SECOND-ORDER STATISTICS BASED DEREVERBERATION BY USING NONSTATIONARITY OF SPEECH

Takuya Yoshioka, Takafumi Hikichi, and Masato Miyoshi

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

ABSTRACT

This paper addresses the problem of speech dereverberation, whose objective is to estimate the inverse filter of an acoustic system in a room (room transfer function). It is well known that using a filter that simply whitens an observed signal deteriorates the characteristics of speech. This is because a speech signal is temporally correlated due to a speech production system (articulatory filter). To avoid this problem, we need to estimate the inverse filter of the room acoustic system separately from that of the speech production system because it is the former that we want to estimate. We recently proposed an algorithm that jointly estimates these inverse filters by exploiting the higher-order statistics of the output. In this paper, we propose an alternative joint estimation based algorithm that uses a criterion involving only the second-order statistics of the output. We present experimental results indicating that the proposed algorithm can estimate the inverse filter of the room acoustic system with a reverberation time of 0.5 seconds from observed signals of 3-5 seconds. Results obtained in the presence of additive noise are also presented showing that the proposed algorithm succeeds in the dereverberation under the noise of 20 dB.

1. INTRODUCTION

Room reverberation often degrades the quality of speech. Hence, speech dereverberation is desired as a preprocessing technique for various speech processing applications. One may consider the speech dereverberation as blind inverse filtering of an acoustic system in a room as follows. Let a clean speech signal at time n be represented by $s(n)$, and let the signal transmission channel from the source to a set of M (≥ 1 in general) microphones be represented by the Kth-order finite impulse response (FIR) system $H = \{\boldsymbol{h}(k)\}_{k=0}^{K} = \{[h_1(k), \cdots, h_M(k)]^T\}_{k=0}^{K}$, where superscript T indicates the transposition of a vector or a matrix [1]. For each i, $H_i = \{h_i(k)\}_{k=0}^{K}$ forms a subchannel of H corresponding to the signal transmission channel from the source to the ith microphone. Observed multichannel signal $\boldsymbol{x}(n) = [x_1(n), \cdots, x_M(n)]^T$ can be described as

$$\boldsymbol{x}(n) = \sum_{k=0}^{K} \boldsymbol{h}(k) s(n-k). \quad (1)$$

Then, the task of the dereverberation is to recover the source signal from the observed signal. This is achieved by convolving an inverse filter of room acoustic system H with observed signal $\boldsymbol{x}(n)$ as

$$y(n) = \sum_{k=0}^{L} \boldsymbol{g}(k)^T \boldsymbol{x}(n-k), \quad (2)$$

where $G = \{\boldsymbol{g}(k)\}_{k=0}^{L} = \{[g_1(k), \cdots, g_M(k)]^T\}_{k=0}^{L}$ is the Lth-order FIR inverse filter of H. Therefore, when the observed signal samples $\{\boldsymbol{x}(n)\}_{n=1}^{N}$ are given, we want to set up each tap $g_m(k)$ of the inverse filter so that the recovered signal $y(n)$ is identical to the source signal, $s(n)$, up to a constant scale and delay.

Speech signal $s(n)$ is produced by an articulatory system, which is widely modeled as a piecewise autoregressive (AR) system, driven by an innovations process [1]. In this model, $s(n)$ is described as

$$s(n) = \sum_{k=1}^{P} b_i(k) s(n-k) + e(n),\ i = \left\lfloor \frac{n-1}{W} + 1 \right\rfloor, \quad (3)$$

where $B_i = \{b_i(k)\}_{k=1}^{P}$ denotes a Pth-order AR system of the ith time frame, $e(n)$ denotes the innovations process, and W is the frame size within which $e(n)$, and therefore $s(n)$, can be regarded as stationary. From (1) and (3), $\boldsymbol{x}(n)$ can be viewed as the output of a composite system of H and B_i driven by $e(n)$. Our objective is to obtain the inverse filter of only room acoustic system H. Therefore, we must identify the inverse filter of H separately from that of B_i under the condition that neither the parameters of H nor that of B_i are available.

[1] For the sake of a simple description, we refer to a set of the signal transmission channel(s) between a source and possibly multiple microphone(s) as a signal transmission channel. The channel between the source and one of the microphones is called as the subchannel. A set of signal(s) observed by the microphone(s) is accordingly refered to as an observed signal.

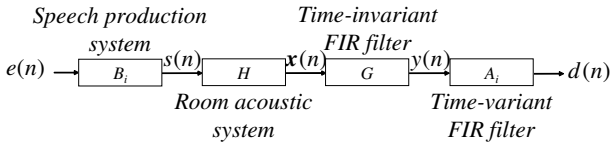

Figure 1: Schematic diagram of overall system.

One major approach is to exploit the diversity between subchannels $H_1, \cdots, H_M$ [2, 3]. However, this approach appears to be sensitive to observation noise. Other methods such as [4, 5] may offer some degree of robustness as regards additive noise. However, they are liable to require a long observation period. We recently proposed a method that jointly estimates the inverse filters of H and B_i [6]. The joint estimation approach is able to provide each of these inverse filters owing to the time-variant nature of speech production system B_i. It was shown that the method could estimate the inverse filters with high accuracy from observed signals of 10-20 seconds.

In this paper, we propose an alternative dereverberation algorithm based on the joint estimation approach in order to further shorten the length of observed signals. The significant difference between the algorithm proposed here and that reported previously is that the former uses a criterion involving the second-order statistics (SOS) of the system output while the latter exploits the higher-order statistics (HOS). Since estimation of the SOS demands a smaller sample size than for estimation of the HOS, the proposed algorithm will be more efficient in terms of observed signal length. Our experimental results showed that the proposed algorithm could work with observed signals of at most five seconds. Moreover, the algorighm is shown to work well in the presence of additive noise.

2. ESTIMATION PRINCIPLE

In this paper, we assume the following conditions.

1) Innovations $\{e(n)\}_{n=1}^{N}$ consists of zero-mean uncorrelated random variables, where N is the number of samples.

2) The number, M, of microphones satisfies $M \geq 2$ and the order, L, of inverse filter G is sufficiently large for the truncation effects to be ignored. The transfer functions of subchannels $H_1, \cdots, H_M$ are assumed to have no common zero.

3) Speech production system B_i has no time invariant pole.

The basic idea we proposed in [6] involves the joint estimation of inverse filters of room acoustic system H and speech production system B_i. Let us consider to filter observed signal $\boldsymbol{x}(n)$ with time-invariant FIR filter G and then with minimum-phase time-variant FIR filter $A_i = \{a_i(k)\}_{k=1}^{P}$ as shown in Fig. 1. The final output $d(n)$ is given as

$$d(n) = y(n) - \sum_{k=1}^{P} a_i(k) y(n-k),\ i = \left\lfloor \frac{n-1}{W} + 1 \right\rfloor, \quad (4)$$

where $y(n)$ is calculated by (2). Under assumptions 2) and 3), we can prove that if $d(n)$ is equalized with $e(n)$ except for the scaling and delay ambiguity and A_i has no time-invariant zero, then G and A_i become the inverse filters of H and of B_i, respectively [6]. Therefore, we have to determine taps $g_m(k)$ and $a_i(k)$ so that $d(n)$ is equalized with $e(n)$.

3. PROPOSED ALGORITHM

3.1. Loss function

Based on assumption 1), it would be natural to estimate $g_m(k)$ and $a_i(k)$ so that the samples, $\{d(n)\}_{n=1}^{N}$, of the output signal are uncorrelated. Let $\mathcal{K}(\cdot)$ be a suitable measure of the correlation between random variables. Then, we would like to minimize $\mathcal{K}(d(1), \cdots, d(N))$ with respect to $a_i(k)$ and $g_m(k)$.

In this paper, we use the measure of correlation proposed in [7]:

$$\mathcal{K}(d(1), \cdots, d(N)) = \sum_{n=1}^{N} \log \upsilon(d(n)) - \log |\det \Sigma(\boldsymbol{d})|, \quad (5)$$

where $\boldsymbol{d} = [d(N), \cdots, d(1)]^T$, $\upsilon(\cdot)$ represents the variance of a random variable, and $\Sigma(\cdot)$ represents the covariance matrix of a multivariate random variable. We believe (5) to be suitable for measuring the degree of correlation because it always takes nonnegative value and it is equal to zero if and only if $d(1), \cdots, d(N)$ are uncorrelated. Under assumption 2), we can use the framework of multichannel linear prediction [8], which means that the first tap of inverse filter G is fixed as

$$[g_1(0), g_2(0), \cdots, g_M(0)] = [1, 0, \cdots, 0] \quad (6)$$

Then, (5) can be simplified as (see Appendix)

$$\mathcal{K}(d(1), \cdots, d(N)) = \sum_{n=1}^{N} \log \upsilon(d(n)) + \text{constant}. \quad (7)$$

Therefore, what we are to solve is finally formulated as

$$\underset{a_i(k),\ g_m(k)}{\text{minimize}} \sum_{n=1}^{N} \log \upsilon(d(n))$$
$$\text{subject to } A_i \text{ is minimum phase.} \quad (8)$$

Problem (8) says that we just have to minimize the logarithmic mean of the variances of $d(1), \cdots, d(N)$. The constraint of (8) is intended to stabilize the estimate, A_i, of speech production system B_i.
Assume that the variance of $d(n)$ is stationary over a whole observation period. The loss function of (8) is then reduced to $N \log \upsilon(d(n))$. Because the logarithmic function increases monotonically, the loss function is further simplified to $N\upsilon(d(n))$, which may be estimated by $\sum_{n=1}^{N} d(n)^2$. Thus, when the variance of $d(n)$ is stationary, the loss function of (8) is equivalent to the traditional least squares (LS) criterion. However, since the variance of the innovations process is globally nonstationary, the loss function proposed here may be more appropriate. This conjecture will be validated experimentally.

3.2. Algorithm

We solve the optimization problem (8) by using an alternating variables method [9]. We optimize the loss function with respect first to the taps $a_i(k)$ for a fixed G, then with respect to the taps $g_m(k)$ for fixed $A_1, \cdots, A_{T-1}$, and A_T, and so on.
First, let us derive the optimization algorithm with respect to $a_i(k)$. The following two points should be noted:

- Because G is fixed here, $y(n)$ is also fixed.
- Output samples $\{d(n)\}_{n=N_i}^{N_i+W-1}$ in the ith time frame depend only on $\{a_i(k)\}_{k=1}^{P}$, where N_i is the first sample number of the ith frame.

The optimization is then realized by minimizing the loss function given as $\sum_{n=N_i}^{N_i+W-1} \log \upsilon(d(n))$ from recovered signal samples $\{y(n)\}_{n=N_i}^{N_i+W+1}$ for each frame number i. Assume that variance $\upsilon(d(n))$ changes with sample number n slowly enough for the variance to be regarded as stationary within a single frame of size W. Then, the loss function is equivalent to $\langle d(n)^2 \rangle_{n=N_i}^{N_i+W-1}$, where $\langle \cdot \rangle_{n=n_1}^{n_2}$ represents an operator taking an average from the n_1-th to n_2-th samples. Then, the loss function can be minimized by applying linear prediction to $\{y(n)\}_{n=N_i}^{N_i+W-1}$. Note that LPC guarantees A_i to be minimum phase when the autocorrelation method is used [1].
Next, we derive the optimization algorithm with respect to $g_m(k)$. By calculating the derivative of the estimate, $\sum_{i=1}^{T} \log \langle d(n)^2 \rangle_{n=N_i}^{N_i+W-1}$, of the loss function, we have the following algorithm based on the gradient method:

$$g_m(k)' = g_m(k) + \delta \sum_{i=1}^{T} \frac{\langle d(n) v_{m,i}(n-k) \rangle_{n=N_i}^{N_i+W-1}}{\langle d(n)^2 \rangle_{n=N_i}^{N_i+W-1}} \tag{9}$$

$$v_{m,i}(n) = x_m(n) - \sum_{k=1}^{P} a_i(k) x_m(n-k), \tag{10}$$

where δ is step size. Note that we again assumed that the variance of $d(n)$ changes slowly.

4. EXPERIMENTAL RESULTS

We conducted experiments to evaluate the performance of the proposed algorithm. Japanese sentences uttered by 10 speakers taken from ASJ-JNAS database were used as the source signals. The signals were sampled at 8 kHz and quantized at 16-bit resolution. The observed signals were simulated by convolving the source signals with impulse responses measured in a room $4.45 \times 3.55 \times 2.5$ m^3 in size. The distance between the loudspeaker and the microphones was about 3.2 m. The reverberation time was around 0.5 seconds.
The following settings were used: $M = 4$, $L = 1000$, $W = 200$, $P = 16$. Estimation variables $a_i(k)$ and $g_m(k)$ were alternated six times. The update equation (9) was employed 50 times.
The dereverberation performance was evaluated by using D_{50} [10], which is a measure related to speech intelligibility. It is defined as

$$D_{50} = \frac{\int_0^{50 \text{ msec.}} f(t)^2 dt}{\int_0^{\infty} f(t)^2 dt} \times 100 \ (\%), \tag{11}$$

where $\{f(t); t \geq 0\}$ denotes an impulse response.
In Fig. 2, we plot the D_{50} score averaged over the 10 speakers' results as a function of the length of the observed signals. We also plot the performance of the algorithm based on the LS criterion, which assumes the stationarity of variance $\upsilon(d(n))$. It can be clearly seen that the D_{50} score was recovered well with observed signals of only 3-5 seconds. By comparing the performance of the proposed and LS-based algorithms, we can also recognize the benefit of taking the nonstationarity of variance $\upsilon(d(n))$ into consideration.
We also tested the case where the observed signals were contaminated by additive noise. We used white Gaussian noise with signal to noise ratios (SNR) of 40, 30, 20, and 10 dB. In Fig. 3, we plot the average D_{50} score as a function of SNR. From this result, we can conclude that the proposed algorithm is robust against additive noise with a SNR of larger than or equal to 20 dB.

5. CONCLUSION

We have described an alternative speech dereverberation algorithm based on the joint estimation approach originally introduced in [6]. The algorithm calculates the inverse filters of a room acoustic system and a speech production system so that the samples of the output signal are

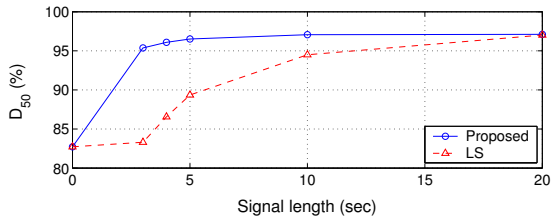

Figure 2: D_{50} as a function of the length of observed signals.

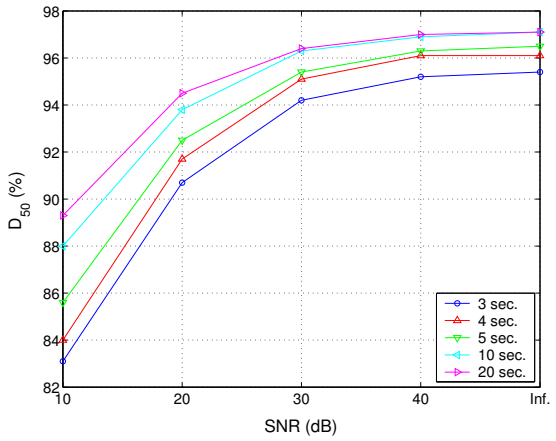

Figure 3: D_{50} obtained in the presence of noise.

uncorrelated. The proposed algorithm shows fast adaptability and robustness against additive noise. This property of the algorithm results from the use of a loss function that can deal with the nonstationarity of signal variance.

6. APPENDIX: DERIVATION

The $\tilde{n}$th sample of $\{d(n)\}_{n=1}^{N}$, $d(\tilde{n})$, is represented as a linear combination of samples $\{s(n)\}_{n=1}^{\tilde{n}}$, and the contribution of $s(\tilde{n})$ is $\sum_{m=1}^{M} h_m(0)g_m(0)$. Therefore, $\boldsymbol{d}$ is written as

$$\boldsymbol{d} = F\boldsymbol{s}, \tag{12}$$

where $\boldsymbol{s} = [s(N), \cdots, s(1)]^T$, and F is an $N \times N$ upper triangular matrix whose diagonal elements are all $\sum_{m=1}^{M} h_m(0)g_m(0)$. From relation (12), we have

$$\log|\det \Sigma(\boldsymbol{d})| = \log|\det \Sigma(\boldsymbol{s})| + 2\log|\det F|. \tag{13}$$

Since the determinant of an upper triangular matrix is the product of its diagonal elements, we obtain under (6)

$$\begin{aligned}\log|\det F| &= N\log\Big|\sum_{m=1}^{M} h_m(0)g_m(0)\Big| \\ &= N\log|h_1(0)| = \text{constant}\end{aligned} \tag{14}$$

when condition (6) holds. (13) and (14) leads to

$$\log|\det \Sigma(\boldsymbol{d})| = \text{constant}, \tag{15}$$

which indicates (7).

7. REFERENCES

[1] L. R. Rabiner and R. W. Schafer, *Digital processing of speech signals*, Prentice Hall, 1983.

[2] S. Gannot and M. Moonen, "Subspace methods for multimicrophone speech dereverberation," *EURASIP J. Applied Signal Processing*, vol. 2003, no. 11, pp. 1074–1090, 2003.

[3] T. Hikichi, M. Delcroix, and M. Miyoshi, "Blind dereverberation based on estimates of signal transmission channels without precise information on channel order," in *Proc. Int'l Conf. Acoust., Speech, Signal Processing*, 2005, pp. 1069–1072.

[4] T. Nakatani and M. Miyoshi, "Blind dereverberation of single channel speech signal based on harmonic structure," in *Proc. Int'l Conf. Acoust., Speech, Signal Processing*, 2003, pp. 92–95.

[5] B. W. Gillespie, H. S. Malvar, and D. A. F. Florêncio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," in *Proc. Int'l Conf. Acoust., Speech, Signal Processing*, 2001, pp. 3701–3704.

[6] T. Yoshioka, T. Hikichi, M. Miyoshi, and H. G. Okuno, "Robust decomposition of inverse filter of channel and prediction error filter of speech signal for dereverberation," in *Proc. European Signal Processing Conference*, 2006, in press.

[7] K. Matsuoka, M. Ohya, and M. Kawamoto, "A neural net for blind separation of nonstationary signals," *Neural Networks*, vol. 8, no. 3, pp. 411–419, 1995.

[8] A. Gorokhov and P. Loubaton, "Blind identification of MIMO-FIR systems: A generalized linear prediction approach," *Signal Processing*, vol. 1, pp. 105–124, 1999.

[9] A. Hyvärinen, "Independent component analysis in the presence of gaussian noise by maximiaing joint likelihood," *Neurocomputing*, vol. 22, pp. 49–67, 1998.

[10] H. Kuttruff, *Room acoustics*, Elsevier Applied Science, third edition, 1991.