Robust Hands-free Human-Robot Communication in Reverberant Environments

Randy Gomez, Keisuke Nakamura, Takeshi Mizumoto and Kazuhiro Nakadai

***Abstract*— Speech-based human-robot interaction is often plagued with issues such as reverberation and changes in speaker position that impacts overall performance. In this paper, we show a method in compensating the joint effects of reverberation and the change in speaker position. The acoustic perturbation caused by these two takes its toll on the Automatic Speech Recognition (ASR) and then the Spoken Language Understanding (SLU). Consequently, these will lead to a failure in the human-robot interaction experience. The proposed method is specifically designed to address the challenging environment condition in which robots are deployed. First, we analyze the impact of reverberation in the form of temporal smearing per change in speaker position. Then, we extract the smearing coefficients that capture the joint dynamics between the speech signal at current position and the room acoustics as observed by the robot. These coefficients are utilized to update the room transfer function (RTF) and the suppression parameters are stored offline. Moreover, all of these processes are optimized in the context of the ASR system for robot application. In the online mode, the reverberant data at an arbitrary position is processed using the parameters pre-computed offline. This effectively compensates the joint effects of reverberation at the arbitrary speaker position. Experimental results using real data gathered in a human-robot communication setting show that the proposed method outperforms existing methods.**

***Index Terms*— Speech Enhancement, Dereverberation, Robustness, Automatic Speech Recognition**

I. Introduction

Reverberation is a phenomenon caused by the reflections of the speech source in an enclosed environment. It is characterized by the smearing effect to the original speech due to the different time delays of arrival of the reflected speech source. The smearing degrades the Automatic Speech Recognition (ASR) system due to mismatch in the Hidden Markov Model (HMM) and impacts the Spoken Language Understanding (SLU) system as well. The overall impact may lead to the failure in human-robot communication experience. To mitigate this, the observed reverberant speech is enhanced through dereverberation. There exists different types of dereverberation methods [1][2] and most of these are originally formulated using human perception criterion and later applied to the ASR system in robot applications. Although this approach works well, the dereverberation method is not optimized for robot environment which is a very challenging task.

We note that in real robot environment, it is very difficult to control the position of the speaker when interacting with the robot as depicted in Fig. 1. For an immersive interaction, the speaker cannot be restricted as to where he initiates the conversation. Moreover, to effectively suppress the effects of the smearing caused by reverberation for robot application, it is important to incorporate the dynamic speaker position in the reverberation model. This means that the changes in speaker positions and dereverberation should be analyzed altogether when addressing the reverberation problem, to be effective in robot application. In this paper, we improve our previous work [4][13] by compensating the changes in speaker position via ASR optimization.

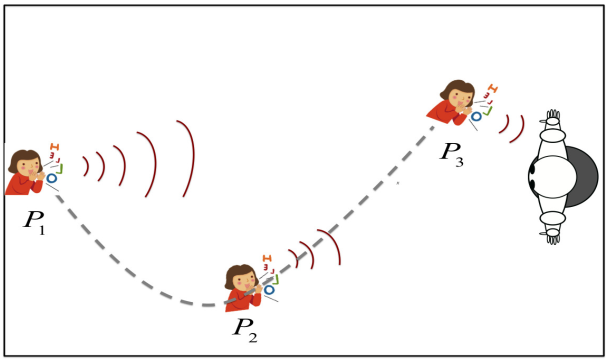

Fig. 1. Problem: Changes in speaker position in an enclosed reverberant room.

Our previous work [4][13] does not take into consideration the joint dynamics of the room characteristics and the change in speaker position. These two were treated independently in our previous work [4][13] but in reality there is a very strong link between the two. In addition, our previous method is more focused on the temporal side of the speech (i.e., waveform) and just stops right there. When dealing with ASR, the temporal representation of speech has its dual in the form of connected symbols which represents the sound units. Each of the sound unit are modelled by the Hidden Markov Models (HMMs). Therefore, it is very important to treat the latter equally likely with the waveform which is not addressed in our previous work [4][13]. In short, there is no mechanism in the previous method to operate in the HMM level. For example, the concept of frame-wise energy may exist but its analysis does not go deeper as to relate it with energy transfer across HMM states. This renders a very coarse treatment of the effect of reverberation when applied to HMM-based ASR, especially in challenging robot environment. In the proposed method, we have expanded the traditional reverberation model to treat jointly the effects of both the waveform and the HMMs. In particular, the effect of acoustic perturbation due to the changes in speaker position is tightly integrated in the dereverberation mechanism

Honda Research Institute Japan Ltd. Co., Honcho Wako-shi, Japan

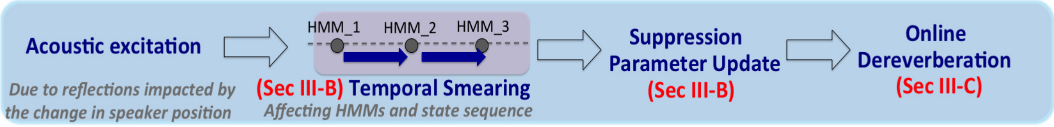
Acoustic excitation
Due to reflections impacted by the change in speaker position
HMM_1
HMM_2
HMM_3
(Sec III-B) Temporal Smearing
Affecting HMMs and state sequence
Suppression Parameter Update
(Sec III-B)
Online Dereverberation
(Sec III-C)

Fig. 2. Overall diagram of the proposed concept.

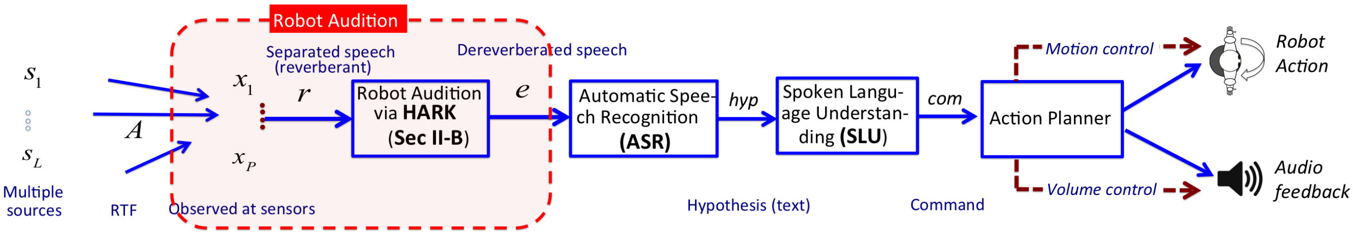

Fig. 3. Block diagram of a voice-based human-robot interactive system.

itself. We implemented a data-driven optimization scheme to extract parameters reflective of the dynamics between the speech and the room characteristics. We note that acoustic dynamics varies as perturbed by the change in position. In addition, the design methodology of the proposed method is hinged to the HMM (i.e., state sequence) which is an integral component of the ASR.

The concept of our approach is shown in Fig. 2. The recognized speech is treated not just a waveform but a sequence of sound units (i.e., phoneme HMMs). It is presumed that each sound unit is characterized by a unique frequency response that behaves differently given an acoustic excitation. In the same figure, it is shown that reflections inside the room drives an acoustic excitation that causes **temporal smearing** acting on the connected HMM sound units and within each HMM. Similarly, at each HMM, temporal smearing spans from one HMM state to the next. The smearing is hypothesized to be caused by the perturbation of the room acoustics due to changes in robot-speaker position. By modelling the smearing effect of reverberation, it is possible to compensate the changes in speaker positions objectively. In this paper we show the method of using a data-driven processing to empirically model the temporal smearing for dereverberation as a function of speaker position. Consequently, we improve the dereverberation performance of our proposed method using the ASR and SLU as metrics.

This paper is organized as follows; in Sec. II, we show the background of the previous reverberation model and the concept of dereverberation. The method in optimizing the temporal smearing for effective dereverberation is discussed in Sec. III. Experimental set-up with actual robot is discussed in Sec. IV followed by results and discussion in Sec. V. Finally, we conclude the paper in Sec. VI.

II. Background

A. Speech Communication-based Human-robot Interaction

Fig. 3 is an example of our human-robot interaction set-up. First, the robot audition framework based on HARK [17] is employed to process the multiple sound sources $S_1, ..., S_L$ as observed at microphones $x_1, ..., x_P$ into separated speech r. Then, enhanced to e via dereverberation. The dereverberated speech is used as input to the ASR system which outputs the hypothesis hyp. Consequently, the SLU system extracts the command com information from hypothesis. Lastly, a robot action is executed which includes motion and/or audio feedback. It is obvious in this figure that we need a robust robot audition system that can support speech communication in adverse conditions. This is vital in achieving a successful robot understanding. Thus, it is imperative that we compensate the dereverberation mechanism per change in speaker position. In reality, reverberation is not the only problem in attaining robust robot audition system. The following are other common problems

- Ego Noise (Noise from within the robot's moving parts)
- Directional Noise (External noise)
- Background Noise (External/Internal but additive in nature)
- Voice Activity Detection (Detecting speech segments)

Most of the problems above are already integrated in HARK, and in this paper we will focus only on the reverberation problem.

B. Robot Audition

Microphone array processing based on beamforming and blind separation described in [9][17] is employed to convert the multi-microphone observed signals $x_1, ..., x_P$ resulting to the separated reverberant signal $r(\omega)$. Moreover, we note that the RTF denoted by $A(\omega)$ is readily available during the microphone array processing [9][17]. However, $A(\omega)$ is assumed to be constant, but in real-world application this may not hold true any more especially when room size is factored in together with the robot and the objects inside the room. More specifically, room acoustics is more likely to change due to the acoustic perturbation caused by the changes in speaker positions. In the end, $A(\omega)$ needs to be updated. In our previous method [4][13], the smearing

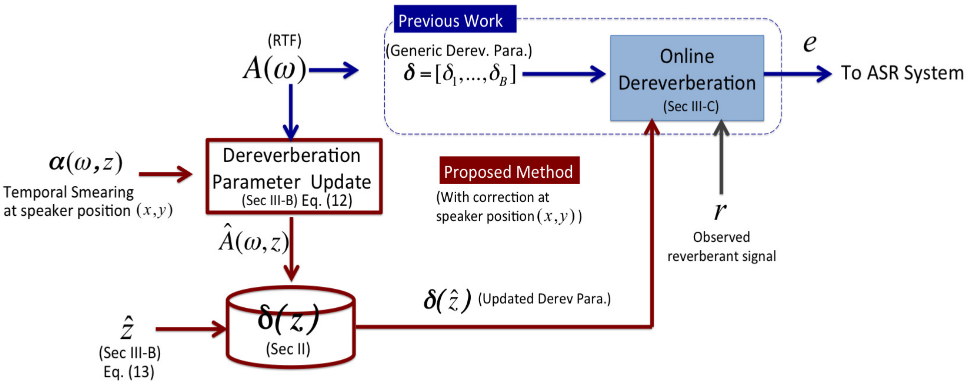

Fig. 4. Block diagram of the proposed method.

effect of reverberation is adopted from [15][5] and is solely dependent on the room transfer function (RTF) given as

$$\begin{array}{rcl} r(\omega) & = & A^{E}(\omega)c(\omega) + A^{L}(\omega)c(\omega) \\ & = & \quad e(\omega) \quad + \quad l(\omega), \end{array} \tag{1}$$

where $r(\omega)$ is the reverberant speech model w.r.t. ω frequency, $c(\omega)$ is the clean speech, $A^{E}(\omega)$ and $A^{L}(\omega)$ are the early and late reflection components extracted from the full RTF $A(\omega)$. Both $A^{E}(\omega)$ and $A^{L}(\omega)$ are experimentally predetermined in [13]. $r(\omega)$ can be treated as the superposition of $e(\omega)$ and $l(\omega)$, known as the early and late reflection, respectively. In [13] we treat $l(\omega)$ as long-period noise which is detrimental to the ASR, and dereverberation is defined as suppressing $l(\omega)$ and recovering $e(\omega)$ estimate. The latter is further processed with Cepstrum Mean Normalization (CMN) during ASR. Eq. (1) simplifies dereverberation into a denoising problem, and through spectral subtraction (SS) [10], the estimate $\hat{e}(\omega)$ in frame-wise manner j is given as

$$|e(\omega,j)|^2 = \begin{cases} |r(\omega,j)|^2 - |l(\omega,j)|^2 \\ \quad \text{if } |r(\omega,j)|^2 - |l(\omega,j)|^2 > 0 \\ \\ \beta|r(\omega,j)|^2 \quad \text{otherwise.} \end{cases} \tag{2}$$

where β is the flooring coefficient. In real condition, $l(\omega,j)$ is unavailable, precluding the power estimate $|l(\omega,j)|^2$. A scheme in [13] shows a workaround to this problem, approximating $l(\omega,j)$ directly from the observed reverberant signal $r(\omega,j)$ through the error

$$E_m = \frac{1}{J}\sum_{j}\sum_{\delta_b \in B_q} |l(\omega,j) - \delta_b(\omega,j)r(\omega,j)|^2. \tag{3}$$

For the given set of bands $\boldsymbol{B} = \{B_1, \ldots, B_Q\}$, the suppression parameter δ_b is determined through minimum mean square error criterion in Eq. (3) via offline training discussed in [4][13]. The multi-band treatment improves error minimization as opposed to single-band. The new estimate $\hat{e}(\omega)$ through the modified SS becomes

$$|e(\omega,j)|^2 = \begin{cases} |r(\omega,j)|^2 - \delta_b|r(\omega,j)|^2 \\ \quad \text{if } |r(\omega,j)|^2 - \delta_b|r(\omega,j)|^2 > 0 \\ \\ \beta|r(\omega,j)|^2 \quad \text{otherwise.} \end{cases} \tag{4}$$

It is obvious that the dereverberation platform in Eq. (4) is dependent on the suppression parameter $\boldsymbol{\delta}$. Consequently, $\boldsymbol{\delta}$ depends on the RTF-centric reverberation model in Eq. (1). Although Eq. (1) is effective for waveform enhancement, it does not have any provision for HMM analysis (i.e., energy transfer in the HMM level) whenever the room acoustics is perturbed. Perturbation exists especially when the sound source (i.e. speaker) changes position relative to the robot. We note that in real scenario the relative position between the human and robot always changes. Thus, it is imperative that the original $A(\omega)$ needs to be updated as we cannot expect that the current acoustic perturbation is still reflective of the original $A(\omega)$. With no update capability, the dereverberation performance is very limited since the suppression parameter is frozen together with the RTF. The simplified block diagram of the previous and proposed methods is shown in Fig. 4. In the proposed method, the suppression parameter can be updated to $\boldsymbol{\delta}(\hat{z})$ depending on the joint dynamics of the room characteristics and the observed reverberant signal as characterized by $\alpha(\hat{z})$. Where $\hat{z}$ is the most probable HMM state sequence. Fig. 4 is explained in detail in the following section.

III. Methods

A. Database

The clean speech database used in the ASR is utilized to generate the reverberant database. The word-level text transcript is converted to phoneme-level transcript. The clean speech database is re-played using a loudspeaker inside a reverberant room and recorded by a microphone located at distance away from the loudspeaker. The newly recorded speech data becomes the reverberant database r. In this paper, we are interested on the basic sound units defined as the phonemes in our application. Thus, when referring to sound units, these basically come from the speech database itself.

B. Optimized Temporal Smearing Coefficients for suppression parameter Update

Suppose that the observed reverberant speech when processed by a filter is given as

$$o[n] = \sum_{m=0}^{M-1} \alpha_m \;\; r[n-m] \tag{5}$$

where r is the observed reverberant data and the temporal smearing filter which is

$$\boldsymbol{\alpha} = [\alpha_0, \alpha_1, ..., \alpha_{M-1}]^T, \tag{6}$$

is unknown. The length of the filter M samples can be indirectly associated to the extent of reverberation (i.e., reverberation time). It is hypothesised that $\boldsymbol{\alpha}$ charcterizes the joint acoustic perturbation due to reverberation and the changes in speaker position. The objective is to estimate $\boldsymbol{\alpha}$ in the context of the ASR system. Thus, the resulting estimate would capture the temporal smearing characteristics associated to the joint dynamics of the room characteristics (RTF) and the actual sound units spoken at an arbitrary position. We assume that $\boldsymbol{\alpha}$ is associated to a change in the speaker position (x, y) but we will drop the position notation (x, y) for simplicity. For now, the actual signal o is immaterial since we are interested with the ASR's output (hypothesis) which is given as

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\text{argmax}} \log \left(P(f^{(o)}(\boldsymbol{\alpha})|\boldsymbol{w})P(\boldsymbol{w})\right. \tag{7}$$

where $f^{(o)}(\boldsymbol{\alpha})$ is the extracted feature vector from the utterance, $\boldsymbol{w}$ is the phoneme-based transcript, $P(f^{(o)}(\boldsymbol{\alpha})|\boldsymbol{w})$ is the acoustic likelihood (i.e., using reverberant acoustic model) and $P(\boldsymbol{w})$ is due to the language (i.e., using language model). The latter can be ignored since phoneme-based transcript $\boldsymbol{w}$ is known, thus, argmax in Eq. (7) acts on $\boldsymbol{\alpha}$ which is rewritten as

$$\hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\text{argmax}} \log P(f^{(o)}(\boldsymbol{\alpha})|\boldsymbol{w}). \tag{8}$$

In ASR, the total log likelihood in Eq. (8) when expanded [14] to include all possible state sequence in conjunction with the length of the smearing template is expressed as

$$\Gamma(\boldsymbol{\alpha}) = \sum_j \log P(f_j^{(o)}(\boldsymbol{\alpha})|\hat{s}_j), \tag{9}$$

where s_j is the state at frame j. Eq. (9) paves the formulation in analyzing the problem based on the HMMs in the form of state sequence. By using the ∇ operator, the total probability is maximized w.r.t the smearing coefficient in Eq. (6), thus,

$$\nabla_{\boldsymbol{\alpha}} \Gamma(\boldsymbol{\alpha}) = \left\{ \frac{\partial\Gamma(\boldsymbol{\alpha})}{\partial\alpha_0}, \frac{\partial\Gamma(\boldsymbol{\alpha})}{\partial\alpha_1}, ..., \frac{\partial\Gamma(\boldsymbol{\alpha})}{\partial\alpha_{M-1}} \right\}. \tag{10}$$

Assuming a Gaussian mixture distribution with mean vector μ_{jv} and diagonal covariance matrix $\boldsymbol{\Sigma}_{jv}^{-1}$, respectively. Eq. (10) can be shown similar to that in [8] as

$$\nabla_{\boldsymbol{\alpha}} \Gamma(\boldsymbol{\alpha}) = -\sum_j \sum_{\mathrm{v}=1}^{\mathrm{V}} \gamma_{jv} \frac{\partial f_j^{(o)}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} \boldsymbol{\Sigma}_{jv}^{-1} (f_j^{(o)}(\boldsymbol{\alpha}) - \mu_{jv}). \tag{11}$$

where γ_{jv} is the posteriori of v mixture and j frame of the most likely HMM state. $\frac{\partial {f_j}^{(o)}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}$ is the Jacobian matrix of the reverberant feature vector. The HMM-optimized smearing coefficients are obtained using [11][12] based on Eq. (11). In general, the HMM can generate Z-best most likely state sequences. Thus, from Eq. (8) we expand $\boldsymbol{\alpha}(z)$ corresponding to the Z possible HMM state sequence. In this manner we can capture the effect in the state sequence caused by the joint dynamics of the room characteristics and the sound excitation as perturbed by the change in speaker position.

The Z-best optimized smearing coefficients are used to update the readily available RTF $A(\omega)$ which is provided in the microphone array processing discussed in Sec. II. The RTF update done for all z is expressed as

$$\hat{A}(\omega, z) = \alpha(\omega, z) A(\omega) \tag{12}$$

where $\alpha(\omega, z)$ is the $z - th$ temporal smearing in frequency domain. Thus, several RTFs are generated using the update in Eq. (13). Then, suppression parameters $\boldsymbol{\delta}(z)$ are computed for each $\hat{A}(\omega, z)$ in the same manner as discussed in Eq. (3) in Sec. II, and these values are kept in the database. In the online mode, the acoustic likelihood of the observed reverberant data is filtered with the pre-computed $\alpha(z)$ for all z templates and the corresponding $\hat{z}$ is selected through

$$\hat{z} = \underset{z}{\text{argmax}} P(f^{(\alpha(\omega,z))*r}|\boldsymbol{w}). \tag{13}$$

$\hat{z}$ signifies that the observed reverberant signal r is a close match to the corresponding $\alpha(\omega, \hat{z})$ in the acoustic likelihood criterion. Thus, its corresponding $\boldsymbol{\delta}(\hat{z})$ is selected as the updated suppression parameter.

C. Online Dereverberation

In the online mode, the system takes in as input the observed reverberant signal and select the optimal $\boldsymbol{\delta}(\hat{z})$ as described in Sec. III-B. $\boldsymbol{\delta}(\hat{z})$ is used as input for dereverberation. Specifically, the spectral subtraction in Eq. (4) is rewritten as

$$|\hat{e}(\omega, j)|^2 = \begin{cases} |r(\omega, j)|^2 - \delta_b(\hat{z})|r(\omega, j)|^2 \\ \quad \text{if } |r(\omega, j)|^2 - \\ \quad \delta_b(\hat{z})|r(\omega, j)|^2 > 0 \\ \\ \beta|r(\omega, j)|^2 \quad \text{otherwise.} \end{cases} \tag{14}$$

where $\delta_b(\hat{z})$ acts on the frame level j.

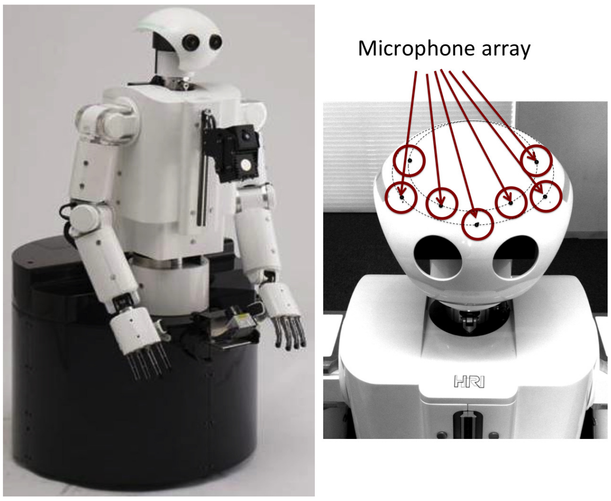

Fig. 5. HRI-JP humanoid robot ”Hearbo”.

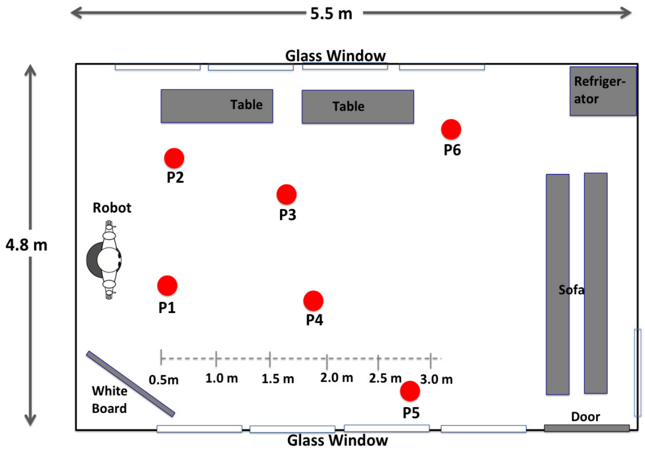

Fig. 6. Room set-up for testing (Room 4).

IV. Experiments with Robot

A. Humanoid Robot: Hearbo

The Honda Research Institute Japan’s (HRI-JP’s) humanoid robot named ”Hearbo” is shown in Fig. 5. It has 20 degrees of freedom and its head is embedded with microphone array arranged in two concentric circles of different diameters. it is equipped with a robot audition software based on HARK [17] which implements microphone array methods for hands-free speech processing.

B. ASR and SLU Systems

The baseline acoustic model is a 3-state HMM based on Gaussian Mixture Models and trained using the World Street Journal corpus. The test data is composed of ten English speakers. Each person utters 20 utterances for each test position in $P1 - P6$ (see Figure 6). Hypothetically speaking, the test speakers may speak in freeform. However, the utterances for the actual testing are scripted to maintain uniformity and to avoid mistakes as these may impact the SLU performance.

The human-robot interaction setting re-enacts a sushi restaurant scene. The customer (speaker) may approach the robot at an unknown position (i.e., P1-P6) and engage via voice communication. In the course of the conversation, the speaker asks the robot questions about the variety of fish used in preparing the traditional Japanese dishes ”Sushi” or ”Sashimi”. Upon recognition via the ASR system, the robot is tasked to translate the English fish name into its Japanese equivalence. Due to reverberation and the acoustic perturbation the observed reverberant speech is processed using our proposed method as shown in Figure 4 prior to ASR. Then the SLU system processes the output of the ASR system hyp to identify the fish name for the possible robot action. An example of the question from the customer would be, ”Hearbo, we had Sweetfish yesterday for dinner. Can you tell me what it is called in Japanese ?”. The robot should be able to identify that the fish in question is ”Sweetfish” and be able to give its corresponding Japanese name. Part of the interaction is that, the robot automatically adjust its volume in accordance to its proximity with the speaker and being able to turn and face towards the speaker direction. In our experiment we provide both the ASR and SLU results to confirm whether the proposed method impacts both the ASR and SLU systems.

C. Room Condition

We conducted our experiment in four different room settings (Room 1- Room 4) with Reverberation Time (RT) of 80 ms., 240 ms., 900 ms. and 940 ms., respectively. Room 1 is the least reverberant while Room 4 exhibits the most effect of reverberation for having the longest RT among the four rooms. In this work, we only focus the effect of reverberation so the background noise has signal to noise ration of 20 dB only. An example of one of the rooms (i.e., Room 4 with RT = 940 ms.) is shown in Fig. 6. Test positions inside the room are denoted as P1-P6. Although the RT is different for each room, the test positions P1-P6 are purposely positioned at the same places for all of the four different rooms for uniformity. Thus, the robot-to-speaker distances are the same.

V. Results and Discussion

The ASR results in terms of word correct are shown in Table 1. The results are averaged over the four different rooms. Method (A) is the result when no enhancement was implemented while method (B) is the result based on Linear Prediction residual approach [1]. By exploiting the characteristics of the vocal chord, it is able to remove the effects of reverberation. The result in method (C) is based on wavelet extrema clustering [2]. Similar to that in [1] except that it operates in the wavelet domain to find and remove the effects of reverberation. Method (D) is based on adaptation by [16], Instead of suppression, this method minimizes the mismatch through adaptation of the feature vector. The method in (E) is the result based on the previous method [13][4] (Eq. (2)) employing the old reverberant model. The proposed method (F) is based on Eq. (14) employing the

TABLE I

ASR RESULTS AVERAGED ACROSS ALL ROOMS (ROOM 1-ROOM 4) IN WORD CORRECT RATE (%)

	P1	*P2*	*P3*	*P4*	*P5*	*P6*
(A) No Enhancement	90.0 %	84.1 %	74.2 %	69.5 %	43.9 %	27.3 %
(B) Based on LP Residuals [1]	90.2 %	86.1 %	77.0 %	72.2 %	58.3 %	42.4 %
(C) Based on Wavelet Extrema [2]	90.4 %	86.3 %	78.1 %	74.5 %	60.6 %	46.2 %
(D) Based on Feature Adaptation [16]	90.7 %	86.5 %	79.3 %	76.2 %	63.4 %	49.8 %
(E) Spectral Subtraction (Previous Reverberation Model) [4][13]	90.8 %	86.9 %	79.6 %	76.5 %	68.3 %	54.3 %
(F) Spectral Subtraction (Proposed Method)	**91.2 %**	**87.7 %**	**82.8 %**	**81.4 %**	**74.7 %**	**66.4 %**

TABLE II

SLU RESULTS AVERAGED ACROSS ALL ROOMS (ROOM 1-ROOM 4) IN CORRECTLY IDENTIFYING THE FISH NAME (%)

	P1	*P2*	*P3*	*P4*	*P5*	*P6*
(A) No Enhancement	100.0 %	94.0 %	83.0 %	78.0 %	35.0 %	10.0 %
(B) Based on LP Residuals [1]	100.0 %	94.0 %	85.0 %	80.0 %	61.0 %	30.0 %
(C) Based on Wavelet Extrema [2]	100.0 %	94.0 %	85.0 %	81.0 %	65.0 %	35.0 %
(D) Based on Feature Adaptation [16]	100.0 %	94.0 %	86.0 %	83.2 %	68.0 %	38.0 %
(E) Spectral Subtraction (Previous Reverberation Model) [4][13]	100.0 %	94.0 %	86.0 %	84.0 %	68.3 %	43.0 %
(F) Spectral Subtraction (Proposed Method)	**100.0 %**	**96.0 %**	**88.0 %**	**86.0 %**	**71.0 %**	**59.0 %**

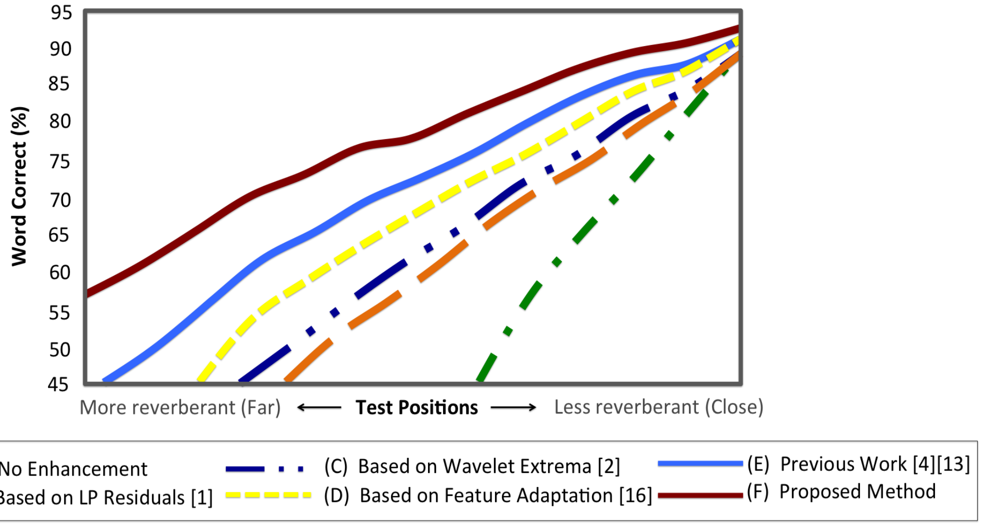

Fig. 7. Sorted ASR results using simulated data across Room 1- Room 5.

current reverberant model analysis that involves the notion of temporal smearing. In this table, we show that the propose method outperforms the existing methods and it is more effective farther distances. The adaptation based approach in [16] is only good in shorter reverberation time but performs poorly at longer reverberation time. This can be attributed to the fact that this method does not actually suppress the effects of dereverberation. We also show in Table 2 the results of the SLU system. This result confirms that the improvement in recognition performance attributed by the proposed method is translated in the machine understanding phase. Thus, the proposed method may positively impact interaction experience. In Fig. 7, we simulated the reverberant data inside the four different room by convolving a known RTF from the database and generate synthetic reverberant data. The purpose of this is to show the overall characteristics of the proposed method with more test data aside from the real recording in Table 1. We note that it is difficult to record different test points and synthetic reverberant data have been used and confirmed to show the same trend as real data. In this figure we concatenate and sort all the results from different rooms. We confirm the effectiveness of our proposed method in addition to that in Table 1.

The possible reasons why the proposed method fares better than the rest of the methods presented in this paper are: (1) the ability to update the suppression parameters reflective of the changes of the acoustic dynamics inside the room. It should be noted that depending on the acoustic excitation, acoustic room dynamics may change. (2) Formulation of the reverberation and optimization problems evolves in the HMM structure which is just proper since the dereverberation task is for the ASR system. This enables the processing of the acoustic waveform to better match the HMM-based ASR system. Lastly, (3) all of the optimization procedures are data-driven which results to a more realistic treatment of the effect of reverberation as opposed to just simply rely on the RTF.

VI. CONCLUSION

In this paper, we have shown the method of analyzing the reverberant model in an effective way that aids dereverberation for improved ASR performance. By analyzing the temporal smearing of the HMMs, we are able to incorporate the acoustic perturbation caused by the change in speaker position. We integrated it in the design process which is centered in the ASR system. This is very important because we have successfully expanded the traditional dereverberation method to environments in which robots are deployed. The proposed method is able to cope with demanding nature or human-robot communication such as the unpredictable change in speaker position. We have confirmed that the proposed method performs well in both real and synthetic data. Lastly, we confirmed the benefit of the proposed method is not just limited to the ASR system, more importantly it is able to improve the SLU performance as well. We note that the latter is a precursor of human-robot interaction experience. In our future work, we will consider the the effects of noise and investigate the prospect of expanding to deep neural networks (DNN).

REFERENCES

[1] B. Yegnanarayana and P. Satyaranyarana, "Enhancement of Reverberant Speech Using LP Residual Signals", *In Proceedings of IEEE Trans. on Audio, Speech and Lang. Proc.*, 2000.

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

[3] K. Kinoshita , T. Nakatani and M. Miyoshi, Spectral Subtraction Steered By Multi-step Forward Linear Prediction For Single Channel Speech Dereverberation, *In Proceedings IEEE Int. Conf. Acoust., Speech, Signal Proc. ICASSP*, 2006.

[4] R. Gomez, K. Nakamura, and K. Nakadai, "Robustness to Speaker Position in Distant-Talking Automatic Speech Recognition" *In Proceedings IEEE Int. Conf. Acoust., Speech, Signal Proc. ICASSP*, 2013.

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

[6] Akinobu Lee, *Multipurpose Large VocabularyContinuous Speech Recognition Engine*, 2001.

[7] S. Vaseghi "Advanced Signal processing and Digital Noise reduction", Wiley and Teubner, 1996.

[8] M. Seltzer, "Speech-Recognizer-Based Optimization for Microphone Array Processing *IEEE Signal Processing Letters*, 2003.

[9] 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.

[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] , "On numerical analysis of conjugate gradient method" *Japan Journal of Industrial and Applied Mathematics*, 1993.

[12] , W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, "Numerical Recipes in C: The Art of Scientific Computing" *Cambridge University Press*, 1988 .

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

[14] "The HTK documentation http://htk.eng.cam.ac.uk/docs/docs.shtml"

[15] E. Habets, "Single and Multi-microphone Speech Dereverberation Using Spectral Enhancement" *Ph.D. Thesis*, June 2007.

[16] 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.

[17] "HARK wiki http://winnie.kuis.kyoto-u.ac.jp/HARK/"