Multi-party Human-Robot Interaction with Distant-Talking Speech Recognition

Randy Gomez
Academic Center for Computing and Media Studies, Kyoto University
Sakyo-ku Kyoto, JAPAN
randy-g@media.kyoto-u.ac.jp

Tatsuya Kawahara
Academic Center for Computing and Media Studies, Kyoto University
Sakyo-ku Kyoto, JAPAN
kawahara@media.kyoto-u.ac.jp

Keisuke Nakamura
Honda Research Institute
Wako-shi, JAPAN
nakamura@hri.jp

Kazuhiro Nakadai
Honda Research Institute
Wako-shi, JAPAN
nakadai@hri.jp

ABSTRACT

Speech is one of the most natural medium for human communication, which makes it vital to human-robot interaction. In real environments where robots are deployed, distant-talking speech recognition is difficult to realize due to the effects of reverberation. This leads to the degradation of speech recognition and understanding, and hinders a seamless human-robot interaction. To minimize this problem, traditional speech enhancement techniques optimized for human perception are adopted to achieve robustness in human-robot interaction. However, human and machine perceive speech differently: An improvement in speech recognition performance may not automatically translate to an improvement in human-robot interaction experience (as perceived by the users). In this paper, we propose a method in optimizing speech enhancement techniques specifically to improve automatic speech recognition (ASR) with emphasis on the human-robot interaction experience. Experimental results using real reverberant data in a multi-party conversation, show that the proposed method improved human-robot interaction experience in severe reverberant conditions compared to the traditional techniques.

Categories and Subject Descriptors

H.5.2 [Information Interfaces and Presentation]: User Interfaces—Voice I/O

General Terms

Algorithms, Design, Human Factors, Performance

1. INTRODUCTION

The notion of robots being a fixture in a manufacturing assembly line is a thing in the past. Recent developments in the use of professional service robots is gaining popularity and is expected to grow dramatically in the next decade. Thus, a potential need for autonomous robots equipped with complex mechanical functions in performing different tasks to work side by side with humans. An example of this is a companion robot in healthcare facilities, hospitals and even at homes. The robot may assist patients performing simple tasks (e.g in taking pills on schedule, pick up newspapers etc.). It can also provide help in accomplishing strenuous activities such as lifting and moving the patient from one place to another. As robot applications are geared toward collaboration with humans in day to day activities, it is safe to assume that its functionality is not just merely mechanical in nature, but it is becoming more personal as well. Given this scenario, it is desirable to have a seamless human-robot interaction where one can communicate with each other at ease. Since speech communication is one of the most natural interface, its use is paramount towards a more effective human-robot interaction. Surveys show that humans are more comfortable in using voice commands when interact-
ing with robots, rather than using a keyboard. In fact, it was found out that ordinary users tend to speak freely when using speech-based systems. User interest is high when the robot recognizes and understands correctly the spoken commands, otherwise, the tendency of becoming dissatisfied is imminent when they have to repeat the spoken command again and again. The latter is most likely to happen in real environments where robots are deployed. Distant-talking speech recognition is a very difficult task due to the effects of reverberation as depicted in Fig. 1. This is a phenomenon attributed by the reflections of the speech signal within the room enclosure. Due to the delays in the time of arrival of the reflected speech signal, the observed signal at the microphone is a smeared version of the actual speech. In a reverberant scenario, the late reflection is considered to be the most detrimental to the ASR. Depending on the speaker-robot distance and the room characteristics, the effects of reverberation may drastically degrade ASR performance. The degradation may be severe that would render the system to be useless for human-robot interaction application. Thus, it is important to enhance the corrupted speech signal prior to speech recognition and understanding. Conventional methods address this problem to improve intelligibility in human perception. This technique is later adopted to achieve robustness in human-robot communication application. Although this works well, the enhancement process is carried out independently from the automatic speech recognition (ASR) system. We note that computers recognize speech differently from humans.

In this paper we will show a method in optimizing human-inspired speech enhancement techniques to be closely linked to the ASR system for improved human-robot interaction experience as depicted in Fig. 2. The objective of the proposed method is to look at the overall impact of the speech enhancement as far as human-robot interaction is concerned. We tailor-fit the speech enhancement technique with the ASR system using maximum likelihood (ML) criterion and evaluate its impact to actual human-robot interaction experience. Although existing methods have been proposed regarding optimization of speech enhancement techniques together with the ASR [2][10][13], these works only tackle the Speech Enhancement->ASR scenario and stops right there. However, there is more to speech enhancement than just recognition performance (ASR figures). We need to investigate whether the speech enhancement translates into improving human-robot interaction experience. Thus, in this paper we consider the complete scenario: Speech Enhancement->ASR->human-robot interaction. Moreover, we evaluate the effectiveness of the proposed approach using real reverberant data in a multi-party conversation scenario which is not addressed in [2][10][13](single-party with no interaction). We used the Honda Research Institute robot platform called “Hearbo” shown in Fig. 3 to interact with humans and answer queries.

The organization of the paper is as follows; in Section 2, we show the background of achieving robustness in a speech-based human-robot interaction system. In Section 3, we present the traditional speech enhancement techniques, followed by the proposed optimization method, linking it with the ASR in Section 4. In Section 5, we discuss the experimental set-up, followed by the results and discussion in Section 6. We will conclude this paper in Section 7.

2. BACKGROUND

A block diagram of a robust speech-based human robot interaction system is shown in Fig. 4. In this figure, a human-robot interaction is initiated from the human side through speech communication. Due to channel effects, the observed signal is smeared with the reflected speech (i.e. late reflection). The two main components of Fig. 4 are summarized as follows:

2.1 Speech Enhancement

Signal processing technique is employed to suppress the effects of late reflection, recovering the original speech. During the processing of the observed speech, artifacts are consequently introduced to the enhanced speech. While, humans tend to be more robust to the presence of artifacts...
The enhanced speech signal is then recognized by evaluating the likelihood \( P(\hat{S}|W; \lambda) \) by the ASR. \( W \) is the language model trained by using text database, and \( \lambda \) is the acoustic model trained using speech database. A mismatch between the training condition of \( \lambda \) and the observed reverberant data \( O \) degrades the ASR performance especially when the latter is trained with clean speech data. Thus, speech enhancement is employed to process \( O \) resulting to \( \hat{S} \). This minimizes the effect of mismatch. The ASR system outputs a hypothesis (in text format) reflecting the recognized speech. This is then processed using conditional random field (CRF) to pick-up the slots that convey meaning for spoken language understanding. In this part here, the unnecessary entries in the sentence are removed, selecting only the slot that conveys meaning to our application. The selected slot is then searched in the database.

3. TRADITIONAL SPEECH ENHANCEMENT

In this paper, we consider three different speech enhancement methods designed for human perception. We adopt the reverberant model in [2] that treats the additive effect of late reflection to the clean speech

\[
O \approx S + R
\]

where \( O, S \) and \( R \) are the observed reverberant speech, clean speech and the late reflection respectively. Enhancement is defined as processing \( O \) and obtaining the estimate \( \hat{S} \).

3.1 Wavelet Thresholding

A technique that operates in the wavelet domain \( (w) \), where the observed speech is enhanced based on thresholding [7]. The enhanced speech estimate is expressed as

\[
\tilde{S}(w) = \begin{cases} 
0 & \text{if } |O(w)| \leq \text{thr} \\
\text{sign}(O(w))(|O(w)| - \text{thr}) & \text{if } |O(w)| > \text{thr}.
\end{cases}
\]

The threshold \( \text{thr} \), is derived using the wavelet coefficients of the late reflection. Any value that goes lesser than this, belongs to the subspace occupied by the late reflection. Thus, based on \( \text{thr} \), Eq. (2) can be interpreted as setting the subspace occupied by late reflection to zero, and implementing a magnitude subtraction in the observed signal subspace. The threshold that defines the late reflection subspace can be calculated [7] as

\[
\text{thr} = \sigma \sqrt{2 \log(L)},
\]

where \( L \) is the length of the late reflection signal \( R(w) \) with variance \( \sigma^2 \).

3.2 Spectral Subtraction

This is a classic speech enhancement technique which operates in the frequency domain \( (f) \) originally proposed for denoising in [3] [4] and later expanded to address reverberation in [5][6]. The short-term Fourier transform estimate of the enhanced speech is given as

\[
\tilde{S}(f) = \begin{cases} 
\sqrt{|S(f)|^2} e^{j\phi(f)} & \text{if } |S(f)|^2 > 0 \\
\beta|O(f)|^2 e^{j\phi(f)} & \text{otherwise},
\end{cases}
\]

where \( |S(f)|^2 \) and \( |O(f)|^2 \) are the short-term power spectra of the clean speech and the observed reverberant signal respectively. \( \beta \) is the constant flooring coefficient derived experimentally. The phase \( \phi(f) \) is derived from the observed signal. Since the clean speech is not available in real scenario, we can approximate its power spectrum by subtracting the power of the observed signal with the power of the late reflection. We note that the latter can be directly estimated as described in [5][6]. Thus,

\[
|S(f)|^2 \approx |O(f)|^2 - |R(f)|^2.
\]

Here, \( |O(f)|^2 \) and \( |R(f)|^2 \) are the short-term power spectra of the observed reverberant signal and late reflection, respectively.

3.3 Wiener Filtering

This technique is based on the Wiener Kinchine theorem, where the observed reverberant signal is filtered with a
Wiener gain for enhancement [11]. Originally developed for denoising, and later expanded to address reverberant condition [10]. The general expression of the Wiener gain [10] is given as

\[ \kappa = \frac{S(w)^2}{S(w)^2 + R(w)^2}, \]  

where \( S(w)^2 \) and \( R(w)^2 \) are the short-term power in the wavelet domain \( w \) of the clean speech and the late reflection respectively. The latter can be directly estimated as described in [5][6], while the former is estimated by subtracting the observed signal with the late reflection power similar to the one described in spectral subtraction. Wiener Filtering in the wavelet domain is carried out by weighting the observed reverberant wavelet coefficients \( O(w) \) with the Wiener gain as,

\[ \tilde{S}(w) = O(w) \cdot \kappa. \]  

In Eq. (7), the Wiener weighting \( \kappa \) dictates the degree of suppression of the late reflection to the observed signal. If the late reflection power estimate is greater than the estimate of the speech power, then \( \kappa \) may be set to zero or a small value, attenuating the effect of the late reflection and enhancing observed reverberant signal to \( \tilde{S}(w) \).

4. PROPOSED ASR-OPTIMIZED SPEECH ENHANCEMENT

Although the traditional speech enhancement techniques used for ASR applications are performing well, robustness can be further achieved when optimizing directly with the ASR. When the reverberant speech is processed, artifacts are introduced to the enhanced speech. The human auditory system together with the brain may be robust to these artifacts but not the ASR. Since the ASR processes speech differently from that with humans, it is logical to closely link the speech enhancement technique with the ASR. In Fig. 5 we show a method of optimizing the speech enhancement technique jointly with the ASR. The details are as follows,

4.1 Optimization Process

- **PHASE 1**
  The reverberant signal \( O \) is synthesized using clean speech \( S \) database and the room impulse response [8] in a method described in [9]. This approach is a common practice, since gathering real reverberant data for training is difficult. Then, \( \delta \) parameter is introduced in the signal processing resulting to the enhanced signal \( \tilde{S}(\delta) \). This parameter is used to control the amount of suppression of the late reflection, and in effect, controls the amount of artifacts introduced during the enhancement process. Consequently, \( \delta \) is optimized iteratively with the ASR using the ML criterion, in which \( \delta_{\text{opt}} \) that maximizes the model likelihood of the ASR is selected. The use of optimal \( \delta_{\text{opt}} \) ensures the ASR to handle the artifacts without performance degradation. The corresponding \( \delta_{\text{opt}} \) when used in enhancing the signal guarantees an improved in ASR performance compared to any other \( \delta \) values.

- **PHASE 2**
  Model is updated with data \( \tilde{S}(\delta_{\text{opt}}) \) that are processed using ASR-optimized enhancement technique. This results to a new model \( \lambda_{\text{opt}} \) which guarantees a better performance than using the initial \( \lambda \) for the ASR.

- **PHASE 3**
  In the actual evaluation, the incoming real reverberant data observed in the microphone is enhanced using
δ opt. Subsequently, we used the optimized model λ_{opt}
for the ASR.

4.2 Expansion of the Enhancement Techniques

Specifically, when the δ parameter is introduced to the
enhancement techniques discussed in Sec. 3, Eq’ns (3), (5)
and (6) become

\[ \text{thr} = \delta \sigma \sqrt{2 \log(L)}, \]  

\[ |S(f)|^2 \approx |O(f)|^2 - \delta |R(f)|^2, \]  

\[ \kappa = \frac{S(w)^2}{S(w)^2 + \delta R(w)^2}. \]

for wavelet enhancement thresholding (Sec. 3.1), spectral
subtraction (Sec. 3.2) and Wiener filtering in the wavelet
domain (Sec. 3.3), respectively. The δ parameter, together
with the optimization method shown in Fig. 5 adjust the
degree in suppressing the late reflection. Consequently con-
trolling the artifacts (by-product of the enhancement pro-
cess), in conjunction with the ASR (for minimum impact).

5. EXPERIMENTAL SET-UP

The multi-party interaction setting is shown in Fig. 6, tak-
ing place at a very reverberant room referred to as Room
5 (reverberation time of 940 ms). The direct distance be-
tween the robot and each speaker is approximately 4 meters.
Other room configurations used in our experiment are shown
in Fig. 7. We used Hearbo as our robot platform. This is
equipped with microphone array processing technology [12].

The topic of the multi-party conversation is about differ-
et fish varieties used in preparing sushi and sashimi (Japanese
traditional dishes). There are two participants in the conver-
sation, Japanese and English language speakers. Although
the conversation is in English, they only know the name of
the fish in their language (i.e. Japanese or English). While
the two are engaged in the conversation process, the robot
is in listening mode. As the conversation progresses, ei-
ther one of the two speakers stumble upon a fish name,
and the other party tries to explain further. Upon realiz-
ing that there is no way the other party would identify the
fish, the speaker would ask Hearbo the Japanese/English
equivalent name of the particular fish. While in listening
mode, Hearbo identifies if a question is being asked and auto-
matically distinguishes whether the Japanese or the English
speaker asked the question. This is done by checking the
acoustic likelihood of the utterance using both Japanese and
English acoustic models. We note that the English speaker asks Hearbo in English while the Japanese speaker asks in Japanese. Then, Hearbo moves its head towards the appropriate speaker and gives the equivalent fish name. Thus, if the question is in Japanese, Hearbo answers the English name and vice versa. Consequently, the two speakers continue to engage in the conversation until one of them asks Hearbo again. The sample of the multi-party conversation is shown in Fig. 8. The actual conversation may be longer than that shown in Fig. 8.

A total of 10 speakers participated in the multi-party human-robot interaction experiment (5 speakers for each language). Each speaker asks 10 fish names to Hearbo. We used both Japanese and English (triphone HMMs) acoustic models trained with the Japanese Newspaper Article Sentence (JNAS) with 20K vocabulary and the Wall Street Journal, respectively. We use a simple dialog system for evaluation, training only a single slot for fish name. We experimented in the condition of reverberation time: Room 1=80 ms, Room 2=240 ms, Room 3=930 ms, Room 4=940 ms, and Room 5 = 940 ms.

Figure 7: Other room configurations used in the experiment.

Figure 8: Multi-party human-robot interaction.
Table 1: Automatic Speech Recognition (ASR): Performance in % Word Accuracy (Japanese/English)

<table>
<thead>
<tr>
<th>Methods</th>
<th>Room 1</th>
<th>Room 2</th>
<th>Room 3</th>
<th>Room 4</th>
<th>Room 5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>No enhancement</strong></td>
<td>75.9 / 72.1</td>
<td>64.3 / 60.2</td>
<td>23.4 / 20.5</td>
<td>20.2 / 17.4</td>
<td>17.1 / 14.2</td>
</tr>
<tr>
<td>Reverb. data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wavelet Thresholding</td>
<td>77.5 / 74.0</td>
<td>67.1 / 63.4</td>
<td>34.8 / 31.8</td>
<td>31.9 / 29.7</td>
<td>30.5 / 29.8</td>
</tr>
<tr>
<td>(Sec. 3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASR-Optimized (Sec. 4)</td>
<td>81.6 / 78.7</td>
<td>73.0 / 70.3</td>
<td>46.3 / 43.6</td>
<td>43.4 / 40.8</td>
<td>42.0 / 40.1</td>
</tr>
<tr>
<td>Spectral Subtraction</td>
<td>78.2 / 75.6</td>
<td>68.7 / 66.0</td>
<td>36.6 / 33.4</td>
<td>34.0 / 31.5</td>
<td>33.2 / 30.7</td>
</tr>
<tr>
<td>(Sec. 3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASR-Optimized (Sec. 4)</td>
<td>82.4 / 79.2</td>
<td>74.5 / 71.3</td>
<td>48.9 / 46.1</td>
<td>46.3 / 43.6</td>
<td>45.2 / 43.0</td>
</tr>
<tr>
<td>Wiener Filtering</td>
<td>79.0 / 76.2</td>
<td>70.2 / 67.5</td>
<td>39.4 / 36.3</td>
<td>36.0 / 33.4</td>
<td>35.1 / 32.7</td>
</tr>
<tr>
<td>(Sec. 3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASR-Optimized (Sec. 4)</td>
<td>83.7 / 80.9</td>
<td>76.7 / 73.6</td>
<td>51.7 / 48.9</td>
<td>48.5 / 46.5</td>
<td>47.7 / 46.0</td>
</tr>
</tbody>
</table>

Table 2: Human-machine interaction: Hearbo responding with a correct fish name % Hits (Japanese/English)

<table>
<thead>
<tr>
<th>Methods</th>
<th>Room 1</th>
<th>Room 2</th>
<th>Room 3</th>
<th>Room 4</th>
<th>Room 5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>No enhancement</strong></td>
<td>90.0 / 86.0</td>
<td>82.0 / 78.0</td>
<td>38.0 / 34.0</td>
<td>34.0 / 30.0</td>
<td>34.0 / 30.0</td>
</tr>
<tr>
<td>Reverb. data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wavelet Thresholding</td>
<td>90.0 / 86.0</td>
<td>84.0 / 82.0</td>
<td>44.0 / 42.0</td>
<td>44.0 / 42.0</td>
<td>44.0 / 42.0</td>
</tr>
<tr>
<td>(Sec. 3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASR-Optimized (Sec. 4)</td>
<td>92.0 / 88.0</td>
<td>86.0 / 84.0</td>
<td>64.0 / 62.0</td>
<td>64.0 / 60.0</td>
<td>64.0 / 60.0</td>
</tr>
<tr>
<td>Spectral Subtraction</td>
<td>90.0 / 86.0</td>
<td>84.0 / 82.0</td>
<td>44.0 / 42.0</td>
<td>44.0 / 42.0</td>
<td>44.0 / 42.0</td>
</tr>
<tr>
<td>(Sec. 3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASR-Optimized (Sec. 4)</td>
<td>92.0 / 88.0</td>
<td>86.0 / 84.0</td>
<td>64.0 / 62.0</td>
<td>64.0 / 60.0</td>
<td>64.0 / 60.0</td>
</tr>
<tr>
<td>Wiener Filtering</td>
<td>92.0 / 88.0</td>
<td>88.0 / 86.0</td>
<td>64.0 / 64.0</td>
<td>64.0 / 60.0</td>
<td>64.0 / 60.0</td>
</tr>
<tr>
<td>(Sec. 3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASR-Optimized (Sec. 4)</td>
<td>92.0 / 88.0</td>
<td>88.0 / 86.0</td>
<td>64.0 / 64.0</td>
<td>64.0 / 60.0</td>
<td>64.0 / 60.0</td>
</tr>
</tbody>
</table>

6. RESULTS AND DISCUSSION

For the quantitative evaluation, we investigated the ASR performance and the actual human-robot interaction experience defined as the percentage in which Hearbo correctly answers the 10 fish names being asked. Moreover, a simple qualitative assessment was conducted by asking the participants’ opinion regarding their overall impression of the proposed method (i.e. satisfied or not satisfied).

6.1 Quantitative Results

Table 1 shows the ASR results in detail for both Japanese and English (referred as “hypothesis”). This result takes into consideration the deletion and insertion errors caused by the ASR in recognizing the utterance. We show the speech recognition performance when no enhancement is employed using a clean acoustic model, together with the results when using traditional speech enhancement techniques. Consequently, the results of both the traditional and the ASR-optimized speech enhancement techniques are compared. A summary is shown in Fig. 9, where the average results for the traditional and ASR-optimized enhancement techniques are presented. It is apparent that all traditional speech enhancement techniques improved speech recognition performance compared to the method that employed no enhancement. Moreover, when these enhancement techniques are optimized in conjunction with the ASR, speech recognition performance is further improved. The improvement of the ASR-optimized enhancement techniques is more evident in extremely reverberant conditions (i.e. Rooms 3, 4 and 5).

In Table 2, we show the detailed results of the actual human-robot interaction experience in which Hearbo correctly answers the fish name (referred as “ACTION” in Fig. 4). This is the result after extracting the slot from the output of the ASR (hypothesis). Table 2 is summarized in Fig. 10. This result shows how the speech enhancement is translated to improving human-robot interaction experience. The proposed method has minimal impact in Room 1 (80 ms) and Room 2 (240 ms). However, almost 100% improvement is achieved in Room 3 (930 ms), Room 4 (940 ms) and Room 5 (940 ms) when compared to the method with no enhancement. The proposed ASR-optimization works well at higher reverberation time. This means that the perception of improved human-robot interaction experience is more evident under severe reverberant conditions. Although ASR recognition performance is insightful, the figures are more dynamic than the actual human-robot interaction. For example, improvement in ASR performance for Room 1 and Room 2 (see Fig. 9) does not necessarily translate to an improvement in human-robot interaction experience as depicted in Fig. 10. Thus, human-robot interaction experiment is preferred when inferring improvement in human-robot interaction experience.

6.2 Qualitative Results

As a result of the survey, the participants were satisfied with the human robot interaction set-up. The perception of satisfaction can be explained further in Fig. 11. This figure shows the percent improvement in human-robot interaction experiment of both the traditional and the proposed ASR-optimized speech enhancement methods as a function of reverberation time. This result impacts user satisfaction experience. In this particular experiment, we synthetically generate data of different reverberation time [9][2] since it is difficult to gather real reverberant data. To synthetically
generate a reverberant data, we need an approximation of the room characteristics in the form of room impulse response (RIR). This can be measured physically or modelled mathematically [9]. Then, the clean speech is convolved with the RIR to generate the synthetic reverberant data. We enhanced the data with both the traditional and the proposed method. Then, we measure the percentage of improvement in terms of Hearbo’s correct response. This figure shows significant improvement under severe reverberant conditions. Further, this validates both user satisfaction and is consistent with our results in Table 2.

7. CONCLUSION

We have presented a method that integrates the traditional speech enhancement techniques in relation with the back-end ASR system for improved human-robot interaction experience. In this paper, we have expanded three independent speech enhancement techniques that operate in different domains and optimized based on the ASR. All of these, show consistent improvement in human-robot experience especially under severe reverberant conditions. In this paper we have dealt more on the effect of seamless human-robot interaction with optimized speech enhancement, rather than solely focusing on ASR results. Currently, we are only dealing with the effects of reverberation. In our future works, we will consider the contribution of noise jointly with reverberation.

8. REFERENCES