Impact of Reverberation to the Energy Transfer of Connected Words

Randy Gomez, Keisuke Nakamura, Takeshi Mizumoto and Kazuhiro Nakadai

Abstract— **In this paper, we present a method in suppressing speech degradation affecting human-robot communication in real reverberant environment condition. The novelty of the proposed method is the mechanism to indirectly incorporate language information in the enhancement process. This is achieved by considering the effects of the acoustic energy transfer of two consecutive words. The proposed method is categorized into two namely, a one-time computation of the smearing coefficients (offline mode) and the actual enhancement (online mode). In the offline mode, word pair smearing coefficients reflective of the inter-word energy transfer within a word pair are calculated and stored, creating a pool of smearing priors database. Then, the online robust enhancement process integrates this information to the conventional framewise enhancement method. In theory, the proposed method outperforms the conventional framewise-only enhancement since it is able to dynamically update the enhancement parameters based on the actual acoustic energy transfer between successive words during testing. Experiments using a humanoid robot inside a reverberant room confirms the effectiveness of the proposed method. Our method renders human-robot speech communication robust to the effect of reverberation as opposed to the conventional method.**

I. Introduction

In recent years, interest towards humanoid robots has gained a dramatic impact in the field of robotics. A humanoid robot is specifically built to resemble the shape of a human body for functional designs replicating human actions such as bipedal movements, arm manipulation, interaction among others. In short, humanoid robots may perform human tasks in manufacturing, assembly line operation or at reception desks to entertain guests. Thus, the notion of a humanoid robot companion assisting humans in day-to-day tasks is not far-fetched. Because of these endless humanoid practical applications, the fascination towards the development of this type of robot is gaining momentum.

Harnessing the potential of a humanoid robot opens a variety of challenging research topics in human-robot interaction. In this paper, we focus on the speech communication interaction with emphasis on robustness in real reverberant environment condition. The very idea that humanoid robot follows the form of a human being makes it more endearing to us, in which the need for it to deliver a more gratifying interaction experience is inevitable. And there is no better interaction experience other than speech communication [1]. The humanoid robot featured in movies that can talk, understand, and execute speech commands is becoming more of a reality than science fiction these days. Recent semiconductor design developments resulting to a fast and more power-efficient processors have significantly improved the robot's ability to process mathematical tasks necessary for speech recognition and understanding. The advancement in computing technology continually benefits the improvement of the speech communication interaction capability of the humanoid robot.

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

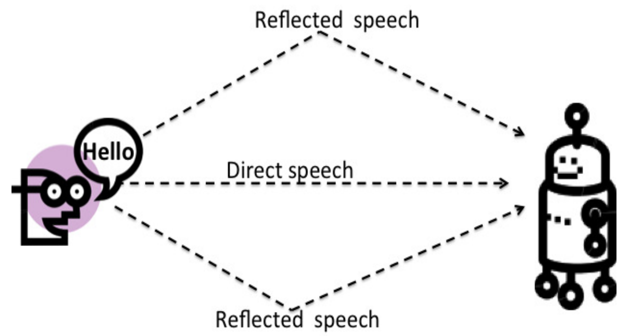

Fig. 1. Reflections of the speech signal inside an enclosed room.

Although system performance using close-talking microphone reaches 90-94 % in recognition accuracy [2], this is only applicable in ideal condition where everything is controlled. In real world however, the pursuit towards a seamless human-machine interaction through speech communication is often plagued with robustness problems [3]. In the real world, humans prefer to communicate with the robot hands-free (not close-talking), which gives the user the freedom to be at some distance away from the robot without being constrained by the microphone [4] as shown in Fig. 1. Hands-free speech communication often use multi-microphone sensors (e.g. microphone array). And as the distance between the user and the microphone array increases, the observed power at the microphones also decreases which makes it more vulnerable to contamination. Although this mode of communication offers more degree of freedom to the user as far as location is concerned, hands-free speech communication is sensitive to the effects of reverberation caused by the reflection of the speech signal in an enclosed environment. In a typical room, the acoustic speech may be reflected on the walls, ceiling, floors, etc. As a result, the speech reflections arrive at different time delays as observed by the microphones mounted on the robot, creating a smearing effect (Fig. 2) to the speech known as reverberation. This phenomenon drastically degrades the performance of the speech recognizer, affecting human-robot interaction experience.

To mitigate the effects of reverberation, speech enhancement is employed. Conventional speech enhancements op-

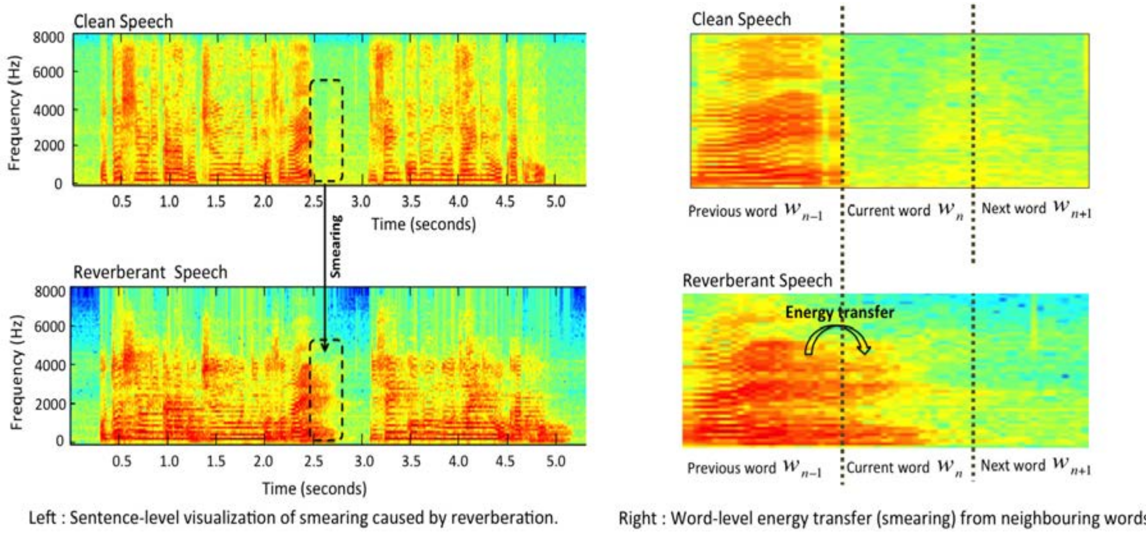

Fig. 2. The smearing effects of reverberation.

erate in framewise manner (e.g. 25 msec. window frame). This design does not distinguish a frame to a word. The immediate concern of speech enhancement is to make the processed speech intelligible to the human ear by focusing on the acoustic quality primarily. Besides, the human brain can infer language from sequence of sounds automatically. In short, the effects of reverberation are considered purely in the acoustics domain. This concept was adopted in speech recognition research. However, speech recognition systems (i.e., model-based systems) are knowledge-based and requires both the acoustic (acoustic waveform) and language (word sequence) information. Thus, simple adoption of the conventional enhancement approach is not sufficient because it only addresses acoustic waveform requirements. In this paper, we propose to indirectly incorporate language information in the speech enhancement process to match it with the speech recognition system. Using the training database, the inter-word relationships among word pairs are analyzed. This enables us to gather prior information of the actual impact of smearing between two successive words. The smearing phenomenon infers the actual energy transfer of two consecutive words due to reverberation. Then, smearing coefficients are calculated for all of the word pairs and a database of smearing coefficients is created. All of these processes are done during offline mode. In the actual testing prior to input to the speech recognizer (online mode), the system dynamically selects the appropriate smearing coefficients and integrate to the framewise enhancement technique for improved recognition performance in real world using a humanoid robot. In our method, the speech enhancement technique can be treated as a black box since the concept is applicable to any conventional speech enhancement platforms employing framewise processing. For simplicity, we focus on Spectral subtraction [5] to explain the proposed concept.

This paper is organized as follows; in Section II, the proposed method is discussed involving both the offline and online procedures. Followed by the experimental setup in Section III. In Section IV, the results from real-world experiment are presented. Finally, we conclude the paper in Section V.

II. METHODS

A. Offline Smearing Coefficient Training

In Fig. 3, the process of obtaining the smearing coefficients of two neighbouring words due to reverberation are explained as follows,

1) Training Database: The clean speech database s is composed of speech recording (waveform) from different speakers using a close-talking microphone. This is a standard database used in speech recognition applications. Consequently, this is transcribed into word-level text transcripts t. Thus, each waveform speech utterance has a corresponding word level text transcription w_n. The clean speech database is re-played using a loudspeaker inside a reverberant room and a microphone embedded on the robot's head, located at a distance away from the loudspeaker is used to capture both the direct and reflected speech. This set up is used to create a realistic reverberant speech database r, which is needed to analyze the actual smearing effect between neighbouring words.

2) Word pair Extraction: Using the training transcripts originally transcribed in words $w_1, w_2, ..., w_{n-1}, w_n$, for $n = 1 : N$ words, word pair tokens are extracted as $w_1w_2, ..., w_{n-1}w_n$. The segmented transcript of a word pair token j is defined as

$$t_j \triangleq w_j w_{j+1} \quad for \quad j = 1 : N-1. \tag{1}$$

Word-pair extraction is applied to all of the transcripts in the database resulting to intermediate word pairs. Using the information from the word and word pair transcripts, the clean waveform database is segmented into word token

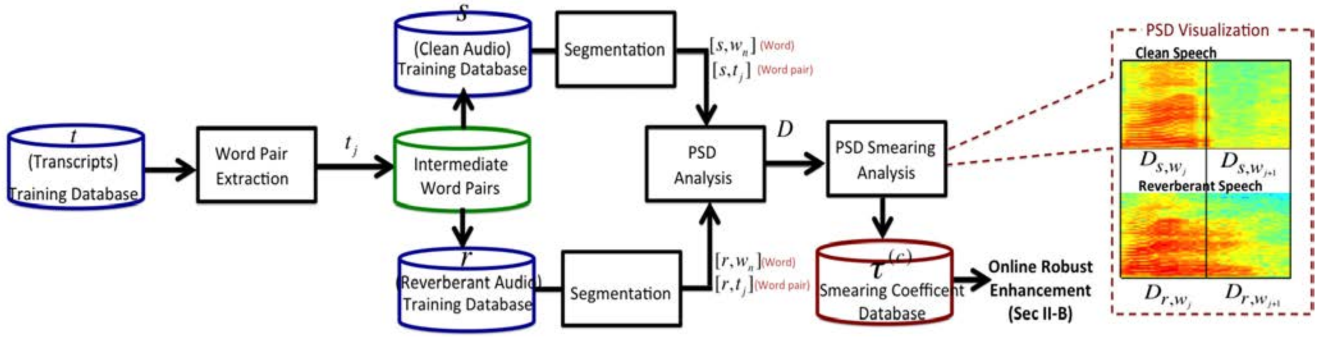

Fig. 3. The training component of the system.

$[s, w_n]$ and word pair token $[s, t_j]$. The segmented reverberant database as $[r, w_n]$ and $[r, t_j]$ for single word and word pair tokens, respectively.

3) Power Spectral Density (PSD) Analysis: The clean and reverberant speech are aligned together, then power spectral density (psd) using the Welch's method is applied to both and calculate the energy distribution in the frequency domain. The Welch's method is preferred since different words vary in duration and the Welch's method is designed to operate by dividing the time domain into successive blocks of periodograms and average across time. The averaging minimizes the impact of the variable word duration. Thus if a word w_n has a total duration of M_{w_n} frames, the PSD of the speech signal s is defined as

$$D_{s,w_n}(\omega) \triangleq \frac{1}{M_{w_n}} \sum_{m=0}^{M_{w_n}-1} P_s(\omega, m) \tag{2}$$

where P_s is the periodogram of the mth frame of the word level segmented speech signal. The word level periodogram of the reverberant signal $D_{r,w_n}(\omega)$ is computed in the same manner as Eq. (2). In addition, the word pair psd is also calculated. From the word level psd in Eq. (2), the psd for word pairs are computed by simply expanding the limit M_{w_n} to M_{t_j} to accommodate both two neighbouring words. Specifically,

$$D_{s,t_j}(\omega) \triangleq \frac{1}{M_{t_j}} \sum_{m=0}^{M_{t_j}-1} P_s(\omega, m)$$

and

$$D_{r,t_j}(\omega) \triangleq \frac{1}{M_{t_j}} \sum_{m=0}^{M_{t_j}-1} P_r(\omega, m)$$

are the word pair psd of the clean speech and reverberant speech, respectively. Word pair psd is used for classification while word level psd is used to calculate for the smearing coefficient.

The intermediate word pairs are divided into two categories, namely the frequently occurring pairs and the infrequently occurring pair (including single pairs) duplicates. The former is referred to as base class $c = 1 : C$ while the latter as infrequent pairs class. Suppose that there are $c = 1 : C$ base classes and $l = 1 : L$ infrequent pairs classes, word pair acoustic similarity is used to re-assign the latter into the base classes. Similarity measure is given as

$$sim^{(c,l)}(\omega) \triangleq \tilde{D}^{(c)*}_{s,t_j}(\omega)\tilde{D}^{(l)}_{s,t_j}(\omega), \tag{3}$$

where $\tilde{D}^{(c)*}_{s,t_j}(\omega)$ and $\tilde{D}^{(l)}_{r,t_j}(\omega)$ are the psd representatives from classes c (base classes) and l (infrequent occurring class), respectively. The objective is to distribute the word-pairs in (l) to the base classes in (c) through the similarity measure in Eq. (3). The word pair in l with corresponding c which results to a maximum value in the similarity measure in Eq. (3) for all base classes $1 : C$ will be assigned to the corresponding base class c. This process is repeated until all entries in l are exhausted and assigned to the base class accordingly.

4) PSD Smearing Analysis: It is important to analyze the effects of smearing in the utterance empirically which is achieved by analyzing the transferred power using the actual sequence of words leading to the computation of the word pair smearing coefficient. This method establishes a direct link between acoustic power transfer impacted by the language itself. In the conventional methods [6][7], only the acoustic contribution is addressed without consideration of the word sequences. Smearing coefficients characterizes the actual energy contribution of preceding word w_{n-1} to the current word w_n. Since we are not dealing with extremely huge rooms that are very echoic, this assumption is valid. The smearing coefficient for two neighbouring words is calculated as follows,

$$\tau^{(c)}_j \triangleq \frac{D_{r,w_{j+1}}(\omega) - D_{s,w_{j+1}}(\omega)}{D_{s,w_j}(\omega)} \quad for \quad j = 1 : N-1, \tag{4}$$

where $D_{r,w_{j+1}}$, $D_{s,w_{j+1}}$ and D_{s,w_j} are the psds of the reverberant and clean speech of the current word in consideration while D_{s,w_j} is the psd of the previous word, respectively.

B. Online Enhancement

Although the proposed robustness method works on different enhancement platforms, the spectral subtraction (SS) [5] enhancement platform is used for simplicity. The SS method is one of the most simple denoising technique for decades

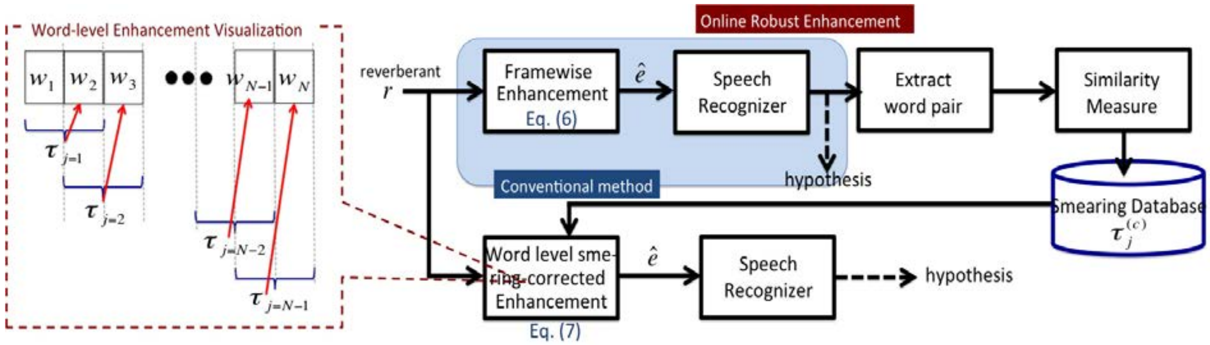

Fig. 4. Robust enhancement in real environment condition.

and has been expanded for dereverberation application [6]. The reverberant speech is modelled as a superimposition of the early and late reflections [8], in which the former is composed mostly of the direct speech signal and the latter is treated as noise. The reverberant speech model in framewise manner m is given as

$$r(\omega,m) = e(\omega,m) + l(\omega,m) \tag{5}$$

where $e(\omega,m)$ and $l(\omega,m)$ are the early and late reflections, respectively. Speech enhancement for speech recognition is defined as suppressing the late reflection and recovering the early reflection [6]. Figure 4 is the block diagram of the the online robust enhancement method.

1) Conventional Method: In our previous work [6], the enhanced speech signal is given as

$$|\hat{e}(\omega,m)|^2 = \begin{cases} |r(\omega,m)|^2 - \delta_b|l(\omega,m)|^2 \\ \quad \text{if } |r(\omega,m)|^2 - \delta_b|l(\omega,m)|^2 > 0 \\ \\ \beta|r(\omega,m)|^2 \quad \text{otherwise.} \end{cases} \tag{6}$$

where β is the flooring coefficient, δ_b is the dereverberation parameters in bands $b = 1,...,B$ and the late reflection power denoted as $|l(\omega,t)|^2$. The process of estimating these parameters is described in detail in [6][7]. Note that Eq. (6) is purely implemented in framewise manner without distinction to the actual word sequence in the actual speech utterance. Thus, the inter-word smearing effect is not accounted for.

2) Proposed Robust Method with Word-level Smearing Compensation: Right after the framewise speech enhancement, the processed reverberant speech is fed into the speech recognition system, the hypothesis from the recognizer is used as a preliminary information to extract the candidate sequence of words. The hypothesis may be imperfect due to acoustic ambiguity in which a framewise processing is unable to address. This is true because reverberation usually spills over several frames rendering framewise processing insufficient. Thus, the hypothesis is further processed to extract word pairs and then calculate similarity measure (see Eq. (3)) against the base class $c = 1 : C$. Then, the corresponding smearing coefficient of the selected base class is used in conjunction with the framewise enhancement to include word level smearing effects. Thus Eq. (6) becomes

$$|\hat{e}(\omega,m,w_j)|^2 = \begin{cases} |r(\omega,m,w_j)|^2 - \delta_b\tau_j^{(c)}|r(\omega,m,w_j)|^2 \\ \quad \text{if } |r(\omega,m,w_j)|^2 - \\ \quad \delta_b\tau_j^{(c)}|r(\omega,m,w_j)|^2 > 0 \\ \\ \beta|r(\omega,m,w_j)|^2 \quad \text{otherwise.} \end{cases} \tag{7}$$

It is obvious that the SS in Eq. (7) is capable of resolving reverberation effects both framewise and word level manner through the introduction of $\tau_j^{(c)}$ which is not possible in the conventional method given in Eq. (6).

III. Experimental Set-up

A. Realistic Environment Condition

In our experiment, the human speaker is positioned in front of the robot. The room set-up is shown in Fig. 5 (right). Two rooms were considered with approximately 240 msec (Room A) and 640 msec. (Room B) of reverberation time (RT). The distances between the robot and the speaker are 0.5 m, 1.0 m., 1.5 m. and 2.0 m., respectively. Occlusions due to refrigerator, board, chairs, etc. are considered during testing to recreate a realistic environment. In the experiment, we use the proprietary humanoid robot of Honda Research Institute named Hearbo as shown in the same figure (left). This experimental robot platform has 8 microphones embedded in a circular fashion along its spherical head. The microphone array technology uses HARK [9] for sound source separation. The separated speech is used as the reverberant speech signal to be enhanced by the proposed method.

B. Speech Recognition

The task is composed of 2000 word vocabulary using continuous speech recognition [2]. The topic of the human-robot interaction is about fish varieties used in preparing sushi and sashimi (Japanese traditional dish). The human-robot interaction is initiated by the speaker by asking the robot questions pertaining a fish and the robot answers back. Due to the effects of reverberation, speech recognition may fail which leads to the failure of the robot to give the correct answer by failing to recognize the fish being asked. Thus, the

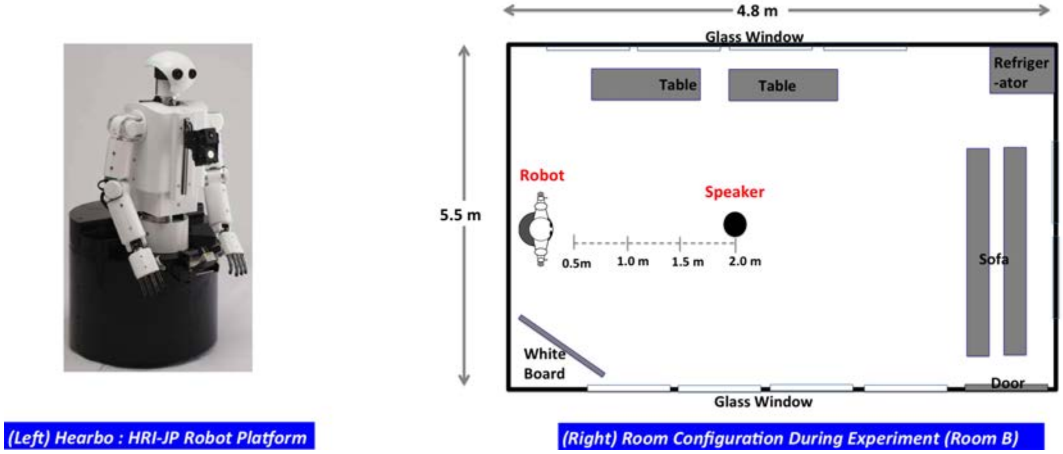

Fig. 5. Room configuration in our experiment

TABLE I

WORD CORRECT RATE IN ROOM A WITH REVERBERATION TIME RT = 240 MSEC..

Methods	*0.5 m*	*1.0 m*	*1.5 m*	*2.0 m*
(A) No Enhancement	90.2 %	84.3 %	74.4 %	69.7 %
(B) Blind Dereverberation [10]	90.6 %	86.4 %	77.3 %	72.6 %
(C) Conventional Wiener Filtering (Framewise only) [7]	90.6 %	86.7 %	78.9 %	75.3 %
(D) Proposed Robust Wiener Filtering (Framewise + Word-level)	**91.0 %**	**87.2 %**	**80.2%**	**78.9 %**
(E) Conventional Spectral Subtraction (Framewise only) [6]	90.6 %	86.9 %	79.0 %	76.2 %
(F) Proposed Spectral Subtraction (Framewise + Word-level)	**91.1 %**	**88.8 %**	**81.9 %**	**79.7 %**

TABLE II

WORD CORRECT RATE IN ROOM B WITH REVERBERATION TIME RT = 640 MSEC..

Methods	*0.5 m*	*1.0 m*	*1.5 m*	*2.0 m*
(A) No Enhancement	81.2 %	65.3 %	44.5 %	28.7 %
(B) Blind Dereverberation [10]	83.6 %	73.5 %	58.1 %	45.3 %
(C) Conventional Wiener Filtering (Framewise only) [7]	84.9 %	76.9 %	60.1 %	48.2 %
(D) Proposed Robust Wiener Filtering (Framewise + Word-level)	**86.7 %**	**83.9 %**	**70.1%**	**61.4 %**
(E) Conventional Spectral Subtraction (Framewise only) [6]	85.9 %	78.6 %	62.3 %	49.3 %
(F) Proposed Spectral Subtraction (Framewise + Word-level)	**87.5 %**	**84.5 %**	**72.4 %**	**63.7 %**

proposed method is used to enhance it. During interaction, the system puts the robot into listening mode while the user is speaking and then switches into speaking mode as soon as it is ready to talk. An example of the conversation is shown as follows,

- *Human>> Hearbo, my friend and I went to a sushi bar yesterday and ordered Sweetfish. Can you give us information of that fish ?*
- *Hearbo>>Sweetfish is common in South East Asia. An edible fish known to its distinctive sweet flavour with melon and cucumber aromas.*
- *Human>> We ate it with maki-sushi. Hearbo, can you give me more information about maki-sushi ?*
- *Hearbo>> Maki-sushi is a rolled rice with other ingredients using a sheet of nori. There are many varieties of maki-sushi like chu-maki, futo-maki, temaki, uramaki among others.*
- *Human>> Hearbo, what is Tororo Kombu ?*
- *Hearbo>> Tororo kombu is made from thinly sliced kombu with vinegar flavour and dried.*

The actual conversation may be longer than the ones above. A total of 20 speakers participated in the human-robot interaction experiment (not included in training). Each speaker asks 10 questions to Hearbo in a freestyle conversation like the ones shown in the conversation example. The only condition is that each question should contain a fish name. We used English triphone Hidden Markov Models (HMMs) acoustic model trained with the Wall Street Journal database.

IV. RESULTS AND DISCUSSION

The ASR results in terms of word correct are shown in Tables 1 and 2 for reverberation times RT=240 msec. and RT=640 msec., respectively. The result in method (A) is when the reverberant speech is not processed prior to input to the speech recognizer. Method (B) is the result using a blind dereverberation approach based on Linear Prediction residual [10]. This is a speech enhancement method that exploits the characteristics of the vocal chords to remove the effects of reverberation. Methods (C) and (E) are the results of two different enhancement platforms based on Wiener filtering [7] and Spectral Subtraction [6] discussed in Eq. (6). Both of these are based on the conventional framewise processing. The methods in (D) and (F) are based on the same platform in methods (C) and (E) but expanded to include the word-level smearing effect using prior information from word pair database. In these results, it is obvious that the proposed robust enhancement method is outperforming the conventional methods. The rate of improvement due to the proposed method is more evident in Table 2 than in Table 1 since the latter is not so reverberant. Thus, the proposed method works well in very reverberant environment. Consequently, since shorter distances have less reverberation effect, the benefit of speech enhancement is not very obvious (i.e., 0.5 m.) as opposed to farther distances (i.e., 2.0 m.). The superior performance of the proposed method as shown in methods (D) and (F) can be attributed to the following,

- More realistic characterization of smearing (transfer of energy) in the word level which is not considered in the conventional method.
- Dynamic update of smearing parameters in the word level enables the system to adapt to the current environment changes as opposed to being constant in the conventional method
- HMM speech recognition system is primarily defined by both the acoustic and language models. The word level sequence treatment in the proposed method creates a synergistic effect to the language model. Note that the language model is derived from word sequences as well.

It is important to note that even though the hypothesis in which the word pair extraction is based may contain wrong pairs due to the erroneous speech recognition results, the system may still recover when applying the the smearing factor since it alters the acoustic characteristics of the word and impacts recognition performance. Bottom line, processing the acoustics of a known misrecognized word is always better than using the same unprocessed acoustics. After all, speech recognition is probabilistic in nature.

V. Conclusion

In this paper, we have indirectly considered the contribution of language (sequence of words) in enhancing the reverberant signal through realistic energy transfer of two consecutive words. By using the smearing prior, we can effectively design and improve any framewise speech enhancement platform for speech recognition application. This is true because reverberation is nothing more of a transfer of sound energy, and framewise processing is not sufficient since smearing affects more than a single frame and sound units impact other sound units differently due to its unique spectral energy characteristics. By knowing before hand the smearing dynamics of two neighbouring words, enhancement can be corrected to reflect the actual energy transfer. Currently we are only limited to two neighbouring words and in the future we will expand this to using more appropriate and well-defined sound unit. Also, in our future works we will further investigate the transference of energy in a more general context.

References

[1] R. Gomez, T. Kawahara, Keisuke Nakamura and Kazuhiro Nakadai, “Multi-party Human Robot Interaction with Distant-talking Speech Recognition” *In Proceedings of IEEE Human Robot Interaction (HRI)*, 2012.

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

[3] S. Vaseghi “Advanced Signal processing and Digital Noise reduction”, Wiley and Teubner, 1996.

[4] 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.

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

[6] 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.

[7] R. Gomez and T. Kawahara , “Optimizing Spectral Subtraction and Wiener Filtering for Robust Speech Recognition in Reverberant and Noisy Conditions” *ICASSP*, 2010.

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

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

[10] 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.