Improving automatic speech recognition performance and speech intelligibility with harmonicity based dereverberation

Keisuke Kinoshita  Tomohiro Nakatani  Masato Miyoshi

NTT Communication Science Laboratories
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
{kino@islab.kecl.ntt.co.jp}  {nak@islab.kecl.ntt.co.jp}  {miyo@islab.kecl.ntt.co.jp}

Abstract
A speech signal captured by a distant microphone is generally smeared by reverberation, that severely degrades both the speech intelligibility and Automatic Speech Recognition (ASR) performance. Previously, we proposed a novel dereverberation method, named “Harmonicity based dEReverBeration (HERB)”, which estimates the inverse filter of an unknown impulse response by utilizing the inherent speech property, harmonics. In this paper, we carry out a formal evaluation of speech intelligibility for dereverberated speech, and further investigate HERB’s possibilities to improve ASR performance. Experimental results show that HERB is able to improve speech intelligibility to the level of clean speech. HERB is also found to be very effective at improving ASR performance, even under unknown severe reverberant environments by being used with MLLR and a multicondition acoustic model.

1. Introduction
The quality of a speech signal captured by a distant microphone is generally degraded by surrounding acoustic interferences such as reverberation and environmental noise. A recorded speech signal can be modelled as

\[ s(n) = x(n) + z(n) \]

where \( s(n) \) refers to clean speech, \( h(n) = [h(0), \ldots, h(M-1), n]^T \) to an M-tap room impulse response, and \( z(n) \) to environmental additive noise. Interference, especially reverberation, is known to severely degrade both Automatic Speech Recognition (ASR) performance and speech intelligibility. Especially, in a reverberant environment with a reverberation time (RT) of more than 0.5 seconds, ASR performance can not be improved even with an acoustic model trained with a matched reverberation condition [1].

Considerable research has been done to deal with additive noise. For example, Spectral Subtraction can greatly reduce the effect of additive noise and result in sufficient improvement of ASR performance and speech intelligibility [2]. In an ASR system, a Parallel Model Combination (PMC) [3] can also be utilized as a good solution to additive noise.

Unlike noise reduction techniques, an efficient dereverberation technique has not yet been proposed, even after all serious efforts in the last decades. The most widely used dereverberation technique to this day is microphone-array [4]. It first estimates the DOAs (direction of arrivals), and steers so-called “nulls” of a microphone array (null beam-forming) to best suppress the reflections. Since the number of reflections are much greater than nulls formed by microphone array, it only works in moderate reverberant environments. Other major approaches attempting to estimate the inverse filter of an unknown impulse response are based on blind equalization methods, such as the Independent Component Analysis (ICA) [5]. This method works effectively if the signals are statistically independent and identically distributed non-Gaussian sequences. However, it cannot appropriately handle speech signals because they have inherent properties, such as periodicity and formants, making the sequence statistically dependent. Another approaches specialized in improving audible quality has been proposed by Nagarayana [6]. Improvement was achieved by attenuating the relative amplitude of LPC residuals where the speech to reverberation ratio is smaller. Even though this method might help to improve speech intelligibility, it does not improve ASR performance because it does not make any changes in spectral features which are essential for ASR.

To achieve an improvement of both ASR performance and speech intelligibility, we have proposed the novel dereverberation technique named “Harmonicity based dEReverBeration (HERB)” [7, 8, 9]. HERB utilizes essential properties of speech, harmonics, and estimates the inverse filter of an unknown impulse response. Therefore, unlike previous studies, it handles the speech signal well. In this paper, we would like to investigate experimentally what kind, and how much, impact HERB has on ASR performance and speech intelligibility. Then, based on a discussion of experimental results, we propose HERB-based ASR system configuration to achieve higher recognition performance in unknown reverberant environments.

2. Harmonicity based dereverberation
In general, both an input (direct sound) and output (reverberant sound) signal are required to calculate the inverse filter of an unknown impulse response, however, the definition of blind dereverberation does not allow us to have any information about an input signal. HERB makes it possible to estimate inverse filter by regarding the harmonic structure of reverberant speech as the voiced portion of a direct sound (see [7]). Let \( X(\omega N) \) be the frequency representation of the reverberant speech \( x(n) \) at frame \( N \), \( S_0(\omega N) \) be the frequency representation of the voiced portion of a direct sound (clean speech) at frame \( N \), and \( F(f_0(n)) \) be the adaptive comb filter of the function of fundamental frequency \( f_0(n) \). Within each frame \( N \), as in equation 2, the adaptive comb filter extracts time-varying harmonic components from \( x(n) \) using estimated fundamental frequencies \( \{f_0\} \), and synthesizes \( S_0(\omega N) \) (see details [9]). \( F[\cdot] \) represents Fourier transform.

\[ S_0(\omega N) = F[F(f_0(n))[x(n)]] \]

Followings our a priori knowledge that voiced speech consists of harmonics in general, \( S_0(\omega N) \) is considered to be approximated to \( S_h(\omega N) \). Using \( S_h(\omega N) \), we can now estimate the inverse filter as is shown in equation 3. \( E[\cdot] \) is the expectation operator. To attenuate the estimation error introduced by unexpectedly extracted harmonics in equation 2, we take the average of \( \frac{S_h(\omega N)}{X(\omega N)} \) over several frames.

\[ W(\omega) = E \left[ \frac{S_h(\omega N)}{X(\omega N)} \right] \]
In [9], \( W(\omega) \) is theoretically shown to be a good estimation of the inverse filter of unknown acoustic environment, and suppress especially the reverberation tail efficiently.

### 2.1. Dereverberation procedure

In this study, HERB is developed in a more possible realistic manner to deal with a continuous speech sequence, in contrast to previous studies which only handled the isolated-word utterance. A block diagram of the dereverberation procedure for continuous speech is illustrated in Fig. 1. First, input speech was divided into 5.5 second frames with rectangular windows each overlapping by 75%. The value \( x_N \) in Fig. 1 denotes the speech signal at frame \( N \). Within each frame, Voiced/UnVoiced (V/UV) segments and \( F_0 \)s are estimated with 30ms overlapping succeeding analysis windows. Utilizing those informations, an adaptive comb filter extracts time-varying harmonic components, as in equation 2. Based on the harmonic components extracted by the adaptive comb filter and \( x_N \), we calculated the initial value for the dereverberation filter in frame \( N \). By averaging the initial estimated value over several frames, as in equation 3, we finally obtained an accurate dereverberation filter for step 1.

At step 2, almost the same procedure is repeated as in step 1, the only difference being that \( F_0 \)s and V/UV segments are estimated based on the dereverberated speech. At step 3, \( F_0 \) and V/UV estimation are done based on the dereverberated speech, as in step 2. The dereverberation filter obtained at step 3 was then applied to the speech that had been dereverberated at step 2.

### 3. Quality of dereverberated speech

In this chapter, we measure the quality of dereverberated speech in comparison with reverberant speech, in terms of speech intelligibility and ASR performance.

Dereverberations were done on 4 impulse responses (RT: 0.1, 0.2, 0.5, 1.0 sec.) \( \times 2 \) genders (female, male). The filter training data corpus is the ATR data set B, with female speaker FKN’s 503 sentences and male speaker MHT’s 503 sentences.

#### 3.1. Improvement of Speech Intelligibility

We objectively measured speech intelligibility of dereverberated speech using \( D_{30} \) and the modulation spectrum. \( D_{30} \) is defined in ISO 3382, and known to be closely related to speech intelligibility [10]. The modulation spectrum is also highly correlated with speech intelligibility. For example, the Speech Transmission Index (STI) calculated from the modulation spectrum is quite often used to evaluate the quality of an acoustic environment [11].

\( D_{30} \) refers to the ratio of early energy relative to the total energy within the impulse response. Thus, the more efficient the dereverberation is, the closer \( D_{30} \) gets to 100%. Table 1 shows \( D_{30} \) at each reverberant environment for each gender. We see the \( D_{30} \)s are improved under all reverberant conditions. The improvements are especially noticeable at a longer reverberation time (i.e. RT=1.0s). This result can be taken as a evidence that HERB did suppress the reverberation tail efficiently, as we expected.

<table>
<thead>
<tr>
<th>Rev. time (s)</th>
<th>1.0</th>
<th>0.5</th>
<th>0.2</th>
<th>0.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Org.</td>
<td>Rev.</td>
<td>87.9</td>
<td>94.6</td>
<td>99.6</td>
</tr>
<tr>
<td>Derev.</td>
<td>98.7</td>
<td>99.4</td>
<td>99.7</td>
<td>99.8</td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Org.</td>
<td>Rev.</td>
<td>87.9</td>
<td>94.6</td>
<td>99.3</td>
</tr>
<tr>
<td>Derev.</td>
<td>96.9</td>
<td>99.4</td>
<td>99.7</td>
<td>99.8</td>
</tr>
</tbody>
</table>

Fig. 2 shows the modulation spectral contours of clean, reverberant (RT=1 sec.) and dereverberated speech for each gender. In the non-reverberant environment, the modulation spectrum exhibits a peak at 4 Hz with appreciable energy between 3 and 8 Hz [12]. This modulation spectral contour changes markedly under reverberant conditions. The peak shifts to lower frequency regions, and the magnitude of this primary locus of energy is appreciably attenuated relative to the clean speech. However, if we apply HERB to reverberant speech, the modulation spectral contour recovers the peak at 4 Hz with appreciable energy around it. This restoration of the modulation spectrum directly indicates that the STI of the dereverberated speech is very close to that of the clean speech, and it implies that HERB has succeeded in recovering a degree of speech intelligibility which is comparable to clean speech.

#### 3.2. Improvement of ASR performance

We now investigate the effect of HERB as a preprocessing algorithm for ASR. In tasks 1 to 4, we explore the possible use of HERB to improve ASR performance. Based on findings from tasks 1 to 4, at task 5, we would then like to apply the method as a possible means of achieving higher recognition performance in unknown reverberant environments. Across the tasks, ASR performance was evaluated in terms of speaker dependent word accuracy. In the acoustic model, we used the following parameters: 12 order MFCCs, 12 order delta MFCCs, 3 state HMMs, and 4 mixture Gaussian distributions.

\[ D_{30} = \frac{\int_{-\infty}^{\infty} \frac{R_{nn}(\omega) - R_{nn}^0(\omega)}{R_{nn}(\omega)}}{\int_{-\infty}^{\infty} R_{nn}(\omega) dx} \times 100 \]

\(^1\)Let \( h(t) \) be the impulse response, then \( D_{30} = \frac{\int_{-\infty}^{\infty} \frac{R_{nn}(\omega) - R_{nn}^0(\omega)}{R_{nn}(\omega)}}{\int_{-\infty}^{\infty} R_{nn}(\omega) dx} \times 100 \)
was used was trained on Japanese newspaper articles from a ten-year period.

3.2.1. HERB only (task 1)

First, we simply employed HERB as a preprocessing algorithm for ASR. The acoustic model was trained on clean speech from the ATR data set B (A)-(H). ATR data set B (J) was used for evaluation.

Fig. 3 illustrates the word accuracy with and without HERB under each reverberant environment. The baseline in Fig. 3 represents the word accuracy of clean speech recognized with the clean acoustic model. Although slight improvements by HERB can be seen in longer reverberant environments, there still remains much room to be improved. From these results, a certain degree of distortion appears to remain in the dereverberated speech after dereverberation, and consequently that residual distortion may cause the mismatch between clean acoustic model and the target dereverberated speech. Since HERB was found to suppress the reverberation tail efficiently, it is very likely that those residual distortions were the result of the remaining early reflections.

3.2.2. HERB + Cepstral Mean Normalization (task 2)

To cope with the residual distortion from the remaining early reflections, we employed Cepstral Mean Normalization (CMN) [13] in addition to HERB. CMN is known to work effectively to reduce relatively shorter convolutional distortion like channel distortion in a telephone channel. The speech data for the acoustic model training and recognition task were the same as the previous task.

Fig. 4 shows the word accuracy under each reverberant environment using CMN with and without HERB. Since we now have HERB to suppress the reverberation tail and CMN to remove early reflections, we obtain relatively larger improvements over reverberant speech, especially in longer reverberant environments. From this result, we found that, the mismatch between the dereverberated speech and clean speech was partially introduced by the difference between their cepstral means.

3.2.3. HERB + supervised MLLR (task 3)

Recently, it was reported that the mismatch between the acoustic model and target speech data could often be greatly reduced by the linear transformation of the Gaussian mean parameters in the acoustic model [14]. This method was found to be effective to deal with the mismatch caused by speaker differences, channel distortion and some background noise. So, in this task, to further investigate the difference between the dereverberated speech and clean speech (or clean acoustic model), we would like to examine if that difference could be compensated by the above linear transformation scheme. To perform the ideal linear transformation for evaluation, we employed a supervised Maximum Likelihood Linear Regression (MLLR). We used the ATR data set B (A)-(H) for the acoustic model training, the ATR data set B (I) for supervised MLLR adaptation, and the ATR data set B (J) for evaluation.

Fig. 5 shows word accuracy in each reverberant environment using a supervised MLLR with and without HERB. Greater improvements were observed in all reverberant environments for both genders by HERB + MLLR. Notice that the recognition performance without HERB could not maintain high level even with supervised MLLR, especially in longer reverberant environments. Therefore, it is reasonably assumed that HERB managed to convert the complicated reverberation problems into relatively accessible ones, which can simply be solved by linear transformation of the Gaussian mean parameters.

3.2.4. HERB + multicondition model (task 4)

In this task, we would like to take a closer look at the characteristics of the residual distortion remaining in the dereverberated speech. More precisely, we would like to examine if the residual distortion after dereverberation is dependent on the reverberant environment. To examine this, we employed an acoustic model similar to multicondition model. Before performing the recognition, we allowed the acoustic model to learn dereverberated speech under various reverberant environments. After that, we used the system to recognize the dereverberated speech of an unknown reverberant environment. In this task, if the residual distortions are solely dependent on HERB and completely independent of the reverberant environment, a much higher word accuracy of dereverberated speech would be expected.

Fig. 6 shows the word accuracy in each reverberant environment using a multicondition model with and without HERB. In most conditions, the word accuracy for dereverberated speech scored very close to baseline performance. These results indicate importantly that, the residual distortions in the dereverberated speech are independent of the reverberant environment, and in addition it can be learned by acoustic model. Thus, if we are allowed to train the acoustic model in advance on dereverberated speech.

2E.g., using an acoustic model trained on RT 0.1s, 0.2s and 0.5s dereverberated speech, we perform the recognition of RT 1.0s dereverberated speech.
Figure 6: Word accuracy using Multicondition acoustic model with (solid line) and without (small broken line) HERB

Figure 7: Word accuracy using a multicondition acoustic model with (solid line) and without (small broken line) HERB

In this paper, we further explored the effect of the proposed dereverberation method “HERB” on speech intelligibility and the possibilities to improve ASR performance. HERB is a novel dereverberation method which estimates the inverse filter of an unknown impulse response accurately by utilizing the inherent speech property, harmonics. Objective measurement of speech intelligibility by the modulation spectrum and $D_{20}$ implicitly showed that HERB is able to improve the speech intelligibility to approximate that of clean speech. HERB was also found to be very effective at improving ASR performance even in unknown severely reverberant environments, by using MLLR and a multicondition acoustic model.

5. Acknowledgements

We would like to express our appreciation to Prof. Dawn Behne and Prof. Peter Svensson of the Norwegian University of Science and Technology for English revision and useful comments on this paper.

6. References