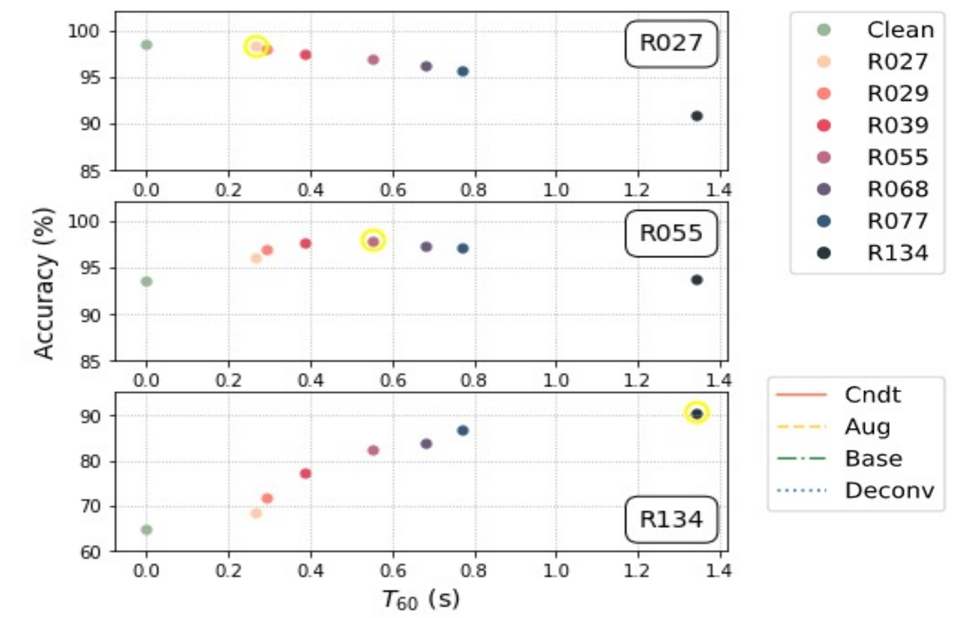


Fig 4. The results of the 'fake conditioning experiment' that gives RIR of other room rather than the same RIR with the input audio. It means that the pair of the input is mismatched intendedly.

Model	Clean	Record1	Record2
Base	98.92	56.17	58.67
Aug	99.42	78.95	72.99
Cndt	99.43	84.38	78.57

Tab 2. Classification accuracy (%) on the recorded test sets.

- If we condition with the RIR that have similar T_{60} of the room, it still works. However as the T_{60} gap between the input audio and conditioning RIR increases, the performance decreases. The proposed method tends to enhance performance with reverberation time-related information. (**Fig 4**)
- The proposed method works not only in the simulated en- vironments but also in the real-world reverberant environments. (**Tab2**)

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.