Dereverberation for Active Human-Robot Communication Robust to Speaker's Face Orientation

Randy Gomez, Levko Ivanchuk, Keisuke Nakamura, Takeshi Mizumoto, and Kazuhiro Nakadai

Honda Research Institute Japan Co., Ltd.

10.21437/Interspeech.2015-80

Abstract

Reverberation poses a problem to the active robot audition system. The change in speaker's face orientation relative to the robot perturbs the room acoustics and alters the reverberation condition at runtime, which degrades the automatic speech recognition (ASR) performance. In this paper, we present a method to mitigate this problem in the context of the ASR. *First*, filter coefficients are derived to **correct the Room Transfer Function (RTF)** per change in face orientation. We treat the change in the face orientation as a filtering mechanism that captures the room acoustics. Then, joint dynamics between the filter and the observed reverberant speech is investigated in consideration with the ASR system. *Second*, we introduce a **gain correction** scheme to compensate the change in power as a function of the face orientation. This scheme is also linked to the ASR, in which gain parameters are derived via the Viterbi algorithm. Experimental results using Hidden Markov Model-Deep Neural Network (HMM-DNN) ASR in a reverberant robot environment, show that proposed method is robust to the change in face orientation and outperforms state-of-the-art dereverberation techniques.

Index Terms: Robust Robot Audition, Speech Enhancement, Dereverberation, Automatic Speech Recognition

1. Introduction

Reverberation is a phenomenon caused by the reflections of the speech signal in an enclosed environment. It smears the original speech due to the different time delays of arrival among the speech reflections. This phenomenon causes mismatch and degrades the ASR performance. To abate the effect of mismatch, the reverberant speech is enhanced, which is referred to as dereverberation. The problem concerning reverberation is further plagued when the room acoustics is perturbed as a result of the change in the speaker's face orientation. This event **alters the RTF resulting to another mismatch at runtime**. Consequently, the change in face orientation **affects the directivity pattern in which the speech is diffused, causing power issues**. There exists different types of dereverberation methods [1][2][13] but most of these have no mechanism in dealing with the acoustic perturbation due to the change in the speaker's face orientation.

In a human-robot communication scenario, the speaker may change its face orientation when communicating to the robot at any given time. Thus, the dereverberation mechanism should be able to cope with this mismatch as well. In this paper, we expand and improve our previous work [3] in mitigating the degradation of the ASR due to the change in the speaker's face orientation. The proposed method employs an **ASR-inspired RTF and gain correction** mechanisms to actively mitigate the changes in the room acoustics and the speech power due to the change in the face orientation. More importantly, the analysis and optimization employed in the proposed method is conducted jointly with the Hidden Markov Models (HMMs) for effective use in ASR application. These HMMs are used in the HMM-DNN ASR evaluation.

In our previous work [3], face direction compensation is achieved through **equalization**. The work in [3] is purely focused on the waveform compensation of the RTF and stops right there without any consideration of the HMMs [3]. Although [3] works well in enhancing the waveform, it has a very coarse treatment of the effect of dereverberation when applied to the HMM-DNN ASR. In contrast, the proposed method takes a HMM-centric approach, in both of the analysis and optimization procedures. In the proposed method, the change in the face orientation is hypothesized to impact the RTF as a filtering mechanism. **Filter coefficients are optimized in the context of the HMMs as per change in the speaker's face orientation**. This process ensures the link between the RTF and the HMMs. Next, we analyze the impact of the change in face the orientation to the power envelope of the speech signal. **Gain values are derived using the dual nature of the speech signal (i.e., acoustic waveform and the hypothesis) to characterize the change in power**. This mechanism links the power correction with the ASR system. Both the filter for RTF correction and the parameters for gain correction are used in the online dereverberation. Hence, the proposed method can adapt to the acoustic perturbation caused by the change in the speaker's face orientation. The derivation of these parameters are linked to the HMMs, a stark contrast from our previous work [3] which focuses purely on waveform enhancement only.

This paper is organized as follows; in Sec. 2, we show the background of the adopted dereverberation platform in our application. The schemes in extracting the filter coefficients, dereverberation parameter update and calculating gain parameters for power correction as per change in face orientation are discussed in Sec. 3. Experimental results and discussion are presented in Sec. 4, and we conclude the paper in Sec. 5.

2. Background

Microphone array processing based on beamforming and blind separation described in [9][17] is employed to convert the multi-microphone observed signals to a separated reverberant signal (single-channel). In our previous method [4][13], the smearing effect of reverberation is adopted from [15][5] and is solely dependent on the room transfer function (RTF) given as

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

where $r(\omega)$ is the separated reverberant speech w.r.t. ω frequency [9][17] and the right side of Eq. (1) is the reverberation model, where $c(\omega)$ 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

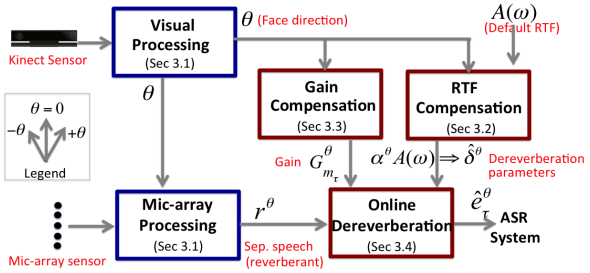

Figure 1: Overall System Structure.

of $e(\omega)$ and $l(\omega)$, known as the early and late reflections, respectively. In this paper, we represent both $A^E(\omega)$ and $A^L(\omega)$ simply as the full RTF $A(\omega)$. We note that the measured $A(\omega)$ is matched with a speaker talking in front of the robot and hypothetically, **a change in the face orientation would require different sets of RTF measurements** which is a cumbersome process. Hence, we **propose a correction method that does not require any measurement**.

In [13] we treat $l(\omega)$ as long-period noise which is detrimental to the ASR, and dereverberation is defined as suppressing $l(\omega)$ while 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 $|l(\omega,j)|^2$. Therefore, the observed reverberant signal $r(\omega,j)$ is used instead of $l(\omega,j)$. This is made possible through a scheme in [13] serving as a workaround to this problem. The scheme introduces a multi-band suppression parameter δ_m optimized via the ASR likelihood criterion given as

$$\delta_m = \arg\max_{\boldsymbol{\delta}_{m,c\Delta}} P(\mathbf{y}^{\boldsymbol{\delta}_{m,c\Delta}}|\mathbf{w};\boldsymbol{\lambda}), \tag{3}$$

where $\boldsymbol{\lambda}$ and $\mathbf{w}$ are the speech acoustic and language models, respectively. $c\Delta$ is the discrete step in the search space while $\boldsymbol{\delta}_{m,c\Delta}$ are the suppression parameter values to be searched upon. For a given set of bands $\boldsymbol{Q} = \{Q_1,\ldots,Q_m,\ldots,,Q_M\}$, in the frequency ω, the dereverberation parameter δ_m dictates the extent of the suppression of the reverberant effects. The new estimate $\hat{e}(\omega,j)$ through the modified SS becomes

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

It is obvious that the dereverberation platform in Eq. (4) is dependent on the dereverberation parameter δ_m. Consequently, δ_m **depends on the RTF** $A(\omega)$ as depicted in the model in Eq. (1) and needs to be corrected depending on the speaker's face orientation. Although Eq. (1) is effective for waveform enhancement, its formulation has no relation with HMM analysis. Thus, dereverberation performance is very limited to the original face orientation. In this paper, we will show the method of effectively correcting $A(\omega)$ as a function of the speaker's face orientation. The simplified block diagram of the proposed method is shown in Fig. 1. In the proposed method, the mechanism for RTF and power correction is implemented via an offline training scheme according to the change in the face orientation θ. The updated suppression parameters $\hat{\delta}_m^\theta$ resulting from RTF compensation with $\alpha^\theta A(\omega)$ and the gain parameters $G_{m_\tau}^\theta$ are stored for online dereverberation use. Details on Fig. 1 are discussed in the following section.

3. Methods

3.1. Microphone-array and Visual Processing

Sound source separation described in [9][17] is used to obtain the separated reverberant signal r^θ, where θ is the speaker's face orientation. It is defined by setting a straight line between the human and the robot (facing each other) as a reference axis. The change in speaker orientation is defined as the angular change θ from the reference axis from the human side. In our work we consider a deviation $-30 \leq \theta \leq 30$, where $\theta = 0$ is the reference angle in which the generic RTF is defined. The angle θ is estimated using the Kinect sensor.

3.2. Room Transfer Function Correction

Suppose that the observed reverberant speech at a particular face orientation θ when processed by a filter is given as

$$x^\theta[h] = \sum_{k=0}^{K-1} \alpha_k^\theta \; r^\theta[h-k], \tag{5}$$

where r^θ and α_k^θ are the observed reverberant speech and the filter coefficients, respectively. We note that the room acoustics information is captured in the observed reverberant speech via reflections on the enclosed space. We use the actual signal r^θ to analyze the reverberation condition as per change in face direction θ through the filter $\boldsymbol{\alpha}^\theta$. The filter of length K is given as

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

The objective is to estimate $\boldsymbol{\alpha}^\theta$ in the context of the ASR. The resulting estimate captures the room acoustics at θ, and later used not just to correct the change in θ but making sure that the correction is more likely to improve the ASR performance. Since we are interested of the ASR's output (hypothesis), the actual signal x is immaterial. The hypothesis is expressed as

$$\hat{\boldsymbol{w}}^\theta = \underset{\boldsymbol{w}}{\text{argmax}} \log\,(P(f^{(x^\theta)}(\boldsymbol{\alpha}^\theta)|\boldsymbol{w})P(\boldsymbol{w}), \tag{7}$$

where $f^{(x^\theta)}(\boldsymbol{\alpha}^\theta)$ is the extracted feature vector from the utterance, $\boldsymbol{w}$ is the phoneme-based transcript, $P(f^{(x^\theta)}(\boldsymbol{\alpha}^\theta)|\boldsymbol{w})$ 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 $\boldsymbol{\alpha}^\theta$ and rewritten as

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

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

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

where s_j is the state at frame j. Eq. (9) heralds the formulation in the context of the HMMs via the state sequence. By using

the ∇ operator, the total probability is maximized w.r.t the filter coefficient in Eq. (6), thus,

$$\nabla_{\boldsymbol{\alpha}^\theta}\ \Gamma(\boldsymbol{\alpha}^\theta)\ = \left\{ \frac{\partial\Gamma(\boldsymbol{\alpha}^\theta)}{\partial\alpha_0^\theta}, \frac{\partial\Gamma(\boldsymbol{\alpha}^\theta)}{\partial\alpha_1^\theta}, ..., \frac{\partial\Gamma(\boldsymbol{\alpha}^\theta)}{\partial\alpha_{K-1}^\theta} \right\}. \quad (10)$$

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

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

where γ_{jv} is the posteriori of v-th mixture and j-th frame of the most likely HMM state. $\frac{\partial f_j{}^{(x^\theta)}(\boldsymbol{\alpha}^\theta)}{\partial\boldsymbol{\alpha}^\theta}$ is the Jacobian matrix of the reverberant feature vector. The filter coefficients are obtained using [11][12] based on Eq. (11). Correcting a generic RTF to the current face orientation θ of the speaker is given as

$$\hat{A}^\theta(\omega) = \alpha^\theta(\omega) A(\omega) \quad (12)$$

where $\alpha^\theta(\omega)$ is the face orientation-compensating filter in the frequency domain. It follows that a new dereverberation parameter can be extracted from the corrected RTF,

$$\hat{A}^\theta(\omega) \Rightarrow \hat{\delta}_m^\theta \quad (13)$$

The updated dereverberation parameters $\hat{\delta}_m^\theta$ are stored for online use in Sec 3.4.

3.3. Speech Power Compensation via Gain Correction

The change in face orientation does not only impact the RTF, but it also affects the power level of the separated signal r^θ. To mitigate the effect of the latter, we employed a power compensation scheme via gain correction. The process of deriving the gain is depicted in Fig. 2. Two sets of reverberant speech database are prepared, one is recorded facing directly the robot θ_A (s.t. $\theta = 0$), and the other set with face orientation θ_B (s.t. $\theta_B \neq 0$). θ_A is the reference face orientation in which θ_B is to be corrected to. The utterances are classified according to the time-duration referred to as template τ. Same duration utterances are grouped together (**time-duration classification**). We note that reverberation is characterized by the smearing phenomenon in which the power of the previous sound frames are carried over to the current frame. In this regard, the effect of reverberation is directly related to the duration of the speech utterance. Hence, it is noteworthy to analyze the impact of both the changes in the face orientation and speech duration, respectively. Consequently, the reverberant utterances are referred to as $r_\tau^{\theta_A}$ and $r_\tau^{\theta_B}$, respectively. Next, we analyze the change in power dynamics per change in face orientation θ_B relative to θ_A.

To effectively establish the correspondence of the sound units (i.e. phonemes) between the two utterances in θ_B and θ_A, the utterances are **aligned via the Viterbi algorithm** using a known acoustic speech model λ. This is a very crucial step because we want to model the change in power similar to the concept of the reverberation phenomenon in which the energy of the current frame is affected by the previous frames. To achieve that, we need to have a correct association of the sound-frames between the speech database A and B. The alignment will guarantee that the particular sound of the current frame of interest in r^{θ_A} likely corresponds the same sound in r^{θ_B}, one-to-one correspondence is achieved. Moreover, the alignment scheme links

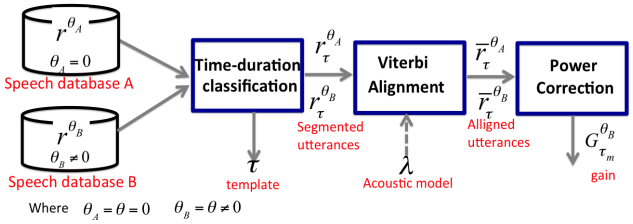

Figure 2: The offline training scheme used to calculate gain parameters for power gain correction.

the power analysis between the acoustic waveform and the hypothesis which are both used by the ASR system.

Frame-wise power spectral analysis is conducted to the aligned utterances $\bar{r}_\tau^{\theta_A}$ and $\bar{r}_\tau^{\theta_B}$ for face orientation θ and the template τ, respectively. The reverberant power of both are compared and analyzed. Then, band coefficients that minimizes the error between the two are extracted. The minimization of the error means minimizing the power mismatch between $\bar{r}_\tau^{\theta_A}$ and $\bar{r}_\tau^{\theta_B}$. For a total of O utterances indexed by o in a template τ, the error to be minimized is given as

$$E_\tau^{\theta_B}(j) = \frac{1}{\mathrm{O}} \sum_{\mathrm{o}} \sum_{\omega \in Q} |\bar{r}_\tau^{\theta_A}(\omega, \mathrm{o}, j) - G_{\tau_m}^{\theta_B}(\omega, \mathrm{o}, j) \bar{r}_\tau^{\theta_B}(\omega, \mathrm{o}, j)|^2, \quad (14)$$

where $G_{\tau_m}^{\theta_B}$ is the gain for the given set of bands $\boldsymbol{Q} = \{Q_1, \ldots, Q_m, \ldots, Q_M\}$ of template τ. $\bar{r}_\tau^{\theta_A}(\omega, \mathrm{o}, j)$ and $\bar{r}_\tau^{\theta_B}(\omega, \mathrm{o}, j)$ are the j-th frame viterbi-aligned utterance o from the speech database A and B, respectively. Since we are interested of the power dynamics for each frame in a given template τ, the summation in Eq. (14) is conducted on the same frame index across O. For a given template τ of j frames, we extract a sequence of multi band m gain values of $[\mathbf{G}_{\tau_m}^\theta(\omega, 1), \ldots, \mathbf{G}_{\tau_m}^\theta(\omega, j), \ldots, \mathbf{G}_{\tau_m}^\theta(\omega, J)]$, for **power correction**. These values are then stored for online use in Sec 3.4.

3.4. Online Dereverberation

In the online mode (see Fig. 1), the visual processing scheme identifies the face orientation θ while the microphone array processing scheme converts the multichannel signal to a single channel separated reverberant signal r^θ. RTF and gain correction due to the change in face orientation θ as discussed in Sec 3.2-3.3 are used for dereverberation. Specifically, the adopted dereverberation platform based on spectral subtraction in Eq. (4) is rewritten as

$$|\hat{e}_\tau^\theta(\omega, j)|^2 = \begin{cases} |r_\tau^\theta(\omega, j)|^2 - \hat{\delta}_m^\theta G_{\tau_m}^\theta(\omega, j) |r_\tau^\theta(\omega, j)|^2 \\ \quad \text{if } |r_\tau^\theta(\omega, j)|^2 - \\ \quad \hat{\delta}_m^\theta G_{\tau_m}^\theta(\omega, j) |r_\tau^\theta(\omega, j)|^2 > 0 \\ \beta |r_\tau^\theta(\omega, j)|^2 \quad \text{otherwise.} \end{cases} \quad (15)$$

Note that $\hat{\delta}_m^\theta$ and $G_{\tau_m}^\theta$ are the pre-stored values discussed in Sec 3.2-3.3 and are selected based on θ as identified through the visual processing scheme.

4. Experimental Results

4.1. Setup

We evaluate the proposed method in large vocabulary continuous speech recognition (LVCSR) based on a HMM-DNN framework. The training database is the Japanese Newspaper Article Sentence (JNAS) corpus with a total of approximately 60 hours of speech. The test set is composed of 200 sentences

Table 1: Recognition performance in word accuracy (%)

Reverberation Time = 940 msec.@ Distance = 2.0 m	$\theta = -30$	$\theta = -15$	$\boldsymbol{\theta = 0}$	$\theta = +15$	$\theta = +30$
(A) No Enhancement	45.5 %	53.0 %	**64.7 %**	54.7 %	48.6 %
(B) Based on Feature Adaptation [16]	55.1 %	62.2 %	**70.0 %**	62.9 %	56.4 %
(C) Based on Wavelet Extrema [2]	57.3 %	63.7 %	**71.8 %**	63.2 %	57.1 %
(D) Based on LP Residuals [1]	59.7 %	65.4 %	**74.2 %**	66.1 %	59.3 %
(E) Based on Equalization (Previous work) [3]	68.1 %	75.9 %	**81.3 %**	76.5 %	69.3 %
(F-a) Proposed Method (RTF Comp. (Sec. 3.2))	74.9 %	77.4 %	**81.3 %**	78.1 %	75.7 %
(F-b) Proposed Method (RTF and gain Comp. (Sec. 3.2 & Sec. 3.3))	76.8 %	79.2 %	**81.3 %**	79.9 %	77.0 %
(G) Dereverberation with θ-matched RTF (Upperlimit) [13]	78.7 %	80.4 %	**81.3 %**	80.7 %	79.3 %
Reverberation Time = 940 msec. @ Distance = 3.0 m	$\theta = -30$	$\theta = -15$	$\boldsymbol{\theta = 0}$	$\theta = +15$	$\theta = +30$
(A) No Enhancement	30.7 %	37.2 %	**52.7 %**	40.5 %	32.1 %
(B) Based on Feature Adaptation [16]	37.0 %	43.4 %	**58.7 %**	44.7 %	36.8 %
(C) Based on Wavelet Extrema [2]	40.5 %	48.7 %	**62.4 %**	49.0 %	42.3 %
(D) Based on LP Residuals [1]	45.2 %	51.3 %	**66.1 %**	52.5 %	45.8 %
(E) Based on Equalization (Previous work) [3]	52.6 %	58.3 %	**73.9 %**	59.1 %	52.1 %
(F-a) Proposed Method (RTF Comp. (Sec. 3.2))	58.0 %	65.2 %	**73.9 %**	66.7 %	59.1 %
(F-b) Proposed Method (RTF and gain Comp. (Sec. 3.2 & Sec. 3.3))	63.8 %	67.3 %	**73.9 %**	68.8 %	64.9 %
(G) Dereverberation with θ-matched RTF (Upperlimit) [13]	65.8 %	69.2 %	**73.9 %**	70.4 %	66.7 %

uttered by 50 speakers. The vocabulary size is 20K and the language model is a standard word trigram model. Speech is processed using 25ms-frame with 10 msec shift. The fBank features of 40 dimensions. The HMM-DNN has 6 layers with 2048 nodes. The reverberation time is approximately 940 msec., and testing is conducted at 2.0 m and 3.0 m distances, respectively. Speaker face orientation θ is defined in degree. The generic RTF matching that of the model training is at $\theta = 0$, in which the speaker is directly facing the robot. The test speakers' face orientation deviates at $\theta = -30, -15, +15, +30$, respectively. Key to evaluating the results of the different methods is the robustness of the recognition performance as θ deviates from $\theta = 0$ (matched condition) to $-30 \leq \theta \leq +30$ (mismatched conditions). The test data are recorded at $\theta = -30, -15, +15, +30$. This is done by re-playing the clean test database using a loudspeaker at angle θ and distances 2.0m and 3.0m, respectively. Hence, we use real reverberant speech.

4.2. ASR Performance

The ASR results are shown in Table 1. Method (A) is when no enhancement is employed while method (B) is the result based on feature adaptation by [16]. Instead of suppression, method [16], minimizes the reverberant mismatch through adaptation of the feature vector. The result in method (C) is based on wavelet extrema clustering [2], which operates in the wavelet domain to remove the effects of reverberation. Method (D) is based on the Linear Prediction residual approach [1]. By exploiting the characteristics of the vocal chord, it is able to remove the effects of reverberation. The method in (E) is based on our previous work [3] which employs an equalization technique to mitigate the change in face orientation. The proposed method (F-a) is evaluated when only the RTF compensation is in effect (Sec. 3.2); and (F-b) when both the RTF and gain compensation are employed (Sec. 3.2 and Sec 3.3), respectively. In method (G), the result of using a θ-matched RTF is shown; RTF are measured for each microphone and for each change in θ. The result in method (G) serves as the upperlimit for the adopted dereverberation platform. We note that methods (E)-(G) use the same dereverberation platform and differs only in the mitigation of the change in the face orientation. Therefore, methods (E)-(G) have the same performance at $\theta = 0$.

Table 1 shows that the proposed method outperforms the existing methods and the previous work [3]. The recognition performance is robust to degradation when face orientation changes relative to the original condition $\theta = 0$. Moreover, it outperforms the previous work in method (E) [3]. This is because the proposed method is linked to the ASR system. The formulation to mitigate the change in the face orientation (i.e., RTF and gain corrections) evolves within the HMM construct. This hinged the optimization procedure to the ASR system itself. In contrast, the previous work and the rest of the methods are focused primarily on the waveform enhancement only.

5. Conclusion

In this paper, we have shown the method of analyzing the impact of the change in the face orientation through the alteration of both the RTF and power. These two creates a mismatch that degrades ASR performance when using the dereverberation framework. Moreover, we compensate its impact to the RTF by correcting it using optimized filter coefficients, specifically derived in the context of ASR. Also, the impact in power is corrected as per change in face orientation. Considerable amount of time is needed when measuring new RTFs. In the proposed method, the re-measurement of the RTF as a function of the face orientation can be avoided, this allows the robot to actively mitigate its impact online. We have compared our results with existing dereverberation methods, our previous work and the method when using a matched RTF.

Currently, our work is limited to the definition of the change in face orientation based on our experiment. In real world, the face orientation is more unpredictable resulting to unsymmetrical face orientation relative to the robot. In our future work, we will improve the current system to include random face directions. Although the proposed method involves the concept HMM in deriving the dereverberation and gain parameters, we did not consider actual model adaptation in this work. Hence, the latter will be part of our future work as well.

6. 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] R. Gomez, K. Nakamura, T. Mizumoto and K. Nakadai, "Dereverberation Robust to Speaker's Azimuthal Orientation in Multi-channel Human-Robot Communication" *In Proceedings IEEE Intelligent Robots and Systems IROS*, 2013.

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