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.ait.kyushu-u.ac.jp,
arakawa@ait.kyushu-u.ac.jp
[2] Future University Hakodate, Hokkaido 041–8655, Japan
ish@fun.ac.jp

Abstract. In this study, we propose a 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 experience 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 this problem, we utilized a room layout to estimate the control target. A user implicitly aims to control devices in a room where they are. Therefore, we 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. In this study, we present the design of a room-layout estimator as the first step toward room-layout-based appliance control. The experimental evaluations conducted in our 1-bedroom smart house revealed 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

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

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

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.

T. Hara and H. Yamaguchi (Eds.): MobiQuitous 2021, LNICST 419, pp. 24–39, 2022.
https://doi.org/10.1007/978-3-030-94822-1_2

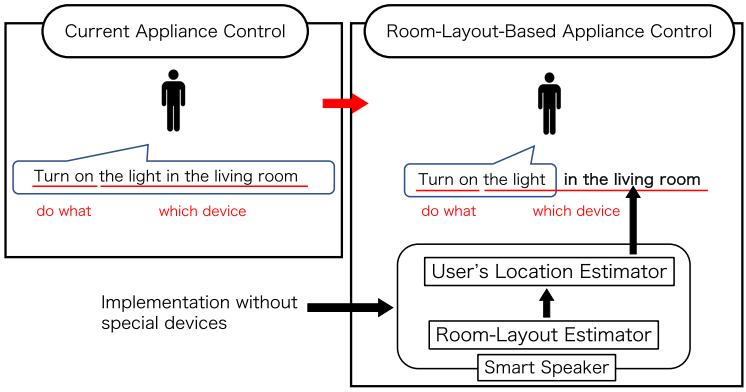


Fig. 1. Concept of room-layout-based appliance control for smart speakers

the *living room* because there are lights in every room. To uniquely specify the target device, we often use room names that are configured to a smart speaker before using the smart speaker.

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

Another cause of the ambiguous target problem is the ambiguity in the room names. Different names are often used to specify rooms. For example, we might attempt 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 the user's context. However, user context estimation requires sensors and a machine-learning model pre-trained with the user's previous behaviors.

In this study, we propose a new approach, room-layout-based appliance control, as shown in Fig. 1. In practical situations, ambiguous control commands are often used, such as *turn on the light*. When a user makes an ambiguous command, such as *turn on the light*, we assume that the user aims to order *turn on the light in the room they are in*. A smart speaker, therefore, estimates the room where the user is located using a user's location estimator. The room layout, which comprises 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 toward this goal, in this study, we present the design of a room-layout estimator for smart speakers. Our assumption here is that smart speakers are equipped with a couple of microphones to estimate the user location. Analyzing the sound source direction, the room-layout estimator first estimates the direction of the rooms. The type of the rooms is then estimated based on

the activity sound, such as faucet sound, dish sound, and TV sounds, derived from the room direction. Although smart speakers on the market have a single microphone, we believe that in the near future, smart speakers will be equipped with multiple microphones to improve robustness to noise and to improve users' voice separation performance.

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 the 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 the reflected sound to distinguish different rooms.
- We show the basic performance of our room-layout estimator through experimental evaluations. We collected the home activity sound data from 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 remainder 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, the paper is concluded in Sect. 5.

2 Related Work

To the best of our knowledge, this is the first attempt to estimate a room layout using a microphone array rather than sound sources.

Sound source localization, which estimates the location of sound sources using a microphone array, has been widely studied and includes 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 the cross-correlations between microphones [7,12,15]. 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 the orthogonality of signal and noise components in the spatial correlation matrix of microphone array signals to estimate the location of sound sources [4,11].

Numerous studies on sound source localization have attempted 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 considers the amplitude of a direct sound signal, the sound signal that first reaches the microphones, and masks the reflected sound signals that reach subsequent to the direct

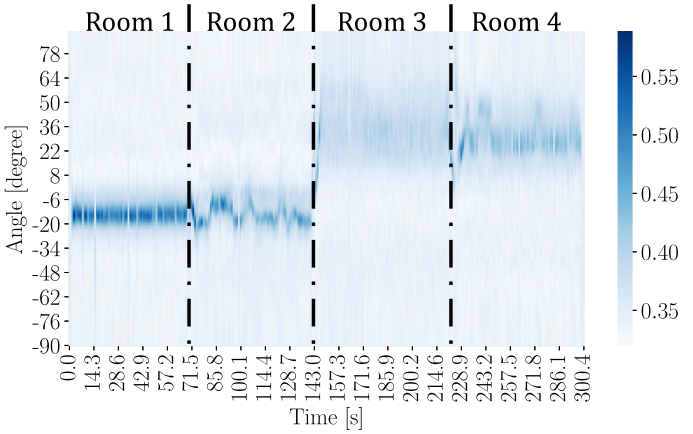


Fig. 2. Example of a sound density map with a single sound source moving in 4 rooms

sound [13]. Okamoto et al. applies a spatial averaging method to a 3-dimensional space model by dividing a microphone array into multiple sub-arrays and averaging the spatial matrices of each subarray [8].

Ishi et al. estimates the locations of multiple sound sources using a spatial model and a ceiling-mounted microphone array comprising 16 microphones [5]. A 3-dimensional space model is utilized to estimate the influence of the 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, for example, 16 microphones. 3-dimensional space modeling is a novel approach, where high computational resources or considerable human effort are required to construct the space model. The estimation of the room layout using a resource-limited smart speaker with a limited number of microphones is associated with numerous unsolved problems.

3 Room-Layout Estimator for Smart Speaker

3.1 Approach

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

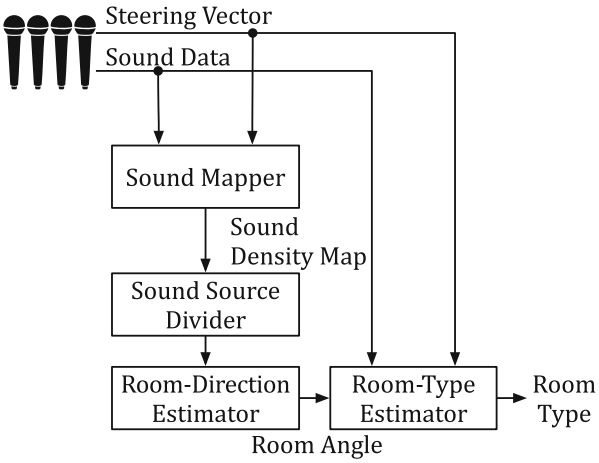


Fig. 3. Overview of room-layout estimator for smart speaker

Figure 2 shows an example of a sound density map with a single sound source, this is, a vacuum cleaner, moving in four rooms. We installed a microphone array in a 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 one room to the next room at the time indicated by the dashed lines. We can confirm that the width and fluctuation of the band appearing on the sound density map are dependent on the location of the sound source.

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 the sound sources and then group them by estimating the room where the sound source was located, by unsupervised learning with features extracted from a sound density map.

3.2 Assumptions

We assume that our method, that is, the 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 located adjacent to each other with doors separating them. A smart speaker with a microphone array was installed in one of the rooms. Our goal is to estimate the room layout of rooms connected via a door to a room where the smart speaker is installed. In these rooms, multiple people live together. They might make living noises at different locations simultaneously. The number of rooms next to the room where the smart speaker is installed is given before the room-layout estimation.

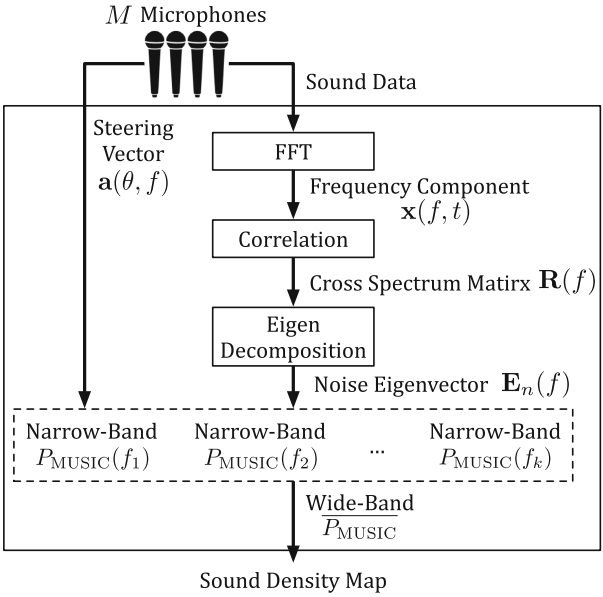


Fig. 4. Overview of sound mapper

3.3 Design Overview

Figure 3 shows an overview of the room-layout estimator for a smart speaker. The room-layout estimator comprises a sound mapper, sound source divider, room-direction estimator, and room-type estimator. The sound mapper retrieves sound data using a microphone array and calculates the sound power distribution at 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 the 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 for each angle as a function of time. The MUSIC method has a high angle estimation resolution and is useful for extracting reverberation features as fluctuations in the sound arrival direction.

Figure 4 shows an 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 the phase differences of sound signals on each microphone, was also calculated from the physical arrangement of the microphone array.

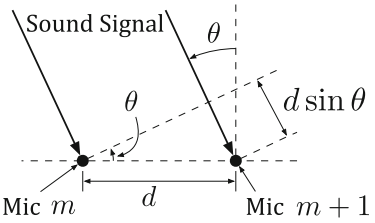


Fig. 5. Difference in sound traveling distance between two linearly aligned microphones

The collected sound signals are segmented using a fixed time-length window for the fast Fourier transform (FFT). Let $\mathbf{x}(f,t)$ be an M-dimensional column vector of sound frequency components of frequency f at time t, where M is the number of microphones in the microphone array. The sound mapper calculates the cross-spectrum matrix $\mathbf{R}(f)$ as follows:

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

where $\mathbf{z}^H$ denotes the Hermitian transpose of a vector $\mathbf{z}$, and $E[\ ]$ denotes an averaging process. We then calculated the eigenvalues and eigenvectors of the cross-spectrum matrix $\mathbf{R}(f)$. The number of signal and noise components was 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.

The steering vector was calculated from the physical arrangement of the microphone array. As shown in Fig. 5, the difference in the 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 a 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 calculated as

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

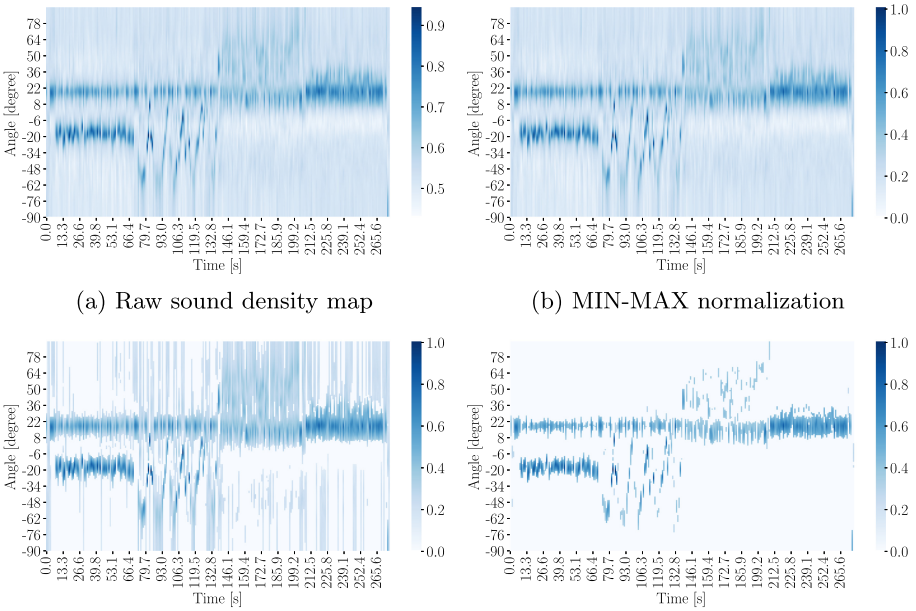
where $\phi = 2\pi f d\sin\theta/c$, and ^{T} denotes the transpose operation. Although we used an example of linearly aligned microphones, the same idea can be used to calculate the steering vector for a different microphone setup.

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

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



(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

Here, we assume that there are k frequency components in the FFT results. $\overline{P_{\mathrm{MUSIC}}}$ has peaks at the angles corresponding to the sound sources. We drew $\overline{P_{\mathrm{MUSIC}}}$ as a function of angle and time to derive a *sound density map*.

We apply a filtering process to the sound density map because the raw sound density map includes sound power information corresponding to the noise components. Figure 6 shows an overview of the filtering process. We first apply a MIN-MAX normalization process (Fig. 6b) and extract the top half points at each time t (Fig. 6c). The top 10% points were 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 does not move quickly, resulting in a continuous band on the sound density map. Multiple bands can be observed on a sound density map when there are multiple sound sources. We apply DBSCAN, a density-based clustering method, to a sound density map to group points on a sound density map into sound sources.

Figure 7 presents an overview of the sound source divider. The clustering process comprises two steps.

In the first step, we extract sound density map points corresponding to sound sources in the room where the smart speaker, that is, the microphone array, is

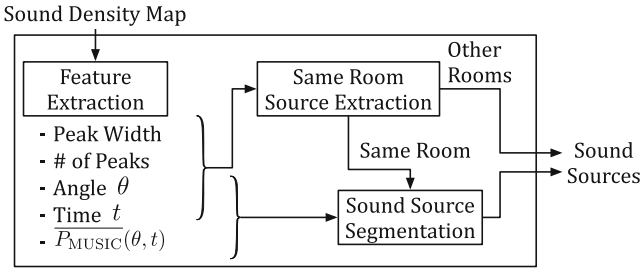


Fig. 7. Overview of sound source divider

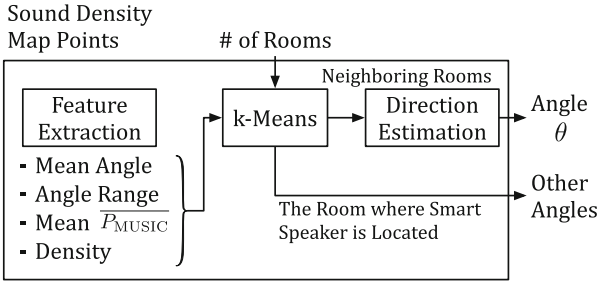


Fig. 8. Overview of room-direction estimator

installed. We perform 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 sound density map points corresponding to a single sound source.

In the second step, the extracted sound density map points, corresponding to sound signals from the room where the smart speaker is installed, are more grouped into sound sources. Sound signals from sound sources in the same room as a 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 $\overline{P_{\mathrm{MUSIC}}}(\theta, t)$. Sound sources in the same room as the smart speaker were 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 can arrive 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, which first groups sound sources into rooms where the sound source is located using k-means clustering. The k-means clustering utilizes four types of features calculated for each sound source: mean angle, the range of angle, mean sound power $\overline{P_{\mathrm{MUSIC}}}(\theta, t)$,

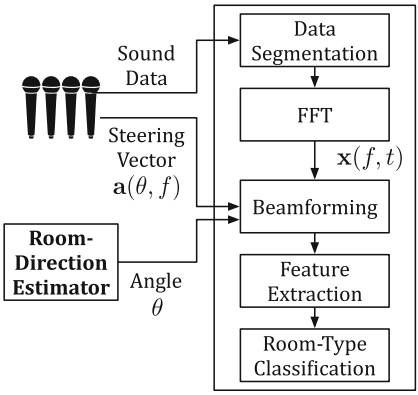


Fig. 9. Overview of room-type estimator

and the density of the sound density map points. The density is the ratio of the number of sound density map points to the area size of the rectangle where the sound density map points are located; k is set as 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 there. The smart speaker co-located room was easily estimated based on the angle variance.

3.7 Room-Type Estimator

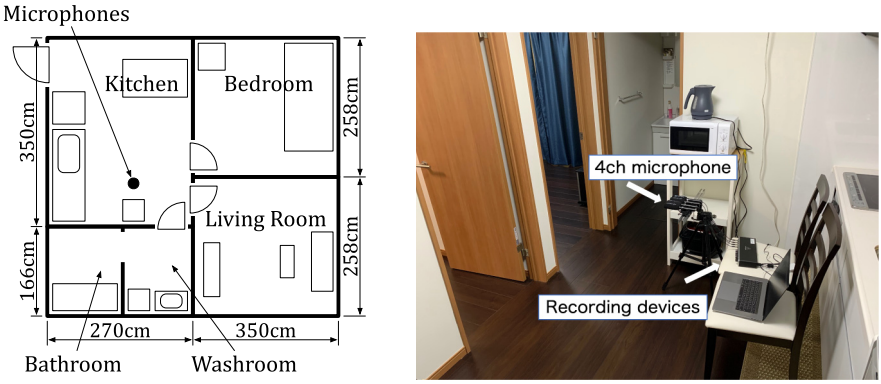
Figure 9 shows an overview of the room-type estimator, which synthesizes the sound signals in the same room and estimates the room-type using supervised learning. The room-type estimator first calculates the frequency components $\mathbf{x}(f,t)$, which is the same process as in the sound mapper. The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

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

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 calculated the basic statistics (i.e., mean, maximum, minimum, and variance) of six types of 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



(a) Room layout (b) Actual microphone setup

Fig. 10. Experiment setup

3. RMS: root mean square of sound signals
4. Spectral flatness: a measure of noise-like sounds noise-like [6,9]
5. Spectral centroid: barycenter of the spectrum [9]
6. Spectral roll-off: frequency such that the major components of the sound energy are contained below this frequency [9]

The room-type estimator finally classifies the room type of each room using the 100-dimensional feature vectors. We did not limit the classifier algorithm. We used a random forest classifier in this study as an example. The classifier model is trained in advance using sound data collected in a typical residential environment and is 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 the 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 the actual microphone setup in our smart house. A 4-channel microphone array, that is, 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 a height of 0.7 m, as shown in Fig. 10a. Sound data were collected using a Behringer UMC404HD USB audio interface connected to a laptop at a sampling rate of 44.1 kHz with a code length of 16 bits.

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)

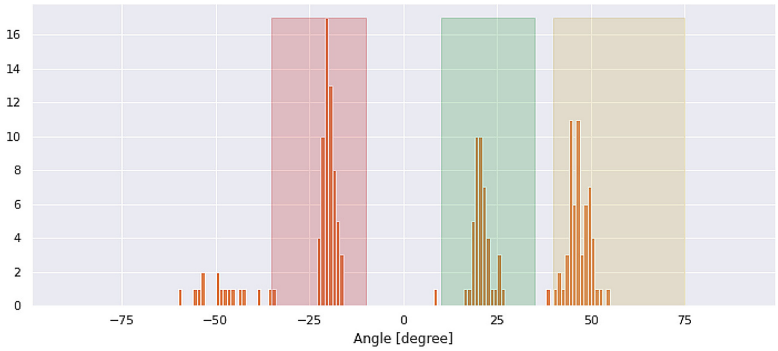


Fig. 11. Histogram of room-direction estimation results

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

Room-direction estimation performance was evaluated with respect to 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 evaluation of clustering performance. Room direction accuracy was evaluated using the rate of the number of trials that correctly estimated the room direction. As shown in Fig. 10, the microphone array was installed in the kitchen, which was next to the living room, bedroom, and bathroom; k in the room-direction estimator, that is, the number of clusters for the k-means, was therefore set to 4.

Figure 11 shows a histogram of the room-direction estimation results. The red, green, and 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, that is, the rate of the number of trials in the red, green, and yellow rectangles in Fig. 11, was 0.850. We can confirm that

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

our room-direction estimator successfully estimated the room direction with no training data.

For reference, the direction estimation accuracy was increased to 0.875 when the trials with ARI were greater than 0.9. Room-direction estimation performance relies heavily on the accuracy of sound source clustering in rooms.

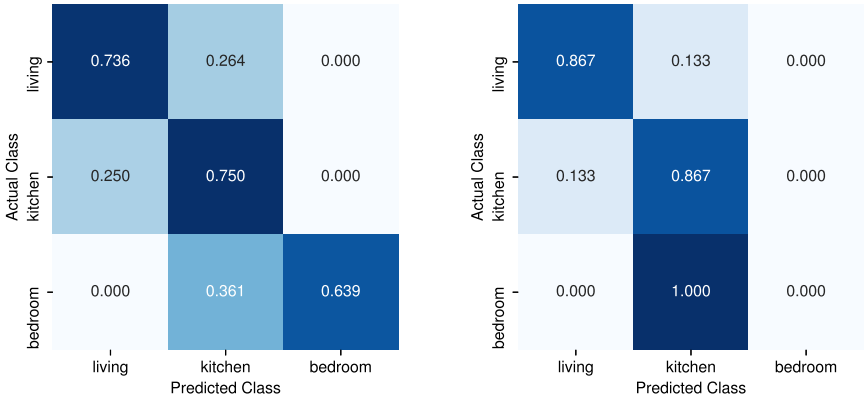
We also compared the performance of room-direction estimation for each dataset. Table 2 shows the mean ARI and direction estimation accuracies for each dataset. From the table, we can observe that the performance with the kitchen dataset was significantly lower than that with other datasets. As shown in Table 1, the kitchen dataset includes the sound signals of subject A in the kitchen, where microphones were installed. Because the sound signals from the room where the microphones were installed can reach the microphones from any direction, the sound source segmentation described in Sect. 3.5 was highly unsuccessful, resulting in s 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 the subject stayed in each room. We installed a microphone at the same height and location, as indicated in Fig. 10, and collected sound data for activity in each room. The sound data were collected both in a

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

controlled environment where the subject performed a specific activity and in a free environment where the subject stayed in a specific room doing activities freely. In the controlled environment, the sound data of each activity shown in Table 3a were collected for 120 s. In the free environment, we collected sound data for 30 min in each room. We emphasize that we gave no instructions for activity during the stay in the free environment. The actual activities during the 30 min are shown in Table 3b.

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

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 the sound data with 10-s windows and calculated the features using the 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 study, 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 the confusion matrices for the controlled and free environments, respectively. The mean accuracies in the controlled and free environments of the room-type estimation were 0.714 and 0.536, respectively. Even though the room type was estimated using the model trained with the data collected in a different environment, we derived a 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 that in other rooms. We can easily guess that the sound power of the bedroom activity shown in Table 3 is relatively low compared to the other room activities, resulting in low estimation accuracy. The room-type estimation accuracy in the bedroom in a controlled environment was 0.639. We believe that a sufficient amount of training data improves the accuracy of room-type estimation.

5 Conclusion

In this study, we presented the design of a room-layout estimator for smart speakers. The room layout, that is, the direction and type of the adjacent rooms, is 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 were grouped into rooms where the sound source was located by unsupervised learning to estimate the room direction. The room type was finally estimated by supervised learning using 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. In 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 large objects, such as furniture, and verify our method for 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. AM '15, Association for Computing Machinery, pp. 1–7. New York, NY, USA (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 Syst. Appl. **75**, 63–79 (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 (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 (2013). https://doi.org/10.1109/IROS.2013.6696919
6. Johnston, J.D.: Transform coding of audio signals using perceptual noise criteria. IEEE J. Selected Areas Commun. **6**(2) (1988). https://doi.org/10.1109/49.608
7. Knapp, C., Carter, G.: The generalized correlation method for estimation of time delay. IEEE Trans. Acoustics Speech Signal Process. **24**(4), 320–327 (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 Sci. Technol. **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 (2004). recherche.ircam.fr/equipes/analyse/synthese/peeters/ARTICLES/Peeters/2003/cuidadoaudiofeatures.pdf
10. Ribeiro, F., Zhang, C., Florencio, D.A., Ba, D.E.: Using reverberation to improve range and elevation discrimination for small array sound source localization. IEEE Trans. Audio Speech Lang. Proc. **18**(7), 1781–1792 (2010). https://doi.org/10.1109/TASL.2010.2052250
11. Schmidt, R.: Multiple emitter location and signal parameter estimation. IEEE Trans. Antennas Propagation **34**(3), 276–280 (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. HLT '90, Association for Computational Linguistics, pp. 151–156. USA (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. J. Acoust. Society 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 (1997). https://doi.org/10.1109/ICASSP.1997.599595
16. Warsitz, E., Haeb-Umbach, R.: Blind acoustic beamforming based on generalized eigenvalue decomposition. IEEE Trans. Audio Speech Lang. Process. **15**(5), 1529–1539 (2007). https://doi.org/10.1109/TASL.2007.898454