IMPLEMENTATION AND EFFECTS OF SINGLE CHANNEL DEREVERBERATION BASED ON THE HARMONIC STRUCTURE OF SPEECH

Tomohiro Nakatani   Masato Miyoshi   Keisuke Kinoshita

Speech Open Lab., NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237 Japan
{nak,miyo,kinoshita}@cslab.kecl.ntt.co.jp

ABSTRACT

Speech signals recorded with a distant microphone usually contain reverberation, which degrades speech recognition performance. This is especially severe when the reverberation time exceeds 0.5 sec because the recognition performance cannot be improved sufficiently even if we use an acoustic model trained under matched reverberation conditions. Therefore, reverberation in speech signals should be removed prior to recognition. In this paper, we propose a single channel dereverberation method based on the harmonic structure of speech. Experimental results show that the present method can effectively suppress reverberation when the reverberation time is longer than 0.1 sec.

1. INTRODUCTION

Many researchers have tackled robust automatic speech recognition (ASR) for real world application, however, long reverberation is still a serious problem that severely degrades the ASR performance [1]. One simple way to overcome this problem is to dereverberate the speech signals prior to ASR, but this is also a challenging problem, especially when using a single microphone.

Past research approaches include blind equalization methods, such as independent component analysis (ICA); these methods can estimate the inverse filter of an unknown impulse response convolved with target signals when the signals are statistically independent and identically distributed non-Gaussian sequences [2]. However, they cannot appropriately deal with speech signals because speech signals have inherent properties, such as periodicity and formant structure, making their sequences statistically dependent. This approach inevitably destroys these essential properties of speech. A different approach uses the properties of speech signals [3]. The basic idea involves adaptively detecting time regions in which signal-to-reverberation ratios become small, and attenuating speech signals in those regions. However, the precise separation of signal and reverberation durations is difficult, so this approach has achieved only moderate results so far.

To overcome this problem, we proposed a new dereverberation principle that uses an essential property of speech signals, namely its harmonic structure, as a clue [4]. In this method, the harmonic structure of target speech is estimated directly from the reverberant signals, and then an inverse filter, referred to as a dereverberation operator, is estimated by calculating the average ratio of the estimated harmonic structure to the reverberant signal in the frequency domain. This paper presents a theoretical analysis of the estimated dereverberation operator: the estimated filter can dereverberate both the harmonic and non-harmonic components of speech signals with no prior knowledge of the target signals, even though only the harmonic components of the signals are used for operator estimation. It also details an implementation of the dereverberation process. The effectiveness of the method is shown in terms of reverberation curves and ASR performances.

2. HARMONICS BASED DEREVERBERATION

2.1. Speech model

In real acoustical environments, a sound source signal, $X(\omega)$, reaching a microphone contains some reverberation. The reverberant signal, Y, is the product of X and the transfer function, H, of the environment as in eq. (1). This transfer function can also be divided into two functions, D and R. The former transforms X to the direct signal, DX, and the latter to the reverberation part, RX, as shown in eq. (2).

$$\begin{aligned} Y(\omega) &= H(\omega)X(\omega), \qquad (1)\\ &= D(\omega)X(\omega) + R(\omega)X(\omega). \qquad (2)\end{aligned}$$

In this paper, direct signal, $X'(\omega) = D(\omega)X(\omega)$, is treated as the target signal that should result from dereverberation. X' can be obtained by subtracting RX from Y (eq. (2)), or by estimating an inverse transfer function, $1/H'(\omega) = D(\omega)/H(\omega)$, and multiplying it to Y (eq. (4)).

$$\begin{aligned} Y(\omega) &= H'(\omega)X'(\omega), \qquad (3)\\ X'(\omega) &= (1/H'(\omega))Y(\omega), \qquad (4)\end{aligned}$$

Speech signal, X, can be modeled as the sum of harmonic components, X_h, and noisy components, X_n, (eq. (5)). The reverberant signal, Y, is then represented by the product of H and $(X_h + X_n)$ (eq. (6)), and also by the sum of the direct signal of the harmonic components, DX_h, and the other components (eq. (7)).

$$\begin{aligned} X &= X_h + X_n, \qquad (5)\\ Y &= H(X_h + X_n), \qquad (6)\\ &= DX_h + (RX_h + HX_n). \qquad (7)\end{aligned}$$

Of these components, the direct harmonic components, $X'_h = DX_h$ can approximately be extracted from Y by harmonics filtering. Although the frequencies of harmonic components change dynamically according to the changes in their fundamental frequency (F_0), their reverberation remains unchanged at the same frequency. Therefore, X'_h, can be enhanced by extracting frequency components located at multiples of F_0. These approximated direct harmonic components $\hat{X}'_h$ can be modeled as follows:

$$\hat{X}'_h = DX_h + (\hat{R}X_h + \hat{N}), \qquad (8)$$

where $\hat{R}X_h$ and $\hat{N}$ are a part of RX_h and a part of HX_n, respectively, which unexpectedly remain in $\hat{X}'_h$ after harmonics filtering[1]. We assume all estimation errors in $\hat{X}'_h$ are caused by $\hat{R}X_h$ and $\hat{N}$ in eq. (8).

2.2. Dereverberation principle

We refer to $\mathcal{O}(\hat{R}) = (D+\hat{R})/H$ as the "dereverberation operator" because signal $DX + \hat{R}X$, obtained by multiplying $\mathcal{O}(\hat{R})$ by Y, becomes in a sense a dereverberated signal.

$$\mathcal{O}(\hat{R})Y = DX + \hat{R}X, \tag{9}$$

where $DX + \hat{R}X$ is composed of direct signal DX and certain parts of the reverberation, $\hat{R}X$. The rest of the reverberation included in $Y(= DX + RX)$, or $(R - \hat{R})X$, is eliminated by the dereverberation operator. Note that eq. (9) holds even when the reverberant signal, Y, contains both harmonic components and non-harmonic components.

To estimate the dereverberation operator, we use the approximated direct signals, $\hat{X}'_h$. Suppose a number of Y values are obtained and $\hat{X}'_h$ values are calculated from individual Y values. The dereverberation operator is then approximated as the average of $\hat{X}'_h/Y$, or $E(\hat{X}'_h/Y)$. $E(\hat{X}'_h/Y)$ is shown to be a good estimate of $\mathcal{O}(\hat{R})$ by substituting $E(\hat{X}'_h/Y)$ for eqs. (6) and (8).

$$E(\frac{\hat{X}'_h}{Y}) = \mathcal{O}(\hat{R})E(\frac{1}{1+\frac{X_n}{X_h}}) + E(\frac{1}{1+\frac{Y-\hat{N}}{\hat{N}}}). \tag{10}$$

The arguments of the two average functions in eq. (10) have the form of a complex function, $f(z) = 1/(1+z)$. $E(f(z))$ is easily proven to equal $P(|z| < 1)$, where $P(\cdot)$ is a probability function, using the residue theorem if it is assumed that the phase of z is uniformly distributed, the phases of z and $|z|$ are independent, and $|z| \neq 1$. Based on this property, the second term of eq. (10) approximately equals zero because $\hat{N}$ is a noisy component that the harmonics filter unexpectedly extracts and thus the magnitude of $\hat{N}$ almost always has a smaller value than $(Y - \hat{N})$ if a sufficiently long analysis window is used. Therefore, $E(\hat{X}'_h/Y)$ can be approximated by eq. (11).

$$E(\frac{\hat{X}'_h(\omega)}{Y(\omega)}) \simeq \mathcal{O}(\hat{R}(\omega))P(|X_h(\omega)| > |X_n(\omega)|), \tag{11}$$

i.e., $E(\hat{X}'_h/Y)$ approximately equals the product of the dereverberation operator and the probability that the harmonic components of speech have larger magnitude than the noisy components.

For a speech signal, the magnitudes of the noisy components tend to increase as the frequency range increases. Therefore, the $P(|X_h(\omega)| > |X_n(\omega)|)$ value becomes smaller as ω increases, and thus, the gain of $E(\hat{X}'_h/Y)$ tends to decrease. To compensate for this fall in gain, it may be useful to use the average attributes of speech on the probability, $P(|X_h(\omega)| > |X_n(\omega)|)$. In our experiment, the $E(\hat{X}'_h/Y)$ itself was treated as the dereverberation operator without any compensation.

3. IMPLEMENTATION

In order to evaluate the fundamental performance of our harmonic-structure-based speech dereverberation scheme, we designed a benchmark test and implemented a prototype system.

3.1. Task: dereverberation of isolated word utterances

The task of the benchmark test is dereverberation of isolated word utterances, in which each reverberant utterance is separable without overlapping other utterances. The test is executed in batch style, that is, all the reverberant signals are given in advance, and the dereverberation operator is estimated using all the signals at once. The benchmark test is summarized as follows:

1. A number of word utterances convolved with a time-invariant impulse response are given in advance as reverberant observed signals, $y(n)$.
2. The dereverberation operator is estimated from the observed signals based on the proposed method. The dereverberated signals are obtained by convolving the dereverberation operator with the reverberant signals, $y(n)$.
3. The performance of the dereverberation is evaluated using the dereverberated signals.

3.2. Prototype system

The dereverberation method of the prototype system is mainly composed of the following subprocedures:

1. **F_0 estimation:** voiced durations and their F_0s are estimated from each reverberant word utterance, $y(n)$.
2. **Harmonics filtering:** harmonic components, $\hat{x}'_h(n)$, included in $y(n)$ are estimated by means of waveforms based on adaptive harmonics filtering.
3. **Dereverberation operator estimation:** each pair of $y(n)$ and $\hat{x}'_h(n)$ is transformed into frequency domain signals, as Y and $\hat{X}'_h$, by discrete Fourier transformation (DFT). The dereverberation operator, $\mathcal{O}(\hat{R})$, is estimated by calculating the average $\hat{X}'_h/Y$ value for pairs of them.
4. **Dereverberation:** the dereverberated signals are obtained by multiplying $\mathcal{O}(\hat{R})$ by Y.

3.2.1. Processing flow

Figure 1 illustrates the complete processing flow of our prototype system. The dereverberation procedures are composed of three processing steps, aiming at gradually improved dereverberation performance in each step. All subprocedures described above are employed in each step. They are summarized as follows:

STEP 1 F_0s, voiced segments, and harmonic components are estimated from reverberant observed signals, $y(n)$, therefore, they may contain many errors in the estimated values.

STEP 2 F_0s and voiced segments are estimated from signals dereverberated by the previous step, and harmonic components are estimated from reverberant observed signals, $y(n)$. Because the estimated F_0s and voiced segments are expected to have improved, harmonic components estimated based on them are also expected to have improved.

STEP 3 All above values are estimated from signals dereverberated by the previous step. Because reverberant components, $\hat{R}X_h$, inevitably included in eq. (8) can further be reduced, more effective dereverberation is expected to be achieved.

[1] Strictly speaking, $\hat{R}$ cannot be represented in linear transformation form because the reverberation included in $\hat{X}_h$ depends on the time pattern of $\hat{X}'$. We introduce this approximation for simplicity.

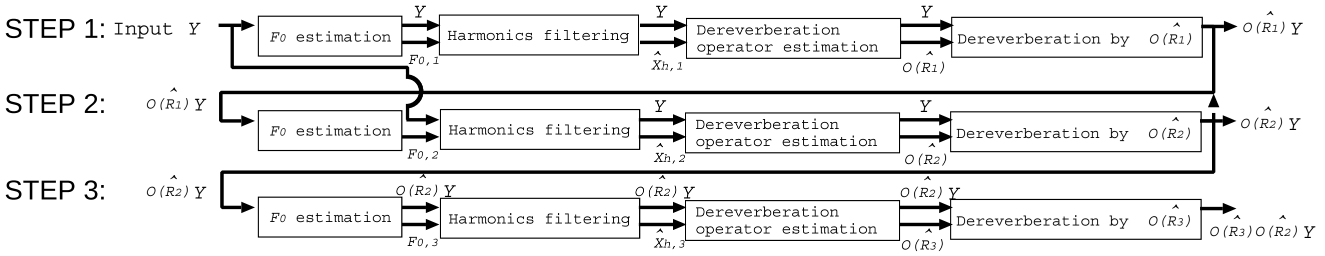

Fig. 1. Processing flow of dereverberation.

In our preliminary experiments, the estimation of F_0 and voiced segments gradually improved when STEP 2 was repeated. By contrast, repeating STEP 3 did not always improve the quality of dereverberated signals. This is because estimation errors in dereverberation operators accumulated in the dereverberated signals when the signals are multiplied by more than one dereverberation operator. Our benchmark test employed all of these steps without repeating any of them.

3.2.2. *Dereverberation operator calculation*

In our prototype system, the estimated dereverberation operator results in a delayed inverse filter. This means that the prototype system is applicable even to reverberant signals that include non-minimum phases. To estimate the delayed inverse filter precisely, we calculate the DFT of Y and $\hat{X}'_h$ in estimating the dereverberation operator with a time window whose length is long enough to contain each whole word utterance with zero padding. In other words, each pair of Y and $\hat{X}'_h$ values is extracted from a whole word utterance. Unlike using a shorter time window, all the direct signals and their reverberation are included in a frame, which avoids the estimation errors caused by superfluous reverberation continuing from the preceding frames and those caused by insufficient reverberation continuing to the following frames.

In addition, the average $\hat{X}'_h/Y$ value in eq. (10) is weighted by the amplitude spectrum $|\hat{X}_h(\omega)|$. Based on this weighted normalization, the influence of noisy components is expected to be further reduced while that of dominant harmonic components is expected to be enhanced.

3.2.3. *F_0 estimation, V/UV classification, and harmonics filtering*

Accurate F_0 estimation is very important to achieve effective dereverberation in our method. Although we employed a robust F_0 estimation method, i.e., the normalized power spectrum, $R(f)$, based method described in [5], it is not sufficiently robust, especially for speech signals with long reverberation. To cope with this problem, we introduced two types of preprocessing to F_0 estimation: one using a simple filter that reduces sound that continues at the same frequency [4], and the other using the dereverberation operator, $\mathcal{O}(\hat{R})$, itself. The effectiveness of these filters was confirmed in our preliminary experiments. The dereverberation operator based method is most effective because the reverberation of the speech can be directly reduced by the operator. This mechanism is included in step 2 and 3 of the dereverberation procedures, so a more accurate F_0 can be obtained in steps 2 and 3 than in step 1.

As regards voiced/unvoiced segments (V/UV) classification, a method based on a magnitude of harmonic components relative to the other components is employed with the above preprocessing. In this method, for each discrete frequency, f, in a frequency region at each time frame, n, a power sum, $H_n(f)$, of frequency components corresponding to multiples of the frequency is first calculated based on the normalized power spectrum, $R_n(f)$, as eq. (12). The relative magnitude of the harmonic components, $V(n)$, whose F_0 is $f_0(n)$ is then determined as eq. (13).

$$H_n(f) \;=\; \sum_k R_n(kf), \qquad (12)$$

$$V(n) \;=\; M_p\left(\frac{H_n(f_0(n)) - E(H_n(f))}{\sigma(H_n(f))}\right), \qquad (13)$$

where $E(\cdot)$ and $\sigma(\cdot)$ are functions yielding the average and standard deviation of $H_n(f)$ over frequencies, respectively, and $M_p(\cdot)$ is a function that extracts a median value over p time frames. A frame is determined as voiced if $V(n)$ is larger than a fixed threshold value. Because $V(n)$ is a value normalized with the standard deviation, this threshold can be set independently of the signal level.

We employed a time-varying linear filter for harmonics filtering since it precisely preserves the phase and amplitude of each harmonic component. The filter is designed as follows:

$$\hat{x}'_{h,n_0}(n) \;=\; y(n) * (g_1(n) \sum_k e^{j2\pi k f_0(n_0) n / f_s}), \qquad (14)$$

$$\hat{x}'_h(n) \;=\; \sum_{n_0} g_2(n-n_0) \mathrm{Re}\{\hat{x}'_{h,n_0}(n)\}, \qquad (15)$$

where n_0 is the center time of each frame, $f_0(n_0)$ is F_0 of the signal at the frame, k is a harmonics index, $g_1(n)$ and $g_2(n)$ are analysis window functions, and f_s is the sampling frequency.

3.2.4. *Analysis conditions*

Signals digitized with 12 kHz sampling frequency were used in the test. In F_0 estimation and V/UV classification, a 42 msec hanning window and 1 msec window shift were adopted. The length of the median filter was $p = 60$, or 60 msec. For $g_1(n)$ and $g_2(n)$ of harmonics filtering in eq. (15), 42 msec and 4 msec hanning windows were used, respectively, with 1 msec window shift. The length of the dereverberation filter was 131,072 taps; that is, a 10.9 sec rectangle window was used for Y and $\hat{X}'_h$ calculation.

4. EXPERIMENTAL RESULTS

We examined the performance of the prototype system in terms of reverberation curves and ASR. We used 5240 Japanese word utterances provided by a male and a female speaker (MAU and FKM) included in the ATR database as source signals, $x(n)$. We used

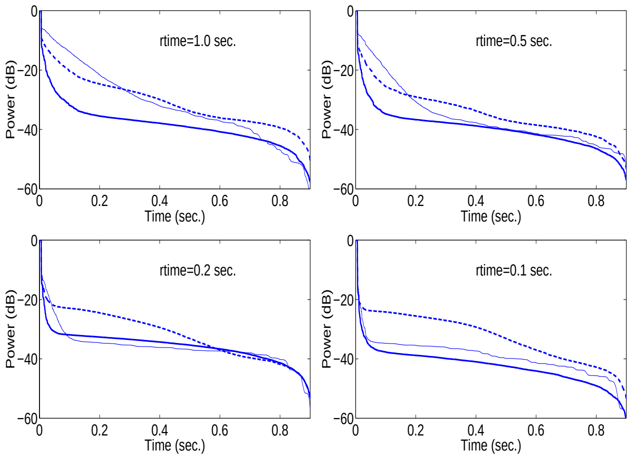

Fig. 2. Reverberation curves of the original impulse responses (thin line) and dereverberated impulse responses (male: thick dashed line, female: thick solid line) for different reverberation times (rtime).

four impulse responses measured in a reverberant room whose reverberation times were about 0.1, 0.2, 0.5, and 1.0 sec, respectively. Reverberant signals, $y(n)$, were obtained by convolving $x(n)$ with the impulse responses.

4.1. Evaluation based on reverberation curves

Figure 2 depicts the reverberation curves of the original and dereverberated impulse responses. The figure shows that the proposed method could effectively reduce the reverberation in the impulse responses for the female speaker when the reverberation time (rtime) was longer than 0.1 sec. For the male speaker, the reverberation effect in the lower time region was also effectively reduced. This means that strong reverberant components were eliminated, and so the intelligibility of the target speech could be expected to be improved [6]. Although the reverberation effect in the higher time region for the male speaker was increased when rtime is 0.1 or 0.2 sec, the sound quality as a whole is expected to improve when rtime is 0.2 sec because the earlier reverberation that is much stronger than the later one was eliminated. This can be easily confirmed by listening to dereverberated signals as demonstrated on our www page [7].

4.2. Evaluation based on word recognition rates

Speaker dependent word recognition rates (WRRs) of reverberant and dereverberated speech signals were examined. Three types of acoustic monophone models trained with reverberant speech signals, dereverberated speech signals, or clean speech signals were prepared. The first two models, referred to as matched condition models, were used to recognize reverberant signals and dereverberated signals, respectively. The last model, referred to as a clean model, was used to recognize both signals. 4740 words randomly selected from 5240 words were used as training data, and the remaining 500 words were used as testing data. 12 order MFCCs, 12 order delta MFCCs, three state HMMs, five mixture Gaussian distributions, 25 msec frame length, and 5 msec frame shift were adopted as the analysis conditions. The results are shown in Fig. 3. The left panel shows WRRs with the clean speech model. The average WRRs for the dereverberated signals (thick dotted line) are much better than those for the reverberant signals (thin dotted line), but the WRRs were at most 55 %. By contrast, the right panel shows the WRRs with matched condition models. The WRRs for dereverberated signals kept more than 95 % even when the reverberation time was 1.0 sec while those of the reverberant signals degraded as the reverberation time exceeded 0.5 sec. These results mean that the prototype system succeeded in reducing the reverberation effects without loosing speech features essential for ASR.

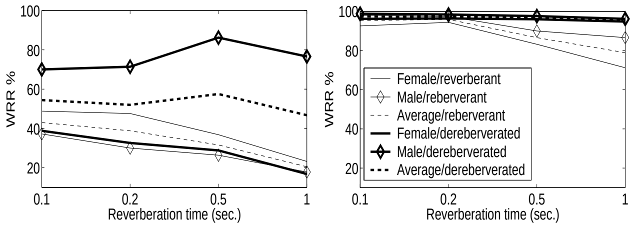

Fig. 3. Word recognition rates (WRRs) of reverberant and dereverberated signals when using clean speech model (left panel) and matched condition model (right panel) under different reverberation time conditions.

5. CONCLUSION

This paper proposed a new blind dereverberation method for speech signals captured by a single microphone. The dereverberation operator is calculated as the average ratio of the approximated direct harmonic signal to the reverberant signal, and is shown to provide a good estimate of the inverse transfer function that can be used for dereverberation. Experimental results showed that the dereverberation operator trained with reverberant speech signals composed of 5240 Japanese word utterances could effectively suppress reverberation from speech. More than 95 % WRRs were achieved using dereverberated speech signals with matched condition models even when the reverberation time was 1.0 sec. Future work will include an investigation of how such high quality speech dereverberation can be achieved with fewer speech data and extending the dereverberation method to include adaptive filtering techniques.

REFERENCES

[1] Baba, A., Lee, A., Saruwatari, H., and Shikano, K., “Speech recognition by reverberation adapted acoustic model,” *Proc. of ASJ*, pp. 27–28, Akita, Japan, Sep., 2002.

[2] Amari, S., Douglas, S. C., Cichocki, A., and Yang, H. H., “Multichannel blind deconvolution and equalization using the natural gradient,” *Proc. IEEE Workshop on Signal Processing Advances in Wireless Communications*, Paris, pp. 101–104, April 1997.

[3] Yegnanarayana, B., and Murthy, P. S., “Enhancement of reverberant speech using LP residual signal,” *IEEE Trans. SAP* vol. 8, no. 3, pp. 267–281, 2000.

[4] Nakatani, T., and Miyoshi, M, “Blind dereverberation of single channel speech signal based on harmonic structure,” *Proc. ICASSP-2003*, vol. 1, pp. 92–95, Apr., 2003.

[5] Nakatani, T., and Irino, T., “Robust fundamental frequency estimation against background noise and spectral distortion,” *Proc. ICSLP-2002*, vol. 3, pp. 1733–1736, Denver, Sep., 2002.

[6] Yegnanarayana, B., and Ramakrishna, B. S., “Intelligibility of speech under nonexponential decay conditions,” *JASA*, vol. 58, pp. 853–857, Oct. 1975.

[7] http://www.kecl.ntt.co.jp/icl/signal/nakatani/sound-demos/dm/derev-demos.html