Design of Room Layout Estimator Using Smart Speaker*

Tomoki Joya[1], Shigemi Ishida[2], Yudai Mitsukude[1], and Yutaka Arakawa[1]

[1] ISEE, Kyushu University, Fukuoka, 819-0395 JAPAN
joya.tomoki@arakawa-lab.com, {mitsukude@f.,arakawa@}ait.kyushu-u.ac.jp
[2] Future University Hakodate, Hokkaido, 041-8655 JAPAN
ish@fun.ac.jp

Abstract. In this paper, we propose room-layout-based appliance control for voice user interfaces (VUIs) such as smart speakers. VUI-based appliance control requires a control command including *which device* to *do what*. However, we often suffer from an ambiguous target problem: the control target device in a control command is ambiguous because an ambiguous room name and demonstrative words are frequently used to specify the target device. To address the ambiguous target problem, we utilize room layout to estimate the control target. A user implicitly aims to control devices in the room where the user is. We therefore estimate the room where the user is now based on the room layout, which is estimated on a smart speaker, to determine the control target. As a first step toward room-layout-based appliance control, this paper presents the design of a room layout estimator. The experimental evaluations conducted in our 1-bedroom smart house reveal that our room layout estimator estimates room directions and room types with accuracies of 0.850 and 0.714, respectively.

Keywords: Voice User Interface (VUI) · acoustic sensing · room direction and type estimation.

1 Introduction

Recent advances in wireless communication technologies and Internet of Things (IoT) related technologies, smart home appliances are becoming prevalent nowadays. Using smart speakers working as a voice user interface (VUI) such as Google Home and Amazon Alexa, we can control smart home appliances by our voice.

On VUI-based control, we need to specify *which device* to *do what*. For example, we can turn on a light by ordering *turn on the light in the living room* to a smart speaker. In this example, we need to explicitly specify the light in the *living room* because there is a light in every room. To uniquely specify the

* This work was supported in part by the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Numbers JP21K11847, JP20KK0258, and JP19KT0020 as well as the Cooperative Research Project Program of RIEC, Tohoku University.

© 2021 Joya Tomoki, Shigemi Ishida, Yudai Mitsukude, and Yutaka Arakawa
This is the author's version of the work presented in EAI Mobiquitous 2021. Personal use of this material is permitted, but distribution of this material is strictly prohibited.

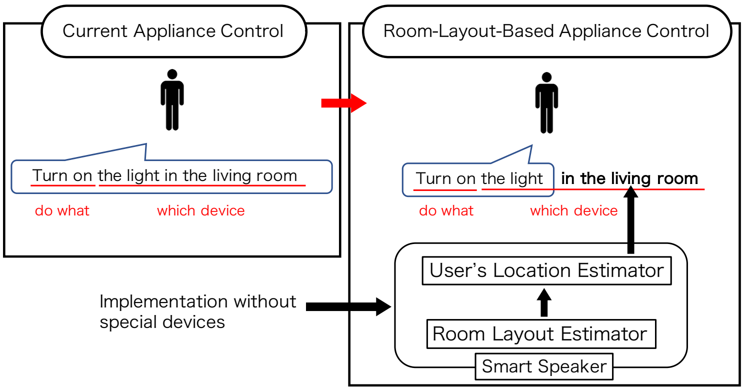

Current Appliance Control
Turn on the light in the living room
do what
which device
Room-Layout-Based Appliance Control
Turn on the light
in the living room
do what
which device
User's Location Estimator
Room Layout Estimator
Smart Speaker
Implementation without special devices

Fig. 1. Concept of room-layout-based appliance control for smart speakers

target device, we often use room names, which are configured to a smart speaker before using the smart speaker.

However, smart speakers often suffer from an ambiguous target problem. We often forget to specify a room name because we implicitly aim to control devices in the room where we are now. A target device specified by demonstrative words such as *this light* also causes the ambiguous target problem.

Another cause of the ambiguous target problem is ambiguous room names. We often use different names to specify a room. For example, we might try to turn on the light in the living room by ordering *turn on the light in the drawing room* or *turn on the light in the front room*.

To address the ambiguous target problem, context-aware decision-making has been proposed [2,3]. In context-aware decision-making approaches, the control target is estimated based on user's context. The user context estimation, however, requires sensors and a machine-learning model pre-trained with the user's previous behaviors.

In this paper, we propose a new approach: room-layout-based appliance control, as shown in Fig. 1. In practical situations, we often use ambiguous control commands such as *turn on the light*. When a user orders an ambiguous command *turn on the light*, we assume that the user aims to order *turn on the light in this room*. A smart speaker therefore estimates the room where the user is located using a user's location estimator. Room layout, which consists of room directions and types such as a living room and bedroom, is also estimated by a smart speaker using a room layout estimator to determine the room name where the control target is located.

As a first step of the above goal, this paper presents the design of the room layout estimator for smart speakers. Our assumption here is that smart speakers are equipped with a couple of microphones to estimate users' location. Analyzing sound source direction, the room layout estimator first estimates the direction of rooms. The type of the rooms is then estimated based on the activity sound, such as faucet sound, dish sound, and TV sound, derived from the room direction.

Although smart speakers on the market have a single microphone, we believe that smart speakers would be equipped with multiple microphones to improve robustness to noise and to improve users' voice separation performance.

Specifically, our main contributions are as follows:

- We propose a room-layout-based appliance control method for smart speakers. To the best of our knowledge, this is a first attempt to utilize the layout of rooms estimated on smart speakers to determine the control target appliance.
- We present the design of a room layout estimator for smart speakers equipped with multiple microphones. In contrast to existing sound source localization technologies, our approach for the room layout estimation utilizes the room-specific characteristics of reflected sound to distinguish different rooms.
- We show the basic performance of our room layout estimator by experimental evaluations. We collected home activity sound data at two different houses. The experimental evaluations demonstrated that the room direction estimation accuracy and room type estimation accuracy were 0.850 and 0.714, respectively.

The rest of this paper is organized as follows. Section 2 describes related work on sound source localization in indoor environments. In Sect. 3, we present the design of our room layout estimator that utilizes multiple microphones on a smart speaker, followed by experimental evaluations in Sect. 4. Finally, Sect. 5 concludes the paper.

2 Related Work

To the best of our knowledge, this is a first attempt to estimate room layout, not sound sources, using a microphone array.

Sound source localization, which estimates the location of sound sources using a microphone array, has been widely studied such as time delay estimation, beamforming, and subspace-based methods. Typical time delay estimators are cross-correlation-based methods where sound sources' locations are estimated by calculating cross-correlations between microphones [12,15,7]. The beamforming methods are represented by delay-and-sum beamformers, which combine sound signals on multiple microphones with phase compensation [14,16]. The representative subspace-based method is the MUSIC method that utilizes orthogonality of signal and noise components in the spatial correlation matrix of microphone array signals to estimate sound sources' location [11,4].

Many studies on sound source localization try to reduce the influence of reflected sound signals in indoor environments, where the sound localization performance degrades because of reverberation.

Suzuki et al. presented a sub-band peak hold process, which takes the amplitude of a direct sound signal, i.e., the sound signal firstly arrived at microphones, and masks the reflected sound signals that arrive after the direct sound [13]. Okamoto et al. applies a spatial averaging method on a 3-dimensional space

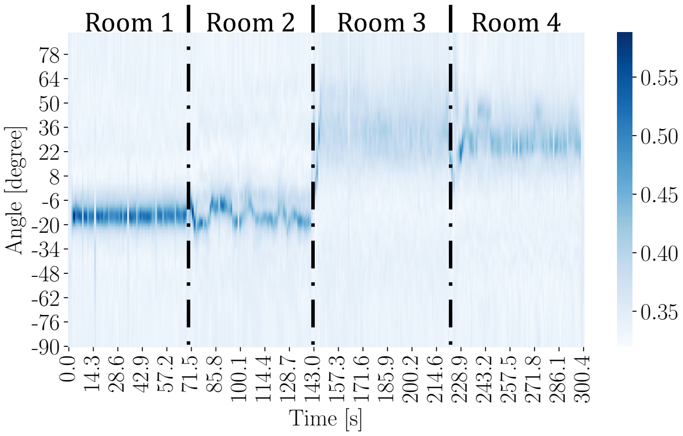


Fig. 2. Example of a sound density map with a single sound source moving in 4 rooms

model by dividing a microphone array into multiple sub-arrays and averaging the spatial matrices of each sub-array [8].

Ishi et al. estimates the locations of multiple sound sources using a spatial model and a ceiling-mounted microphones array consisting of 16 microphones [5]. The 3-dimensional space model is estimated, which is used to estimate the influence of reflected sound signals. Ribeiro et al. also reported a sound source localization robust to reflected signals relying on an actual 3-dimensional space model [10].

However, these methods require a large number of microphones, e.g., 16 microphones. The 3-dimensional space modeling is a novel approach, while high computational resources or much human effort are required to construct the space model. Many issues remain to be solved to estimate the room layout using a resource-limited smart speaker with a limited number of microphones.

3 Room Layout Estimator for Smart Speaker

3.1 Approach

Our primary approach to estimate the room layout is to extract the reverberation features using a *sound density map*. The sound density map is a map of sound power distribution as a function of time on each angle. We found that the sound signals from different rooms have different reverberation features because of the difference in size, wall locations, and diffraction objects. The difference in reverberation features appears as a difference of a *band* on the sound density map. We therefore distinguish sound signals from different rooms based on the features of bands on a sound density map by unsupervised learning algorithms.

Figure 2 shows an example of a sound density map with a single sound source, i.e., a vacuum cleaner, moving in 4 rooms. We installed a microphone array in a

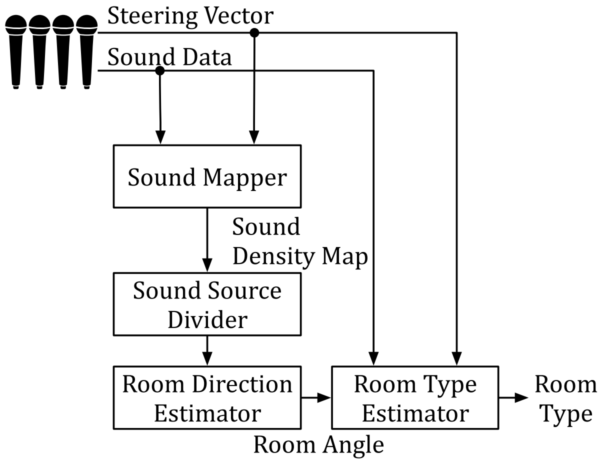


Fig. 3. Overview of room layout estimator for smart speaker

room of a 1-bedroom smart house and collected sound signals to draw a sound density map using the MUSIC method [11]. In Fig. 2, the moving sound source moves from a room to the next room at the time indicated by the dashed lines. We can confirm that the width and fluctuation of the band appear on the sound density map are dependent on the room where the sound source is located.

There are multiple sound sources in a practical environment, resulting in multiple bands corresponding to the sound sources on a sound density map. We first divide sound sources and then group the sound sources by estimating the room where the sound source is located by unsupervised learning with features extracted from a sound density map.

3.2 Assumptions

We assume that our method, i.e., room layout estimator for a smart speaker, is used in a residential environment such as a 2-bedroom house where multiple rooms are on the same floor and are adjacent to each other via doors. A smart speaker with a microphone array is installed in one of the rooms. Our goal is to estimate the room layout of rooms connected via a door to the room where the smart speaker is installed. In these rooms, multiple people are living together. They might make living noises at different locations at the same time. The number of rooms next to the room where the smart speaker is installed is given prior to the room layout estimation.

3.3 Design Overview

Figure 3 shows the overview of our room layout estimator for a smart speaker. The room layout estimator consists of a sound mapper, sound source divider,

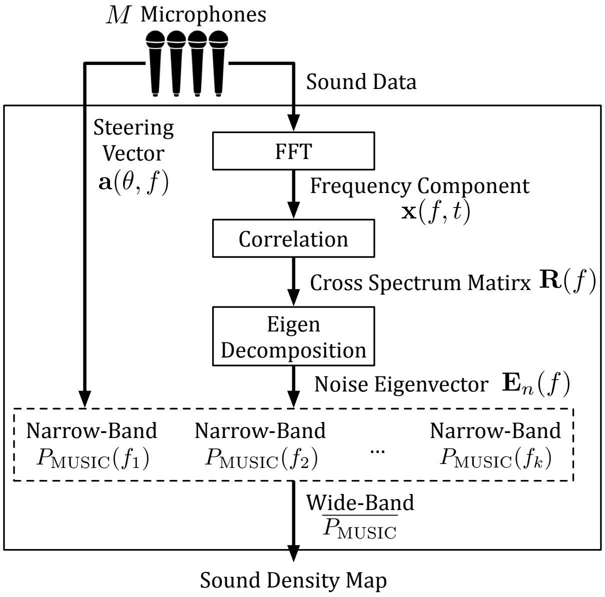

M Microphones
Sound Data
Steering Vector $\mathbf{a}(\theta, f)$
FFT
Frequency Component $\mathbf{x}(f, t)$
Correlation
Cross Spectrum Matirx $\mathbf{R}(f)$
Eigen Decomposition
Noise Eigenvector $\mathbf{E}_n(f)$
Narrow-Band $P_{\mathrm{MUSIC}}(f_1)$
Narrow-Band $P_{\mathrm{MUSIC}}(f_2)$
...
Narrow-Band $P_{\mathrm{MUSIC}}(f_k)$
Wide-Band $\overline{P_{\mathrm{MUSIC}}}$
Sound Density Map

Fig. 4. Overview of sound mapper

room direction estimator, and room type estimator. The sound mapper retrieves sound data using a microphone array and calculates sound power distribution on each angle using the MUSIC method to draw a sound density map. The sound source divider groups sound density map points into sound sources, which are more grouped into rooms where the sound source is located in the room direction estimator to estimate the room direction. The room type is finally estimated by the room type estimator using supervised learning with features extracted from sound signals of each room.

The following sections describe the details of each component.

3.4 Sound Mapper

The sound mapper performs the MUSIC to draw a sound density map, which is a map of the sound power distribution on each angle as a function of time. The MUSIC method has a high angle estimation resolution and is useful to extract reverberation features as fluctuations of the sound arrival direction.

Figure 4 shows the overview of the sound mapper. As shown in Fig. 4, the sound mapper collects sound data using a microphone array. A steering vector, a vector describing phase differences of sound signals on each microphone, is also calculated from the physical arrangement of the microphone array.

The collected sound signals are segmented by a fixed time-length window for fast Fourier transform (FFT). Let $\mathbf{x}(f, t)$ be a vector of sound frequency

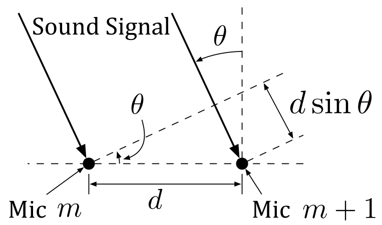


Fig. 5. Difference in sound traveling distance between two linearly aligned microphones

components of frequency f at time t. $\mathbf{x}(f,t)$ is a M-dimensional vertical vector, where M is the number of microphones in the microphone array.

The sound mapper calculates cross spectrum matrix $\mathbf{R}(f)$ as

$$\mathbf{R}(f) = E\left[\mathbf{x}^H(f,t)\,\mathbf{x}(f,t)\right], \tag{1}$$

where $\mathbf{z}^H$ denotes the Hermitian transpose of a vector $\mathbf{z}$ and $E[\ ]$ denotes an averaging process. We then calculate the eigenvalues and eigenvectors of the cross spectrum matrix $\mathbf{R}(f)$. The number of signal and noise components is estimated based on the distribution of the magnitudes of the eigenvalues over multiple windows. Assuming that we have $N(< M)$ signal components, we obtain the noise eigenvectors $\mathbf{E}_n(f)$ corresponding to the remaining $M-N$ eigenvalues.

A steering vector is calculated from the physical arrangement of the microphone array. As shown in Fig. 5, the difference in sound traveling distance between two linearly aligned microphones separated by distance d is $d\sin\theta$, where θ is the sound arrival angle. $d\sin\theta$ corresponds to the phase difference of $2\pi f d\sin\theta/c$, where c is the speed of sound in air. The steering vector $\mathbf{a}(\theta,f)$ of M linearly aligned microphones is therefore calculated to be

$$\mathbf{a}(\theta,f) = \left[1\ e^{-j\phi}\ e^{-j2\phi}\ \cdots\ e^{-j(M-1)\phi}\right]^T \tag{2}$$

where $\phi = 2\pi f d\sin\theta/c$ and ^{T} denotes the transpose operation. Although we used an example of linearly aligned microphones, the same idea can be used to calculate the steering vector for a different microphone setup.

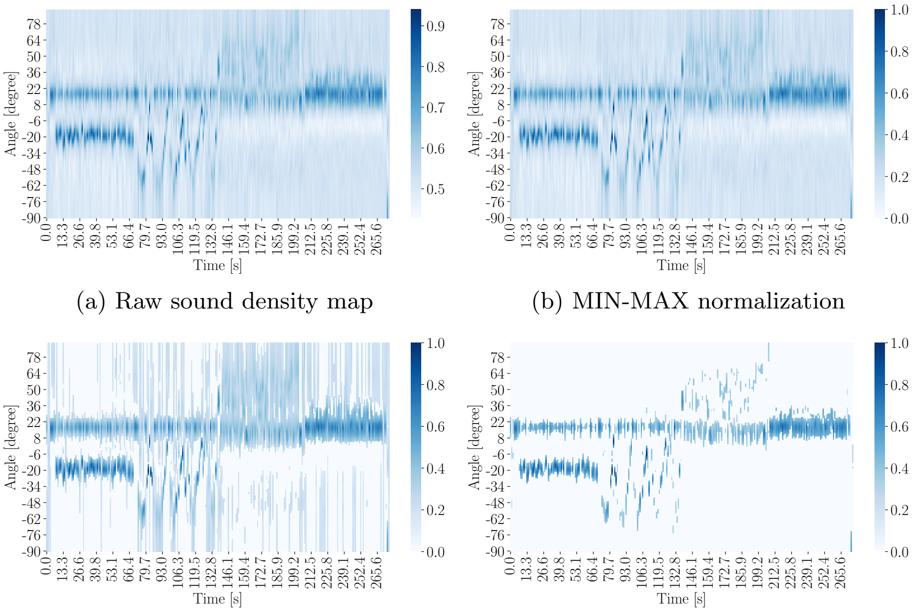
Using the eigenvectors $\mathbf{E}_n(f)$ and the steering vector $\mathbf{a}(\theta,f)$, we derive narrow-band sound power distribution $P_{\mathrm{MUSIC}}(\theta,f)$ as

$$P_{\mathrm{MUSIC}}(\theta,f) = \frac{1}{\mathbf{a}^H(\theta,f)\,\mathbf{E}_n(f)\,{\mathbf{E}_n}^H(f)\,\mathbf{a}(\theta,f)}. \tag{3}$$

We finally derive wide-band sound power distribution $\overline{P_{\mathrm{MUSIC}}}$ as

$$\overline{P_{\mathrm{MUSIC}}} = \frac{1}{k}\sum_f P_{\mathrm{MUSIC}}(\theta,f). \tag{4}$$

Here we assume there are k frequency components in FFT results. $\overline{P_{\mathrm{MUSIC}}}$ has peaks on the angle of sound sources. We draw $\overline{P_{\mathrm{MUSIC}}}$ as a function of angle and time, deriving a *sound density map*.



(a) Raw sound density map

(b) MIN-MAX normalization

(c) Extract top half at each time

(d) Extract top 10-percent at each time

Fig. 6. Overview of filtering on sound density map

We apply a filtering process to a sound density map because a raw sound density map includes sound power information corresponding to noise components. Figure 6 shows the overview of the filtering process. We first apply a MIN-MAX normalization process (Fig. 6b) and extract top half points at each time t (Fig. 6c). Top 10-percent points are finally extracted at each time t (Fig. 6d).

3.5 Sound Source Divider

The sound source divider groups the points on a sound density map into sound sources. A sound source moves not so quickly, resulting in a continuous band on a sound density map. When there are multiple sound sources, we can observe multiple bands on a sound density map. We apply the DBSCAN, a density-based clustering method, to a sound density map to group points on a sound density map into sound sources.

Figure 7 shows an overview of the sound source divider. The clustering process consists of two steps.

In the first step, we extract sound density map points corresponding to sound sources in the room where the smart speaker, i.e., the microphone array, is installed. We perform the DBSCAN clustering with four features: peak width at time t, the number of peaks at time t, angle θ, and time t. Each cluster is a set of points on a sound density map corresponding to a single sound source.

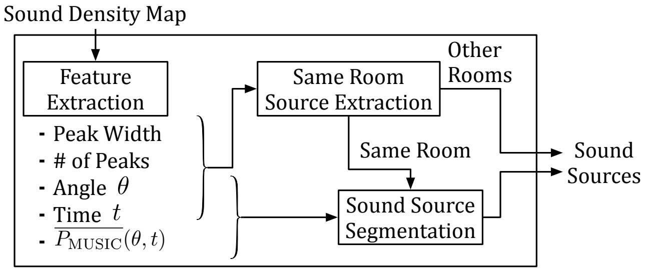


Fig. 7. Overview of sound source divider

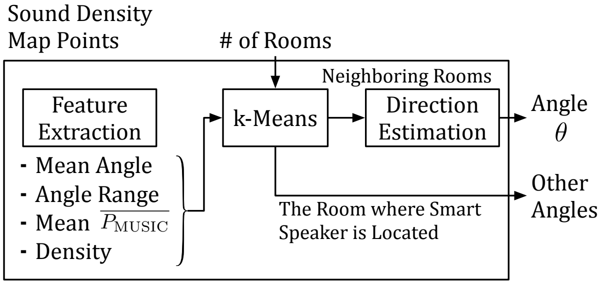


Fig. 8. Overview of room direction estimator

The second step more divides sound density map points corresponding to sound signals from the room where the smart speaker is installed. Sound signals from sound sources in the same room as the smart speaker show specific features. The second step utilizes the DBSCAN clustering with three features, different from the first step: angle θ, time t, and wide-band sound power information $\overline{P_{\text{MUSIC}}}(\theta, t)$. Sound sources in the same room as the smart speaker are estimated based on the angle variance of points in clusters divided in the first step. The cluster that has the largest angle variance is estimated as the sound source in the same room as the smart speaker because the sound signal arrives from any direction in the room.

Finally, all the clustering results are merged to complete the sound source segmentation.

3.6 Room Direction Estimator

Figure 8 shows an overview of the room direction estimator. The room direction estimator first groups sound sources into the rooms where the sound source is located using the k-means clustering. The k-means clustering utilizes four kinds of features calculated for each sound source: mean angle, the range of angle, mean sound power $\overline{P_{\text{MUSIC}}}(\theta, t)$, and the density of sound density map points. The density is a ratio of the number of sound density map points to the area

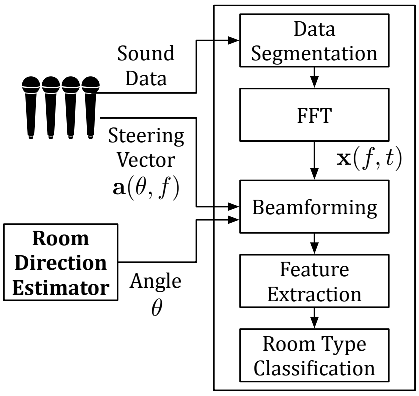


Fig. 9. Overview of room type estimator

size of the rectangle where the sound density map points are located. The k is set to the number of rooms, as is assumed to be given.

The room direction estimator then calculates the most frequent sound arrival angle in each cluster, estimating the room direction. The room where the smart speaker is located is excluded from the room direction estimation because the room direction cannot be defined. The smart speaker co-located room is easily estimated based on the angle variance.

3.7 Room Type Estimator

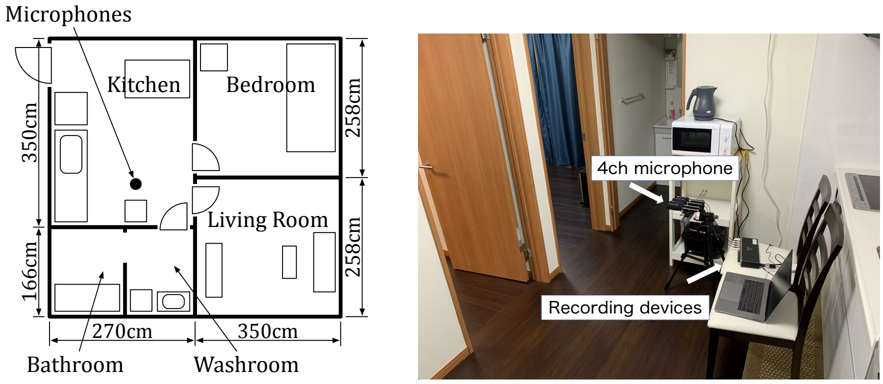
Figure 9 shows an overview of the room type estimator. The room type estimator synthesizes the sound signals in the same room and estimates the room type by supervised learning. The room type estimator first calculates frequency components $\mathbf{x}(f, t)$, which is the same process as in the sound mapper. The synthesized sound signal is then calculated as

$$y(f, t) = \mathbf{a}^T(\theta, f)\,\mathbf{x}(f, t), \tag{5}$$

where $\mathbf{a}(\theta, f)$ is the steering vector given in Sect. 3.4.

The synthesized sound signal $y(f, t)$ is divided by a fixed time-length window to extract features for supervised learning. We calculate basic statistics, i.e., mean, maximum, minimum, and variance, of six kinds 25 metrics below in each window as features, referring to [1], resulting in a 100-dimensional feature vector.

1. MFCCs: 20 Mel frequency cepstrum coefficients (MFCCs)
2. Zero crossing rate: the rate at which the positive and negative amplitudes are switched in the time-domain waveform
3. RMS: root mean square of sound signals
4. Spectral flatness: a measure of how the sound is noise-like [6,9]
5. Spectral centroid: barycenter of the spectrum [9]



(a) Room layout

(b) Actual microphone setup

Fig. 10. Experiment setup

6. Spectral roll-off: frequency so that major components of the sound energy is contained below this frequency [9]

The room type estimator finally classifies the room type of each room using the 100-dimensional feature vectors. We don't limit the classifier algorithm. We use a Random Forest classifier in this paper as an example. The classifier model is trained in advance using sound data collected in a typical residential environment, not limited to the actual smart speaker location.

4 Evaluation

We conducted initial evaluations using sound data collected in our 1-bedroom smart house. We also collected sound data of specific daily activities in a normal house, which is used for training of room type estimator. We separately evaluated two tasks in our room layout estimation: room direction and room type estimations.

4.1 Experiment Setup

Figure 10 shows the room layout and actual microphone setup in our smart house. A 4-channel microphone array, i.e., four AZDEN SGM-990 microphones separated by 50 mm, was installed in the living room on a tripod 1 m away from the walls at the height of 0.7 m, as shown in Fig. 10a. Sound data was collected using a Behringer UMC404HD USB audio interface connected to a laptop at a sampling rate of 44.1 kHz with code length 16 bits.

4.2 Room Direction Estimation Performance

To evaluate the room direction estimation performance, we collected sound data while two subjects A and B were talking and walking in our smart house, creating

Table 1. Dataset used for evaluation of room direction estimation

Dataset (40 s × 20)	Sound source 1 (Subject A voice)	Sound source 2 (Subject B voice)
Bedroom DS	Bedroom	Bedroom (10 s)
Kitchen DS	Kitchen	→ kitchen (10 s)
Washroom DS	Washroom	→ washroom (10 s)
Living DS	Living room	→ living (10 s)

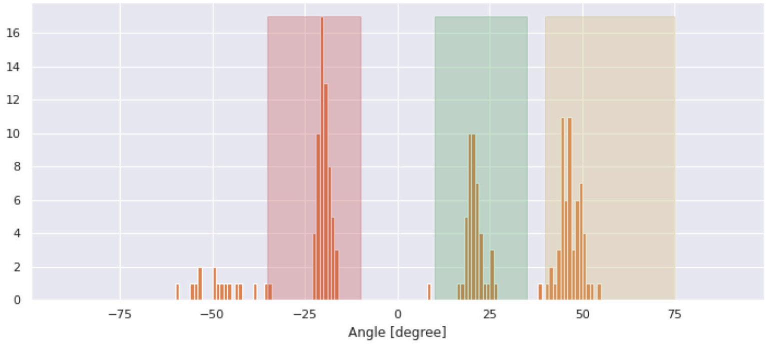


Fig. 11. Histogram of room direction estimation results

four datasets shown in Table 1. Each dataset consists of 20 40-second recordings. Each recording is sound data collected while the subject A was freely walking in a room indicated in Table 1. The subject B was freely walking in a room for 10 seconds and moved to another room, as indicated in Table 1.

The room direction estimation performance was evaluated in two aspects: the room-based sound source clustering performance and room direction accuracy. The room-based sound source clustering performance was evaluated using the adjusted Rand index (ARI), which is a commonly used metric for evaluations of clustering performance. The room direction accuracy was evaluated using the rate of the number of trials correctly estimated room direction. As shown in Fig .10, the microphone array was installed in the kitchen, which is next to the living room, bedroom, and bathroom. k in the room direction estimator, i.e., the number of clusters for the k-means, was therefore set to 4.

Figure 11 shows the histogram of the room direction estimation results. Red, green, yellow rectangles represent the correct room directions of the bedroom, living room, and washroom, respectively. The mean ARI was 0.725. The direction estimation accuracy, i.e., the rate of the number of trials in red, green, and yellow rectangles in Fig. 11, was 0.850. We can confirm that our room direction estimator successfully estimated room direction with no training data.

For reference, the direction estimation accuracy was increased to 0.875 when we use trials with ARI greater than 0.9. The room direction estimation performance highly relies on the accuracy of sound source clustering into rooms.

Table 2. Room direction estimation performance for each dataset

Dataset	Mean ARI	Direction estimation accuracy
Bedroom DS	0.897	0.783
Kitchen DS	0.327	0.683
Washroom DS	0.746	0.967
Living DS	0.925	0.967

Table 3. Activities used in room type estimation evaluation

(a) Specific activity

Room type	Activities
Living room	Watching TV, talking on phone
Kitchen	Tidying up dishes, washing dishes, opening/closing fridge and kitchen cabinet doors
Bedroom	turning over in bed, sleeping

(b) Free activity

Room type	Activities
Living room	Watching TV
Kitchen	Washing dishes, eating, using microwave
Bedroom	Using smartphone on a bed, sleeping

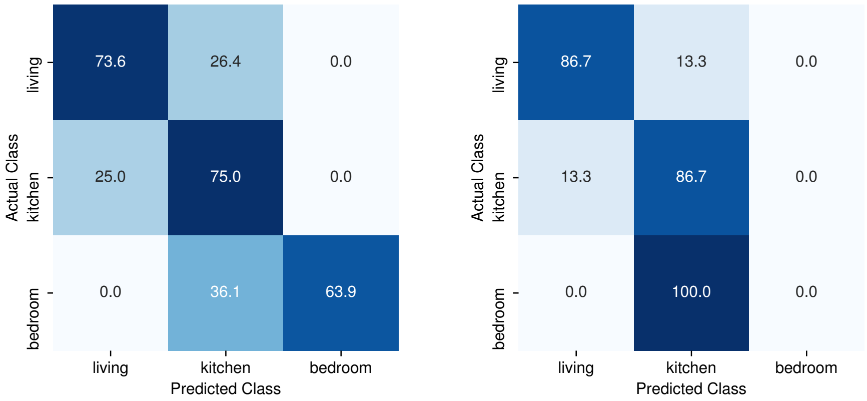
We also compared the performance of room direction estimation for each dataset. Table 2 shows the mean ARI and direction estimation accuracy for each dataset. From Table 2, we can see that the performance with the kitchen dataset was significantly lower than that with other datasets. As shown in Table 1, the kitchen dataset includes sound signals of subject A in the kitchen, where microphones were installed. Because the sound signals from the room where microphones were installed can reach the microphones from any direction, sound source segmentation described in Sect. 3.5 was highly unsuccessful, which resulted in the significant degradation in performance.

4.3 Room Type Estimation Performance

To evaluate the room type estimation performance, we collected sound data in our smart house while a subject stays in each room. We installed a microphone at the same height and location as indicated in Fig. 10 and collected sound data for each room activity. The sound data is collected both in the controlled environment where the subject did a specific activity and in the free environment where the subject stayed in a specific room doing free activities. In the controlled environment, sound data of each activity shown in Table 3a was collected for 120 seconds. In the free environment, we collected sound data for 30 minutes for each room. We emphasize that we gave no instruction for activity during the stay in the free environment. The actual activities during the 30 minutes are shown in Table 3b.

Table 4. Activities for training of room type estimator

Room type	Activities
Living room	Watching TV, talking
Kitchen	Cutting, frying, eating, washing dishes, using microwave
Bedroom	Sleeping



(a) Controlled environment

(b) Free environment

Fig. 12. Confusion matrices of room type estimation results

We also collected training data for room type estimation because the room type estimator uses supervised learning. The training data was collected in a normal house while a subject did an activity shown in Table 4. We used a Sony PCM-D100 recorder with an embedded microphone to evaluate the influence of microphone and environment differences. Each activity sound was recorded for 120 seconds.

We evaluated the room type classification accuracy both in the controlled and in the free environments using the room type estimation model trained with the data collected in the normal house. We divided sound data by a 10-second window and calculated features for each windowed data, which were used as input to the room type estimator. Note that we did not perform the sound signal synthesis described in Sect. 3.7 as an initial evaluation in this paper, evaluating the raw room type estimation performance to validate the feasibility of our proposal.

Figure 12 shows the confusion matrices of the room type estimation results. Figures 12a and 12b show confusion matrices for the controlled and free environments, respectively. The mean accuracies in the controlled and free environments of room type estimation were 0.714 and 0.536, respectively. Even though the room type is estimated using the model trained with the data collected in a different environment, we derived high estimation accuracy. We can conclude that the room type estimation can be realized using the estimation model trained in advance with data collected in a normal house environment.

However, in the bedroom, the room type estimation accuracy was lower than the accuracy in other rooms. We can easily guess that the sound power of bedroom activity shown in Table 3 is relatively low compared to the other room activities, which resulted in the low estimation accuracy. The room type estimation accuracy in the bedroom in the controlled environment was 63.9%. We believe that a sufficient amount of training data improves the room type estimation accuracy.

5 Conclusion

In this paper, we presented the design of a room layout estimator for smart speakers. The room layout, i.e., the direction and type of next rooms, are estimated using reverberation features that are extracted from a sound density map, which is a map of sound power distribution as a function of time. The sound sources are grouped into rooms where the sound source is located by unsupervised learning to estimate the room direction. The room type is finally estimated by supervised learning with a pre-trained model. We conducted experimental evaluations and demonstrated that our room type estimator successfully estimated room directions and room types with accuracies of 0.850 and 0.714, respectively. As our future work, we plan to improve the accuracy of room type estimation in the bedroom by introducing novel features. We also plan to study the influence of the location of big objects such as furniture, and verify our method in different room layouts.

References

1. Bountourakis, V., Vrysis, L., Papanikolaou, G.: Machine Learning Algorithms for Environmental Sound Recognition: Towards Soundscape Semantics. In: Proceedings of the Audio Mostly 2015 on Interaction With Sound. pp. 1–7. AM '15, Association for Computing Machinery, New York, NY, USA (Oct 2015). https://doi.org/10.1145/2814895.2814905
2. Chahuara, P., Portet, F., Vacher, M.: Making Context Aware Decision from Uncertain Information in a Smart Home: A Markov Logic Network Approach. In: Ambient Intelligence, vol. 8309, pp. 78–93. Springer International Publishing, Cham (2013). https://doi.org/10.1007/978-3-319-03647-2_6
3. Chahuara, P., Portet, F., Vacher, M.: Context-aware decision making under uncertainty for voice-based control of smart home. Expert Systems with Applications **75**, 63–79 (Jun 2017). https://doi.org/10.1016/j.eswa.2017.01.014
4. Danès, P., Bonnal, J.: Information-theoretic detection of broadband sources in a coherent beamspace MUSIC scheme. In: 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 1976–1981. IEEE, Taipei (Oct 2010). https://doi.org/10.1109/IROS.2010.5651249
5. Ishi, C.T., Even, J., Hagita, N.: Using multiple microphone arrays and reflections for 3D localization of sound sources. In: 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 3937–3942 (Nov 2013). https://doi.org/10.1109/IROS.2013.6696919
6. Johnston, J.D.: Transform coding of audio signals using perceptual noise criteria. IEEE Journal on Selected Areas in Communications **6**(2) (Feb 1988). https://doi.org/10.1109/49.608
7. Knapp, C., Carter, G.: The generalized correlation method for estimation of time delay. IEEE Transactions on Acoustics, Speech, and Signal Processing **24**(4), 320–327 (Aug 1976). https://doi.org/10.1109/TASSP.1976.1162830
8. Okamoto, T., Nishimura, R., Iwaya, Y.: Estimation of sound source positions using a surrounding microphone array. Acoustical Science and Technology **28**(3), 181–189 (2007). https://doi.org/10.1250/ast.28.181
9. Peeters, G.: A large set of audio features for sound description (similarity and classification) in the CUIDADO project (Jan 2004), http://recherche.ircam.fr/equipes/analyse-synthese/peeters/ARTICLES/Peeters_2003_cuidadoaudiofeatures.pdf
10. Ribeiro, F., Zhang, C., Florêncio, D.A., Ba, D.E.: Using Reverberation to Improve Range and Elevation Discrimination for Small Array Sound Source Localization. IEEE Transactions on Audio, Speech, and Language Processing **18**(7), 1781–1792 (Sep 2010). https://doi.org/10.1109/TASL.2010.2052250
11. Schmidt, R.: Multiple emitter location and signal parameter estimation. IEEE Transactions on Antennas and Propagation **34**(3), 276–280 (Mar 1986). https://doi.org/10.1109/TAP.1986.1143830
12. Silverman, H.F.: An algorithm for determining talker location using a linear microphone array and optimal hyperbolic fit. In: Proceedings of the Workshop on Speech and Natural Language. pp. 151–156. HLT '90, Association for Computational Linguistics, USA (Jun 1990). https://doi.org/10.3115/116580.116632
13. Suzuki, T., Kaneda, Y.: Improving the robustness of multiple signal classification (MUSIC) method to reflected sounds by sub-band peak-hold processing. Acoust. Sci. & Tech. **30**(5), 387–389 (2009). https://doi.org/10.1250/ast.30.387

14. Tanaka, M., Kaneda, Y.: Performance of sound source direction estimation methods under reverberant conditions. Journal of the Acoustical Society of Japan (E) **14**(4), 291–292 (1993). https://doi.org/10.1250/ast.14.291

15. Wang, H., Chu, P.: Voice source localization for automatic camera pointing system in videoconferencing. In: 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing. vol. 1, pp. 187–190 (Apr 1997). https://doi.org/10.1109/ICASSP.1997.599595

16. Warsitz, E., Haeb-Umbach, R.: Blind Acoustic Beamforming Based on Generalized Eigenvalue Decomposition. IEEE Transactions on Audio, Speech, and Language Processing **15**(5), 1529–1539 (Jul 2007). https://doi.org/10.1109/TASL.2007.898454