2013 IEEE/RSJ International Conference on
Intelligent Robots and Systems (IROS)
November 3-7, 2013. Tokyo, Japan

Dereverberation Robust to Speaker's Azimuthal Orientation in Multi-channel Human-Robot Communication

Randy Gomez, Keisuke Nakamura and Kazuhiro Nakadai

Abstract— **The acoustical dynamics of reverberation in an enclosed environment poses a problem to human-robot communication. Any change in the azimuthal orientation of the speaker contributes to unpredictable acoustical activity resulting in a degradation in the performance of the automatic speech recognition (ASR) system. Thus, dereverberation techniques need to address this issue prior to ASR. Dereverberation in multi-channel applications primarily evolves in the adoption of a suitable reverberant model that results to a computationally feasible solution and at the same time yields an accurate estimate of the harmful reflections (i.e., late reflection) for effective suppression. In this paper we address this problem by introducing a hybrid method based on multi-channel processing on a single-channel reverberant model platform. The proposed method is capable of accurate signal estimation, a property inherent to a multi-channel system, and at the same time bears the computational efficiency derived from single-channel reverberant model approach. The proposed method is summarized as follows; First, multi-channel sound-source processing is employed to obtain the full reverberant and the late reflection signal estimates. Then, equalization is employed to update the late reflection estimate reflective of the change in azimuth prior to dereverberation. The equalization parameters for azimuthal change are obtained through an offline optimization procedure. Experimental evaluation in an actual human-robot communication environment shows that the proposed method outperforms existing methods in terms of robustness in the ASR performance.**

I. Introduction

Research in speech-based human-robot interaction has advanced rapidly nowadays. After all, speech is one of a human's most preferred mode of communication, it is just fitting to harness the speech modality in interacting with robots. Consequently, the pursuit of achieving a seamless human-robot communication is coupled with the very challenging issue on robustness. Prior to machine understanding, the system has to recognize the spoken utterance and the speech recognition system has to deal with contamination issues attributed by the environment to the acoustic speech signal. It is extremely difficult to control the environment (e.g. room) where the human-robot communication takes place. The speech signal is reflected onto the walls, ceilings, obstructions, etc. as it travels in free space and arrives at the microphones with different time delays. This results in a phenomenon called reverberation which manifests a smearing effect to the clean speech. The effect of reverberation is very detrimental to the model-based speech recognition system as it causes mismatches to the original model condition which is usually trained using a clean speech database. To minimize the effects of mismatches caused by reverberation, waveform enhancement referred to as dereverberation is employed [1][2].

Reverberation problems become more complicated due to the active nature of the parties involved. In a practical scenario as shown in Fig. 1, it is impossible to fixate the azimuthal orientation of the user when conversing with a robot and the changes in the azimuth (i.e., θ_1,...,θ_g,...,θ_G) often lead to the degradation in ASR performance. It is important to stress that when a person changes face direction, the directivity of the speech is also changed which impacts the reflection of speech. Moreover, the asymmetrical configuration of the room setup due to the mobile nature of both the robot and the user further contributes to the problem. In general, it is prudent to assume the existence of a very dynamic acoustic activity in scenarios involving robots.

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

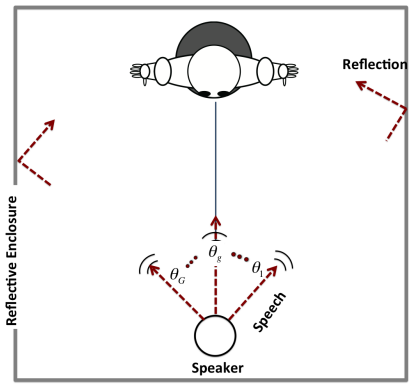

Fig. 1. Acoustic dynamics inside a reverberant room as a function of variable azimuth orientation θ.

Multi-channel analysis using microphone sensors embedded on a robot have been proven effective in capturing these dynamics. However, dereverberation based on a true multi-channel model is a computationally daunting task. On the other hand, a purely single-channel dereverberation platform may be computationally feasible but ineffective in capturing the acoustical dynamics inside the room. We have extended the single-channel reverberant model to a multi-channel system in our previous work [3]. Although this approach works, most of its methodologies are based on a single-channel processing concept (i.e., late reflection estimate) in which the full potential of microphone array processing is not utilized. The main problem of [3] is that, computational efficiency is favoured over effectiveness and accuracy.

In this paper we address the problem in [3] through a hybrid approach. We propose a combination of multi-channel signal analysis and a single-channel dereverberation model. First, multi-channel signals are transformed to a single-channel source (separated signal) via microphone array sound source separation. This results in an accurate late reflection signal and full reverberant signal estimates. After this, the reverberation problem is addressed via the single-channel model dereverberation platform. In this manner, optimization and dereverberation is performed on the reverberant separated signal (single-channel), instead of performing to each channel independently. Thus, more computationally tractable and faster implementation is achieved. To simulate the effects of the changes in the azimuthal orientation of the speaker, equalization is employed to the late reflection. This scheme mitigates the need of using multiple RTF measurements that match the corresponding azimuth change which is required in our previous work [3]. In effect, the equalization scheme digitally steers the signal simulating the actual physical change in azimuthal orientation. Although, a

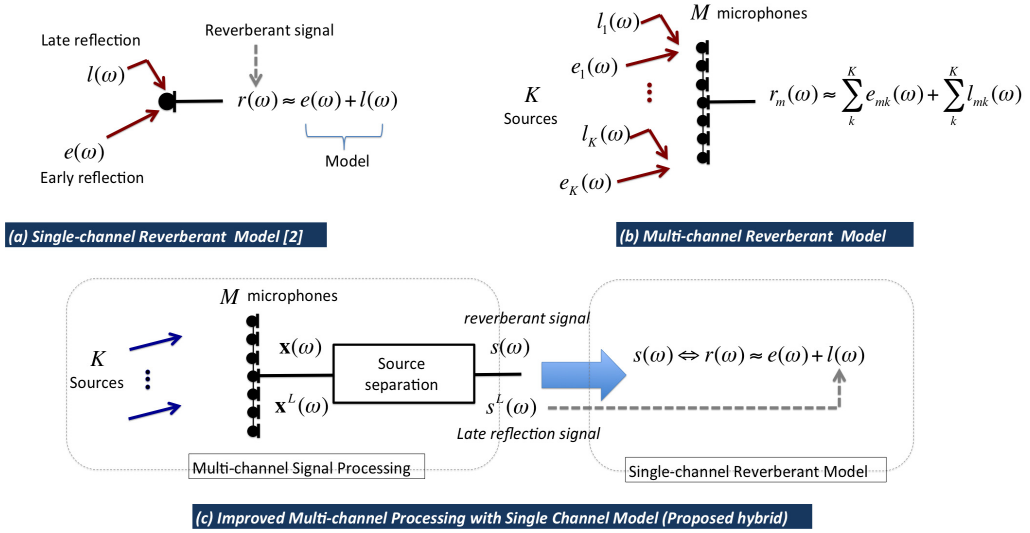

Fig. 2. Different reverberant speech models.

number of dereverberation techniques utilize pre-measured RTFs [4][5], a specific measurement for every possible azimuthal change may be impractical to realize.

This paper is organized as follows; in Section II, the background of dereverberation is discussed. We show the proposed method in Section III and in Section IV, we discuss the experimental setup, followed by ASR recognition results using real reverberant data collected in a human-robot communication environment in Section V. Finally, we conclude the paper in Section VI.

II. BACKGROUND

A. Single-Channel Reverberant Model

The reverberant speech model $r(\omega)$ as shown in Fig. 2(a) can be expressed as

$$r(\omega) = \quad e(\omega) \quad + \quad l(\omega), \tag{1}$$

where $e(\omega)$ and $l(\omega)$ are the early and late reflections. The model in Eq. (1) is adopted from [6][7] with the assumption that both are uncorrelated and independent. If the single channel room transfer function (RTF) $A(\omega)$ is available and the boundary attributed for the early and late reflections are identified, Eq. (1) can be expressed as

$$r(\omega) = A^E(\omega)c(\omega) + A^L(\omega)c(\omega), \tag{2}$$

where $c(\omega)$ denotes the clean speech, $A^E(\omega)$ and $A^L(\omega)$ are the early and late reflection coefficients of the full RTF $A(\omega)$ which are experimentally pre-determined in [1][2]. The late reflection which is considered harmful to the ASR can be treated as noise [1][2], and dereverberation is defined by suppressing $l(\omega)$ while recovering $e(\omega)$ estimate. The early reflection is further processed with Cepstrum Mean Normalization (CMN) within the ASR system. By inspection, Eq. (1) resembles a denoising problem, thus dereverberation can be treated as such. Specifically, through spectral subtraction (SS) [8], the estimate $\hat{e}(\omega)$ in frame-wise manner t is expressed as

$$|\hat{e}(\omega,t)|^2 = \begin{cases} |r(\omega,t)|^2 - |l(\omega,t)|^2 \\ \quad \text{if } |r(\omega,t)|^2 - |l(\omega,t)|^2 > 0 \\ \\ \beta |r(\omega,t)|^2 \quad \text{otherwise.} \end{cases} \tag{3}$$

where β is the flooring coefficient. In real condition, $l(\omega,t)$ is not accessible, precluding the power estimate $|l(\omega,t)|^2$. Thus, an approximation scheme in [1][2] is employed, estimating $|l(\omega,t)|^2$ directly from the observed reverberant signal $r(\omega,t)$ through the minimization of the error:

$$E_p = \frac{1}{T} \sum_t \sum_{\delta_p \in B_p} |l(\omega,t) - \delta_p(\omega,t) r(\omega,t)|^2, \tag{4}$$

and for the given set of bands $\boldsymbol{B} = \{B_1, \ldots, B_P\}$, the weighting parameter δ_p is determined through minimum mean square error criterion via offline training [9]. The new early reflection estimate $\hat{e}(\omega)$ becomes

$$|\hat{e}(\omega,t)|^2 = \begin{cases} |r(\omega,t)|^2 - \delta_p |r(\omega,t)|^2 \\ \quad \text{if } |r(\omega,t)|^2 - \delta_p |r(\omega,t)|^2 > 0 \\ \\ \beta |r(\omega,t)|^2 \quad \text{otherwise.} \end{cases} \tag{5}$$

The details of the single channel model-based dereverberation is discussed in [1][2].

B. Multi-Channel Reverberant Model

Let M and K be the number of microphones and sound sources, respectively such that M ($\geq K$) as shown in the multi-channel setup in Fig. 2(b). The multi-channel reverberant model can be expressed in the same manner as that of the single channel in Fig. 2(a) except for a total of K reflections (early and late) which are observable (i.e., $r_m(\omega)$) for each microphone m. In this case, all of the processes discussed in Sec. II-A will be implemented across M channels which is exhaustive in nature. To illustrate the complexity of this model, let us examine the optimization of the multi-band scaling parameters δ_p [9] expanded across m microphone channels, Eq. (4) becomes

$$E_{mp} = \frac{1}{M} \frac{1}{T} \sum_m \sum_t \sum_{\delta_p \in B_p} |l(\omega,t,m) - \delta_p(\omega,t,m) r(\omega,t,m)|^2. \tag{6}$$

This optimization is executed all throughout the speech utterances in the training database. Thus, the true multi-channel reverberant model would definitely strain the system both at training and at runtime, precluding realtime processing. We note that due to battery issues, robots are often equipped with an onboard computer having limited computational power. The very limited computing resources are often shared by several critical processes, not just by the ASR system.

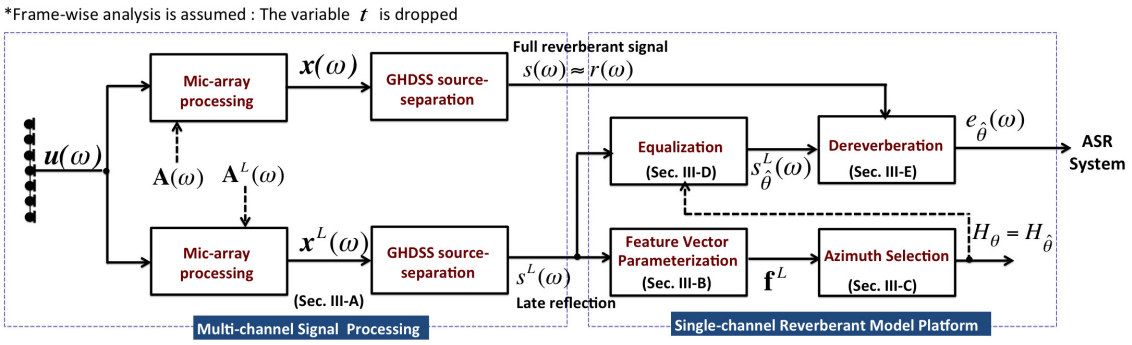

Fig. 3. Proposed hybrid dereverberation method.

C. Hybrid: Multi-Channel Signal Processing in Single-Channel Reverberant Model Platform

A temporary solution is presented in our previous work [3] but this approach requires multiple unique RTF measurements for each microphone, to match each possible change in the azimuth since RTF interpolation works only for changes in radial position. Multiple RTF measurements may be physically impractical but the most compromising attribute of our previous work presented in [3] is the inaccurate estimation of the late reflection based on single-channel processing. The method [3] is dependent on δ and its optimization (Eq. 6) is computationally expensive when done in multi-channels, thus, the method [3] is constrained to a single-channel based late reflection estimation, wasting the more accurate multi-channel processing advantage. In general, [3] can be viewed more of an engineering approach which renders the dereverberation task to be computationally feasible with less consideration on performance. Thus, we propose a hybrid scheme shown in Fig. 2(c), built on multi-channel signal processing analysis of superior signal estimation, having the efficiency of a single-channel model dereverberation platform. First, the multi-channel signals, as observed at the microphones $\boldsymbol{x}$, are resolved via sound-source separation into $s(\omega)$ (full reverberation) and $s^L(\omega)$ (late reflection), respectively. Then single-channel reverberant model analysis is applied to $s(\omega)$ decomposing $s(\omega)$ into early and late reflections where the latter is equivalent to the estimate $s^L(\omega)$. This hybrid mechanism is possible due to the elimination of the scaling parameters δ which will be explained in Sec III.

III. METHODS

The proposed hybrid dereverberation method is shown in Fig. 3. Microphone array processing and source-separation is employed resulting to single-channel reverberant signal $s(\omega)$ and the late reflection signal $s^L(\omega)$ estimates. Feature parameters $\mathbf{f}^L$ are extracted from the latter which are then used to identify the azimuthal information used to select the appropriate equalization parameter $H_{\hat{\theta}}$. Then, the late reflection is equalized to $s^L_{\hat{\theta}}(\omega)$. Finally, dereverberation is applied to the reverberant signal prior to the ASR system.

A. Multi-channel Processing

Let $\boldsymbol{u}(\omega)$ be the vector that consists of K sources as $\boldsymbol{u}(\omega) = [u_1(\omega), \cdots, u_K(\omega)]^T$, where T represents the transpose operator. And the vector containing the observed signals by the M microphones is

$$\boldsymbol{x}(\omega) = [x_1(\omega), \cdots, x_M(\omega)]^T \tag{7}$$

is the vector containing the signals received by M microphones. Suppose that the multi-channel RTFs in matrix form between the sources and the microphones are given as $\boldsymbol{A}(\omega) \in \mathbb{C}^{M \times K}$, Eq. (7) can be expressed as

$$\boldsymbol{x}(\omega) = \boldsymbol{A}(\omega)\boldsymbol{u}(\omega), \tag{8}$$

Assuming that the late reflection coefficients $\boldsymbol{A}^L$ of the full RTF $\boldsymbol{A}(\omega)$ is identified in advance as described in [2], the observed late reflection is given as

$$\boldsymbol{x}^L(\omega) = \boldsymbol{A}^L(\omega)\boldsymbol{u}(\omega). \tag{9}$$

In our method, the *Geometrically constrained High-order Decorrelation based Source Separation (GHDSS)* which is a combination of beamforming and blind source separation is employed for spatial separation of the multi-channel signals [10][11]. The separated full reverberant and late reflection estimates are expressed as

$$s(\omega) = \text{GHDSS}[\boldsymbol{x}(\omega)] \tag{10}$$

and

$$s^L(\omega) = \text{GHDSS}[\boldsymbol{x}^L(\omega)], \tag{11}$$

respectively.

B. Feature Parameterization

The late reflection $s^L(\omega)$ contains redundant information in the time domain. Thus, we extract only relevant information that best describes the signal characteristics of the late reflection. It was verified in our experiments that 12-order MFCCs, 12-order Δ MFCCs and 1-order Δ Energy are sufficient to effectively represent the late reflection characteristics. The parameterization process is expressed as

$$\mathbf{f}^L = F[s^L(\omega)], \tag{12}$$

where F denotes the feature extraction procedure resulting to the vector of features $\mathbf{f}^L$.

C. Azimuth Selection

The feature vectors $\mathbf{f}^L$ are evaluated based on likelihood score given as

$$\hat{\theta} = \arg\max_{\theta_g} p(\mathbf{f}^L | \mu_{\theta_g}), \tag{13}$$

where μ_{θ_g} is the probabilistic model for $\{\theta_1,...,\theta_g,...,\theta_G\}$ azimuthal orientations. The corresponding θ_g that maximizes the likelihood score is used to select the appropriate equalizer parameter $H_{\hat{\theta}}$ for equalization (Sec. III-D). We note that Eq. (13) is conducted at runtime, and requires an offline training procedure to train the probabilistic models prior to classification.

In the offline training of μ_{θ_g}, a synthetic late reflection signal is generated $s^L = A(\omega)u(\omega)$ similar to that in Eq. (11) except that we are operating in a single-channel. Then, the synthetic

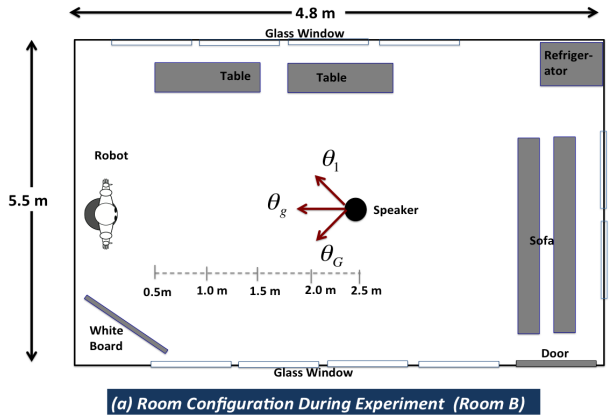

Fig. 4. Room configuration in our experiment

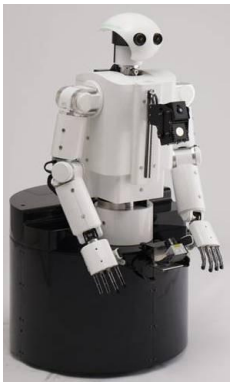

Fig. 5. Hearbo: The Honda Research Institute Robot platform.

signal is passed to an equalization process (Sec III-D) with various equalization templates H_θ which is then parameterized to

$$\mathbf{f}_\theta^L = F[s^L(\omega)H_\theta], \tag{14}$$

where H_θ is the equalization process discussed in Sec III-D. Consequently, models are trained for $\{\theta_1,\ldots,\theta_g,\ldots,\theta_G\}$ using the training vectors in Eq. (14) as

$$\hat{\mu}_\theta = \arg\max_{\mu} \prod_{i=1}^{I} \max_{\theta} P(\mathbf{f}_{\theta^i};\mu), \tag{15}$$

where μ is unknown model parameters, $\mathbf{f}_{\theta^i}$ is the i-th late reflection training vector which is equalized with H_θ. We note that as G is increased in θ_1,...,θ_g,...,θ_G, the azimuthal resolution is improved and may positively affect system performance as discussed in Fig. 9 in Sec. V.

D. 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 ones in our experiment as verified in Sec IV-B. 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 $s^L(\omega)$ instead of the multi-channel RTF characteristics. This scheme simplifies the supposed complicated analysis of the effect of the azimuthal orientation with respect to the multi-channel RTFs into simple single channel filtering. The equalized late reflection signal becomes

$$s_\theta^L(\omega) = s^L(\omega)H_\theta. \tag{16}$$

where $s^L(\omega)$ 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 $s^L_{\boldsymbol{A}_\theta}(\omega)$ 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 $s^L_{\boldsymbol{A}_\theta}(\omega)$ and H_θ for $\{\theta_1,\ldots,\theta_g,\ldots,\theta_G\}$ are derived by properly positioning the poles to achieve the target response $s^L_{\boldsymbol{A}_\theta}(\omega)$ [14]. Note that the target response $s^L_{\boldsymbol{A}_\theta}(\omega)$ was preprocessed via smoothing to avoid direct inversion problems [14].

At runtime, after the selection of the optimal $H_{\hat{\theta}}$ in Sec IV-C, Eq. (16) is used to correct the separated late reflection through equalization without requiring the matched RTFs $\boldsymbol{A}_\theta(\omega)$. With an effective θ selection procedure as discussed in Sec III-C, the equalization process virtually transforms the supposed multi-channel analysis into simple single channel filtering.

E. Dereverberation

After multi-channel signal processing analysis in which $s(\omega)$ and $s^L(\omega)$ are estimated, the single channel reverberant model is used as the dereverberation platform (see Fig. 2(a)). We note that the scaling parameters δ in Eq. (4) were employed to correct the late reflection estimate for the single channel model. Since late reflection is accurately estimated in the proposed method through multi-channel processing, δ is eliminated and so is its optimization process in Eq. (4). The effects of δ can be easily absorbed by the equalization process which also includes azimuthal correction as described in Sec III-D. The SS in Eq. (5) is modified to

$$|e_{\hat{\theta}}((\omega,t)|^2 = \begin{cases} |s(\omega,t)|^2 - H_{\hat{\theta}}(\omega)|s^L_{\hat{\theta}}(\omega,t)|^2 \\ \quad \text{if } |s(\omega,t)|^2 - H_{\hat{\theta}}(\omega)|s^L_{\hat{\theta}}(\omega,t)|^2 > 0 \\ \beta|s(\omega,t)|^2 \quad \text{otherwise.} \end{cases} \tag{17}$$

where $s|(\omega,t)|^2$ is the power of the separated reverberant signal ($s|(\omega,t)|^2 \approx r|(\omega,t)|^2$) and $s^L|(\omega,t)|^2$ 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. (17). In our previous method [3], 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.

IV. Experimental Set-up

A. Speech Database for ASR

The Japanese Newspaper Article Sentence (JNAS) corpus is used as the training database. The open test set is composed of 200 utterances from 24 speakers. The language model is a standard word trigram model while the acoustic model is a phonetically tied mixture (PTM) of Hidden Markov Models (HMMs) with 8256 Gaussians in total. Recognition evaluation is conducted on a 20K vocabulary Japanese dictation task in a human-robot communication setup as shown in Fig 4. Occlusions from tables, chairs, etc. are considered during testing to recreate a realistic environment. The proprietary humanoid robot, Hearbo of Honda Research Institute

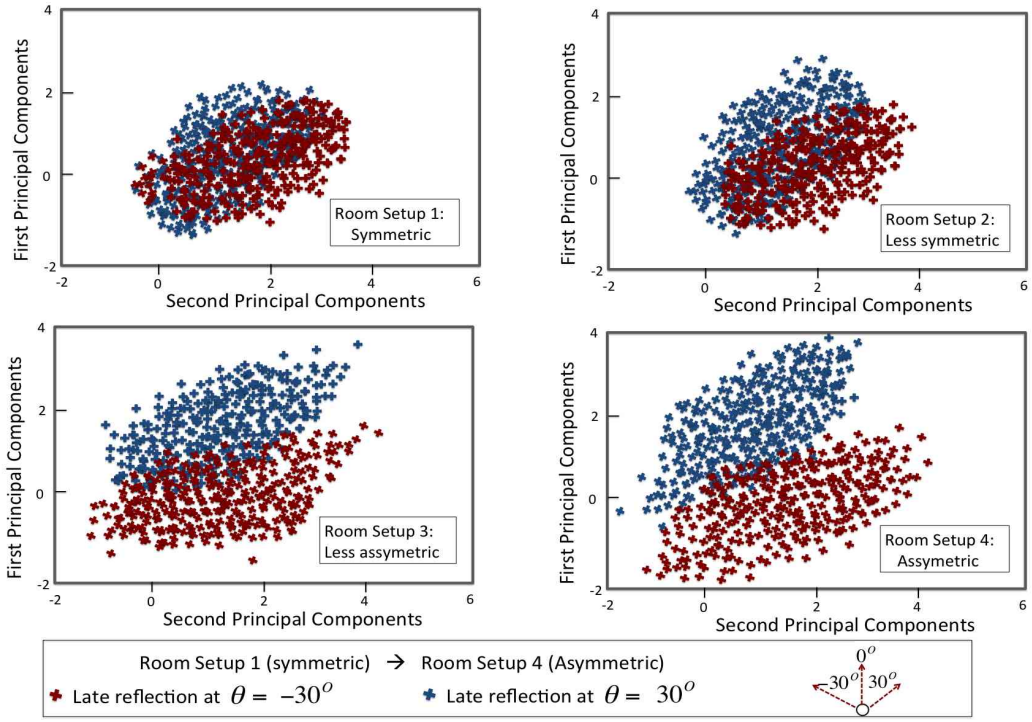

Fig. 6. Principal Component Analysis.

as shown in Fig. 5 is used as the experimental robot platform in which a microphone array is embedded on the head of the robot.

Real reverberant data are recorded inside two different reverberant rooms (Room A and Room B) with reverberation time (RT) of 240 msec. and 640 msec., respectively. Room B (Fig 4) is more acoustically challenging than Room A. There are six different location points $\{0.5\text{m}, 1.0\text{m}, 1.5\text{m}, 2.0\text{m}, 2.5\text{m}\}$ and five azimuths are considered, $\{\theta_1 = 30^o, \theta_2 = 15^o, \theta_3 = 0^o, \theta_4 = -15^o, \theta_5 = -30^o\}$ where 0^o is the reference axis perpendicular to the robot. Thus, each location point has a combined total of 1000 test utterances (200 utt. x 5 = 1000 utt.) for all of the five azimuth values. These utterances are then processed with different methods for comparison. We note that in the proposed method, we use the same room RTF for all the different test locations and angles as discussed in Sec III-D (no multiple RTFs are required).

B. Late Reflection Investigation

One of the contentions of this paper is that the late reflection varies with respect to the change of the azimuth θ_g. To effectively estimate the late reflection is to use multiple RTFs matched to the θ_g or through the proposed method using a generic RTF with equalization. In Fig. 6 we extract the principal components of the feature vectors of the late reflection at $\theta = 30^o$ and $\theta = -30^o$ (0^o is the reference axis perpendicular to the robot). The room setup is altered to simulate occlusions and symmetry. There are 4 levels of symmetry (setup 1 - setup 4) from a very symmetric room setup 1 towards an asymmetric room setup 4. It is shown in this figure that the late reflection can be approximated to be similar at either azimuth values when the room is symmetric (setup 1) as shown by the overlapping concentration of the red and blue crosses. However, this presumption fails as the room tends to become asymmetric (setup 4).

TABLE I
CLASSIFICATION ACCURACY OF AZIMUTH ($\hat{\theta}$) SELECTION.

	Position 1	**Position 2**	**Position 3**
Room setup 1 (Sym.)	100 %	100 %	100 %
Room setup 2 (Less sym.)	98 %	99 %	98 %
Room setup 3 (Less asym.)	94 %	95 %	95 %
Room setup 4 (Asym.)	89 %	88 %	90 %

C. Effectiveness of Azimuth Classification

The proper identification of the azimuth $\hat{\theta}$ is instrumental in selecting the optimal equalization parameter $H_{\hat{\theta}}$. Thus the overall performance of the proposed method depends on the correct identification of $\hat{\theta}$. We show the effectiveness of correctly classifying the azimuthal orientation using the scheme discussed in Sec. III-C in four different room setups used in Sec IV-B. In each room setup, three random positions (Positions 1-3) are selected in which the azimuth classification experiment is conducted. The classification accuracy is shown in Table 1 and it is apparent that azimuth classification performance is best when the room is more symmetric. Although, the level of performance achieved in an asymmetric room (setup 4) is slightly lesser than that achieved in symmetric rooms, the values are sufficient for the overall method to work well which is verified in the recognition performance in the following section.

V. RESULTS AND DISCUSSION

The ASR results in terms of word accuracy are shown in Fig. 7 for Room A with RT = 240 msec. and Fig. 8 for Room B with RT = 640 msec., respectively. Recognition results for all of the different azimuths $\{\theta_1 = 30^o, \theta_2 = 15^o, \theta_3 = 0^o, \theta_4 = -15^o, \theta_5 = -30^o\}$ are averaged in each location points $\{0.5\text{m}, 1.0\text{m}, 1.5\text{m}, 2.0\text{m}, 2.5\text{m}\}$.

The methods (A) and (B) are respective results of using single and multi-channel microphones with no dereverberation. In the latter, sound-source separation is employed. In the methods (C)-(G), multi-channel processing is employed, in which the resulting

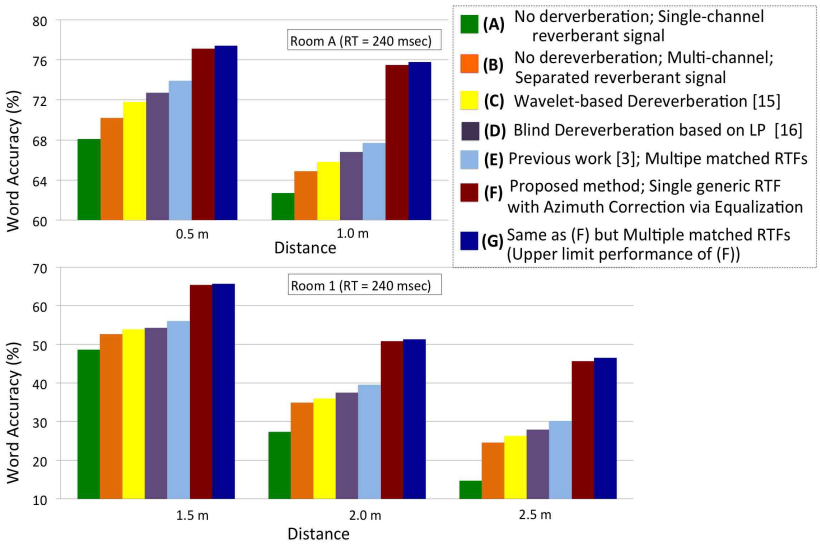

Fig. 7. Recognition results for room A.

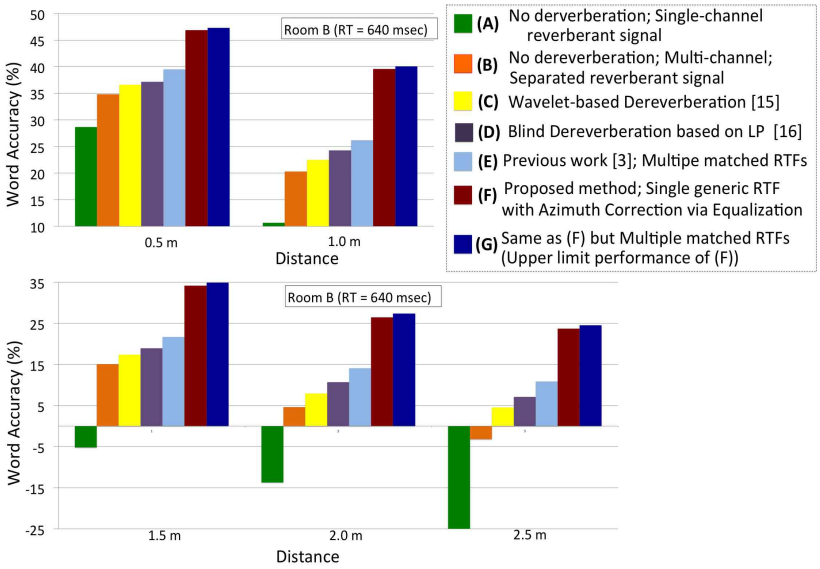

Fig. 8. Recognition results for room B.

separated reverberant data from method (B) is processed with different dereverberation methods, and will be compared in this section. The method (C) is a wavelet-based dereverberation technique [15] using a human speech production model in suppressing reverberant effects. An improvement in accuracy is achieved when suing the method (D) based on blind dereverberation [16]. Our previous work [3] is shown in method (E) while the proposed method based on a hybrid approach is shown in method (F). Method (G) is the same as that in method (F) except that method (G) does not employ equalization, instead it uses the actual multiple RTFs matching each θ_g (upper limit).

The results show that the proposed method (F) outperforms methods (A)-(E) and this can be attributed to the fact that the proposed method is accurate enough in capturing the acoustical dynamics of the late reflection as a function of azimuthal change θ_g. We note that in Fig. 6, the late reflection cannot be assumed to be the same for all θ_g which is a common assumption in the rest of the methods. The proposed method addresses this problem through effective late reflection estimation via multi-channel signal analysis and through equalization. Moreover, there is insignificant change in performance for matched RTF in method (G) as compared to the proposed method (F). We note that it is impractical to measure different RTFs as the speaker changes azimuthal orientation. The proposed method on the other hand only uses a single RTF, the same RTF used by the microphone array processing in sound-source separation.

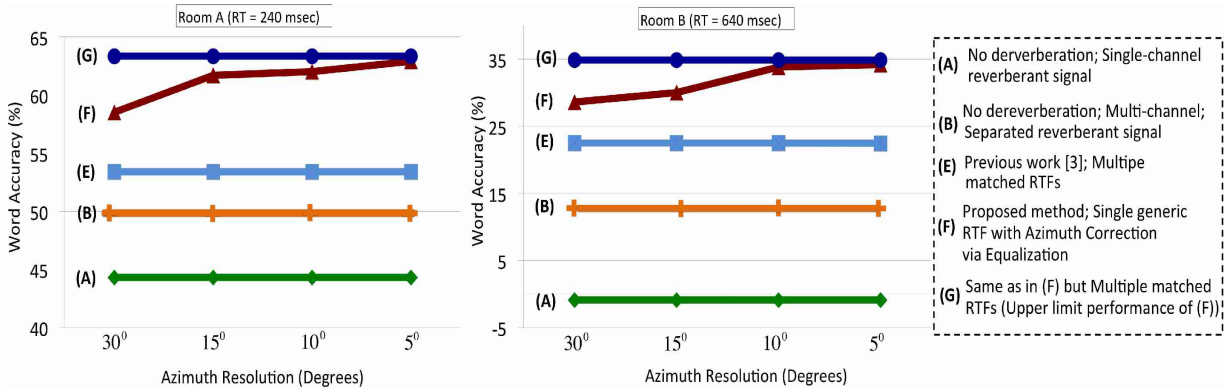

Fig. 9. Performance of the proposed method as a function of azimuth resolution.

In Fig. 9 we show the performance of the proposed method (F) while increasing azimuth resolution from 30^o-5^o. In this figure, (A): Single-channel no-dereverberation, (B): Multi-channel no-dereverberation, (E): Previous work [3] and (G): Upper limit. Only the proposed method (F) is a function of the variable resolution, the rest are constant. The result is averaged for all location points $\{0.5\text{m}, 1.0\text{m}, 1.5\text{m}, 2.0\text{m}, 2.5\text{m}\}$. In this figure, it is apparent that the performance of the proposed method improves as the resolution is increased and it saturates at around 5^o resolution. Moreover, it is verified that even with a very coarse resolution of 30^o, the proposed method still outperforms the previous work [3] in method (E). We note that both the methods (E) and (G) require multi-channel RTFs matched to the corresponding azimuth change while the proposed method uses only a single generic RTF.

Due to the different room setups (i.e., different RT), the recognition performance between Fig. 7 and Fig. 8 is disparate. Room A has lesser occlusions and shorter RT = 240 msec. while the latter is more reverberant (RT = 640 msec.) and with more occlusions. In addition, we stress that the experimental evaluation was conducted on a large vocabulary continuous dictation task. Unlike an isolated word recognition task, continuous dictation tends to be more susceptible to the effects of reverberation due to long-duration utterances. Consequently, the latter considers insertion errors, deletion errors, etc., and vocabulary size is at least 100 times more than isolated word recognition. Thus, recognition for a continuous dictation task is always lower than the isolated word task. The negative recognition values in Fig. 8 are attributed to insertion and deletion errors.

VI. CONCLUSION

We have shown a hybrid method by combining multi-channel signal processing analysis and a single-channel dereverberation model. This results in an effective late reflection estimate utilizing all of the microphones and at the same time, a simplified dereverberation procedure reminiscent of a single-channel reverberant model. The synergy of the proposed hybrid method renders the recognition performance of the system to be robust to the azimuthal change through a simple equalization in the late reflection signal. Traditionally, robustness to the change in azimuth is achieved by matching the RTFs for all of the microphones in accordance to the change in the azimuth. The proposed method mitigated this through a simple equalization without the need of dealing with multiple channels. We have shown that the the difference in recognition performance between the matched RTF and the proposed method is negligible and yet the latter is more simple and convenient. Since we are currently dealing with static occlusions inside a room, in the future, we will investigate the effects of moving occlusions towards a more realistic human-robot communication environment.

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

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

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

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

[9] R. Gomez, J. Even, H. Saruwatari, and K. Shikano, "Fast Dereverberation for Hands-Free Speech Recognition" *IEEE Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2008.

[10] 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.

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

[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] S. Griebel and M. Brandstein, "Wavelet Transform Extrema Clustering for Multi-channel Speech Dereverberation"

[16] 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.