

PAPER

Blind dereverberation algorithm for speech signals based on multi-channel linear prediction

Marc Delcroix^{1,2,*}, Takafumi Hikichi^{1,†} and Masato Miyoshi^{1,2,‡}

¹*NTT Communication Science Laboratories, NTT Corporation,
2-4, Hikaridai, Seika-cho, "Keihanna Science City," Kyoto, 619-0237 Japan*

²*Graduate School of Information Science and Technology, Hokkaido University,
Kita 14, Nishi 9, Kita-ku, Sapporo, 060-0814 Japan*

(Received 5 July 2004, Accepted for publication 31 January 2005)

Abstract: This paper proposes an algorithm for the blind dereverberation of speech signals based on multi-channel linear prediction. Traditional dereverberation methods usually perform well when the input signal is white noise. However, when dealing with colored signals generated by an autoregressive process such as speech, the generating autoregressive process is deconvolved causing excessive whitening of the signal. We overcome this whitening problem by estimating the generating autoregressive process based on multichannel linear prediction and applying this estimated process to the whitened signal so that the input signal can be recovered. Simulation results show the good potential of the proposed method.

Keywords: Blind dereverberation, Multi-channel, Linear prediction, Prediction filters, Autoregressive process

PACS number: 43.60.Pt, 43.72.Ew [DOI: 10.1250/ast.26.432]

1. INTRODUCTION

The effect of reverberations in a room on speech signals is a critical problem in many speech applications. For example, it is important to eliminate reverberations if we are to achieve robust automatic speech recognition (ASR) in real environments. Reverberation in rooms severely changes signal characteristics, and thus degrades recognition performance. Much effort has been devoted to the dereverberation problem using both single and multi-channel based techniques [1–11], but no satisfactory method has been found yet.

As for single channel dereverberation, a technique has been developed for estimating the inverse filter of a room transfer function using the harmonic structure of speech [1,2]. It works well for long reverberation times, but practical use is still limited due to the large amount of speech data required. Another single microphone method [3] proposes enhancing speech regions where direct speech components are dominant compared with the reverberant parts of the signal. The difficulty of determining those

regions, however, limits the current success of this method.

Microphone-array methods have also been investigated. A typical technique uses Direction Of Arrival (DOA) [5–7] to enhance the target signal. However, when a small number of microphones are used, DOA methods can only be employed if there are few reflections. Other methods are based on a calculation of the inverse filters of room acoustics. If the room transfer functions between one source and two microphones are known, exact inverse filtering can be achieved [8]. For independent and identically distributed (i.i.d.) sequences, such inverse filters can be blindly calculated [9–11]. However, because a speech signal is not i.i.d. it is known [12] that such dereverberation methods also deconvolve the speech-generating autoregressive (AR) process causing excessive whitening of the signal. The whitening changes the signal characteristics and may lead to problems in the speech recognition task.

In this paper we propose a two-microphone dereverberation method that blindly recovers an original signal without suffering from this whitening problem. We use two-channel linear prediction to calculate the prediction filter set and estimate the generating AR process [13,14]. By applying the estimated AR process to the filtered signal, speech is completely recovered.

This paper is organized as follows: In Section 2 we

*e-mail: marc.delcroix@cslab.kecl.ntt.co.jp

†e-mail: hikichi@cslab.kecl.ntt.co.jp

‡e-mail: miyo@cslab.kecl.ntt.co.jp

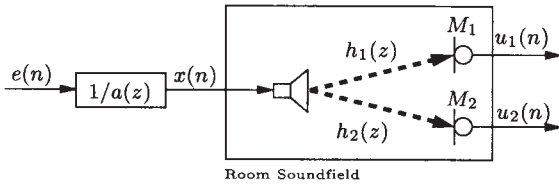


Fig. 1 Schema of room. The input signal $x(n)$ is generated by an AR process on white noise $e(n)$. $a(z)$ is the AR polynomial. $u_1(n)$ and $u_2(n)$ are the signals received at microphones M_1 and M_2 , respectively. We call $h_1(n)$ and $h_2(n)$ the room impulse responses.

formulate the problem and explain how linear prediction can be used to solve it. We summarize the developments and provide a blind dereverberation algorithm in Section 3. Section 4 describes the simulation conditions and presents our results. Sections 5 and 6 contain some remarks and our conclusion.

2. PRINCIPLE

We consider a room soundfield with a sound source and two microphones as shown in Figure 1. Although the developments are presented for the particular case of two-microphones, the method could be extended to a more general multi-microphone situation.

The objective of blind dereverberation is to cancel out the effects of a room's reverberation on the input signal, based only on signals received at the microphones.

We construct the following hypotheses:

- First, we assume that input signal $x(n)$ is generated from a finite AR process applied on white noise $e(n)$. The AR polynomial is

$$a(z) = 1 - \{a_1 z^{-1} + \dots + a_N z^{-N}\}. \quad (1)$$

- We also assume that room transfer functions $H_1(z)$ and $H_2(z)$, modeled by polynomials, are time-invariant and have no common zeros.

$$H_i(z) = \sum_{k=0}^m h_i(k) z^{-k} \triangleq h_{i,0} + h_{i,1} z^{-1} + \dots + h_{i,m} z^{-m}, \quad i = 1, 2. \quad (2)$$

Let us call the signals received at the microphones M_1 and M_2 , $u_1(n)$ and $u_2(n)$ respectively. They are obtained by

filtering $x(n)$ with the room transfer functions. The blind dereverberation problem thus consists in recovering input signal $x(n)$ from microphone signals $u_1(n)$ and $u_2(n)$.

The proposed method solves the problem by first calculating the prediction filters that cancel out the reverberation effects of the room transfer functions. As those filters also whiten the signal, we estimate the AR process that recovers input signal $x(n)$. Figure 2 shows a schematic diagram of the dereverberation system.

2.1. Prediction Filters

The linear prediction formalism can be used to calculate the prediction filters. Indeed, the impulse response of the prediction filters $w_1(n)$, $w_2(n)$ (i.e. whitening filters) can be obtained by minimizing the mean square value of the prediction error $\hat{e}(n)$:

$$\hat{e}(n) = u_1(n) - (w_1(n) * u_1(n-1) + w_2(n) * u_2(n-1)), \quad (3)$$

$$= h_1(n) * x(n) - \{w_1(n) * (h_1(n) * x(n-1)) + w_2(n) * (h_2(n) * x(n-1))\}, \quad (4)$$

where $*$ denotes the convolution operator.

We can reformulate Eq. (4) using matrix notations as:

$$\hat{e}(n) = \mathbf{x}_n^T \mathbf{h}_1 - \mathbf{x}_{n-1}^T \mathbf{H} \mathbf{w} \quad (5)$$

where:

$$\mathbf{x}_n = [x(n), \dots, x(n - (m + L))]^T,$$

$$\mathbf{h}_1 = [h_{1,0}, \dots, h_{1,m}, 0, \dots, 0]^T,$$

\mathbf{H} is a full row-rank matrix of size $(m + L) \times 2L$ and $2L \geq m + L$ [8,13],

$$\mathbf{H} = [\mathbf{H}_1, \mathbf{H}_2],$$

\mathbf{H}_i is a $(m + L) \times L$ convolution matrix expressed as

$$\mathbf{H}_i = \begin{pmatrix} h_{i,0} & 0 & \dots & 0 \\ h_{i,1} & h_{i,0} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \\ h_{i,m} & & & 0 \\ 0 & h_{i,m} & & h_{i,0} \\ \vdots & & \ddots & \vdots \\ 0 & \dots & 0 & h_{i,m} \end{pmatrix}, \quad i = 1, 2,$$

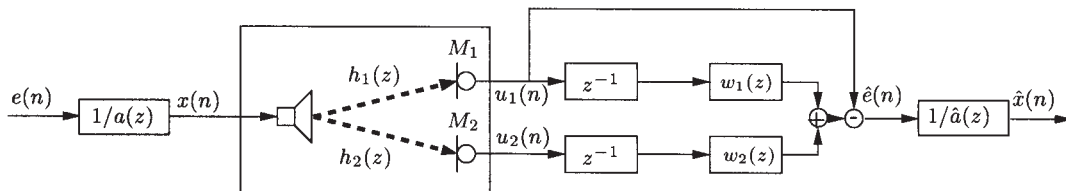


Fig. 2 Schematic diagram of dereverberation system. To recover the input signal from the microphone signals, we first calculate the prediction error, then filter it with estimated AR process $\hat{a}(z)$. The combination of the prediction filters ($w_1(n)$, $w_2(n)$) and the estimated AR process forms the dereverberation process.

\mathbf{w} is the prediction filter set,

$$\mathbf{w} = [\mathbf{w}_1^T, \mathbf{w}_2^T]^T,$$

$$\mathbf{w}_i = [w_{i,0}, \dots, w_{i,L-1}]^T, w_{i,k} \triangleq w_i(k), i = 1, 2.$$

Minimizing the mean square value of the prediction error gives us:

$$\mathbf{w} = (\mathbf{H}^T E\{\mathbf{x}_{n-1}\mathbf{x}_{n-1}^T\}\mathbf{H})^+ \mathbf{H}^T E\{\mathbf{x}_{n-1}\mathbf{x}_n^T\}\mathbf{h}_1 \quad (6)$$

where A^+ is the Moore-Penrose generalized inverse of matrix A [15], and $E\{\}$ is an expectation operator. If we replace the column vector \mathbf{h}_1 with matrix \mathbf{H} , we can define matrix \mathbf{Q} as:

$$\mathbf{Q} \triangleq (\mathbf{H}^T E\{\mathbf{x}_{n-1}\mathbf{x}_{n-1}^T\}\mathbf{H})^+ \mathbf{H}^T E\{\mathbf{x}_{n-1}\mathbf{x}_n^T\}\mathbf{H}. \quad (7)$$

As the input signal is generated by an AR process, we can write [16]:

$$\mathbf{x}_n = \mathbf{C}^T \mathbf{x}_{n-1} + \mathbf{e}_n, \quad (8)$$

where:

\mathbf{C} is the companion matrix defined as:

$$\mathbf{C} = \begin{pmatrix} a_1 & 1 & 0 & \dots & 0 \\ a_2 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & & \ddots & 1 \\ a_N & 0 & \dots & \dots & 0 \end{pmatrix}, \quad (9)$$

and $\mathbf{e}_n = [e(n), 0, \dots, 0]^T$.

We then have:

$$E\{\mathbf{x}_{n-1}\mathbf{x}_n^T\} = E\{\mathbf{x}_{n-1}\mathbf{x}_{n-1}^T\}\mathbf{C}. \quad (10)$$

Assuming that $E\{\mathbf{x}_{n-1}\mathbf{x}_{n-1}^T\}$ is positive definite, we can replace it with $\mathbf{X}^T \mathbf{X}$ where \mathbf{X} is a matrix. Matrix \mathbf{Q} is thus expressed as:

$$\begin{aligned} \mathbf{Q} &= (\mathbf{H}^T \mathbf{X}^T \mathbf{X} \mathbf{H})^+ \mathbf{H}^T \mathbf{X}^T \mathbf{X} \mathbf{C} \mathbf{H} \\ &= (\mathbf{X} \mathbf{H})^+ \mathbf{X} \mathbf{C} \mathbf{H} \\ &= \mathbf{H}^T (\mathbf{H} \mathbf{H}^T)^{-1} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \mathbf{C} \mathbf{H} \\ &= \mathbf{H}^T (\mathbf{H} \mathbf{H}^T)^{-1} \mathbf{C} \mathbf{H}, \end{aligned} \quad (11)$$

By definition, the first column of \mathbf{Q} gives us the prediction filter set,

$$\mathbf{w} = \mathbf{H}^T (\mathbf{H} \mathbf{H}^T)^{-1} \mathbf{C} \mathbf{h}_1. \quad (12)$$

The prediction error is thus:

$$\begin{aligned} \hat{e}(n) &= \mathbf{x}_n^T \mathbf{h}_1 - \mathbf{x}_{n-1}^T \mathbf{H} \mathbf{w} \\ &= \mathbf{x}_n^T \mathbf{h}_1 - \mathbf{x}_{n-1}^T \mathbf{H} \mathbf{H}^T (\mathbf{H} \mathbf{H}^T)^{-1} \mathbf{C} \mathbf{h}_1 \\ &= (\mathbf{x}_n^T - \mathbf{x}_{n-1}^T \mathbf{C}) \mathbf{h}_1 \\ &= \mathbf{e}_n^T \mathbf{h}_1 \\ &= h_{1,0} e(n). \end{aligned} \quad (13)$$

Equation (13) shows that the prediction error is proportional to white noise $e(n)$. The effect of room reverberation is thus canceled out but the signal is whitened. To recover the signal $x(n)$ we still need to estimate the AR polynomial $a(z)$ as defined in equation (1). By filtering the prediction error with the estimated AR process $1/\hat{a}(z)$, we will recover the input signal $x(n)$.

2.2. Estimated AR Process

Let us first recall the expression of the characteristic polynomial of the companion matrix \mathbf{C} defined in Eq. (9):

$$\begin{aligned} f_c(\mathbf{C}, \lambda) &= -\lambda^N + a_1 \lambda^{N-1} + \dots + a_N \\ &= -\lambda^N \{1 - (a_1 \lambda^{-1} + \dots + a_N \lambda^{-N})\}, \end{aligned} \quad (14)$$

where $f_c(\mathbf{A}, \lambda) = \det(\mathbf{A} - \lambda \mathbf{I})$ is the characteristic polynomial of matrix \mathbf{A} . From Eqs. (14) and (1) we note that the coefficients of the polynomial $a(z)$ are equivalent to the characteristic polynomial coefficients of matrix \mathbf{C} .

Let us now consider the non-zero eigenvalues of matrix \mathbf{Q} [17]:

$$\begin{aligned} \lambda(\mathbf{Q}) &= \lambda(\mathbf{H}^T (\mathbf{H} \mathbf{H}^T)^{-1} \mathbf{C} \mathbf{H}) \\ &= \lambda(\mathbf{H} \mathbf{H}^T (\mathbf{H} \mathbf{H}^T)^{-1} \mathbf{C}) \\ &= \lambda(\mathbf{C}). \end{aligned} \quad (15)$$

We can thus derive the following relation:

$$f_c(\mathbf{Q}, \lambda) = f_c(\mathbf{C}, \lambda) \quad (16)$$

From Eq. (16) we deduce that the estimated AR polynomial, $\hat{a}(z)$, can be obtained from the characteristic polynomial of matrix \mathbf{Q} . By filtering the prediction error with the inverse of the estimated AR process, $1/\hat{a}(z)$, we obtain $\hat{x}(n)$, the recovered input signal.

2.3. Calculation of Matrix \mathbf{Q}

The algorithm is “blind” because the dereverberation is achieved without prior knowledge of the room transfer functions. Indeed, we only need to calculate matrix \mathbf{Q} in order to recover the input signal, and matrix \mathbf{Q} can be calculated with the signals received at the microphones. Using the matrix notation defined previously, the microphone signals can be expressed as:

$$\mathbf{u}_n = \mathbf{H}^T \mathbf{x}_n \quad (17)$$

where $\mathbf{u}_n = [u_1(n) \dots u_1(n-L), u_2(n) \dots u_2(n-L)]^T$. Using relation (7) and (17), we can express matrix \mathbf{Q} as a function of the microphone signals:

$$\mathbf{Q} = (E\{\mathbf{u}_{n-1}\mathbf{u}_{n-1}^T\})^+ E\{\mathbf{u}_{n-1}\mathbf{u}_n^T\}. \quad (18)$$

Equation (18) is used in practice to calculate \mathbf{Q} .

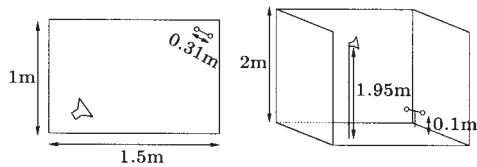


Fig. 3 Simulated soundfield. We chose high reflection coefficients for the walls (0.8) and placed the microphones close to the corner to obtain a non-minimum phase transfer function.

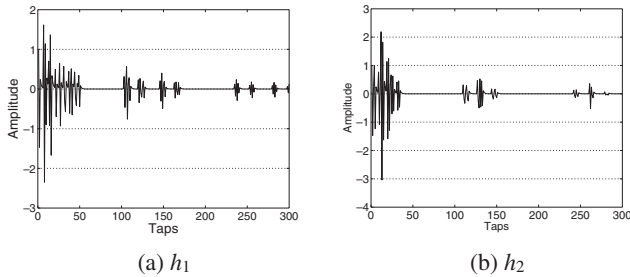


Fig. 4 Room transfer functions. The transfer functions are simulated by the image method and truncated to 300 taps. They are non-minimum phase.

3. ALGORITHM

We can summarize the dereverberation algorithm as follows:

- (1) First we calculate matrix \mathbf{Q} with the two signals received at the microphones using Eq. (18).
- (2) The first column of matrix \mathbf{Q} gives us the prediction filter set, \mathbf{w}_1 and \mathbf{w}_2 .
- (3) The prediction error is calculated using formula (3).
- (4) The estimated AR parameters are obtained by the characteristic polynomial of \mathbf{Q} .
- (5) The input signal is recovered by filtering the prediction error with the estimated AR parameters.

4. SIMULATIONS

We conducted simulations to test the described method. We carried out two types of simulation. Each simulation was undertaken for the ideal case of a noise free environment. With the first type the input signals were generated by a time-invariant AR process applied on white noise. With the second type the input signal was speech. The latter case corresponds to a time-variant AR process.

4.1. Time-Invariant AR Process

This first kind of simulation is very close to the principle described above and was carried out to prove its validity.

4.1.1. Simulation conditions

Room transfer functions were simulated using the image method [18]. We simulated a typical space environ-

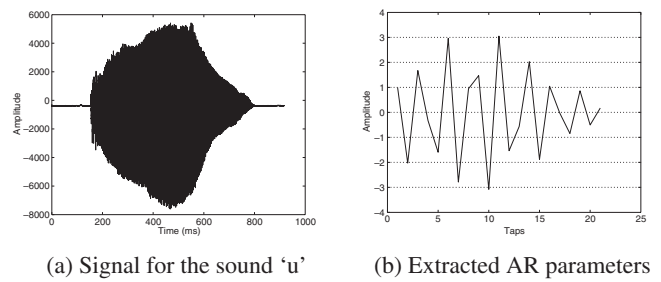


Fig. 5 AR process of input signal. Linear prediction was applied to a vowel sound signal to extract a generating AR process. The length of the AR process was set at 21 taps.

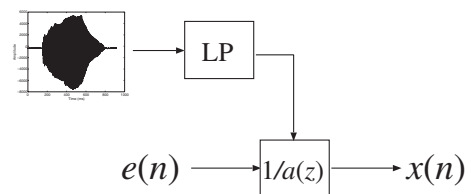


Fig. 6 Generation of input signal $x(n)$. Polynomial $a(z)$ is extracted from the vowel signal using linear prediction (LP). The input signal $x(n)$ is obtained by filtering white noise $e(n)$ with $1/a(z)$.

Table 1 Simulation conditions.

Length of impulse response	300 taps
Number of input signal samples	50,000
Length of generating AR process	21 taps
Sampling frequency	16 kHz
Length of prediction filters	300 taps
Length of estimated AR process	601 taps

ment found in offices where a desk is surrounded by three walls (Fig. 3). The room impulse responses were truncated to 300 taps corresponding to a short duration of 18.75 ms, the sampling frequency being 16 kHz. The actual reverberation time calculated with Sabine formula [19] is around 70 ms. Figure 4(a) and (b) show the room impulse responses. We used non-minimum phase transfer functions to show that the method works in general cases.

We generated the input signal $x(n)$ by filtering white noise with an AR process as described in Section 2. AR polynomial $a(z)$ was extracted by linear prediction applied to a speech signal. Figure 5(a) and (b) show the speech signal and the derived AR parameters. Figure 6 shows how we generated the input signal.

The simulation conditions are summarized in Table 1.

4.1.2. Results

Table 2 shows the results we obtained for three different AR polynomials corresponding to the sounds 'a', 'i' and 'u'. In each case, the input signals are generated by filtering a white noise signal with the AR process. The

Table 2 Simulation results.

	SDR_{Before} [dB]	SDR_{After} [dB]
Vowel 'a'	-3.68	101
Vowel 'i'	-2.69	107
Vowel 'u'	-2.78	97.7

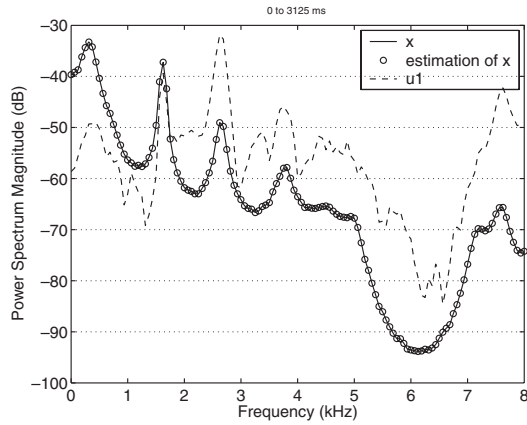


Fig. 7 Power spectrum of the input signal $x(n)$, the estimation of $x(n)$ and the signal received at the microphone, $u_1(n)$. The dashed line represents the power spectrum of the microphone signal, $u_1(n)$, which clearly suffers from the effect of the room transfer function. The circles represent the recovered signal's power spectrum, which precisely fits the input signal's power spectrum.

evaluation of the results were made using SDR's (Signal to Distortion Ratio) defined in Eqs. (19) and (20):

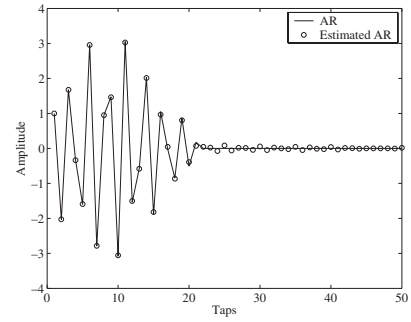
$$SDR_{\text{Before}} = 10 \log_{10} \left(\frac{\sum |x(n)|^2}{\sum |x(n) - u_1(n)|^2} \right), \quad (19)$$

$$SDR_{\text{After}} = 10 \log_{10} \left(\frac{\sum |x(n)|^2}{\sum |x(n) - \hat{x}(n)|^2} \right), \quad (20)$$

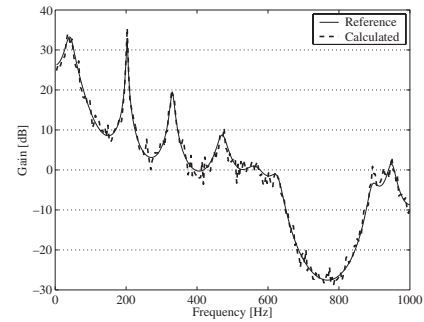
where $x(n)$ is the input signal, $u_1(n)$ is the signal received at the microphone $M1$, and $\hat{x}(n)$ is the estimated signal. The first column shows the SDR obtained at the microphone (before processing) as defined in Equation (19). The second column shows the SDR after applying the de-reverberation method (after processing) as defined in Equation (20). The method performs well since the input signal can be recovered with a high SNR.

Figure 7 shows the power spectrum of the input signal $x(n)$, the signal at the microphone $u_1(n)$, and the recovered signal $\hat{x}(n)$. We see clearly that the effect of room's impulse response is completely removed from signal $u_1(n)$.

Figure 8(a) shows the AR polynomial and the estimated AR polynomial. The two are in good agreement. However, the two polynomials have different lengths. The



(a) AR polynomials



(b) AR polynomial spectrum

Fig. 8 (a) Coefficients of the AR polynomials. Circles represent the estimated AR parameters calculated with the proposed method. They are very close to the generating AR parameters of speech shown by the solid line. The actual length of the estimated AR polynomial is 601 taps. (b) Spectra of the generating and estimated AR processes.

length of the estimated AR polynomial is given by the size of matrix \mathbf{Q} which depends on the order of the transfer function. To obtain a precisely estimated AR process, we used Leverrier-Faddeev's algorithm [20] to calculate \mathbf{Q} 's characteristic polynomial.

4.2. Time-Variant AR Process

The second type of simulation confirms the applicability of the proposed method to speech signals.

4.2.1. Simulation conditions

In this case, the input signals were Japanese sentences, taken from ATR's speech database [21]. The room transfer functions are the same as those used for the time-invariant AR simulations.

The simulation conditions are summarized in Table 3.

4.2.2. Results

Table 4 shows the SDR's as defined in Eqs. (19) and (20) for three different sentences for both male and female speakers. The algorithm enables very precise estimation of the input signal since on average we obtained an SDR after processing of 106.5 dB for the female speakers and 108.4 dB for the male speakers.

Furthermore, for the first sentence pronounced by the female speaker, we plotted the power spectrum of the input

Table 3 Simulation conditions.

Length of impulse response	300 taps
Duration of speech signals	<5 s
Sampling frequency	16 kHz
Length of prediction filters	300 taps
Length of estimated AR process	601 taps

Table 4 Experimental results for speech signals.

	Female [dB]		Male [dB]	
	SDR_{Before}	SDR_{After}	SDR_{Before}	SDR_{After}
Sentence 1	-2.76	110.9	-2.75	104.1
Sentence 2	-2.71	107.2	-2.68	106.5
Sentence 3	-2.59	101.4	-2.88	114.5
Average	-2.69	106.5	-2.77	108.4

signal $x(n)$, the signal at the microphone, $u_1(n)$, and the recovered signal $\hat{x}(n)$, for two different time frames of 30 ms in Fig. 9. Even though the room impulse responses that we used are short, the distortion of the microphone signal is large as seen in Fig. 9. Our method works very well since the distortion is totally removed from the estimated signal.

5. DISCUSSION

5.1. Estimated AR Process

The proposed method was developed for the time invariant generating AR process as explained in Section 2. However, the simulation results show that the same algorithm can also be applied to such input signals originating from a time-variant AR process as speech. In both cases, the prediction filters and the estimated AR parameters are calculated for the whole signals and are static. Indeed, the room transfer functions are assumed to be time-invariant.

For time-invariant generating AR processes $1/a(z)$, the prediction filters deconvolve this process and the signal is whitened as shown in Eq. (13). The prediction filters thus intrinsically contain the effect of the generating AR polynomial $a(z)$. Here the static estimated AR polynomial $\hat{a}(z)$ corresponds simply to the static generating AR polynomial $a(z)$ as explained in Section 2 and shown in Fig. 8(a).

When the generating AR process is time-variant, the prediction filters also whiten the signal. However, static prediction filters cannot contain the dynamic generating AR process. In this case, the information contained by the prediction filters is expected to be an average AR process, equivalent to linear prediction coefficient calculated for a long time frame. The average AR polynomial is calculated by taking the characteristic polynomial of matrix \mathbf{Q} and then used to cancel out the whitening effect of the

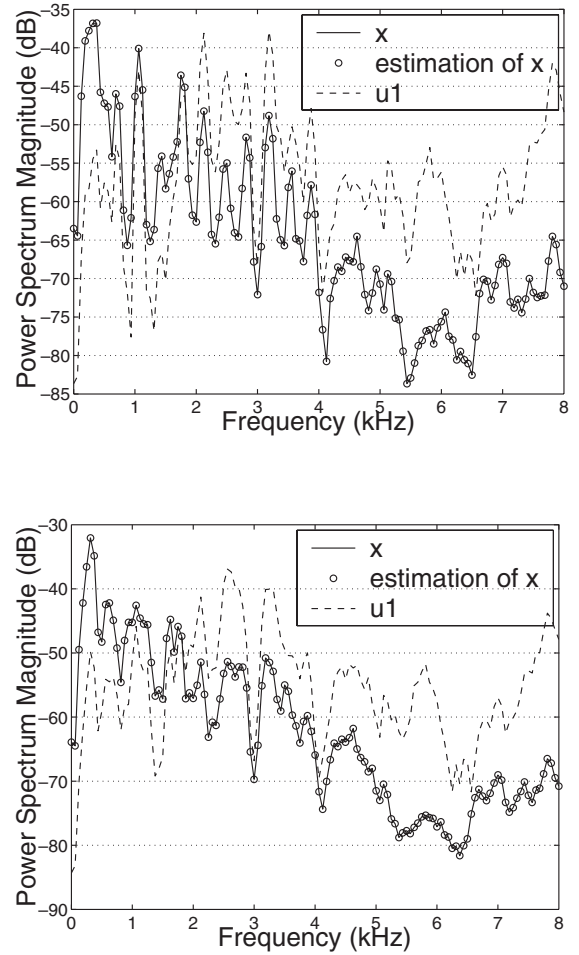


Fig. 9 Power spectrum of $x(n)$, $\hat{x}(n)$, and $u_1(n)$ for two different time frames of 30 ms. The dashed line represents the power spectrum of microphone signal, $u_1(n)$, which clearly suffers from room reverberation. The circles represent the power spectrum of the recovered signal, which precisely fits that of the input signal.

prediction filters. To illustrate this, Fig. 10 plots the estimated AR parameters and the linear prediction coefficients of the whole input signal (duration of around 5 s) representing speech signal average AR parameters. The figure shows that the two AR parameters are close, proving the validity of our interpretation.

5.2. Toward Semi-Batch Dereverberation

In our method, we use a speech time frame to calculate the prediction filters and the averaged AR process. The calculated filters $w_1(n)$ and $w_2(n)$ deconvolve the room impulse response but simultaneously degrade the characteristic of the signal because they contain the inverse of the average AR process. The estimated AR process is used to compensate the degradation of the filters. Combination of the filters and the estimated AR process assure the deconvolution of the transfer functions without degradation. Consequently, the same filters and estimated AR

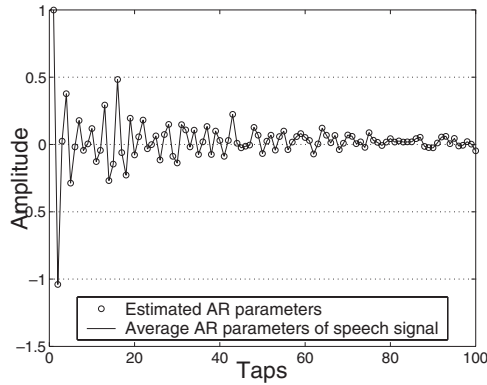


Fig. 10 AR polynomials. Circles represent the estimated AR parameters calculated by the proposed method. They are very close to the average AR parameters of speech signal shown by the solid line. Note that for clarity, only the first hundred AR parameters are plotted.

parameters can be used to dereverberate the following frames of the speech. In that sense, we believe that our dereverberation algorithm have potential for semi-batch implementations.

In the current simulations, we used one whole sentence to calculate the prediction filters and the averaged AR process. This corresponds to a time frame of around five seconds. We believe however that shorter time frames could be used.

5.3. Length of Prediction Filters

According to the theory, the prediction filters should be longer than the order of the transfer function. Indeed, matrix \mathbf{H} must have more columns than rows.

$$\begin{aligned} 2L &\geq m + L, \\ L &\geq m. \end{aligned} \quad (21)$$

Relation (21) gives us a threshold value for the length of the prediction filters. In the simulation, we fixed the length of the prediction filters manually knowing the order of the room transfer function. However, in a real case, we have no prior knowledge of the room transfer function and thus it could be difficult to determine the optimal length of the prediction filters. However, as shown in Fig. 11, precise knowledge of the order of the transfer function is not necessary since the dereverberation performance is stable for prediction filters longer than the threshold value.

6. CONCLUSION

We presented an algorithm for speech dereverberation that uses two-channel linear prediction. The method enables the precise recovery of speech signals suffering from room reverberation. In particular, the output signal is not whitened as found with traditional dereverberation

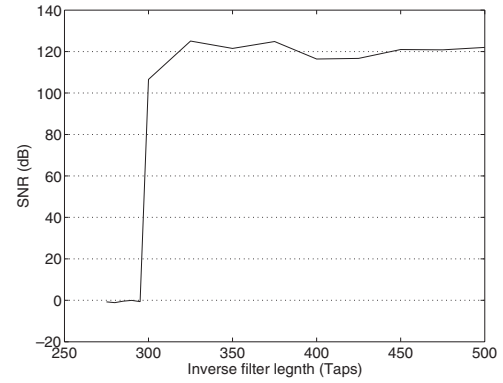


Fig. 11 SDR as a function of the prediction filter length. The performance of the dereverberation is plotted as a function of the length of the prediction filters. For a filter length longer than 300 taps (length of the transfer function), the performance is good and stable.

techniques. The excellent simulation results show the potential of the method and prove its solid theoretical background. However, the current method suffers from several limitations. First, we are currently limited to short room impulse responses. Indeed, a longer room impulse response would require longer prediction filters and thus a larger matrix \mathbf{Q} . In this case, computational time and accuracy would become an issue. One major reason for this problem may be that the two transfer functions have numerically common zeros. Moreover, the current results were obtained for a noise free environment, which is quite unrealistic. However, in theory if the hypotheses are satisfied, the method could be extended. Future work will thus consist in improving the method to cope with longer room impulse responses and noisy environments.

REFERENCES

- [1] T. Nakatani, M. Miyoshi and K. Kinoshita, "One microphone blind dereverberation based on quasi-periodicity of speech signals," in *Advances in Neural Information Processing Systems 16 (NIPS16)* (to appear), (MIT Press, Cambridge, Mass., 2004).
- [2] T. Nakatani and M. Miyoshi, "Blind dereverberation of single channel speech signal based on harmonic structure," *Proc. ICASSP '03*, Vol. 1, pp. 92–95 (2003).
- [3] B. Yegnanarayana and P. S. Murthy, "Enhancement of reverberant speech using LP residual signal," *IEEE Trans. Speech Audio Process.*, **8**, 267–281 (2000).
- [4] M. Unoki, M. Furukawa, K. Sakata and M. Akagi, "A method based on the MTF concept for dereverberating the power envelope from the reverberant signal," *Proc. ICASSP '03*, Vol. 1, pp. 840–843 (2003).
- [5] R. Roy and T. Kailath, "ESPRIT: Estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust. Speech Signal Process.*, **37**, 984–995 (1989).
- [6] O. R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, **34**, 276–280 (1986).
- [7] H. L. Van Trees, *Optimum Array Processing, Part IV of*

Detection, Estimation, and Modulation Theory (John Wiley & Sons, New York, 2002).

- [8] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust. Speech Signal Process.*, **36**, 145–152 (1988).
- [9] S. Amari, S. C. Douglas, A. Cichocki and H. H. Yang, "Multichannel blind deconvolution and equalization using the natural gradient," *Proc. IEEE Workshop on Signal Processing in Advances in Wireless Communications*, Paris, pp. 101–104, April (1997).
- [10] X. Sun and S. C. Douglas, "A natural gradient convolutive blind source separation algorithm for speech mixtures," *Proc. ICA '01*, pp. 59–64 (2001).
- [11] R. Aichner, S. Araki and S. Makino, "Time domain blind source separation of non-stationary convolved signals by utilizing geometric beamforming," *Proc. NNSP '02*, pp. 445–454 (2002).
- [12] S. Haykin, *Adaptive Filter Theory*, 3rd ed. (Prentice-Hall, Englewood Cliffs, N.J., 1996).
- [13] M. Miyoshi, "Estimating AR parameter-sets for linear-recurrent signals in convolutive mixtures," *Proc. ICA '03*, pp. 585–589 (2003).
- [14] T. Hikichi and M. Miyoshi, "Blind algorithm for calculating the common poles based on linear prediction," *Proc. ICASSP '04*, Vol. 4, pp. 89–92 (2004).
- [15] J. L. Stensby, "Analytical and computational methods," <http://www.eb.uah.edu/ece/courses/ee448/>.
- [16] T. Kailath, A. H. Sayed and B. Hassidi, *Linear Estimation* (Prentice Hall, Upper Saddle River, N.J., 2000).
- [17] D. A. Harville, *Matrix Algebra from a Statistician's Perspective* (Springer-Verlag, New York, 1997).
- [18] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, **65**, 943–950 (1979).
- [19] H. Kuttruff, *Room Acoustics*, 4th ed. (Spon Press, London, 2000).
- [20] S. H. Hou, "A simple proof of the Leverrier-Faddeev characteristic polynomial algorithm," *SIAM Rev.*, **40**, 706–709 (1998).
- [21] ATR International, "Speech database," <http://www.red.atr.co.jp/database.page/digdb.html>.



IEEE and ISCA.

Marc Delcroix was born in Brussels in 1980. He received the Master of Engineering from the Free University of Brussels and Ecole Centrale Paris in 2003. He is currently doing his Ph.D. at the Graduate School of Information Science and Technology of Hokkaido University. He is doing his research on speech dereverberation in collaboration with NTT Communication Science Laboratories. He is a member of



Information Science, Nagoya University. His research interests include physical modeling of musical instruments, room acoustic modeling, and signal processing for speech enhancement and dereverberation. He was honored to receive the Kiyoshi-Awaya Incentive Awards from the ASJ in 2000. He is a member of IEEE, ASA, ASJ, IEICE, and IPSJ.

Takafumi Hikichi was born in Nagoya, in 1970. He received his Bachelor and Master of Electrical Engineering degrees from Nagoya University in 1993 and 1995, respectively. In 1995, he joined the Basic Research Laboratories of NTT. He is currently working at the Signal Processing Research Group of the Communication Science Laboratories, NTT. He is a visiting associate professor of the Graduate School of



School of Information Science and Technology, Hokkaido University. He was honored to receive the 1988 IEEE ASSP Senior Awards, the 1989 ASJ Kiyoshi-Awaya Incentive Awards, and the 1990 ASJ Satoh Paper Awards. He also received the Ph. D. from Doshisha Univ. in 1991. He is a member of IEEE, AES, ASJ and IEICE.

Masato Miyoshi received the M.E. from Doshisha University in Kyoto in 1983. Since joining NTT as a researcher that year, he has been engaging on the research and development of acoustic signal processing technologies. Currently, he is a group leader of the Media Information Laboratory of NTT Communication Science Laboratories in Kyoto. He is also a visiting associate professor of the Graduate