Distant-talking speaker identification using a reverberation model with various artificial room impulse responses

Longbiao Wang*, Zhaofeng Zhang†, Atsuhiko Kai† and Yoshiki Kishi‡
* Nagaoka University of Technology, Japan
E-mail: wang@vos.nagaokaut.ac.jp
† Shizuoka University, Japan
E-mail: zhang@spa.sys.eng.shizuoka.ac.jp kai@sys.eng.shizuoka.ac.jp
‡ NS Solutions Kansai Corp., Japan

Abstract—**In this paper, we propose a distant-talking speaker recognition method using a reverberation model with various artificial room impulse responses. These artificial room impulse responses with different speaker and microphone positions, room sizes, and reflection coefficients of walls and convoluted with clean speech are used to train an artificial reverberation speaker model. This artificial reverberation model is also combined with a reverberation speaker model trained with room impulse responses measured in real environments. Speaker identification performance using a combination of the two reverberation speaker models achieved a relative error reduction rate of 50.0% and 78.4% compared with that using a reverberation model trained with real-world room impulse responses and a clean speech model, respectively.**

I. Introduction

Distant-talking speaker identification and verification [1], [2] have recently received increased attention. However, in a distant environment, channel distortion may drastically degrade speaker recognition performance. This is mostly due to the mismatch between the real and training environments. Nevertheless, few studies have analyzed the effects of environmental differences on distant-talking speaker recognition. In [3], the effect of far-field speaker identification under various environments, simulated by artificial room impulse responses (RIR), was analyzed. An analysis of the effect on speaker recognition of different 2D sound source and microphone positions, room sizes, heights of the sound source and microphone, and reflection coefficients of walls, was carried out.

In this paper, various artificial room impulse responses corresponding to a variety of environments (that is, different 2D speaker and microphone positions, room sizes, and reflection coefficients of walls), and which significantly affect speaker identification performance were used to train the artificial reverberation model. The likelihood of this artificial reverberation model was also linearly coupled with that of the reverberation speaker model trained using room impulse responses measured in real environments.

II. Speaker model training under reverberant environments

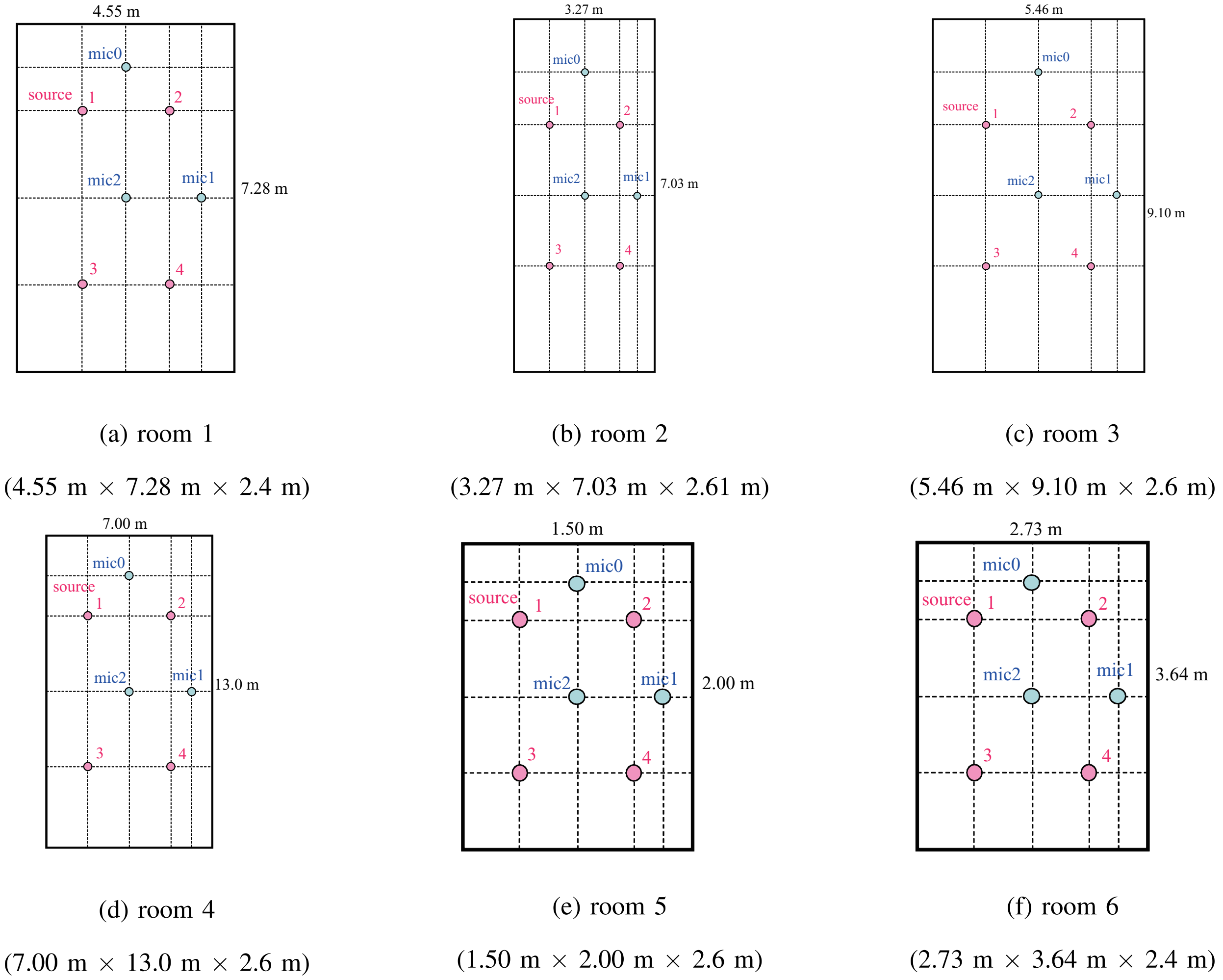
A. Speaker model training using artificial room impulse responses

In this section, we explain how the speaker models were trained using artificial reverberant speech, which was simulated by convolving clean speech with artificial room impulse responses generated by software based on the image method [4], [5]. Room impulse responses were generated from six artificial rooms with different speaker and microphone positions, room sizes, and reflection coefficients of walls as shown in Figs. 1 (a) $\sim$ (f). The heights of the sound source and microphone were 1.72 m and 1.66 m, respectively. The reflection coefficients of the surrounding walls and ceiling were set as 0.80, 0.82, 0.84, 0.86, 0.89, 0.92, 0.94 and 0.95, while the reflection coefficient of the floor was set as 0.8. A variety of rooms from a small bathroom (Fig. 1 (e)) to a relatively large meeting room (Fig. 1 (d)) were simulated.

B. Speaker model training using real-world room impulse responses

In this paper, artificial reverberant speech simulated by convolving four impulse responses measured in real environments from the Real World Computing Partnership (RWCP) database [6] with clean speech were used as test data to evaluate our proposed method. Table I lists the details of the four recording conditions. A single channel signal was taken from a circular + linear microphone array (30 channels). Impulse responses were measured at several positions, 2 m from the microphone array.

For comparison with the speaker model trained using artificial room impulse responses described in Section II-A, reverberant speech simulated by convolving seven impulse



(a) room 1

(4.55 m × 7.28 m × 2.4 m)

(b) room 2

(3.27 m × 7.03 m × 2.61 m)

(c) room 3

(5.46 m × 9.10 m × 2.6 m)

(d) room 4

(7.00 m × 13.0 m × 2.6 m)

(e) room 5

(1.50 m × 2.00 m × 2.6 m)

(f) room 6

(2.73 m × 3.64 m × 2.4 m)

Fig. 1. Configuration of artificial reverberant environments.

responses measured in real environments from the CENSREC-4 database [7] with clean speech were used to train the reverberation speaker model as well. Table II lists the details of the seven recording conditions. For the CENSREC-4 database, a single channel signal was taken from a linear microphone array (7 channels). Impulse responses were measured at several positions, 0.5 m from the microphone array.

III. Experiments

A. Experimental setup

1) Clean speech database: Twenty male speakers from the Tohoku University and Panasonic isolated spoken word database [8], each with a close-up microphone, uttered 200 isolated words. The average time for each utterance was about 0.6 s. From the utterances of each speaker, the first 100 words were used as test data, while the rest were used to train the speaker-specific Gaussian mixture model (GMM) for each speaker. The sampling frequency was 12 kHz. The frame length was 21.3 ms and the frame shift was 8 ms with a 256 point Hamming window. The feature space for a speaker-specific GMM with diagonal matrices was composed of ten mel-frequency cepstral coefficients (MFCCs) with cepstral mean normalization. First-order derivatives of the cepstra plus first derivatives of the power component were also included. The mixture number of the GMMs was 128.

It is difficult to identify the correct speaker using only a single utterance because the average time (about 0.6 s) for each utterance is too short. Therefore, in this paper the combined likelihood of seven utterances (about 4.2 s) was used to identify the speaker.

2) Speaker models: To evaluate the effect of the artificial RIR-based multi-condition training for distant-talking speaker identification under reverberant environments, we created the following speaker models, which differ in that either single clean speech data or multi-condition artificial reverberant speech data were used for training. The multiple artificial reverberant speech data were generated by convolving either the artificial RIRs or the RIRs measured in real environments as described in Section II.

TABLE I

DETAILED RECORDING CONDITIONS FOR IMPULSE RESPONSE MEASUREMENT ON RWCP DATABASE. “RT60 (S)”: REVERBERATION TIME IN ROOM; “S”: SMALL; “L”: LARGE.

#	room	RT60 (s)
1	Tatami-floored room (S)	0.47
2	Tatami-floored room (L)	0.60
3	Conference room	0.78
4	Echo room (panel)	1.30

TABLE II

DETAILED RECORDING CONDITIONS FOR IMPULSE RESPONSE MEASUREMENT ON CENSREC-4 DATABASE.

#	Room	Room size	RT60 (s)
1	Japanese style room	3.5 × 2.5 m	0.40
2	Lounge	11.5 × 27.0 m	0.50
3	Japanese style bath room	1.5 × 1.0 m	0.60
4	Living room	7.0 × 3.0 m	0.65
5	Meeting room	7.0 × 8.5 m	0.65
6	Office	9.0 × 6.0 m	0.25
7	Elevator hall	11.5 × 6.5 m	0.75

- **Clean speech model**: A GMM-based speaker model trained on clean speech data.
- **Artificial RIR-based simulated reverberant speech model**: A GMM-based speaker model trained on multi-condition simulated speech data, generated by convolving the clean speech data with various RIRs of different artificial environments (speaker and microphone positions, room sizes, and sound reflection coefficients of walls).
- **Real-world RIR-based simulated reverberant speech models**: Twelve different GMM-based speaker models trained on multi-condition simulated speech data generated by convolving the clean speech data with various RIRs measured in real environments. The twelve speaker models differ with respect to how many kinds of simulated reverberant speech training data, generated by convolving one of the seven RIRs in the CENSREC-4 database, were used for training the GMMs. The configuration of each model is detailed in Table III.

B. Experimental results

Artificial reverberant speech simulated by convolving four real-world impulse responses from the RWCP database [6] with clean speech were used as test data to evaluate our proposed method.

TABLE III

DETAILS OF CENSREC-4 REVERBERANT ENVIRONMENTS USED FOR REVERBERATION MODEL TRAINING.

number of reverberant environments	details of RIRs
1	one of the 7 RIRs
2	RIRs (1,3) or (6,7)
3	RIRs (2,6,7)
5	RIRs (1,3,4,6,7)
7	all RIRs

TABLE IV

COMPARISON OF SPEAKER RECOGNITION PERFORMANCE FOR ARTIFICIAL REVERBERATION SPEAKER MODEL AND CLEAN SPEECH SPEAKER MODEL (%)

# of reverberant environments	1	2	3	4	Ave.
Clean speech model	73.3	68.2	67.1	60.3	67.2
Artificial RIR-based SRSM	88.5	85.6	83.8	81.2	84.8

1) Comparison between clean speech model and reverberant speech models based on artificial RIRs and real-world RIRs: The speaker identification results for the clean speech model and the artificial RIR-based simulated reverberant speech model (SRSM) are shown in Table IV. “# of reverberant environments” in Table IV corresponds to “#” in Table I. The proposed artificial RIR-based SRSM simulates various reverberant environments, and is thus able to mitigate the degradation of speaker recognition performance on distant-talking speech. The proposed method outperforms the clean speech model for all reverberant environments and reduces speaker identification errors by more than half.

The speaker identification results for the artificial RIR-based SRSM and the real-world RIR-based SRSM are compared in Fig. 2. The real-world RIR-based SRSM was trained on simulated reverberant speech using a single CENSREC-4 room impulse response as shown in the second line of Table III. The recognition results for the artificial RIR-based SRSM (84.8%) were similar to the average results for the real-world RIR-based SRSM (85.9%), but were better than those for the real-world RIR-based SRSMs trained on the 3rd or 6th CENSREC-4 environment.

2) Distant-talking speaker recognition by combining artificial RIR-based SRSM with real-world RIR-based SRSM: Combining the artificial RIR-based SRSM and the real-world RIR-based SRSM is proposed and evaluated in this section. The likelihoods of the above two simulated reverberant speech models are linearly coupled to produce a new score L^n_{comb}

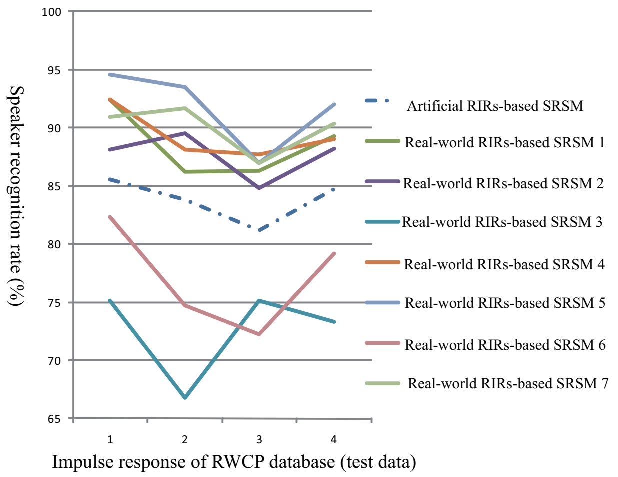


Fig. 2. Comparison of speaker recognition for artificial and real-world RIR-based SRSMs.

given by

$$L^n_{comb} = (1-\alpha)L^n_{artificial} + \alpha L^n_{real-world}, n = 1, 2, \cdots, N, \tag{1}$$

where $L^n_{artificial}$ and $L^n_{real-world}$ are the likelihood products of the *n-th* artificial RIR-based SRSM and real-world RIR-based SRSM, respectively. N is the number of registered speakers and α denotes the weighting coefficients. The speaker with the maximum likelihood was considered as the target speaker.

The distant-talking speaker identification results of the combination method are shown in Fig. 3. The result on the far left of Fig. 3 is that for the artificial RIR-based SRSM only. This result shows that the greater the number of real-world impulse responses is, the better is the speaker recognition performance. However, the cost of measuring various real-world impulse responses is greater than that of artificial impulse responses. When using only one real-world impulse response, the combination method (92.9%) achieves an average relative error reduction rate of 77.1% compared to the clean speech model (67.2%), 50.7% compared to the artificial RIR-based SRSM (84.8%), and 50.0% compared to the real-world RIR-based SRSM (85.9%).

IV. Conclusions and Future Works

In this paper, we proposed a distant-talking speaker recognition method using an artificial RIR-based simulated reverberant speech model. Various artificial room impulse responses with different speaker and microphone positions, room sizes, and reflection coefficients of walls were convolved with clean speech to simulate the different reverberant environments. The likelihood of the artificial RIR-based SRSM was also linearly coupled with that of the real-world RIR-based SRSM. Speaker identification performance based on the combination method achieved a relative error reduction rate of 50.0% compared with that using the real-world RIR-based SRSM only.

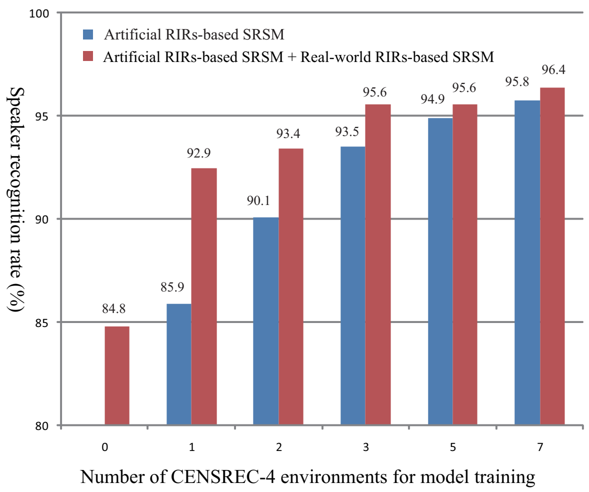


Fig. 3. Speaker identification performance based on the combination of the artificial RIR-based SRSM and real-world RIR-based SRSM.

In the future, we intend extending our proposed methods to include real-world speech data and to perform dereverberation before feature extraction.

References

[1] L. Wang, N. Kitaoka and S. Nakagawa, "Robust distant speaker recognition based on position-dependent CMN by combining speaker-specific GMM with speaker-adapted HMM," *Speech Communication*, Vol. 49, No.6, pp. 501–513, June 2007.

[2] Q. Jin, T. Schultz and A. Waibel, "Far-field speaker recognition," *IEEE Trans. ASLP*, Vol. 15, No. 7, pp. 2023–2032, 2007.

[3] K. Yoshiki, L. Wang and A. Kai, "Effect Analysis of Environmental Differences on Hands-free Speaker Recognition by Using Artificial Room Impulse Response", Proc. of the 2011 Spring Meeting of the Acoustical Society of Japan, 2-P-20, Mar. 2011 (in Japanese).

[4] Jont Allen and David Berkley, "Image Method for Efficiently Simulating Small Room Acoustics", Journal of the Acoustic Society of America, pp. 912-915, April 1979.

[5] S. G. McGovern, "A model for room acoustics", http://2pi.us/rir.html

[6] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura and T. Yamada, "Acoustical Sound Database in Real Environments for Sound Scene Understanding and Hands-Free Speech Recognition, " Proc. of LREC2000, pp. 965-968, May. 2000.

[7] T. Nishiura et al., "Evaluation Framework for Distant-talking Speech Recognition under Reverberant Environments, " Proc. of INTERSPEECH-2008, pp. 968-971, Sep. 2008.

[8] S. Makino, K. Niyada, Y. Mafune and K. Kido, "Tohoku University and Panasonic isolated spoken word database", Journal of the Acoustical Society of Japan, Vol. 48, No. 12, pp. 899–905, Dec. 1992. (in Japanese)