BLIND DEREVERBERATION OF SINGLE CHANNEL SPEECH SIGNAL BASED ON HARMONIC STRUCTURE

Tomohiro Nakatani *Masato Miyoshi*

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

ABSTRACT

This paper presents a new method for dereverberation of speech signals with a single microphone. For applications such as speech recognition, reverberant speech causes serious problems when a distant microphone is used in recording. This is especially severe when the reverberation time exceeds 0.5 of a second. We propose a method which uses the fundamental frequency (F_0) of target speech as the primary feature for dereverberation. This method initially estimates F_0 and harmonic structure of the speech signal and then obtains a dereverberation operator. This operator transforms the reverberant signal to its direct signal based on an inverse filtering operation. Dereverberation is achieved with prior knowledge of neither room acoustics nor the target speech. Experimental results showed that the dereverberation operator estimated from 5240 Japanese word utterances could effectively reduce the reverberation when the reverberation time is longer than 0.1 of a second.

1. INTRODUCTION

Speech signals recorded with a distant microphone in a usual room contains certain reverberant quality; this often causes severe degradation in automatic speech recognition performance. Although a number of adaptive recognition methods have been proposed [1], it was reported that the recognition performance cannot sufficiently be improved when the reverberation time is longer than 0.5 sec even if they use acoustic models trained with a matched reverberation condition [2]. Therefore, reverberation in speech signals should be removed prior to recognition.

To this end, microphone-array technologies are frequently exploited. A typical technique is to estimate the DOAs (direction of arrival) of the speech signals reflected on the walls, and direct "nulls" at the signals (null beam-forming). There have been a number of proposed methods for DOA estimation such as MUSIC [3] and ESPRIT [4]. Since the number of nulls produced is less than that of reflected signals, the null beam-forming may be effective only for suppressing a relatively few reflections. Another technique based on inverse-filtering may, on the other hand, suppresses reflections using a small number of microphones as compared to the beam-forming. Theoretically, the reflections are completely eliminated by arranging microphones so that the transfer-functions from N signal sources to $N+1$ microphones have no-common zeros [5], where $N = 1, 2, \ldots$. The inverse-filtering is, however, very sensitive to the locations of signal sources.

This paper presents a new dereverberation method for speech signals with a single microphone. We consider this method as a blind processing technique because the dereverberation is achieved with prior knowledge of neither room acoustics nor the target speech. Only reverberant signals and general attributes of speech are used. In our method, fundamental frequency (F_0) of target speech and its harmonic structure are estimated directly from the reverberant signals, and then a dereverberation operator is obtained by calculating the average ratio of the estimated harmonic structure to the reverberant signal in frequency domain. Because we consider F_0 is a robust feature that can be estimated from a reverberant sound, we adopted it as the primary feature for the dereverberation.

In the rest of this paper, the new dereverberation method is described in section 2, experimental results are given in section 3, and concluding remarks are summarized in section 4.

2. HARMONICS BASED DEREVERBERATION

2.1. Concept

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

$$\begin{aligned} Y(\omega) &= H(\omega)X(\omega), && (1)\\ &= (D(\omega)+R(\omega))X(\omega), && (2)\\ &= (1+R(\omega)/D(\omega))(D(\omega)X(\omega)), && (3)\\ &= H'(\omega)X'(\omega), && (4)\\ X'(\omega) &= Y(\omega)/H'(\omega). && (5) \end{aligned}$$

It is very difficult to estimate X only from the reverberant signal, therefore, $X'(= DX)$ is treated as target signal that should be obtained by dereverberation in this paper. This can be done by estimating $H'(= H/D)$ and by dividing Y by H' as shown in Eq. (5).

In our approach, we first approximate the direct sound X' as $\hat{X}'$ from reverberant sound Y based on harmonic structure[2]. An initial estimation of the transfer function $\hat{H}'$ is obtained by $Y/\hat{X}'$, and H' is then estimated by calculating the average value of $\hat{H}'$. Once H' is obtained, we can easily extract X' by Y/H'.

[1] In this paper, a capitalized variable means a frequency domain signal while a non-capitalized variable means a time domain signal. "(ω)" is often omitted for a frequency domain signal.

[2] Estimated value is denoted by " $\hat{}$ " in this paper.

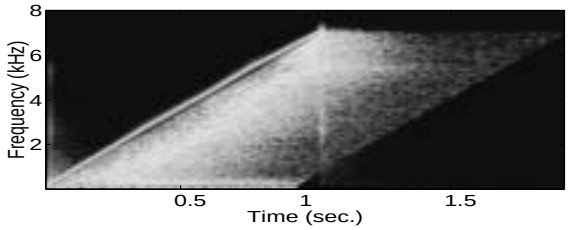

Fig. 1. Spectrogram of reverberant sweeping sound.

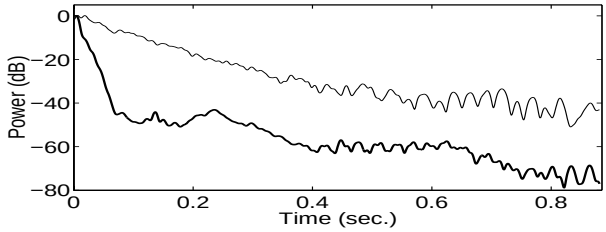

Fig. 2. Reverberation curves of original impulse response (thin line), and dereverberated impulse response (thick line).

We use a sinusoidal representation [6, 7] to approximate direct sound in a reverberant environment. Let us consider a very simple case to explain the concept of our method. Suppose that a sound $y(t)$ whose frequency sweeps from 100 to 7000 Hz is convolved with an impulse response measured in a reverberant room as shown in Fig. 1. The direct sound can be approximated by tracking frequency $\hat{\omega}(n)$ of a dominant sinusoidal component in the reverberant sound at time frame n, and by extracting its amplitude $\hat{a}(n)$ and phase $\hat{\phi}(n)$. Then, the estimated direct sound $\hat{x}'(t)$ can be modeled as follows:

$$\hat{x}'(t) = \sum_n w(t-\tau_n)\hat{a}(n)\sin(\hat{\omega}(n)t + \hat{\phi}(n)), \quad (6)$$

where $w(t)$ is a window function and τ_n is the time corresponding to frame n. In this simple case, accuracy of direct sound estimation is rather high because there is no other sound occurring simultaneously when the direct sound occurs, and thus the estimated transfer function $\hat{H}' = Y/\hat{X}'$ is expected to be very close to H'. Figure 2 shows the reverberation curves [8] of the transfer function H' in a reverberant room and that of the dereverberated transfer function $H'/\hat{H}'$ based on this method. This demonstrates the effective reduction of the reverberation.

2.2. The speech model

We extend the dereverberation method discussed in the previous section to cope with speech signals. For this purpose, we first present a model that assumes speech signal is composed of harmonic components $x_h(t)$ and noisy components $x_n(t)$ in Eq. (7), and the harmonic components can be represented by a sum of sinusoidal components. Then, the speech signal that is recorded in a reverberant room, referred to as the reverberant signal, is modeled by multiplication with the transfer function H in Eq. (8), or with $H'(= H/D)$ in Eq. (9).

$$\begin{aligned} x(t) &= x_h(t) + x_n(t), & (7)\\ Y &= H(X_h + X_n), & (8)\\ &= H'(X'_h + X'_n). & (9) \end{aligned}$$

An adaptive harmonic filter is used in order to approximate direct sound of the harmonic components in the reverberant signal. This filter is expected to reduce reverberant components as well as noisy components by extracting dominant harmonic components in the reverberant signal. With this filter, F_0 of speech, that is, $\hat{\omega}_0(n)$ at time frame n, is first calculated. Then, frequency of each harmonic component is estimated as a multiple of F_0, that is, $k\hat{\omega}_0(n)$. Finally, amplitude $\hat{a}_k(n)$ and phase $\hat{\phi}_k(n)$ of each harmonic component k are estimated based on those of the reverberant sound, $Y(k\hat{\omega}_0(t))$, and the estimated direct sound is represented in Eq. (10).

$$\hat{x}'_h(t) = \sum_n \sum_k w(t-\tau_n)\hat{a}_k(n)\sin(k\hat{\omega}_0(n)t + \hat{\phi}_k(n)). \quad (10)$$

Here, not as the case of a sweeping sinusoid discussed in the previous section, this estimation is affected by other sounds occurring simultaneously, that is, reverberation of preceding harmonic sound and certain noisy components of speech often overlap $Y(k\hat{\omega}_0(t))$. Therefore, the estimated harmonic sound inevitably includes certain parts of reverberant and noisy components. This approximated direct sound $\hat{X}'_h$ can also be modeled[3] by Eq. (12).

$$\begin{aligned} \hat{X}'_h &= \hat{D}X'_h + \hat{N}', & (11)\\ &= (D + \hat{R})X_h + \hat{N}', & (12) \end{aligned}$$

where $\hat{R}X_h$ and $\hat{N}'$ are the parts of reverberation of X_h and those of direct sound and reverberation of X_n included in $\hat{x}'_h(t)$. In Eq. (12), we assume all estimation error in $\hat{D}X'_h$ is caused by $\hat{R}$.

2.3. Dereverberation of speech

We call $\mathcal{O}(\hat{R}) = (D + \hat{R})/H$ a "dereverberation operator" because the sound $(D + \hat{R})X$ that can be obtained by multiplying $\mathcal{O}(\hat{R})$ to Y becomes in a sense a dereverberated sound.

$$(D + \hat{R})X = \mathcal{O}(\hat{R})Y, \quad (13)$$

where $(D + \hat{R})X$ is composed of direct sound, DX, and certain parts of reverberation, $\hat{R}X$. The rest of reverberation included in $Y(= DX + RX)$, or $(R - \hat{R})X$, is considered as eliminated by the dereverberation operator.

A method to estimate this dereverberation operator is introduced. For this purpose, we first calculate an initial estimate of the transfer function $\hat{H}'$ in the form of an inverse transfer function, $1/\hat{H}'$, in Eq. (15) based on the models of Y and $\hat{X}'_h$.

$$\begin{aligned} \frac{1}{\hat{H}'} &= \frac{\hat{X}'_h}{Y}, & (14)\\ &= \mathcal{O}(\hat{R})\frac{X'_h}{X'_h + X'_n} + \frac{\hat{N}'}{Y}. & (15) \end{aligned}$$

In order to extract the dereverberation operator from Eq. (15), average value of the initial estimate $1/\hat{H}'$, or $E(1/\hat{H}')$ is calculated over time, where $E(\cdot)$ is a function representing this average

[3] To be strict, $\hat{R}$ cannot be represented in a form of linear transformation because reverberation included in $\hat{x}'_h(t)$ depends on time pattern of $\hat{x}'(t)$. We introduce this approximation for simplification.

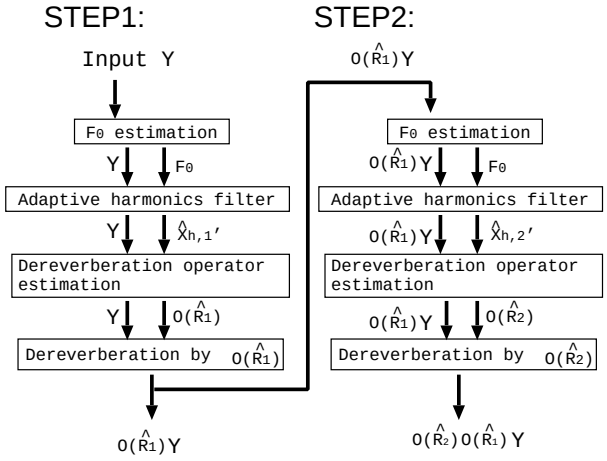

Fig. 3. Processing flow of dereverberation.

value.

$$E(\frac{1}{\hat{H}'}) = E(\mathcal{O}(\hat{R})\frac{X'_h}{X'_h + X'_n}) + E(\frac{\hat{N}'}{Y}), \quad (16)$$

$$= \mathcal{O}(\hat{R})E(\frac{1}{1+\frac{X'_n}{X'_h}}) + E(\frac{\frac{\hat{N}'}{Y-\hat{N}'}}{1+\frac{\hat{N}'}{Y-\hat{N}'}}). \quad (17)$$

Here, as the number of reverberant speech data sets increases, $E(\frac{1}{1+\frac{X'_n}{X'_h}})$ and $E(\frac{\frac{\hat{N}'}{Y-\hat{N}'}}{1+\frac{\hat{N}'}{Y-\hat{N}'}})$ are expected to converge to 1 and 0, respectively if $\frac{X'_n}{X'_h}$ and $\frac{\hat{N}'}{Y-\hat{N}'}$ are assumed to be complex random values whose average values are 0, and satisfy $|\frac{X'_n}{X'_h}| < 1$ and $|\frac{\hat{N}'}{Y-\hat{N}'}| < 1$. Although some of these assumptions are not always satisfied with speech signals such as in fricatives, the effectiveness of this approach is shown in the experiments of section 3.

2.4. Processing flow

The dereverberation algorithm is composed of two steps as shown in Fig. 3.

1. In the first step, F_0 is first estimated from the reverberant signal, Y. Then harmonic components included in Y is estimated as $\hat{X}'_{h,1}$ based on an adaptive harmonic filtering. The dereverberation operator $\mathcal{O}(\hat{R}_1)$ is estimated by calculating the average of $\hat{X}'_{h,1}/Y$ on a number of reverberant speech data. Finally, the dereverberated sound is obtained by multiplying $\mathcal{O}(\hat{R}_1)$ with Y.
2. In the second step, almost the same procedures as the first step are applied except that the speech data dereverberated by the first step is used as the input signal. Because reverberant components, $\hat{R}_2 X_{h,2}$, inevitably included in Eq. (12) is expected to be reduced using $\mathcal{O}(\hat{R}_1)Y$, further dereverberation performance is achieved in step 2.

In both steps, average of $1/\hat{H}'(\omega)$ in Eq. (17) 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.

2.5. Robust F_0 estimation

Accurate F_0 estimation is very important to achieve effective dereverberation in our method. However, this is a difficult task especially for speech with long reverberation using some of the existing F_0 estimators [9].

To cope with this problem, we designed a simple filter reducing sound that continues at the same frequency, and applied it as a preprocessor to the F_0 estimation. Because our method [9] estimates F_0 from the amplitude spectrum, the filter is applied to time series of amplitude spectrum, $|Y(\omega, n)|$. We designed it as $f(n) = (1, -r, -r, \ldots - r)$, where the length of $f(n)$ is m, and r is a positive constant smaller than $1/(m-1)$. The resulting amplitude spectrum is obtained by $|Y^*(\omega, l)| = \sum_n f(n)|Y(\omega, l-n)|$. Although the optimum filter parameters depend on the reverberation condition, we chose a set of parameters that were applicable to all conditions. The effectiveness of this filter has been shown in our preliminary experiments.

In addition to the filter described above, the dereverberation operator, $\mathcal{O}(\hat{R}_1)$, itself is a very effective preprocessor of a F_0 estimator because the reverberation of the speech can be directly reduced by the operator. This mechanism is already included in step 2 of the dereverberation procedure, that is, F_0 estimation is applied to $\mathcal{O}(\hat{R}_1)Y$. Therefore, more accurate F_0 can be obtained in step 2 than in step 1.

3. EVALUATION

We examined the performance of the proposed dereverberation method in terms of reverberation curves [8] and quality of dereverberated sound. For this purpose, we used 5240 Japanese word utterances of a male and a female speakers (MAU and FKM, 12 kHz sampling) included in the ATR database as sound source signal, X. Four impulse responses measured in a reverberant room whose reverberation times are about 0.1, 0.2, 0.5, and 1.0 sec, respectively, are used. The reverberation times are measured based on the method described in the Japanese Industrial Standard [10]. Then, reverberant sounds, Y, are obtained by convolving the word utterances with the impulse responses.

3.1. Reverberation curves

Figure 4 depicts the reverberation curves of original and dereverberated impulse responses. It shows the proposed method could effectively reduce the reverberation in the impulse responses for the female speaker when the reverberation time (r-time) is longer than 0.1 sec. For 0.1 sec r-time, the reverberation effect was neither reduced nor increased. This may be the shortest limitation of the reverberation time in our method. For the male speaker, the reverberation effect in the lower time region was also effectively reduced. This means that strong reverberant components were eliminated from the impulse response, and therefore, the intelligibility of the target speech was expected to improve [8]. Although the reverberation effect in the higher time region for the male speaker was increased when r-time is 0.1 or 0.2 sec, the sound quality as a whole is expected to improve when r-time is 0.2 sec because the earlier reverberation that is much stronger than the later one was eliminated.

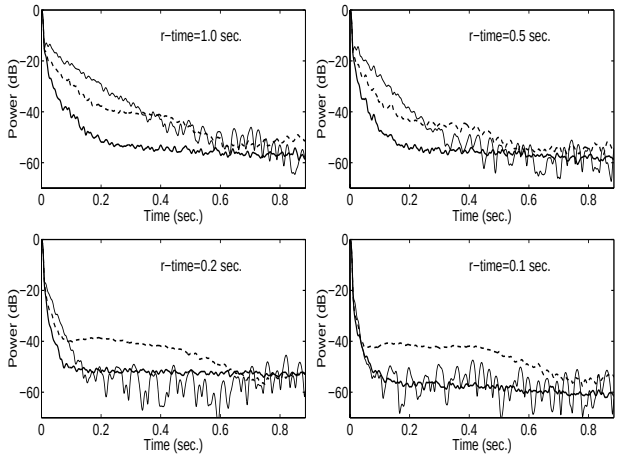

Fig. 4. Reverberation curves of the original impulse responses (thin line) and dereverberated impulse responses (male: thick dashed line, female: thick solid line) for different reverberation time (r-time).

3.2. Sound quality

Figure 5 shows the spectrogram of reverberant and dereverberated speech when r-time is 1.0 sec. As shown in the figure, not only reverberation of speech was effectively reduced, but also the formant structure of the speech was restored. Similar spectrogram features were observed with other reverberation condition when r-time is longer than 0.1 sec. Also, the improvement of sound quality could clearly be recognized by listening to the dereverberated sounds [11].

4. CONCLUSION

This paper proposed a new blind dereverberation method of speech signal with a single microphone. Harmonic structure is estimated and used to approximate the direct sound in a reverberant sound. The dereverberation operator is calculated as the average ratio of the approximated direct harmonic sound to the reverberant sound, and is shown to give the estimate of the inverse transfer function that can be used for the dereverberation. For accurate estimation of the dereverberation operator, an F_0 estimation method that is robust in the presence of long reverberation was also presented. Experimental results showed that the dereverberation operator trained on reverberant speech signals composed of 5240 Japanese word utterances could effectively reduce the reverberation of speech as well as clearly restore the formant structure of speech. Moreover, the results were especially good in the case of the female speaker. Future work includes improving the dereverberation performance for a male speaker, and decreasing the data size required for accurate estimation of the dereverberation operator.

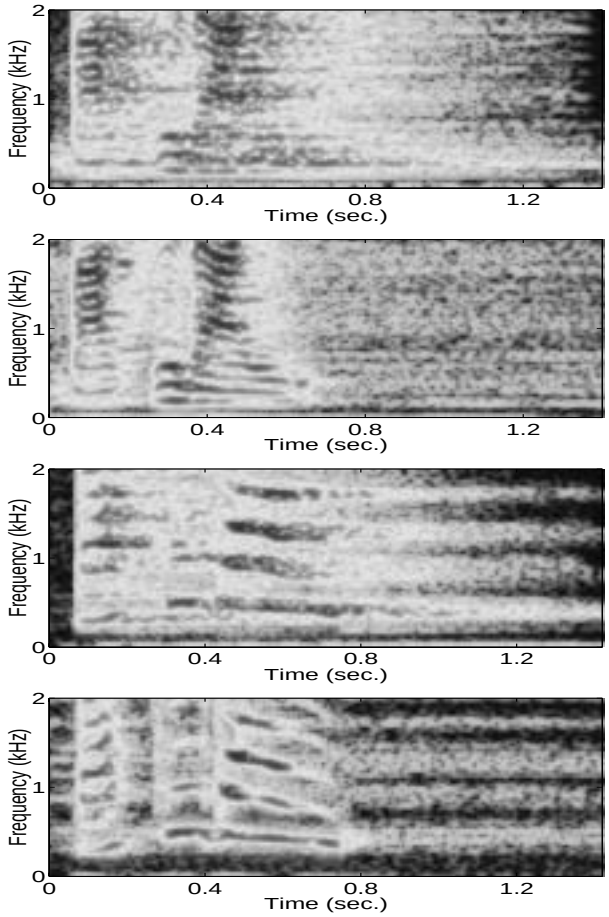

Fig. 5. Spectrogram of reverberant (1st and 3rd panels) and dereverberated (2nd and 4th panels) speech of male (upper two panels) and female (lower two panels) speakers uttering “ba-ku-da-i”.

REFERENCES

[1] Takiguchi, T., Nakamura, S., Huo, Q., and Shikano, K., “Model adaptation based on HMM decomposition for reverberant speech recognition,” *Proc. ICASSP-97*, vol. 2, pp. 827–830, 1997.

[2] Baba, A., Lee, A., Saruwatari, H., and Shikano, K., “Speech recognition by reverberation adapted acoustic model,” *Proc. of ASJ*, 1-9-14, pp. 27–28, Akita, Sep., 2002.

[3] Schmidt, R. O., “Multiple emitter location and signal parameter estimation,” *IEEE Trans. AP*, 34(3), pp. 276–280, 1986.

[4] Roy R., and Kailath T., “ESPRIT: Estimation of signal parameters via rotational invariance techniques,” *IEEE Trans. ASSP*, 37(7), pp. 984–995, 1989.

[5] Miyoshi, M., and Kaneda, Y., “Inverse filtering of room acoustics,” *IEEE Trans. ASSP*, 36(2), pp. 145–152, 1988.

[6] McAulay, R. J., and Quatieri, T. F., “Speech analysis/synthesis based on a sinusoidal representation,” *IEEE Trans. ASSP*, 34, No. 4, pp. 744-754, 1986.

[7] Nakatani, T., “Computational auditory scene analysis based on residue-driven architecture and its application to mixed speech recognition,” Ph.D. thesis, Dept. of Applied Analysis & Complex Dynamical Systems, Kyoto Univ., Mar., 2002.

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

[9] 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.

[10] Japanese Industrial Standard, “Method for measurement of sound absorption coefficients in a reverberation room,” JIS A 1409:1998.

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