Hands-free Speech Recognition Robust to distance and Azimuth in Robot Application

Randy Gomez, Keisuke Nakamura, Takeshi Mizumoto and Kazuhiro Nakadai
Honda Research Institute Co. Ltd., Japan

***Abstract*— In this paper we present two methods in addressing the changes in radial position and azimuth, respectively, relative to the robot and speaker. In the case of the former, room transfer function (RTF) estimation is employed via waveform-level compensation to reflect the change in power caused by the change of radial position to the RTF. In addition, acoustic model-level compensation is also used to complement the effect of robustness to changes in radial position. Finally, equalization is utilized to mitigate the effects of the change in the azimuth. All of the processes in the two methods are in accordance to maximizing the automatic speech recognition (ASR) performance for effective human-robot communication. Experimental evaluation in real environment condition confirms the robustness in recognition performance when used in hands-free human-robot communication.**

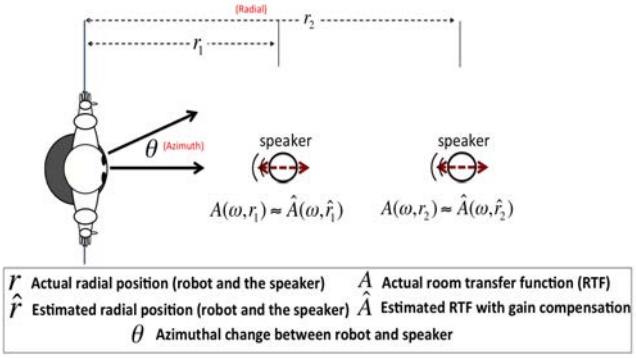

Fig. 1. Room scenario: Change in speaker's radial and azimuthal position.

I. Introduction

Automatic speech recognition (ASR) is one of the most important component in hands-free human-robot communication. Before robots execute any speech-based command, the speech acoustic signal has to be converted into text using the ASR and then, processed by an intelligent system for machine understanding. As the acoustic speech signal travels through free space inside an enclosed room, the observed signal at the microphone is distorted due to reflections known as reverberation. Late reverberation leads to a significant deterioration of the ASR performance. Dereverberation is therefore needed in a robust ASR system. In robot applications, this problem is further complicated since we cannot control the position of the speaker. When a speaker changes its radial position relative to the robot (r_1 to r_2 and vice versa) as shown in Fig. 1, the reverberant speech power observed at the microphones also changes, resulting in a mismatch to the acoustic model used in the ASR system. Thus, the dereverberation algorithm should adapt by estimating the change in location $\hat{r}$ and RTF $\hat{A}$ to minimize the effect of mismatch.

We have previously proposed an enhancement algorithm that automatically detects the reverberation time (RT) inside the room, then synthetically generate a new RTF from the pre-measured one [1][2]. The estimated RTF is used to enhance the reverberant speech by removing its late reflection component in conjunction with our ASR-based dereverberation scheme [1][2]. Although this method works, the assumption is very conservative. First, it assumes that only RT plays a significant role in describing the RTF. This assumption is only valid in symmetric rooms with no occlusions (i.e., chairs, tables, etc.). In real environments where robots are dispatched, it is fair assumption that the room will be filled with objects and speakers. Second, it does not take into consideration the change in speech power. It assumes that the speaker is stationary, which is unrealistic because we cannot control the speaker's radial position. Lastly, it is not immune to changes in the angular position (azimuth) of the speaker relative to the robot. These causes a major problem as the ASR system is very sensitive to the change in speech power (mismatch). We define radial position as the absolute distance between the speaker and the robot while the azimuth is the angle from a reference position (robot) to the speaker.

First, we will show a method that significantly improves the RTF estimation of our previous work [1][2]. The new method is capable of compensating the change in speaker's radial position through room RTF compensation. In effect, the variation in speech power which is crucial in ASR is also considered in the new RTF estimate. As a result, we can achieve a robust ASR performance in realistic environments where robots are deployed. The proposed compensation is done in two-way synergetic processes, accounting for both the waveform and acoustic model aspects that affect ASR performance. In the waveform level, the RTF is compensated in a manner that reflects possible changes in speaker's radial position. This focuses on the impact of the speaker's power variation to the RTF waveform. Consequently, the acoustic model-level compensation connects the waveform-level compensation to the ASR, by adopting the criterion used by the ASR system in estimating the RTF. This guarantees that the estimated RTF translates to ASR performance improvement, when used in conjunction with our ASR-based dereverberation scheme. Secondly, we will show the method in addressing changes in the azimuth via equalization.

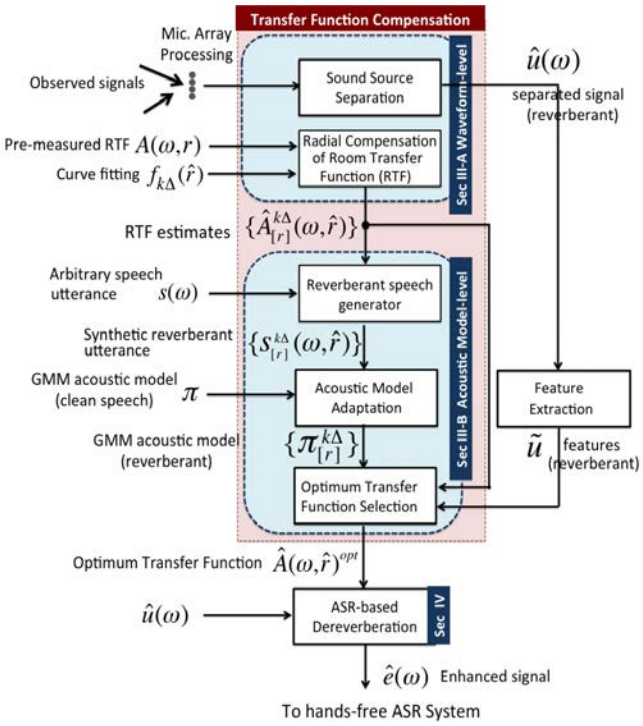

Fig. 2. Proposed RTF estimation for ASR-based dereverberation.

The organization of the paper is as follows; in Section II, the background of RTF estimation for ASR application is introduced. Then, in Section III, we show the method of our proposed RTF estimation with radial compensation, followed by equalization scheme for azimuthal change in Section IV. In Section V, we discuss the experimental set-up, together with recognition results using real reverberant data collected in a human-robot communication environment in Section VI. We will conclude this paper in Section VII.

II. BACKGROUND

A number of dereverberation approaches use readily available RTF through room impulse response measurement [3] [4]. Due to the dynamics inside the room, the generalization of the RTF becomes unrealistic in real environments. This arises the need of physically measuring several RTFs inside the room, which is impractical. Thus, RTF estimation becomes an interesting research topic. There are a number of RTF estimation techniques focusing on waveform accuracy. Although this is important, the requirement for RTF estimation is different when used in hands-free ASR applications. For example, when RTF is used in dereverberation for ASR, we are not interested in accurately modeling the reverberant speech but in estimating the late reflections, in which RTF is not the sole-determining factor. In ASR, we convert speech waveform to models (i.e., Hidden Markov Models (HMMs)). This process requires a conversion of a rich signal information to a more watered-down representation. Each phone HMM represents a short speech segment with a duration of 30-100 msec, and each state captures information about a distribution of spectral parameters. With this perspective, the HMMs' description of speech is of low resolution, compared to the RTF, with respect to time and frequency. Thus, for ASR application, it may be sufficient to use an RTF estimate instead of the accurate RTF [6]. In our previous work [1][2] we adopted an RTF estimator based on the premise that the multiple reflections of sound can be described by a decaying acoustical energy [5] given as,

$$A^2(l) \approx e^{(6\ \ln(10)/RT)\ l}, \tag{1}$$

where l is the discrete time sample, and RT is the reverberation time. From Eq. (1) we can easily derive the RTF's frequency domain equivalence $\hat{A}(\omega)$, where ω denotes frequency domain. We note that this RTF estimate does not take into consideration the distance between the speaker and the robot. This model is applicable only if there is not much perturbation inside the room. Moreover, it fails when the speaker moves along the radial axis or changes angular position relative to the robot because it does not model the variation in speech power as a function of position change, which the ASR is very sensitive to. Thus, we need a new RTF estimation method that takes care of the change in the power of the reverberant speech to minimize mismatch in the ASR system.

There is no guarantee whether an estimated RTF (even using the most accurate ones) would result to an improvement in the ASR performance when used in conjunction with any dereverberation scheme for hands-free human-robot communication [1][2]. RTF estimation should be based on the practical requirements for which it is to be utilized. In the case of hand-free human-robot ASR application, the design criterion should not be based on the waveform accuracy of the RTF estimate. It is prudent to adopt the criterion used by the ASR system (i.e., acoustic model likelihood criterion) as part of the RTF estimation criterion, which is the fundamental objective of this paper.

III. ROBUSTNESS TO RADIAL POSITION

The proposed RTF estimation method is shown in Fig. 2. First, the microphone array signals are processed resulting in $\hat{u}(\omega)$. Then, RTF is estimated by compensating the pre-measured RTF $A(\omega,r)$ together with the curve fitting function derived offline $f_{k\Delta}(\hat{r})$. The step-size increment $k\Delta$ generates a set of RTFs $\{A_{[r]}^{k\Delta}(\omega,\hat{r})\}$. Using these RTFs, together with an arbitrary clean speech utterance $s(\omega)$, a set of synthetic reverberant data are generated $\{s_{[r]}^{k\Delta}(\omega,\hat{r})\}$. Consequently, the clean acoustic model π (Gaussian Mixture Model (GMM)) is adapted using the generated reverberant data resulting to adapted models $\{\pi_{[r]}^{k\Delta}\}$. Then, the optimum RTF is selected by evaluating the likelihood scores using the features $\tilde{u}$ of the separated reverberant signal $\hat{u}(\omega)$. The corresponding k that maximizes the likelihood score is used to select among $\{A_{[r]}^{k\Delta}(\omega,\hat{r})\}$ the optimal RTF $\hat{A}(\omega,\hat{r})^{opt}$. This optimal RTF is then used in conjunction with the ASR-based dereverberation scheme.

A. Acoustic Waveform Compensation

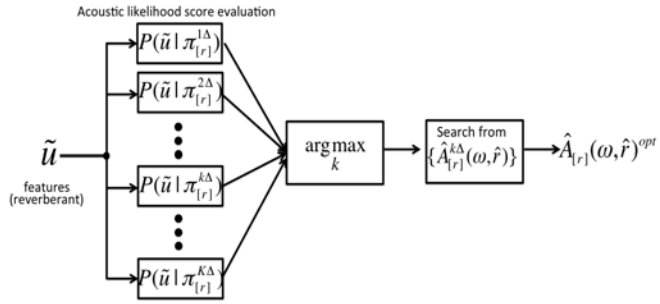

Fig. 3. Optimum RTF selection based on maximum likelihood estimation.

1) Microphone Array Sound Separation: Suppose that there are N sources and M ($\geq N$) microphones. Let $\boldsymbol{u}(\omega)$ denote the input acoustic signals of N sources in frequency domain, described as $\boldsymbol{u}(\omega) = [u_1(\omega), \cdots, u_N(\omega)]^T$, where T represents the transpose operator. $\boldsymbol{x}(\omega) = [x_1(\omega), \cdots, x_M(\omega)]^T$ is the vector containing the signals received by M microphones. The model for microphone array signal processing is described as follows:

$$\boldsymbol{x}(\omega) = \boldsymbol{A}(\omega)\boldsymbol{u}(\omega) + \boldsymbol{n}(\omega), \tag{2}$$

where $\boldsymbol{A}(\omega) \in \mathbb{C}^{M \times N}$ is a *Transfer Function (TF)* matrix between a microphone array and sound sources; $\boldsymbol{n}(\omega)$ denotes an additive noise, which is assumed to be statistically independent of $\boldsymbol{u}(\omega)$.

The sound sources are spatially separated by a hybrid algorithm of beamforming and blind separation called *Geometrically constrained High-order Decorrelation based Source Separation (GHDSS)*. The input vector $\boldsymbol{x}(\omega)$, $\hat{\boldsymbol{u}}(\omega)$ is used to define by $\hat{\boldsymbol{u}}(\omega) = \boldsymbol{V}(\omega)\boldsymbol{x}(\omega)$ in frequency domain, where $\boldsymbol{V}(\omega)$ stands for the separation matrix. GHDSS updates $\boldsymbol{V}(\omega)$ so that it can correctly estimate $\boldsymbol{u}(\omega)$ in Eq. (2) by $\hat{\boldsymbol{u}}(\omega)$. In order to estimate $\boldsymbol{V}(\omega)$, GHDSS introduces two cost functions, that is, separation sharpness (J_{SS}) and geometric constraints (J_{GC}):

$$J_{SS}(\boldsymbol{V}(\omega)) = \|\phi(\hat{\boldsymbol{u}}(\omega))\hat{\boldsymbol{u}}^H(\omega) - \mathrm{diag}[\phi(\hat{\boldsymbol{u}}(\omega))\hat{\boldsymbol{u}}^H(\omega)]\|^2$$
$$J_{GC}(\boldsymbol{V}(\omega)) = \|\mathrm{diag}[\boldsymbol{V}(\omega)\boldsymbol{A}(\omega) - \boldsymbol{I}]\|^2$$

where $\|\cdot\|^2$ indicates the Frobenius norm, diag$[\cdot]$ is the diagonal operator, and H represents the conjugate transpose operator. For a nonlinear function, $\phi(\hat{\boldsymbol{u}}(\omega))$, we selected a hyperbolic-tangent-based function [7] in this paper. Since the best $\boldsymbol{V}(\omega)$ is always changing in the real world, $\boldsymbol{V}(\omega)$ is adaptively updated as described in [8]. Consequently, the separated signal $\hat{u}(\omega)$ that satisfies all the criterion is achieved from the vector $\hat{\boldsymbol{u}}(\omega)$ in the same manner in [8].

2) Radial Compensation of Room Transfer Function: Let $A(\omega, r)$ denote pre-measured RTF between any of the microphones in the array and sound sources. Here, r is the distance between a microphone in the array and the sound sources. Our objective is to estimate $A(\omega, \hat{r})$ by radial compensation, where $\hat{r}$ is the radial distance of an estimated point. The following conditions are assumed:

a) The sound sources are in a far field. This implies that the phase of the RTF is static with respect to the change in distance.
b) The amplitude of $|A(\omega, \hat{r})|$ decays exponentially with $\hat{r}$.

Under these assumptions, RTF of unknown distance can be estimated as follows:

$$\hat{A}_{[r]}(\omega, \hat{r}) = f(\hat{r})A(\omega, r) \ , \tag{3}$$

where $\hat{A}_{[r]}(\omega, \hat{r})$ is an estimated RTF at $\hat{r}$ using pre-measured $A(\omega, r)$. $f(\hat{r}) \in \mathbb{R}$ is the exponential gain function of $\hat{r}$. Since $f(\hat{r})$ is unknown, $f(\hat{r})$ is obtained as a priori information based on a nonlinear curve fitting using measured RTFs. Specifically, $f(\hat{r})$ is described by

$$f(\hat{r}) = \frac{\alpha_1}{\hat{r}} + \alpha_2, \tag{4}$$

where α_1 and α_2 are the estimated fitting parameters.
The steps for the radial compensation are as follows:

1) Measure a limited amount of RTFs along the radial axis with the microphone array with different r, denoted as $A(\omega, r_{[i]})$, where $r_{[i]}$ is a measured point, and i_r is the number of measured points.
2) Obtain mean amplitude of RTFs over frequency bins of $A(\omega, r_{[i]})$ by

$$\bar{A}(r_{[i]}) = \frac{1}{p_h - p_l + 1} \sum_{p=p_l}^{p_h} \left| A(\omega_{[p]}, r_{[i]}) \right| \ , \tag{5}$$

where p_h and p_l are the indexes of the maximum and minimum frequencies respectively.
3) Obtain α_1 and α_2 through nonlinear curve fitting as follows:

$$\boldsymbol{F}_x = \begin{bmatrix} \frac{1}{r_{[1]}} & 1 \\ \vdots & \vdots \\ \frac{1}{r_{[i_r]}} & 1 \end{bmatrix}, \boldsymbol{F}_y = \begin{bmatrix} \bar{A}_m(r_{[1]}) \\ \vdots \\ \bar{A}_m(r_{[i_r]}) \end{bmatrix},$$
$$[\alpha_1, \alpha_2]^T = \left(\boldsymbol{F}_y^T \boldsymbol{F}_y\right)^{-1} \boldsymbol{F}_y^T \boldsymbol{F}_y \tag{6}$$

4) Select a reference RTF $A_m(\omega, r)$ among the pre-measured RTFs.
5) $\hat{A}_{[r]}(\omega, \hat{r})$ is estimated by Eq. (3) with α_1 and α_2 in Eq. (6).

For practical reasons, it is desirable to allow more degrees of freedom for the compensated RTF, Eq. (4) is allowed to deviate within a close neighbourhood in a discrete step-wise manner $k\Delta$ for $k = 1, .., K$. This covers the uncertainty of the RTF estimate which proved to be effective in our previous work [1][2]. Thus, Eq. (4) becomes

$$f_{k\Delta}(\hat{r}) = k\Delta \ \frac{\alpha_1}{\hat{r}} + \alpha_2,$$

where k is an integer and Δ is a constant value derived experimentally. Thus, Eq. (3) is rewritten as

$$\hat{A}^{k\Delta}_{[r]}(\omega, \hat{r}) = f_{k\Delta}(\hat{r})A(\omega, r) \ .$$

By introducing $k\Delta$, we are able to generate a set of RTF estimates $\{\hat{A}^{k\Delta}_{[r]}(\omega, \hat{r})\}$ within a close neighbourhood $\hat{A}_{[r]}(\omega, \hat{r})$.

The selection of the most probable (optimum) RTF is done through acoustic model likelihood score evaluation discussed in Sec III-B-3.

B. Acoustic Model-level Compensation

1) Reverberant Speech Generator: Using an arbitrary speech utterance $s(\omega)$ and transfer function $\hat{A}^{k\Delta}_{[r]}(\omega,\hat{r})$, we synthesize reverberant speech data

$$s^{k\Delta}_{[r]}(\omega,\hat{r}) = s(\omega)\hat{A}^{k\Delta}_{[r]}(\omega,\hat{r}). \tag{7}$$

For a set of transfer functions $\{\hat{A}^{k\Delta}_{[r]}(\omega,\hat{r})\}$, we generate the corresponding set of synthetic reverberant utterances $\{s^{k\Delta}_{[r]}(\omega,\hat{r})\}$. The synthesized reverberant speech is used for acoustic model adaptation.

2) Acoustic Model Adaptation: Prior to acoustic model adaptation, a Gaussian Mixture Model (GMM) π is trained using a clean speech database (i.e., the database used in our ASR). The GMM is composed of four states which guarantee that the temporal smearing caused by the reverberation is captured by the model (i.e., in the states). Since we are only interested in the reverberant effects of the sound and not its phonetic meaning, the GMM is designed to be phoneme, speaker and gender independent. Thus, we can use any arbitrary utterance. Model adaptation is implemented through Maximum Likelihood Linear Regression (MLLR) which is effective when dealing with small amount of data [9]. MLLR can adapt the means and covariance of π. Using the synthesized reverberant speech $\{s^{k\Delta}_{[r]}(\omega,\hat{r})\}$ as adaptation data, the MLLR adapted mean for every r is expressed as

$$\boldsymbol{\mu}^{k\Delta}_{m_c} = \boldsymbol{W}_{m_c}\boldsymbol{\zeta}_{m_c}\,, \tag{8}$$

where $\boldsymbol{W}_{m_c}$ is the transformation matrix while $\boldsymbol{\zeta}_{m_c}$ refers to the extended mean vector. The subscript m_c denotes the mixture m at class c. The adapted covariance is given as

$$\boldsymbol{\Sigma}^{k\Delta}_{m_c} = \boldsymbol{B}^{T}_{m_c}\boldsymbol{H}_c\boldsymbol{B}_{m_c}\,, \tag{9}$$

where $\boldsymbol{B}_{m_c}$ is the inverse of the Choleski factor

$$\boldsymbol{B}_{m_c} = \boldsymbol{C}^{-1}_{m_c}\,. \tag{10}$$

Both the transformation matrix $\boldsymbol{W}_{m_c}$ and the linear transformation $\boldsymbol{H}_c$ are calculated using the adaptation data $s^{k\Delta}_{[r]}(\omega,\hat{r})$ discussed in Sec. III-B-1. The adapted GMM $\pi^{k\Delta}_{[r]}$ has the corresponding means $\boldsymbol{\mu}^{k\Delta}_{m_c}$ and variance $\boldsymbol{\Sigma}^{k\Delta}_{m_c}$, respectively.

3) Optimum Transfer Function Selection: This process shown in Fig. 3 is crucial as it takes into consideration the overall contribution of the different variables/processes operating in the entire system in the selection of the optimal RTF. It connects the different modules, such as the waveform-level compensation, the ASR mechanism and the reverberant signal $\hat{u}(\omega)$ altogether. As a result, the RTF choice guarantees optimal ASR performance when used in conjunction with the ASR-based dereverberation scheme. Using the GMM adapted models $\pi^{k\Delta}_{[r]}$ discussed in Sec III-B-2, we identify the corresponding k associated with each model that best matches the actual reverberant signal $\hat{u}(\omega)$. The spectral features $\tilde{u}$ of $\hat{u}(\omega)$ is extracted and used to evaluate the likelihood scores based on the acoustic likelihood criterion, which is identical to the one used in the ASR system. The corresponding k that results in the highest likelihood score is used to select the optimal RTF.

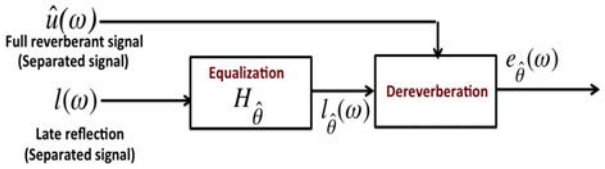

Fig. 4. Robustness to azimuthal change via equalization.

The ASR-based dereverberation in [1][2] is adopted. In the offline training mode (left figure), we replaced the previous RTF estimator based on RT (see Sec II) with the proposed method discussed in Sec III. After selecting $\hat{A}(\omega,\hat{r})^{opt}$, we extract the late reflection component of the RTF $\hat{A}_l(\omega,\hat{r})^{opt}$ [1][2]. Then, the reverberant signal $\hat{u}(\omega)$ is simulated using $\hat{A}(\omega,\hat{r})^{opt}$ and the clean speech database $c(\omega)$ in which we extract the late reflection approximation $\hat{l}(\omega)$. In the same manner, using $\hat{A}_l(\omega,\hat{r})^{opt}$ and $c(\omega)$ we can simulate the real late reflection $l(\omega)$. Consequently, $\{\delta_1,...,\delta_B\}$ is optimized with the objective of minimizing the error between $l(\omega)$ and $\hat{l}(\omega)$. The optimized weighting parameter $\{\delta_1,...,\delta_B\}^{opt}$ guarantees the actual dereverberation (right figure) to work even when using only the $\hat{l}(\omega)$ and not the actual $l(\omega)$. Dereverberation through the modified SS [2] is obtained through

$$|e(\omega)|^2 = \begin{cases} |\hat{u}(\omega)|^2 - \delta_b|\hat{l}(\omega)|^2 \\ \quad \text{if } |\hat{u}(\omega)|^2 - \delta_b|\hat{l}(\omega)|^2 > 0 \\ \\ \beta|\hat{u}(\omega)|^2 \quad \text{otherwise,} \end{cases} \tag{11}$$

where δ_b is the multi-band weighting parameters optimized through an offline training scheme. The multi-band treatment improves error minimization as opposed to single-band. In the actual dereverberation, these parameters are used together with $\hat{u}(\omega)$ to recover $\hat{e}(\omega)$ for ASR.

IV. Robustness to Azimuth Change Via Late Reflection Equalization

In theory, multiple unique RTFs are needed to match the corresponding change in azimuthal orientation θ for each channel (i.e., $\boldsymbol{A}_\theta(\omega)$). This is because when θ changes, the acoustical dynamics inside the room is perturbed as the concentration of speech power changes as a function of θ. In short, the late reflection also varies with θ in reverberant environments like the one in our work in [15]. However, it is impractical to measure all possible θ variations since it requires a corresponding RTF measurement for all M microphones. To mitigate this, we employed an equalization scheme, by dealing with the source-separated late reflection $l(\omega)$ instead of the multi-channel RTF characteristics. This scheme simplifies the supposed complicated analysis of

TABLE I

AVERAGED WORD RECOGNITION RATE (%) (RT= 240 MSEC AND ROOM 2 RT = 640 MSEC)

Methods (Room 1)	0.5 m	1.0 m	1.5 m	2.0m	2.5 m
A. No processing	79.1%	73.2%	57.6%	35.3%	20.6%
B. Blind Dereverberation	80.3%	76.9%	65.6%	49.6%	37.7%
C. Previous Method (Sec II)	**81.2%**	**78.3%**	**71.3%**	**55.7%**	**46.1%**
D. Waveform Compensation (Sec III-A)	81.6%	79.4%	73.1%	57.2%	50.8%
E. Waveform and Acoustic Model Comp. (Sec III-A&B)	82.3%	81.2%	75.8%	60.7%	55.4%
F. Waveform and Acoustic Model Comp. + Equalization (Sec III-A&B + Sec IV)	**82.9%**	**82.3%**	**77.5%**	**62.7%**	**57.8%**
Methods (Room 2)	0.5 m	1.0 m	1.5 m	2.0m	2.5 m
A. No processing	31.8%	15.6%	0.40%	-8.10%	-20.2%
B. Blind Dereverberation	41.9%	33.4%	20.6%	10.0%	0.90%
C. Previous Method (Sec II)	**45.0%**	**38.3%**	**26.9%**	**16.1%**	**7.40%**
D. Waveform Compensation (Sec III-A)	46.3%	41.1%	32.5%	23.5%	16.5%
E. Waveform and Acoustic Model Compensation (Sec III-A&B)	48.4%	43.8%	36.7%	28.4%	22.1%
F. Waveform and Acoustic Model Comp. + Equalization (Sec III-A&B + Sec IV)	**50.1%**	**46.4%**	**38.9%**	**30.7%**	**25.9%**

the effect of the azimuthal orientation with respect to the multi-channel RTFs into simple single channel filtering. The equalized late reflection signal becomes

$$l_\theta(\omega) = l(\omega)H_\theta. \tag{12}$$

where $l(\omega)$ is the separated late reflection using a generic (unmatched) RTF while H_θ is the equalizer.

H_θ is a filter derived experimentally during the offline mode by analyzing the response of the late reflection as a function of the actual azimuthal change θ. Suppose that $l_\theta(\omega)$ is the actual late reflection with a corresponding multi-channel RTF $\boldsymbol{A}_\theta(\omega)$. The filter design involves the poles positioning method on a logarithmic frequency grid based on [12][13]. The target response is set to $l_\theta(\omega)$ and H_θ for $\{\theta_1,\ldots,\theta_g,\ldots,\theta_G\}$ are derived by properly positioning the poles to achieve the target response $l_\theta(\omega)$ [14]. Note that the target response $l_\theta(\omega)$ was preprocessed via smoothing to avoid direct inversion problems [14]. With an effective θ selection procedure similar to that in Fig. 3 the equalization process [15], azimuthal change is compensated via filtering.

Dereverberation based on [2] is given as

$$|e_{\hat{\theta}}(\omega)|^2 = \begin{cases} |\hat{u}(\omega)|^2 - H_{\hat{\theta}}(\omega)|l_{\hat{\theta}}(\omega)|^2 \\ \quad \text{if } |\hat{u}(\omega)|^2 - H_{\hat{\theta}}(\omega)|l_{\hat{\theta}}(\omega)|^2 > 0 \\ \beta|\hat{u}(\omega)|^2 \quad \text{otherwise.} \end{cases} \tag{13}$$

where $\hat{u}|(\omega,t)|^2$ is the power of the separated reverberant signal ($\hat{u}|(\omega,t)|^2 \approx r|(\omega,t)|^2$) and $l_{\hat{\theta}}|(\omega,t)|^2$ is the separated late reflection power. We note that the equalization process is key to the hybrid approach as it eliminate δ in the Eq. (13). In our previous method [16], the dependence on the δ parameter was the stumbling block towards the utilization of multi-channel processing since the optimization of δ is computationally expensive for multi-channel signals. This limitation is rectified in the proposed method.

V. EXPERIMENTAL SET-UP

A. Training and Testing Database for ASR

The training database is from the Japanese Newspaper Article Sentence (JNAS) corpus. The open test set consists of 200 test utterances coming from 24 speakers. Recognition experiments are carried out on the Japanese dictation task with 20K-word vocabulary . The language model is a standard word trigram model. The acoustic model is a phonetically tied mixture (PTM) HMMs with 8256 Gaussians in total. Test experiment is conducted using actual human-robot communication set-up. The microphone array is embedded on the head of the robot. In the experiment, we used different occlusions such as table, chairs, etc. (real environment setting).

Real reverberant data are recorded inside two different reverberant rooms (Room 1 and Room 2) with RT of 240 ms and 640 ms respectively. Six different radial axes at different azimuth θ selected randomly. We considered five radial location points $r_{[i]} = \{0.5\text{m}, 1.0\text{m}, 1.5\text{m}, 2.0\text{m}, 2.5\text{m}\}$, $1 \le i \le i_r = 5$ are for testing. Each location point consists of 200 test utterances. These are then processed with our proposed methods (i.e. Sec. III and Sec. IV).

VI. ASR RECOGNITION RESULTS

In Table I, the method (A) shows the performance when the reverberant test data are not processed (no dereverberation), together with an acoustic model matched on the test data condition. Method (B) shows the performance using a blind dereverberation scheme that does not require any RTF estimation to carry out dereverberation [11]. The method (C) is the performance when using our previous RTF estimation [1][2]. (D) is the result when using the proposed RTF estimation method discussed in Sec III-A where only the waveform-level compensation is in effect. The method (E) shows the result when both the waveform-level and acoustic model-level compensation are in effect in estimating the RTF (both Sec III-A&B). Lastly, the method in (F) shows the performance when equalization in Sec. IV is implemented on top of the radial compensation method in Sec. III.

Although the method (B) performs better than method (A), it is less effective than methods that use RTF information. We note that (B) operates blindly (no RTF is required). It is confirmed in (D) that the proposed waveform-level RTF estimation outperforms the previous method in (C). This is due to the fact that in the proposed method (Sec III-A),

we are able to address the variation of speech power as a function of the speaker's position. Moreover, the proposed RTF estimator performs best when optimal RTF selector through acoustic model likelihood criterion (Sec III-B) is employed together with the waveform-level compensation (Sec III-A) in (E).

The recognition performance disparity between Room 1 and Room 2 is attributed to the different RT. The latter has larger RTF with more occlusions inside. Also, we emphasize that we experiment on a large vocabulary continuous dictation task and not on isolated word recognition. Unlike the latter, continuous dictation is more susceptible to reverberation due to long-duration utterances which generates more reflections. Moreover, isolated word recognition task has a very small vocabulary (hundred words) and does not consider insertion and deletion errors. In our case, we used 20K-words. Thus, recognition rate is always higher in isolated word recognition task compared to the continuous dictation task. The negative recognition values are attributed to the insertion and deletion errors.

VII. CONCLUSION

We have presented a method that compensates the variation of the speaker's radial position and azimuth relative to the robot, respectively. In the case of compensation the effect of the change in the radial position, we have shown the effect of designing the RTF estimator in conjunction with the ASR system. This results in an RTF estimate that is more tailored to achieving optimal ASR performance. In the case of azimuthal change, we have shown a method based on equalization which further mitigates the effect of mismatch. Currently, these two methods are implemented independently and in the future, we will consider the joint optimization of the two methods.

REFERENCES

[1] R. Gomez, T. Kawahara, "Optimization of Dereverberation Parameters based on Likelihood of Speech Recognizer" *In Proceedings of Interspeech*, 2009.

[2] R. Gomez and T. Kawahara, "Robust Speech Recognition based on Dereverberation Parameter Optimization using Acoustic Model Likelihood" *In Proceedings IEEE Transactions Speech and Acoustics Processing*, 2010

[3] P. Naylor and N. Gaubitch, "Speech Dereverberation" *In Proceedings IWAENC*, 2005

[4] Y. Huang, J. Benesty, and J. Chen, "Speech acquisition and enhancement in a reverberant, cocktail-party-like environment" *In Proceedings IEEE Int. Conf. Acoust., Speech, Signal Proc. ICASSP*, 2008

[5] H. Kuttruff, "Room Acoustics" *Spon Press*, 2000

[6] H.-G. Hirsch and H. Finster, "A new approach for the adaptation of HMMs to reverberation and background noise" *Speech Communication*, pp 244-263, 2008.

[7] H. Sawada *et al.*,"Polar coordinate based nonlinear function for frequency-domain blind source separation," in *Proc. of ICASSP 2002*, 2002.

[8] H. Nakajima, K. Nakadai, Y. Hasegawa and H. Tsujino, "Adaptive Step-size Parameter Control for real World Blind Source Separation" *In Proceedings IEEE Int. Conf. Acoust., Speech, Signal Proc. ICASSP*, 2008.

[9] C.J.Leggeter and P.C. Woodland,"Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models" *In Proceedings of Computer Speech and Language*, 1995

[10] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction" *In Proceedings IEEE Int. Conf. Acoust., Speech, Signal Proc. ICASSP* ,1979.

[11] S. Griebel and M. Brandstein, "Wavelet Transform Extrema Clustering for Multi-channel Speech Dereverberation' *IEEE Workshop on Acoustic Echo and Noise Control*, 1999.
'

[12] B. Bank, "Direct Design of Parallel Second-order Filters for Instrument Body Modeling", *In Proceedings of the International Computer Music Conference*, 2007.

[13] J. Laroche and J-L. Meillier, "Multichannel Excitation/Filter Modeling of Percussive Sounds with Application to the Piano" *In Proceedings IEEE Transactions Speech and Audio Processing*, 1994.

[14] B. Bank, G de Poli and L. Sujbert, "A Multi-rate Approach to Instrument Body Modeling for Real-time Sound Synthesis Spplications" *In Proceedings of 112th AES Conventiona*, 2002.

[15] R. Gomez, K. Nakamura and K. Nakadai, "Dereverberation Robust to Speaker's Azimuthal Orientation in Multi-channel Human-Robot Communication" *In Proceedings of the IEEE IROS*, 20013.

[16] R. Gomez, K. Nakamura and K. Nakadai "Hands-free Human-Robot Communication Robust to Speaker's Radial Positition" *In Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, 2013.