# Acoustic Simulation in Dynamic Environments for Robot Audition

Zhaofeng Zhang*, Kazuhiro Nakadai†, Hirofumi Nakajima‡ and Naoaki Sumida†
*Honda R&D Co., Ltd., Japan, zhaofeng_zhang@n.w.rd.honda.co.jp
†Honda Research Institute Japan Co., Ltd., Japan, {nakadai, naoaki.sumida}@jp.honda-ri.com
‡Kogakuin University, Japan, nakajima@cc.kogakuin.ac.jp

***Abstract*—This paper addresses acoustic simulation in dynamic environments for robot audition. For such environments, we consider three cases, that is, a moving microphone, a moving sound source, and a combination of the two. The proposed method simulates a dynamic environment by assuming that a motion trajectory of a microphone and/or a speaker can be discretized. We validated the proposed method through the accuracy of the simulated signals in terms of frequency and volume, and the performance of automatic speech recognition (ASR) with an acoustic model trained by simulated speech signals. The experimental results showed that the proposed method can simulate the sound properties of volume and frequency in dynamic environments well. The performance of ASR is improved with the acoustic model trained with the simulated speech signals.**

***Index Terms*—acoustic simulation, robot audition, moving sound source, moving microphone, dynamic environment, robust automatic speech recognition**

## I. Introduction

Robot audition has been studied for many years [1]. However, most studies assumed that environments were stationary. In practice, most environments are dynamic. Hereafter, a scene in which there is speaker and/or microphone in motion will be referred to as a dynamic environment. Automatic speech recognition (ASR) is an important function of robot audition because a robot should verbally communicate with people. In dynamic environments where such a robot is operated, ASR becomes a challenging issue. An effective way for this issue is to train an acoustic model using a large amount of speech data which is recorded in similar acoustic conditions [2]. However, collecting such a large amount of data in dynamic environments is time consuming. Simulating such kind of speech is considered a feasible way to solve this problem.

Due to the continuous change in the spatial relationship between the sound source and the microphone, the characteristics of the recorded sound signal also change. Room Impulse Response (RIR) describes the transfer function between the sound source and the microphone [3]. There are several methods available for the estimation of RIR [4]–[8] in a stationary case. By using the information of RIR and the trajectory of motion, an acoustic simulation method for a moving sound source has been discussed by Matsumoto et al. [9] and Nakajima et al. [10]. However, according to the authors' knowledge, how to simulate both a sound source and a microphone in motion has not been studied yet.

In this study, we propose an acoustic simulation method for simulating recordings in a dynamic environment where the sound source and/or the microphone is in motion. Note that it is difficult to simulate a moving signal in a continuous space, the signal discussed in this study will be sampled in a discrete space. The RIR at each sampling point of the moving trajectories can be calculated by using an open source RIR generation toolkit [4]. We construct an RIR matrix representation for the discrete trajectory. The sound signal captured during the motion can then be simulated.

We validated the proposed method in terms of frequency and volume characteristics. For frequency, we attained a mean error of $0.032$ Hz compared to the ground truth obtained from the Doppler effect. For our volume experiment, we observed that the errors between simulated sounds and recorded sounds maintained to be the same levels even when the motion speed was as high as $1.5$ m/s. In addition, we used our simulated data to train an acoustic model for ASR in dynamic environments. The word error rate (WER) was reduced $18.10\%$ relative to the baseline.

## II. Related work

In robot audition, beamforming [11], sound source localization [12], and tracking [13] techniques have been commonly studied to handle distortion of speech signal processing in dynamic environments. However, motion brings volume and frequency distortion, the above-mentioned methods are difficult to mitigate this type of distortion. Acoustic simulation in dynamic environments will provide a different perspectives to study that type of distortion.

In [14], Chowning discussed a method to simulate a moving sound source. However, the information of room reverberation and Doppler shift should be calculated and explicitly added.

In [10], Nakajima et al. discussed how to simulate a moving sound source by using RIR along the sampling points of moving trajectory. In this method, the moving sound source was treated as discrete sampling points on the moving trajectory. Sound is simulated at each sampling point by assuming that the sound is stationary. Since the RIR simulation method has been well studied, by using a mature RIR generation tool, sound can be simulated to contain features of room reverberation, frequency change, and volume change without using other additional information.

978-9-0827-9703-9/19/$31.00 

In this study, we extend Nakajima's acoustic simulation method to support a moving microphone. Because in the human-robot speech communication, the movement of both the speaker and the microphone should be considered.

## III. Acoustic simulation for dynamic environments

In this section, Nakajima's acoustic simulation method for dynamic environments is firstly introduced [10]. Then we discuss our acoustic simulation method for a moving microphone. Finally, a combination of moving sound source and moving microphone is proposed.

### A. Simulation for a moving sound source

When a sound source $s(t)$ located at $p$, a microphone $m$ located at $q$, and the RIR between the sound source and the microphone $h_{pq}(t)$ are given, the observed signal $x_{pq}(t)$ captured with $m$ is defined by,

$$x_{pq}(t) = \sum_{n=0}^{N-1} h_{pq}(n-t)s(t), \tag{1}$$

where $t$ is the time index, and $N$ is the length of $h_{pq}$.

Eq. (1) works only when the sound source and microphone are stationary. To relax this limitation, by defining the trajectory of the sound source as $\hat{p} = [\hat{p}(0), \cdots, \hat{p}(t), \cdots, \hat{p}(T-1)]^T$, the moving sound source can be treated as many discretized point sound sources located along the trajectory. Each point sound source is expressed by,

$$\begin{aligned}
\boldsymbol{s}_{\hat{p}(0)} &= [s(0), 0, 0, \ldots, 0]^T, \\
\boldsymbol{s}_{\hat{p}(1)} &= [0, s(1), 0, \ldots, 0]^T, \\
&\vdots \quad\quad \vdots \\
\boldsymbol{s}_{\hat{p}(t)} &= [0, 0, \ldots, s(t), \ldots, 0]^T, \\
&\vdots \quad\quad \vdots \\
\boldsymbol{s}_{\hat{p}(T-1)} &= [0, 0, \ldots, s(T-1)]^T.
\end{aligned} \tag{2}$$

The observed signal of these point sound sources, $\boldsymbol{x}_{p(t)}$, can be calculated by using Eq. (1). The observed signal $\boldsymbol{x}_{\hat{p}q}$ is the summation of all $\boldsymbol{x}_{p(t)}$. It is also defined as the form of matrix multiplication,

$$\boldsymbol{x}_{\hat{p}q} = \boldsymbol{H}_{\hat{p}q}\boldsymbol{s}, \tag{3}$$

$$\boldsymbol{s} = [s(0), s(1), s(2), \ldots, s(t), \ldots, s(T-1)]^T,$$

$$\boldsymbol{H}_{\hat{p}q} = \begin{bmatrix}
h_{\hat{p}(0)q}(0) & 0 & \ldots & 0 \\
h_{\hat{p}(0)q}(1) & h_{\hat{p}(1)q}(0) & & \vdots \\
\vdots & h_{\hat{p}(1)q}(1) & & \ddots \\
 & \vdots & & \\
h_{\hat{p}(0)q}(N-1) & h_{\hat{p}(1)q}(N-2) & & \vdots \\
0 & h_{\hat{p}(1)q}(N-1) & & \\
\vdots & \ldots & & h_{\hat{p}(T-1)q}(0) \\
 & & & \vdots \\
0 & 0 & \ldots & h_{\hat{p}(T-1)q}(N-1)
\end{bmatrix},$$

where $h_{\hat{p}(t)q}(n)$ is the $n$-th sample of the impulse response between the sound source located at $\hat{p}(t)$ and a microphone located at $q$.

### B. Simulation for a moving microphone

We propose an acoustic simulation method to support a moving microphone by extending Nakajima's method. We define a trajectory of a moving microphone as $\hat{q} = [\hat{q}(0), \cdots, \hat{q}(t), \cdots, \hat{q}(T+N-2)]^T$, and a sound source is located at $p$. The problem can be defined in a similar discrete form. The recorded signal of the moving microphone at each point of $\hat{q}(t)$ can be calculated by,

$$x_{\hat{q}(t)} = \boldsymbol{H}_{p\hat{q}(t)}\boldsymbol{s}[t], \tag{4}$$

where $\boldsymbol{H}_{p\hat{q}(t)}$ is the impulse response matrix at position $\hat{q}(t)$, $\boldsymbol{s}[t]$ means the $t$-th element of the vector $\boldsymbol{s}$. The recorded signal of the moving microphone can be expressed as,

$$\boldsymbol{x}_{p\hat{q}} = [x_{\hat{q}(0)}, x_{\hat{q}(1)}, \ldots, x_{\hat{q}(k)}, \ldots, x_{\hat{q}(T+N-2)}]^T. \tag{5}$$

$\boldsymbol{x}_{p\hat{q}}$ can also be defined as the form of matrix multiplication.

$$\boldsymbol{x}_{p\hat{q}} = \boldsymbol{H}_{p\hat{q}}\boldsymbol{s}, \tag{6}$$

$$\boldsymbol{H}_{p\hat{q}} = [\boldsymbol{h}_{\hat{q}_0}, \boldsymbol{h}_{\hat{q}_1}, \ldots, \boldsymbol{h}_{\hat{q}_k}, \ldots, \boldsymbol{h}_{\hat{q}_{T+N-2}}]^T.$$

$$\boldsymbol{h}_{\hat{q}_k} = \begin{cases}
\big[h_{\hat{q}(k)p}(k), h_{\hat{q}(k)p}(k-1), \cdots, \\
\quad h_{\hat{q}(k)p}(0), \boldsymbol{0}(T-k-1)\big]^T, & (k < N-1) \\
\big[\boldsymbol{0}(k-N+1), h_{\hat{q}(k)p}(N-1), \\
\quad h_{\hat{q}(k)p}(N-2), \cdots, \\
\quad h_{\hat{q}(k)p}(0), \boldsymbol{0}(T-k-1)\big]^T, & (N-1 \le k < T-1) \\
\big[\boldsymbol{0}(k-N+1), h_{\hat{q}(k)p}(N-1), \cdots, \\
\quad h_{\hat{q}(k)p}(k+1-T)\big]^T, & (k \ge T-1)
\end{cases}$$

where $h_{\hat{q}(k)p}(n)$ is the $n$-th sample of the RIR between microphones at $\hat{q}(k)$ and the sound source located at $p$. $\boldsymbol{0}(Z)$ represents a $Z$-dimentional zero vector

### C. Extension to support both a microphone and a sound source in motion

For this problem. The ideas used in sections III-*A* and III-*B* can be considered as, firstly assuming that the sound source is in motion and the microphone is stationary. Then we introduce the case when the microphone is in motion. By a combination of Eq. (3) and Eq. (6), we define a trajectory of a moving sound source as $\hat{p}$ and a moving microphone as $\hat{q}$. The observed signal with both the microphone and the sound source in motion is defined as,

$$\boldsymbol{x}_{\hat{p}\hat{q}} = \boldsymbol{H}_{\hat{p}\hat{q}}\boldsymbol{s}, \tag{7}$$

$$\boldsymbol{H}_{\hat{p}\hat{q}} = [\boldsymbol{h}_0, \boldsymbol{h}_1, \cdots, \boldsymbol{h}_k, \cdots, \boldsymbol{h}_{T+N-2}]^T,$$

$$\boldsymbol{h}_k = \begin{cases}
\big[h_{\hat{q}(k)\hat{p}(0)}(k), h_{\hat{q}(k)\hat{p}(1)}(k-1), \cdots, \\
\quad h_{\hat{q}(k)\hat{p}(k)}(0), \boldsymbol{0}(T-k-1)\big]^T, & (k < N-1) \\
\big[\boldsymbol{0}(k-N+1), h_{\hat{q}(k)\hat{p}(k-N+1)}(N-1), \\
\quad h_{\hat{q}(k)\hat{p}(k-N+2)}(N-2), \cdots, \\
\quad h_{\hat{q}(k)\hat{p}(k)}(0), \boldsymbol{0}(T-k-1)\big]^T, & (N-1 \le k < T-1) \\
\big[\boldsymbol{0}(k-N+1), h_{\hat{q}(k)\hat{p}(k-N+1)}(N-1), \cdots, \\
\quad h_{\hat{q}(k)\hat{p}(T-1)}(k+1-T)\big]^T, & (k \ge T-1)
\end{cases}$$

where $h_{\hat{q}(k)\hat{p}(k)}(n)$ is the $n$-th sample of the RIR between the microphone at $\hat{q}(k)$ and the sound source located at $\hat{p}(k)$.

## IV. Evaluation

### A. Experimental setting

In this section, the following three terms were evaluated: **frequency**, **volume**, and **ASR**.

The properties of sounds are changed due to movement. The frequency change of sound obeys the Doppler effect. The change in the distance between the sound source and the microphone causes that in volume. Therefore, the proposed methods were evaluated in terms of frequency and volume. For frequency, the signal simulated with a fixed moving speed is compared to the theoretical value under the Doppler effect. For volume, two kinds of distance measures were compared. One was the distance between simulated and recorded speech signals, and the other was the one between the recorded and the original speech signals. We also performed an experiment to evaluate the performance of ASR in dynamic environments.

For all experiments, we prepared seven kinds of sound datasets, all data was stored in wave form and the sample rate was set as $8$ kHz.

- *D1*: Simulated pure tones.
- *D2*: Down-sampled WSJ corpus (evaluation set).
- *D3*: Down-sampled WSJ corpus (30 sentences selected from the evaluation set).
- *D4*: Recorded speech signals for *D3*.
- *D5*: Simulated speech signals for *D4*.
- *D6*: Down-sampled WSJ corpus (training set).
- *D7*: Simulated speech signals for *D6*.

*D1* is used for frequency evaluation. We prepared three kinds of pure tones as the sound sources. The frequency's settings are $250$ Hz, $1$ kHz, and $4$ kHz. The simulation was performed at nine kinds of fixed speeds for each frequency setting. The speed's settings are $0.1, 0.2, 0.5, 1, 2, 5, 10, 20$ and $50$ m/s. We also considered three different movement patterns: a moving sound source, a moving microphone and both on the move. Hence 3 sounds with totally 3 (frequency) $\times$ 9 (speeds) $\times$ 3 (motion patterns) $=$ 81 kinds of motion settings were simulated.

*D3*, *D4* and *D5* are used for volume evaluation. We recorded the speech signal of *D4* played from a loudspeaker in motion in an anechoic room ($6.2\,\text{m} \times 4.8\,\text{m} \times 5.1\,\text{m}$). For the recording, a person held a microphone and another held a speaker. They walked along a straight line. We also used a camera to capture the recording situation, to confirm the location of the moving microphone and the moving sound source. The speech was recorded in three kinds of speed settings. The speeds are around $0.7$ m/s, $1.0$ m/s, and $1.5$ m/s. For each speed setting, ten sentences from the Wall Street Journal (WSJ) speech corpus [15] were selected as the sound sources (*D3*). The microphone and the speaker move as follows:

- Both microphone and speaker were stationary (**st**).
- The speaker was stationary and the microphone was moving (**mic**).
- The microphone was stationary and the speaker was moving (**ss**).
- Both the speaker and the microphone were moving towards the same directions (**both**).
- Both the speaker and the microphone were moving towards opposite directions (**both (oppo)**).

Hence, the number of utterances in *D4* are 3 (speeds) $\times$ 5 (motion patterns) $\times$ 10 (utterances) $=$ 150. We manually obtained the motion trajectory of the microphone and the speaker in *D4* using the captured video. Using the obtained trajectory, we used *D3* to simulate the moving sound sources and microphones. It was *D5*.

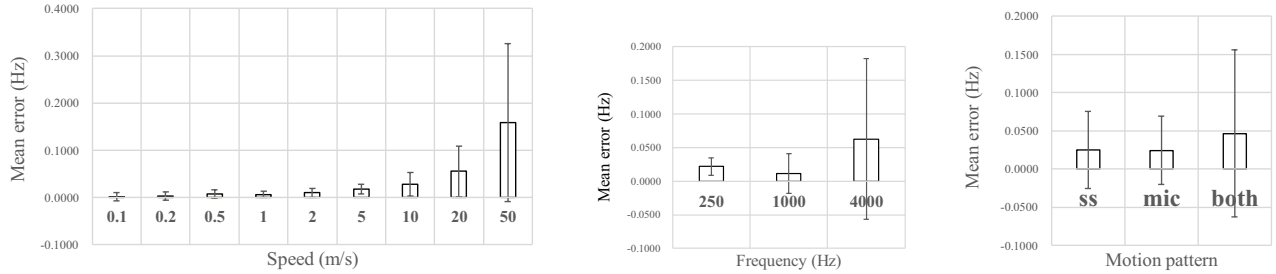
*D2*, *D3*, *D4*, *D6* and *D7* are used for the ASR evaluation. *D7* is generated by down-sampled original training dataset of the WSJ corpus (*D6*), It contains 37,416 utterances. The total amount of training data is approximately 80 hours. For each utterance, one of 10,240 motion patterns was randomly selected. These motion patterns had different directions of movement, speeds and starting position for each microphone and sound source. The details of these motion patterns are illustrated in Fig. 2. *D2*, *D3* and *D4* are used for decoding. *D3* is a subset of *D2*.

For the simulation, we used a machine with a CPU of Xeon E5-2687 and a memory of 256GB (GPU was not used). Audio files sampled at $8k$Hz and $16$ bits were used for the simulation. The simulation algorithm was implemented by MATLAB. Twelves threads were used to generate simulation data, and it took around $70$s to generate a $10$s voice sample with the usage of 23 GB memory.

In every experiment, the length of the RIR was set to $0.256$s which were generated based on [4].

### B. The frequency evaluation

The observed frequency of a moving sound obeys the Doppler effect. If we know the speed and frequency of a sound source or a microphone, the theoretical frequency of the recorded sound can be calculated. In this experiment, we measured the frequency of simulated pure tones in motion in *D1*. The frequency of theoretical value was compared with the simulated one. The result is shown in Fig. 1. We calculated the mean frequency error among different motion patterns, frequency, and moving speeds. The proposed method generally showed high simulation performance, but it still produced a small error. The mean error among all actual and theoretical frequency was $0.032$ Hz, and the standard deviation was $0.074$ Hz. It is observed that as the speed increases, the error of frequency also increases in Fig. 1(a). This phenomenon is considered being caused by two reasons. One is the cancellation of digits and the round-off error during computation. Another is caused by discretization assumed with the proposed method. For higher speed, the distance between adjacent sample points increases, which results in a large error. Increasing the sampling rate of sound may mitigate this error, although it is necessary to consider balancing it with computational cost.




(a) Different speed settings. (b) Different frequency settings. (c) Different motion patterns.

Fig. 1. The frequency evaluation results: mean errors were calculated between theoretical values and measured values of simulated sounds. An error bar shows the standard deviation of each mean error. Figure 1.(a), (b), and (c) tallies the results for speeds, frequencies and motion patterns, respectively. In (c), "ss" means that only a moving sound source, "mic" means that only a moving microphone, and "both" means that both are in motion.

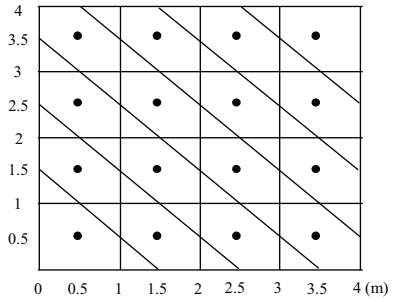



Fig. 2. The motion patterns of *D7*: This figure depicts the position of all lines and points for a microphone and sound source pair. When the microphone moved along one of the lines, the sound source would be stationary at one of the points, and vice versa. Considering computational cost, we do not simulate both a microphone and a sound source in motion. The moving speed was set from 0.3 m/s to 3 m/s in step of 0.3 m/s. The direction of movement could be either of the directions of a line. Relative position of points and lines were measured by the numbers in x-axis and y-axis.

### C. The volume evaluation

We designed another experiment to validate simulation performance in volume. In this experiment, we adopted two kinds of distance measures. One was the distance between a recorded speech signal (*D4*) and their corresponding simulated signal (*D5*, hereafter, "simu-recorded"). The other one was the distance between the original speech signal (*D3*) and the corresponding recorded signal (*D4*, hereafter, "orig-recorded"). In both cases, the distance measures are defined by,

$$D = \frac{1}{FN^2} \sum_{f=1}^{F} \left( \sum_{n=1}^{N} |x_{s/o}(f,n)| - \sum_{n=1}^{N} |x_r(f,n)| \right)^2, \quad (8)$$

where $x_s(f,n)$, $x_o(f,n)$, and $x_r(f,n)$ are amplitude values at the $n$-th sample of the $f$-th frame of the simulated, original, and recorded speech signal, respectively. $N$ is the length of a frame, and $F$ is the number of frames. The smaller $D$ is, the larger the similarity of volume between two signals is.

When calculating $D$, two pre-processing steps were performed. First, the offset of the recorded speech signal was decided to synchronize with the original and simulated speech signals, and then, the amplitude gain of the simulated speech signals was adjusted to be a similar level to the recorded ones. Fig. 3 illustrates the results. Using the two kinds of distance measures for the five motion patterns in three different kinds of speed settings. For all speed settings, it was observed that $D$ of "orig-recorded" became larger when its relative speed increased. It means that the signals were recorded differently from the original signals in a dynamic environment. On the other hand, "simu-recorded" maintains a similar amount of average errors even when its relative speed changes. This indicates that the proposed method can bridge the gap between stationary and dynamic environments in terms of volume.

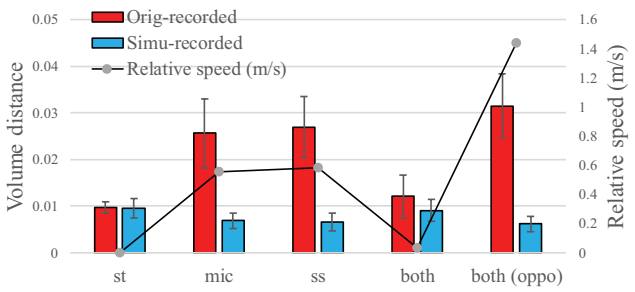
TABLE I
ASR PERFORMACE IN WER: TWO ACOUSTIC MODELS EVALUATED ON TWO KINDS OF DATASETS.

| | Base clean | Base recorded | Proposed recorded |
|---|---|---|---|
| Acoustic model | Clean acoustic model (*D6*) | | Simulated acoustic model (*D7*) |
| Test data | Clean speech (*D3*) | Recorded speech in motion (*D4*) | |
| WER (%) | 17.98 | 32.15 | 26.33 |
| Relative improvement (%, compared with Base reocrded.) | - | - | +18.10 |

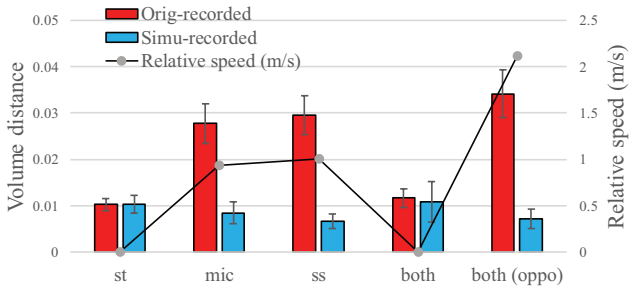
### D. The ASR evaluation

We believe that acoustic models trained in dynamic environments can improve the performance of ASR in similar environments. In this experiment, the proposed method is considered to generate speech data for dynamic environments. We trained two acoustic models. One was trained by using *D6*. The other one was trained by using the *D7*. For evaluation data, the test datasets *D2*, *D3* and *D4* were adapted. "Kaldi" [16] (Version 5.3), an open source ASR toolkit, was employed for ASR evaluation. We selected the recipe of WSJ task without using the big dictionary for language model. It applied sequence-discriminative training of deep neural networks [17] to train an acoustic model. Note that the WER of the down-sampled original training (*D6*) and test (*D2*) data set is 6.33%. Table I shows the result of our ASR evaluation experiment.

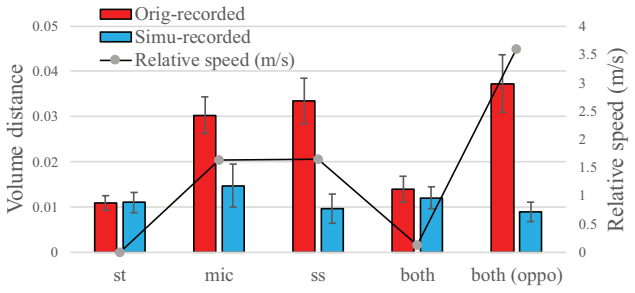
The result of the WSJ task in an original clean environment was shown for comparison (the column of "Base clean" in Table I). The performance in the dynamic environment dropped with acoustic model of original clean speech (the




(a) The speed of about 0.7 m/s.




(b) The speed of about 1.0 m/s.




(c) The speed of about 1.5 m/s.

Fig. 3. The effectiveness of the proposed method in terms of volume. In each figure, the moving speeds for the microphone and sound source pair are set as around 0.7 m/s, 1.0 m/s and 1.5 m/s, respectively. The red and blue bars are the value of $D$ in Eq. (8). The polylines show the relative speed between the sound source and microphone pair. The x-axis refers to the motion patterns of *D4* and *D5*, the y-axis on the left is the scale of $D$, and the y-axis on the right is the scale of relative speed for each motion pattern.

column of "Base recorded"). Using the acoustic model of our simulated speech data the result performed better (the column of "Proposed method"). The relative improvement is 18.10%. It proves that our proposed simulation method could improve the performance of ASR in dynamic environments.

## V. Conclusions and Future Work

We presented an acoustic simulation method for robot audition in dynamic environments. The method can simulate a recorded sound signal where both a microphone and a sound source are in motion. We assume that motions of the microphone and the sound source can be discretized and the sound signal can be constructed with a matrix form using RIRs. We showed that the proposed method can well simulate sound properties in terms of volume and frequency. We also applied our proposed method to the task of ASR. It improved ASR performance in dynamic environments.

For future work, since robot audition usually uses a microphone array for high quality noise reduction, we will extend our research for a microphone array. Besides, algorithms with low computational cost will be discussed.

## Acknowledgment

The authors thank Yuichi Yoshida from Honda research institute Japan, Co,. Ltd. for his help.

## References

[1] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano, "Active Audition for Humanoid," In Proc. of AAAI-2000, pp.832–839, 2000.
[2] L. Wang, Z. Zhang, A. Kai, and Y. Kishi, "Distant-talking speaker identification using a reverberation model with various artificial room impulse responses," In Proc. of APSIPA ASC-2012, pp.1-4, 2012.
[3] H. Kuttruff, "Room acoustics," Crc Press, 2016.
[4] E. A. Habets, "Room impulse response generator," Technische Universiteit Eindhoven, Tech. Rep 2.2.4 (2006): 1, 2006.
[5] E. A. Lehmann, A. M. Johansson, "Diffuse reverberation model for efficient image-source simulation of room impulse responses," IEEE Transactions on Audio, Speech, and Language Processing, 18(6), pp.1429-1439, 2010.
[6] A. B. Jont, D. A. Berkley, "Image method for efficiently simulating small room acoustics," The Journal of the Acoustical Society of America 65.4 (1979): pp.943-950, 1979.
[7] P. M. Peterson, "Simulating the response of multiple microphones to a single acoustic source in a reverberant room," The Journal of the Acoustical Society of America 80.5 (1986): pp.1527-1529, 1986.
[8] D. Murphy, M. Beeson, S. Shelley, A. Moore, et al. "Hybrid room impulse response synthesis in digital waveguide mesh based room acoustics simulation," Proceedings of the 11th International Conference on Digital Audio Effects (DAFx-08) pp.129-136, 2008.
[9] M. Matsumoto, M. Tohyama and H. Yanagawa, "A method of interpolating binaural impulse responses for moving sound images," Acoust. Sci. & Tech. 24, 5, pp.284-292, 2003.
[10] H. Nakajima, K. Nakadai, Y. Hasegawa and H. Tsujino, "Moving Sound Source Extraction by Time-Variant Beamforming," JSAI 2007 Conference and Workshops, pp.47-53, 2007.
[11] J. M. Valin, F. Michaud, J. Rouat, and D. Letourneau, "Robust sound source localization using a microphone array on a mobile robot," In Proc. of IROS 2003, Vol.2, pp.1228-1233, 2003.
[12] K. Nakadai, T. Takahashi, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, "Design and Implementation of Robot Audition System 'HARK' Open Source Software for Listening to Three Simultaneous Speakers," Advanced Robotics, 24(5-6), pp.739-761, 2010.
[13] K. Nakadai, H. G. Okuno, and H. Kitano, "Real-time sound source localization and separation for robot audition," Seventh International Conference on Spoken Language Processing, 2002.
[14] J. M. Chowning, "The simulation of moving sound sources," Computer Music Journal, pp.48-52, 1977.
[15] D. B. Paul, J. M. Baker, "The design for the Wall Street Journal-based CSR corpus." In Proc. of workshop on Speech and Natural Language, pp.357-362, 1992.
[16] D. Povey, et al. "The Kaldi speech recognition toolkit." In IEEE 2011 workshop on automatic speech recognition and understanding (No. EPFL-CONF-192584), IEEE Signal Processing Society, 2011.
[17] K. Vesely, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks." In Proc. of Interspeech, pp.2345-2349, 2013.