RAMDC: Room-Aware Multi-Device Clustering for Large Scale Teleconferencing

Yile (Angela) Zhang*§, Wei-Ting Lai*§, Amy Bastine*, Xingyu Chen†,
Lachlan I. Birnie*, Thushara D. Abhayapala*, and Prasanga N. Samarasinghe*,
* The Australian National University, Australia
Email: wei-ting.lai@anu.edu.au
† University of Technology Sydney, Australia

***Abstract*—In large-scale teleconferencing, many devices can share the same physical room even though each one joins as an independent participant. The resulting loudspeaker-microphone couplings can cause cross-device acoustic echo. To mitigate this issue, we introduce Room-Aware Multi-Device Clustering (RAMDC), an unsupervised initialization step at the start of a teleconference. During initialization, every device simultaneously plays a short chirp, and the aggregate responses are transformed into energy decay curves, embedded by a lightweight Conformer encoder, and clustered with DBSCAN to infer room membership. The resulting cluster labels let each device mute streams from co-located devices which can remove cross-device feedback paths. Evaluated on the GTU-RIR dataset, RAMDC outperforms alternatives in the clustering accuracy, and remains robust even in unseen rooms.**

I. Introduction

Acoustic echo cancellation (AEC) is a fundamental signal processing technique designed to suppress echoes resulting from the acoustic feedback loop between a loudspeaker and a microphone [1], [2]. In modern multi-user scenarios such as video conferencing or multiplayer gaming, it is common for multiple personal devices such as laptops, smartphones, or tablets to operate concurrently within the same space. However, conventional AEC algorithms, whether single-channel [3]–[6] or multichannel [7]–[11], typically assume that all echo paths are confined to individual devices. Although multichannel AEC methods leverage microphone arrays for improved echo suppression, their operation is limited to intra-device processing to cancel far-end echo and cannot address echo arising between devices.

A major challenge in multi-device acoustic environments arises when several devices within the same physical room capture and replay each other's audio. As a result, users perceive both the direct speech and multiple slightly delayed replays from nearby devices. These redundant re-broadcasts act as a form of echo, as users expect to hear only the direct speech. In severe cases, the cross-device feedback loop can induce high-pitched howling when the loop gain exceeds unity. However, conventional AEC algorithms are designed to cancel far-end echo and are not applicable to handle this type of near-end, cross-device interference. To the best of our knowledge, this problem remains largely unexplored, and no prior study has formally defined it. In this paper, we therefore introduce the term *intra-room echo* to describe this phenomenon and aim to propose a method to eliminate it.

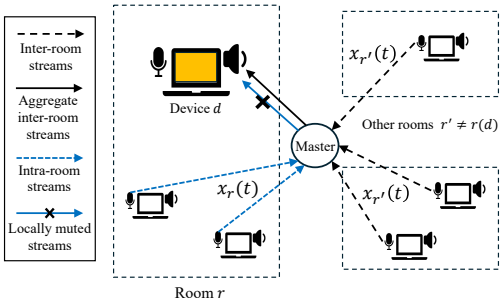

Fig. 1. Illustration of room-aware stream control from device d's perspective. In this example, $R = 3$ rooms contain a total of $D = 6$ devices. A central master controller manages stream aggregation and routing. During operation, the devices within room r only plays the aggregated inter-room signals $x_{r'}(t)$. The intra-room stream $x_r(t)$ is locally muted. Note that although the signal $x_r(t)$ is blocked from being played back within room r, it still contributes to the aggregated stream delivered to other rooms via the master.

One potential solution is to eliminate the problematic feedback paths at the stream routing stage, rather than using signal-processing techniques to remove mixed components from measurements. In a centralized server-based model, as adopted in many teleconferencing platforms, a master controller manages stream aggregation and routing across devices. This design allows each device to selectively mute audio playback from peers in the same physical room, thereby proactively mitigating *intra-room echo* feedback at the master level.

In this context, the core technical challenge shifts from echo cancellation to accurate room-level device clustering, allowing each device to identify its acoustic group and mute intra-room audio streams accordingly. A naive sequential probing scheme, where each device plays a test signal one by one to reveal its room membership, would scale linearly with the number of devices, and is therefore impractical. Thus, developing an efficient device clustering method is necessary.

Room classification has been widely studied in spatial audio research. Prior work exploits reverberation parameters (e.g., reverberation time) [12], MFCCs [13]–[16], or learned embeddings [17], extracted from active probing signals [12], [18], ambient noise [15], [19], or reverberant speech [13], [14], [16], [17], [20]–[22] for identifying or classifying room environments. However, these studies assume a fixed and

§These authors contributed equally to this work.

979-8-3315-7206-8/25/$31.00

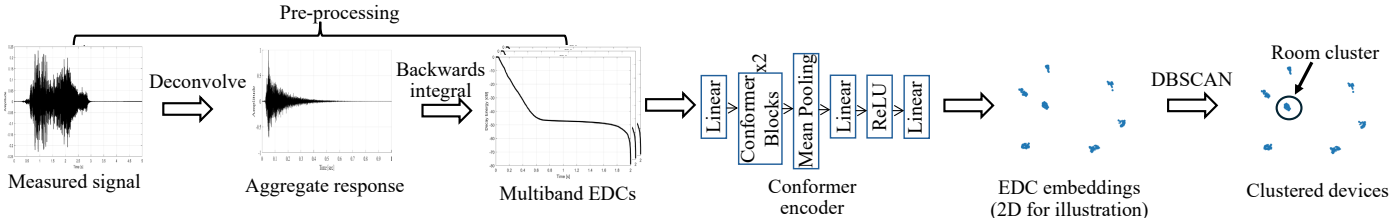

Fig. 2. Steps in the RAMDC framework: Measured probing signals are pre-processed into deconvolved aggregate response and EDCs, which are passed through a Conformer encoder to produce high-dimensional EDC embeddings. Embeddings from devices in the same room lie close together in embedding space, and DBSCAN then groups them into room clusters.

known set of rooms, which limits their applicability to cluster unseen rooms in an unsupervised manner. This closed-set assumption breaks down in ad-hoc meetings, where the number and identity of rooms are unknown.

The less-explored task of room clustering removes this assumption: the system must group devices into previously unseen room-based clusters, without any labels or assumptions on the number of rooms. Recent work has attempted unsupervised device clustering based on reverberation time and noise variance extracted from hand-clap recordings [23]. While promising in controlled settings, performance collapses when acoustically similar rooms share decay trends. Hence, a scalable, device-centric solution should operate in an unsupervised manner, exploit richer high-dimensional acoustic cues, and rely only on passive or minimally structured probe signals.

In this paper, we propose Room-Aware Multi-Device Clustering (RAMDC), an unsupervised framework for room-based clustering of devices based on acoustic similarity. During system initialization, each device records one or more chirps played by itself and co-located devices. From these recordings, we extract Energy Decay Curves (EDCs), which are fed to a Conformer-based [24] encoder trained with self-hard mining triplet loss [25] to learn embeddings. This training objective pulls embeddings of the EDCs from the same room closer together, while pushing apart those from different rooms. DBSCAN [26] and HDBSCAN [27], [28] are then applied to the learned embeddings to discover room clusters. We evaluate RAMDC on the GTU-RIR [29] dataset, where synthetic mixtures are generated by convolving real room impulse responses (RIRs) with chirp signals. Experimental results show that our learned embeddings significantly improve clustering accuracy over raw EDCs and that DBSCAN offers more robust performance than alternative clustering methods.

Once clustered, the system mutes all intra-room signal via stream-level control. Further connection to downstream AEC to control the conventional inter-room far-end signal allows for a comprehensive AEC for multi-device scenarios. The detailed integration with downstream AEC systems is beyond the scope of this paper.

II. Problem Formulation

Consider a teleconferencing scenario which has multiple participants with multiple devices, in an unknown R number of acoustically isolated rooms, as shown in Fig. 1. Let the global device set be $\mathcal{D} = \{1, 2, \ldots, D\}$, where

$$D = \sum_{r=1}^{R} |\mathcal{S}_r|, \quad \mathcal{S}_r = \{d \in \mathcal{D} \,|\, \text{device } d \text{ is located in room } r\}, \tag{1}$$

and $|\mathcal{S}_r|$ is the number of devices in room r, which is unknown in advance. Without loss of generality, each device is equipped with one loudspeaker and one microphone.

During normal operation, a common far-end signal $x(t)$ is routed to and reproduced by all D devices. The microphone signal observed at device $d \in \mathcal{D}$ is modeled as

$$y_d(t) = s_d(t) + e_d(t) + n_d(t), \tag{2}$$

where $s_d(t)$ is the aggregated near-end speech captured by device d's microphone, $e_d(t)$ is the total echo caused by playback of $x(t)$ from all devices in the same room as d, and $n_d(t)$ denotes additive noise. The playback signal is composed of

$$x(t) = x_{r(d)}(t) + \sum_{r' \neq r(d)} x_{r'}(t), \tag{3}$$

where $r(d)$ denotes the room index of device d, $x_{r(d)}(t)$ is the intra-room playback due to signals captured from the same room $r(d)$, and $x_{r'}(t)$ is the inter-room playback due to signals captured from other rooms r'.

The total echo $e_d(t)$ can be decomposed into intra-room and inter-room components as follows:

$$\begin{aligned} e_d(t) &= \sum_{d' \in \mathcal{S}_{r(d)}} g_{d,d'}(t) * x_{r(d)}(t) + \sum_{d' \in \mathcal{S}_{r(d)}} \sum_{r' \neq r(d)} g_{d,d'}(t) * x_{r'}(t) \\ &= e_d^{\text{intra}}(t) + e_d^{\text{inter}}(t), \end{aligned} \tag{4}$$

where $g_{d,d'}(t)$ denotes the acoustic echo path from the loudspeaker of device d' to the microphone of device d. Note that the summation over $d' \in \mathcal{S}_{r(d)}$ includes $d' = d$, since each device also receives its own loudspeaker's playback.

Conventional AEC post-processes the microphone signal $y_d(t)$ to estimate the acoustic path $g_{d,d'}(t)$ and eliminate the composite echo $e_d(t)$. However, even with perfect cancellation, every device in room $r(d)$ still plays the intra-room playback $x_{r(d)}(t)$. Participants therefore hear speech from co-located talkers once directly and as multiple delayed copies from surrounding loudspeakers. If the loop gain exceeds unity, this can even lead to howling. We call this unwanted replay of local speech $x_{r(d)}(t)$ *intra-room echo*, to distinguish it from the far-end echo targeted by conventional AEC.

The objective of this work is to introduce an upstream framework that identifies and suppresses the intra-room playback $x_{r(d)}(t)$, thereby removing the intra-room echo $e_d^{\text{intra}}(t)$ in (4) and leaving only the inter-room component. As a result, the subsequent stage only needs to cancel the inter-room echo $e_d^{\text{inter}}(t)$ using established multi-channel AEC algorithms [7]–[11].

III. Proposed Approach

In this section, we introduce our two-fold strategy to achieve intra-room echo suppression. First, we develop Room-Aware Multi-Device Clustering (RAMDC), an unsupervised room-level clustering method. Second, based on the inferred device clusters, we assume the teleconferencing system supports per-device stream-level control, inspired by voice routing strategies in multiplayer games, where communication is restricted within predefined teams. Under this assumption, the intra-room echo e_d^{intra} can be eliminated at the master level. This stream control approach is illustrated in Fig. 1.

The technical contribution of this paper lies in the RAMDC framework. RAMDC comprises three steps: probe and preprocess signals, encode signals via embeddings and clustering, as shown in Fig. 2 and detailed in the following subsections.

A. Measurements and Preprocessing of Probing Signals

In scenarios where the environment is quiet, we consider an active probing signal necessary to obtain accurate acoustic features of rooms. Our room-aware clustering can be enabled through a short initialization phase:

1) All devices mute the playback of $x(t)$ (i.e. master control is silenced)
2) All loudspeakers simultaneously emit a known probe signal $v(t)$.

During this phase, the observed microphone signal of device d becomes

$$y_d(t) = \mathcal{G}_d(t) * v(t) + n_d(t), \tag{5}$$

where $\mathcal{G}_d(t)$ is the intra-room aggregate response which can be estimated by deconvolution of $y_d(t)$, defined as

$$\mathcal{G}_d(t) = \sum_{d' \in \mathcal{S}_{r(d)}} g_{d,d'}(t), \tag{6}$$

where g is the acoustic paths from device d' loudspeaker to device d microphone.

We compute the energy-decay curve (EDC) of $\mathcal{G}_d(t)$ via

$$\mathcal{G}_{\text{EDC}}^d(t) = \int_t^\infty \mathcal{G}_d(t)^2 \, \mathrm{d}\tau, \tag{7}$$

and feed these EDCs directly into our neural encoder to produce fixed-dimensional embeddings for further clustering.

B. Encoding EDC to Embeddings

The EDCs from different rooms follow a consistent structure consisting of early decay, late decay, and noise floor regions [30]. While this structure provides a reliable basis for room comparison, EDCs from acoustically similar rooms often exhibit nearly indistinguishable shapes, making it difficult to differentiate between them (see Section IV for detailed analysis). To address this, we propose learning a high-dimensional embedding E that emphasize subtle EDC variations to improve room clustering using a Conformer network f_θ as

$$E_d = f_\theta\big(\mathcal{G}_{\text{EDC}}^d\big), \tag{8}$$

and the resulting embeddings are used by the clustering algorithm. We use the triplet loss [25] as the training objective to measure the relative similarity between the output embeddings such that EDCs from the same room, denoted as positive embeddings, are more similar, and those from different rooms, denoted as negative embeddings, are more different, given by

$$\mathcal{L}_{\text{triplet}} = \frac{1}{N} \sum_{i=1}^{N} \big[\|a_i - p_i\|_2 - \|a_i - n_i\|_2 + \alpha \big]_+, \tag{9}$$

where N is the batch size, a_i, p_i, n_i are the anchor, positive, and negative embeddings for the i-th triplet, $d(x, y) = ||x - y||_2$ is the Euclidean distance, α is the margin, and $[\cdot]_+ = \max\{\cdot, 0\}$.

C. Clustering Algorithm

We denote $\mathcal{C}(\cdot)$ as the clustering algorithm, which cluster the embeddings from the previous Conformer model output to infer the room labels for each device

$$\hat{r}(d) = \mathcal{C}(E_d), \quad d \in \mathcal{D}. \tag{10}$$

We experimented with the popular density-based clustering techniques DBSCAN [26] and HDBSCAN [27], as they can handle high-dimensional inputs, identify clusters of arbitrary shapes and sizes, and determine the number of clusters directly from the data without requiring any a priori knowledge [23], [28]. Once the room clustering algorithm in (10) has identified the set $\hat{\mathcal{S}}_{r(d)}$ for each device d, intra-room playback $x_{r(d)}(t)$ can be locally muted and thus no $e_d^{\text{intra}}(t)$. The microphone signal reduces to

$$y_d(t) \;=\; s_d(t) \;+\; e_d^{\text{inter}}(t) \;+\; n_d(t), \tag{11}$$

where only the inter–room echo $e_d^{\text{inter}}(t)$ remains since loudspeakers now play exclusively the far-end signals from other rooms $x_{r'}(t)$.

IV. Experiments

A. Dataset and Experimental Setup

To assess the effectiveness of the proposed RAMDC algorithm, we simulate a multi-user teleconference scenario in which $D \in [3, 20]$ personal devices are randomly distributed across an a priori unknown number of acoustically isolated rooms. Each device emits a 3-second logarithmic chirp probe signal sweeping from 100 Hz to 21 kHz, sampled at 44.1 kHz,

and played at a consistent volume. These probe signals are convolved with real-world RIRs from the GTU-RIR dataset, which contains measurements from 11 distinct rooms, assuming negligible system latency.

We synthesize 50 training scenarios using a subset of 7 rooms from the dataset, and evaluate performance on two test sets: Test 1, comprising 50 scenarios sampled from all 11 rooms, and Test 2, consisting of 50 scenarios sampled from the 4 previously unseen rooms. In each scenario, both the number of rooms and the number of devices per room (up to a maximum of 10) are drawn uniformly at random. Consequently, the total number of rooms R varies across scenarios and remains unknown to the clustering algorithm.

Each recorded signal is deconvolved with the original chirp to estimate the aggregate response, from which we extract eight EDCs. These include one broadband EDC spanning 100 Hz to 5 kHz, and seven octave-band EDCs centered at 125 Hz, 250 Hz, 500 Hz, 1 kHz, 2 kHz, 4 kHz, and 8 kHz. All EDCs are 2 seconds long, downsampled to 256 Hz, and used as input features.

For the learning setup, we use a Conformer-based encoder with two layers, a model dimension of 16, a feed-forward dimension of 32, a convolution kernel size of 7, attention heads of 4, and a dropout rate of 0.4. The model maps the 8-band EDC input to a 64-dimensional embedding space.

Training is performed with a batch size of 8 for 50 epochs using the Adam optimizer with an initial learning rate of 3×10^{-4} and a weight decay of 10^{-5}. A triplet loss with margin $\alpha = 0.2$ is used to enforce the relative distances among anchor, positive, and negative embeddings. Semi-hard negative mining is employed, where 10 negative pairs per anchor-positive sample are selected from the 20 hardest candidates. We exhaustively generate all possible anchor–positive combinations within each room to maximize sampling diversity during training. A learning rate scheduler reduces the learning rate by a factor of 0.8 if the validation loss does not improve for two consecutive epochs.

The model is pre-trained on the 50 training scenarios and validated using the same number of additional scenarios. During evaluation, embeddings are clustered using DBSCAN with cosine distance metric. The distance threshold ϵ is selected on the training set, as detailed in Section IV-C. We report clustering performance on a per-scenario basis using Accuracy (ACC), Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI), following the definitions in [31]. All performance metrics are averaged over 50 test scenarios.

All code and datasets from this work are available online [1].

B. Energy Decay Curve Evaluation

We first evaluate the EDC of each recording to investigate whether room-level similarity can be observed under varying numbers of chirp signals. Since each measurement—regardless of whether it receives one or up to ten chirps—is deconvolved with a single clean chirp, the resulting response does not strictly correspond to a physically meaningful RIR. Therefore, we aim to examine whether the resulting aggregate response still retain room characteristics.

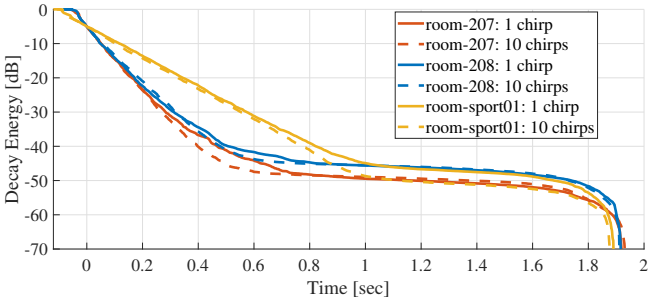

Fig. 3. The EDCs of the intra-room aggregate responses from three rooms, extracted from recordings with one and ten chirps per room, captured using the same microphone in each room.

To this end, we analyze recordings from three rooms in the GTU-RIR dataset: `room-207`, `room-208`, and `room-sport01`. We compare EDCs derived from recordings with one and ten chirps, captured using the same microphone within each room. Notably, `room-207` and `room-208` share similar room geometries, while `room-sport01` exhibits a distinctly different physical structure. As shown in Fig. 3, the late decay region—typically corresponding to the linear portion of the EDC between -5 dB and -35 dB [30]—reveals that `room-sport01` exhibits a significantly flatter slope than `room-207` and `room-208`, regardless of the number of chirps recorded from the measurements. In contrast, `room-207` and `room-208` display similar decay slopes, and variations in the number of chirps recorded make it even more difficult to distinguish between the two. These observations suggest that while raw EDCs can differentiate rooms with distinct reverberation characteristics, they are less effective for acoustically similar rooms. To address this, we consider learning embeddings to better capture minor acoustic differences.

C. Baseline Comparison

We evaluate clustering performance by comparing our learned embeddings with clustering based on raw reverberation features. Specifically, we consider three baselines: (i) 8-band raw EDCs, (ii) 8-band reverberation time T_{30}, and (iii) the parametric feature set proposed in [23], which includes decay rate τ and noise variance σ^2 estimated from late reverberation tails via maximum likelihood estimation. All methods employ DBSCAN clustering with cosine distance metric, where the distance threshold ϵ is manually tuned on the training set by selecting the value that yields the highest ACC and fixed for evaluation on the test scenarios.

As shown in Table I, the proposed RAMDC consistently outperforms all baselines, with the largest margin in Test 1: 11 rooms. A higher number of rooms increases clustering complexity due to more overlapping acoustic characteristics, making the limitations of raw features more apparent. In contrast, our learned embeddings better capture subtle differences, leading to robust performance. Even in the unseen-room-only

[1] https://github.com/AngelaYZhang/RAMDC

TABLE I
CLUSTERING PERFORMANCE (ACC, NMI, ARI) OF RAMDC COMPARED TO BASELINE FEATURES UNDER TEST 1 (T_1: 11 ROOMS) AND TEST 2 (T_2: 4 ROOMS).

Feature	ε	$\mathbf{ACC}_{T1}$	$\mathbf{NMI}_{T1}$	$\mathbf{ARI}_{T1}$	$\mathbf{ACC}_{T2}$	$\mathbf{NMI}_{T2}$	$\mathbf{ARI}_{T2}$
Proposed	7×10^{-2}	**83.98%**	**89.37%**	**0.69**	**78.63%**	**65.72%**	**0.53**
EDCs	2×10^{-3}	70.56%	76.89%	0.38	73.98%	57.67%	0.46
RT_{30}	5×10^{-3}	66.99%	72.07%	0.28	66.24%	52.44%	0.33
τ, σ^2 [23]	3×10^{-6}	55.16%	52.38%	0.14	61.13%	24.26%	0.12

(Test 2), our method maintains the best results, demonstrating strong generalization.

Among the baselines, raw EDCs yield the second-best performance, indicating that decay curves carry more distinctive room information than parametric room features.

D. Ablation Study

We conduct an ablation study to analyze the impact of three key components in RAMDC: (i) the embedding model architecture (Conformer vs. MLP), (ii) the clustering algorithm (DBSCAN vs. HDBSCAN), and (iii) the clustering distance threshold ϵ.

Table II summarizes the clustering performance under different configurations for both Test 1 and Test 2. Overall, we find that each component plays a critical role in the final performance.

Embedding architecture. Across all conditions, the Conformer consistently outperforms the MLP encoder. For instance, with DBSCAN and $\epsilon = 0.07$, the Conformer achieves 83.98% ACC on Test 1 and 78.63% on Test 2, compared to 77.48% and 72.58% with MLP. This highlights the strength of the Conformer in capturing temporal-frequency patterns in EDC features.

Clustering algorithm. DBSCAN generally yields better clustering results than HDBSCAN across all embedding types and test conditions. We attribute this to the fixed-density assumption in DBSCAN, which aligns well with the relatively uniform room distributions in teleconferencing scenarios. In contrast, HDBSCAN, while more flexible, may under-cluster or merge acoustically similar rooms.

Effect of ϵ. While varying the clustering threshold ϵ slightly affects performance, both Conformer and MLP embeddings consistently outperform raw reverberation features under all settings. This indicates that the learned embeddings are more robust to clustering threshold and better capture room-level acoustic characteristics.

In summary, the best performance is achieved using the Conformer encoder with DBSCAN clustering and $\epsilon = 0.07$, which we adopt as our final configuration. This configuration enables effective room-aware device grouping based on acoustic similarity.

V. CONCLUSION

In this paper, we introduced Room-Aware Multi-Device Clustering (RAMDC), an unsupervised framework for grouping devices into room clusters. To address intra-room echo in multi-user teleconferencing scenarios, we design an initialization step prior to activating AEC, where all devices simultaneously emit a logarithmic chirp. The calculated aggregate response are processed into EDCs, embedded using a Conformer-based encoder trained with self-hard triplet loss, and clustered via DBSCAN. Simulation results demonstrate that RAMDC outperforms other acoustic reverberation features and remains robust under unseen-room conditions. Once clusters are determined, each device locally mutes intra-room far-end signals. This simplifies downstream AEC, which only needs to suppress inter-room echoes using shorter and more efficient filters. Future work includes integrating RAMDC with downstream AEC systems and validating overall performance through experimental recordings, exploring alternative measurements such as speech or ambient noise for enhanced practicality.

TABLE II
ABLATION STUDY OF EMBEDDING MODELS, CLUSTERING ALGORITHMS, AND DISTANCE THRESHOLDS (ϵ) UNDER TEST 1 AND TEST 2.

Method	Clustering	ϵ	Clustering Metrics		
			ACC	NMI	ARI
Test 1: 11 rooms (7 seen + 4 unseen)					
Conformer	DBSCAN	0.07	**83.98%**	**89.37%**	**0.69**
	DBSCAN	0.10	81.67%	86.94%	0.62
	HDBSCAN	0.07	62.92%	59.34%	0.38
	HDBSCAN	0.10	61.47%	57.32%	0.36
MLP	DBSCAN	0.07	77.48%	82.79%	0.48
	DBSCAN	0.10	75.18%	80.75%	0.48
	HDBSCAN	0.07	59.47%	56.53%	0.33
	HDBSCAN	0.10	58.91%	56.15%	0.32
Test 2: 4 unseen rooms only					
Conformer	DBSCAN	0.07	78.63%	**65.72%**	0.53
	DBSCAN	0.10	**78.76%**	65.52%	**0.56**
	HDBSCAN	0.07	76.63%	57.45%	0.47
	HDBSCAN	0.10	74.80%	55.70%	0.44
MLP	DBSCAN	0.07	72.58%	58.28%	0.47
	DBSCAN	0.10	72.58%	57.86%	0.46
	HDBSCAN	0.07	73.41%	52.08%	0.41
	HDBSCAN	0.10	73.27%	52.06%	0.41

REFERENCES

[1] E. Hänsler and G. Schmidt, *Acoustic echo and noise control: a practical approach*. John Wiley & Sons, 2005.

[2] G. Enzner and P. Vary, "Frequency-domain adaptive Kalman filter for acoustic echo control in hands-free telephones," *Signal Process.*, vol. 86, no. 6, pp. 1140–1156, 2006.

[3] F. Kuech, E. Mabande, and G. Enzner, "State-space architecture of the partitioned-block-based acoustic echo controller," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 1295–1299.

[4] A. Ivry, I. Cohen, and B. Berdugo, "Deep adaptation control for acoustic echo cancellation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 741–745.

[5] J. Casebeer, J. Wu, and P. Smaragdis, "Meta-AF echo cancellation for improved keyword spotting," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2024, pp. 676–680.

[6] S. S. Shetu, N. K. Desiraju, J. M. M. Aponte, E. Habets, and E. Mabande, "A hybrid approach for low-complexity joint acoustic echo and noise reduction," in *Proc. Int. Workshop Acoust. Signal Enhanc.*, 2024, pp. 349–353.

[7] J. Benesty and D. R. Morgan, "Frequency-domain adaptive filtering revisited, generalization to the multi-channel case, and application to acoustic echo cancellation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, 2000, pp. II789–II792.

[8] H. Zhang and D. Wang, "Neural cascade architecture for multi-channel acoustic echo suppression," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 2326–2336, 2022.

[9] Y. Konforti, I. Cohen, and B. Berdugo, "Multichannel acoustic echo cancellation with beamforming in dynamic environments," *IEEE Open J. Signal Process.*, vol. 4, pp. 479–488, 2023.

[10] M. M. Halimeh and W. Kellermann, "Efficient multi-channel nonlinear acoustic echo cancellation based on a cooperative strategy," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 461–465.

[11] A. Cohen, A. Barnov, S. Markovich-Golan, and P. Kroon, "Joint beamforming and echo cancellation combining QRD based multichannel AEC and MVDR for reducing noise and non-linear echo," in *Proc. Eur. Signal Process. Conf.*, 2018, pp. 6–10.

[12] H. T. T. Truong, J. Toivonen, T. D. Nguyen, C. Soriente, S. Tarkoma, and N. Asokan, "DoubleEcho: Mitigating context-manipulation attacks in copresence verification," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun.*, 2019, pp. 1–9.

[13] M. Mascia, A. Canclini, F. Antonacci, M. Tagliasacchi, A. Sarti, and S. Tubaro, "Forensic and anti-forensic analysis of indoor/outdoor classifiers based on acoustic clues," in *Proc. Eur. Signal Process. Conf.*, 2015, pp. 2072–2076.

[14] H. Malik and H. Zhao, "Recording environment identification using acoustic reverberation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2012, pp. 1833–1836.

[15] L. Ma, B. Milner, and D. Smith, "Acoustic environment classification," *ACM Trans. Speech Lang. Process.*, vol. 3, no. 2, pp. 1–22, 2006.

[16] N. Peters, H. Lei, and G. Friedland, "Name that room: Room identification using acoustic features in a recording," in *Proc. ACM Int. Conf. Multimedia*, 2012, pp. 841–844.

[17] C. Papayiannis, C. Evers, and P. A. Naylor, "End-to-end classification of reverberant rooms using DNNs," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 3010–3017, 2020.

[18] L. Treybig, J. Höbel-Müller, S. Werner, and A. Nürnberger, "Acoustic inter- and intra-room similarity based on room acoustic parameters," in *Engineering for a Changing World: 60th ISC, Ilmenau Scientific Colloquium, Technische Universität Ilmenau, September 4–8, 2023: Proceedings*, 2023, p. 5.2.136.

[19] N. Karapanos, C. Marforio, C. Soriente, and S. Capkun, "Sound-Proof: Usable two-factor authentication based on ambient sound," in *Proc. USENIX Secur. Symp.*, 2015, pp. 483–498.

[20] M. Baum, L. Cuccovillo, A. Yaroshchuk, and P. Aichroth, "Environment classification via blind room-prints estimation," in *Proc. IEEE Int. Workshop Inf. Forensics Secur.*, 2022, pp. 1–6.

[21] N. Shabtai, B. Rafaely, and Y. Zigel, "Room volume classification from reverberant speech," in *Proc. Int. Workshop Acoust. Signal Enhanc.*, 2010.

[22] J. Bitterman, D. Levi, H. H. Diamandi, S. Gannot, and T. Rosenwein, "RevRIR: Joint reverberant speech and room impulse response embedding using contrastive learning with application to room shape classification," in *Proc. Interspeech*, 2024, pp. 3280–3284.

[23] H. Malik and H. Mahmood, "Acoustic environment identification using unsupervised learning," *Secur. Inform.*, vol. 3, pp. 1–17, 2014.

[24] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech*, 2020, pp. 5036–5040.

[25] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 815–823.

[26] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN," *ACM Trans. Database Syst.*, vol. 42, no. 3, pp. 1–21, 2017.

[27] L. McInnes, J. Healy, S. Astels, *et al.*, "HDBSCAN: Hierarchical density based clustering.," *J. Open Source Softw.*, vol. 2, no. 11, p. 205, 2017.

[28] P. Best, S. Paris, H. Glotin, and R. Marxer, "Deep audio embeddings for vocalisation clustering," *PLoS One*, vol. 18, no. 7, pp. 1–18, Jul. 2023.

[29] M. Pekmezci and Y. Genc, "Evaluation of SSIM loss function in RIR generator GANs," *Digit. Signal Process.*, vol. 154, p. 104 685, 2024.

[30] H. Kuttruff and M. Vorländer, *Room acoustics*. Crc Press, 2024.

[31] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Is a correction for chance necessary?" In *Proc. Int. Conf. Mach. Learn.*, Montreal, QC, Canada: ACM, 2009, pp. 1073–1080.