REMOVING REFLECTED WAVES USING TEMPORAL AND SPECTAL SUBTRACTION WITHOUT PRIOR KNOWLEDGE

Kenko Ohta[*], Leandro Di Persia[†] and Masuzo Yanagida[*]

[*]Faculty of Engineering, Doshisha University
1-3, Tatara-Miyakodani, Kyo-Tanabe, Kyoto, 610-0394, Japan
E-mail:dtd0736@mail4.doshisha.ac.jp myanagid@mail.doshisha.ac.jp
[†]Universidad de Entre Rios
Parana,Entre Rios, Argentina
E-mail:ldpersia@ciudad.com.ar

***Abstract*- Proposed method is a method of removing reflected waves from a mixed wave consisting of a source signal and reflected waves. The method is a kind of waveform subtraction referring to auto-correlation functions(ACFs) of multi channel speech signals. We assume that a reflected wave has two parameters; path amplitude and delay time. The method estimates these parameters based on ACFs of signals received by microphones. The delay time of a particular path is estimated as the time lag that gives the maximum difference between the ACF of the channel in concern and the average ACF of the other rest channels. The delayed wave is subtracted from the received wave using an estimated delay only for vocalic segments and spectral subtraction is applied to non-speech segments. The rate of waveform subtraction, or the path amplitude of the reflected wave, is estimated by minimizing the difference between the ACF of the signal in concern and the average ACF of the rests at the time delay attributed to the reflection in concern. The proposed method can be realized without prior knowledge about room characteristics or the target speech. Speech recognition rate for the signals picked up with 3 microphones in a reverberant environment is improved about 8% employing the proposed method.**

I. INTRODUCTION

Recently, performance of automatic speech recognition has reached a practical level in case a close contact microphone is used in quiet environments. In real situations, however, recognition rate degrades miserably due to environmental noises, reflected waves and so forth. There are two approaches for improving the recognition rate in real environments. One is signal manipulation on the input signals and the other is introducing adaptation in recognition. Proposed method is an approach that classified to the former.

Inverse filtering of transfer function is generally employed for suppressing the effects of reflected waves [1], but this can't be applied to cases where the transfer functions from the source to microphones are not given or time variant. Several methods of spectral subtraction have been proposed to solve the problem [2].

Unoki et al. proposed a method based on Modulation Transfer Function (MTF), which doesn't need measuring transfer functions [3]. A source signal and transfer characteristics are modeled by MTF and are used to recover the power envelope from the reverberant speech. Although the method proposes a procedure for estimating the reverberation time and path amplitude, no proper method for determining them has been developed yet.

Takiguchi et al. proposed an adaptive recognition method as an alternative which doesn't need transfer functions [4]. The method can't give sufficient improvement in case the reverberant time is long even if acoustic models trained on site are employed [5]. Nakatani et al. proposed a method based on harmonic structure to solve the problem mentioned above [6]. This method develops the inverse filter based on a lot of reverberant speech data. However, a lot of time is required to develop the accurate inverse filter. Hence, it is difficult to put this method into practical use.

We have proposed a method for removing reflected waves based on ACFs without measuring the transfer function [7]. The method, however, cannot sufficiently improve the recognition rate because of the following three reasons: delay time estimation is unreliable, the received wave is used in place of the source wave in the iterative procedure, and fricative and nasal segments were over-suppressed.

These problems mentioned above are almost solved by the method proposed here. The performance of delay time estimation is improved by introducing majority decision on ACFs. Partitioning and classifying the non-vocalic segments into three categories, fricative, nasal and non-speech segments, solve the problem of over-suppression.

The current paper describes the effects of the proposed method on improving recognition rate.

II. Brief Description Of The Proposed Method

A. Basis of Removing Reflected Waves

In this paper, we assume quiet environments having only reverberations from walls, floors, ceiling and so forth but no noise source. The signal $r_i(t)$ received by microphone #*i* consists of waves from the source and $r_i(t)$ is represented by convolution of source signal $s(t)$ and the impulse response of a set of paths from the source to microphones. The signal $r_i(t)$ received by microphone #*i* is expressed by the following equation.

$$r_i(t) = s(t) * h_i(t) \qquad (1)$$

where * denotes convolution, $h_i(t)= \sum_{j=0}^{J} h_{ij}(t)$ and $h_{ij}(t)$ represents the impulse response of *j*th path from the source to microphone #*i* including the direct path(*j*=0).

The reflected signal along *j*th path is assumed to have amplitude of a constant rate α_{ij} (between -1 and 1) and a certain amount of delay time l_{ij}, that is, the reflected signal is expressed as $\alpha_{ij} s(t - l_{ij})$. Here, we assume that major reflection waves consist of single reflection waves, and multiple reflection waves decay much more compared with single reflection waves. That is, by removing single reflection waves, the effect of reflection waves can be reduced. Based on this assumption, an equation for estimating the direct wave $\alpha_{i0} s(t - l_{i0})$ is expressed as follow:

$$\alpha_{i0} s(t - l_{i0}) = r_i(t) - \sum_{j=1}^{J} \alpha_{ij} s(t - l_{ij}) \qquad (2)$$

where *J* denotes the effective number of reflection waves.

Now, $s(t-l_{ij})$ is replaced by its approximate wave because we don't know $s(t-l_{ij})$. We use the received signal that delay l_{ij}, $r(t - l_{ij})$ instead of $s(t-l_{ij})$.

$$\alpha_{i0} s(t - l_{i0}) \cong r_i(t) - \sum_{j=1}^{J} \alpha_{ij} r(t - l_{ij}) \qquad (3)$$

In discrete form we have

$$\alpha_{i0} s(k - l_{i0}) \cong r_i(k) - \sum_{j=1}^{J} \alpha_{ij} r_i(k - l_{ij}) \qquad (4)$$

where *k* denotes *k*th sampling point and *j* denotes the *j*th path. Here, we simply use l_{ij} to represent time delay counted by the sampling interval. The frequency characteristics of reflective surfaces are assumed to be flat [8]. Then, in iteration form we have

$$r_i^{(j)}(k) \cong r_i^{(j-1)}(k) - \alpha_{ij} r_i^{(j-1)}(k - l_{ij}) \quad j = 1 \cdots J$$
$$\hat{s}_i(k) \cong r_i^{(J)}(k) \qquad (5)$$

where $\alpha_{i0} = 1$ and $l_{i0} = 0$.

B. Estimation of the Delay Time Based on ACFs

In case a speech signal is picked up with a distant microphone in real environments, reflected signals jointly enter the received signal. Therefore, the ACF of the received signal would show a large value at the time lag corresponding to the time required for reflection than ACFs of the signals received by other microphone at the same time lag. The ACF of the source signal itself is not flat even if it is not affected by reflection. Hence, the ACF of the received signal, consisting of a source signal and reflected waves, would show local peaks at the time lags corresponding to both reflection and the local peaks of the ACF of the source signal itself. That is, even if the ACF of the received signal shows a local peak at a particular time lag, the time lag cannot be definitely determined as a delay due to reflection.

The proposed method solves the problem using a certain number of microphones. Figure 1 shows how to solve the problem for a case of using three microphones.

Assume that we want to remove a principal reflection from signal r_1, received by microphone #1. First, the average ACF $\overline{R}_1^*$ is calculated by averaging ACFs of the signals received by microphones #2 and #3. $\overline{R}_1^*$ is regarded to approximate the ACF R_s of the source signal. Furthermore, $\overline{R}_1^*$ is used later as the reference to estimate the path amplitude α_{1j} of the reflected wave along *j*th path.

Next, the difference between R_1, ACF of signal r_1, and $\overline{R}_1^*$ is calculated. The amount

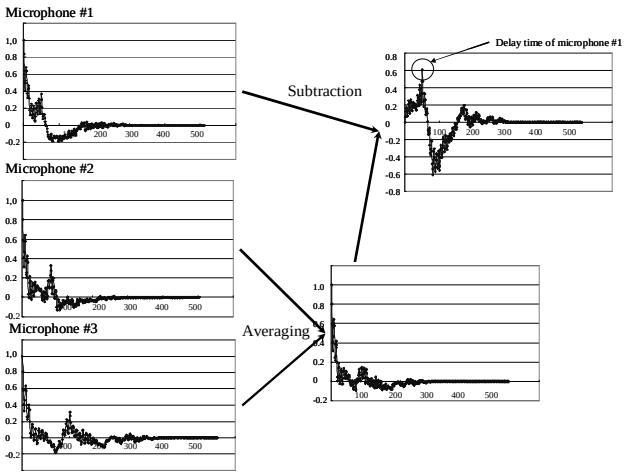

Fig. 1. Robust estimation of the delay.

of delay due to reflection would be dependent on the relative position of microphones and walls. So, the time lag at which $R_1 - \overline{R}_1$ gets large is assumed to be the time difference between the *j*th reflection path and the direct path. The delay time l_{1j} of *j*th path to microphone #1 is estimated by detecting the positive maximal value of the difference $R_1 - \overline{R}_1$.

C. Estimation of Path Amplitudes α_{ij}

The proposed method requires estimation of two parameters.

One is delay time and the other is path amplitude. Explained in this section is how to estimate the path amplitude. The delay estimated as in the subsection *B* and the average ACF are used to estimate the path amplitude. $\overline{R}_i^*$ is assumed to approximate R_s, so it is used as the reference in a recursive procedure to estimate path amplitudes. Figure 2 shows the algorithm of estimating path amplitudes.

First, the initial value for path amplitude α_{ij} is set to be zero. Then, an ACF of the estimated signal, $\hat{R}_i^{(j)}$, is calculated from $r_i^{(j)}$ that defined by (5).

Next, $\Delta R_i(l_{ij}) = \hat{R}_i^{(j)}(l_{ij}) - \overline{R}_i^{*(j-1)}(l_{ij})$ is calculated. Where $\hat{R}_i(l_{ij})$ and $\overline{R}_i^*(l_{ij})$ are the value of $\hat{R}_i(l_{ij})$ and $\overline{R}_i^*(l_{ij})$ at the delay time l_{ij} respectively. If the difference $\Delta R(\alpha_{ij})$ is less than 10^{-6}, then take the current α_{ij} to be the path amplitude for *j*th path, then go to the next step. Otherwise, the path amplitude is modified further. Figure 3 explains the reducing scheme on ACFs. Analysis of this estimation algorithm is shown in the next section.

D. Segmentation of the received signal

In our previous method, reflected waves are subtracted from the received wave in the time domain. The method, however, has difficulties that fricative and nasal portions are over-suppressed because of relatively low power compared to other sounds. Over reduction may occurs by waveform subtraction for removing waves in case received wave $r_i(k)$ is used instead of source wave $s_i(k)$ even if both delay time and path amplitude are exactly estimated. So, recognition rate for speech signals picked up in reverberant circumstances cannot be satisfactory even if our previous method is employed.

The proposed method detects fricative-like and nasal-like portions in input signals and leaves them as they are in order not to over-suppress them. They are detectable as fricatives and nasals show power concentration in the high and low frequency regions, respectively. To suppress reflected waves sufficiently, the input signal is segmented into speech or non-speech portions. Temporal waveform subtraction is employed only for speech portion except fricative-like and nasal-like portions, and spectral subtraction is used for the non-speech portions.

Explained below is how to partition a received signal into speech and non-speech portions and then, further segmentation of speech portion into fricative-like or nasal-like portions and the rest. First, a received signal is partitioned into the segment (A) whose amplitude is larger than a threshold that has been used to detect the utterance initial and the segment (B) whose amplitude is smaller than that. If the power spectrum of the segment (B) shows maximum value at frequency over 4kHz (where the sampling rate is 16ksamples/sec), the segment is classified into a fricative-like segment, or segment (C). Then, a segment having its spectral peak at a low frequency region below 1kHz with fundamental frequency less than 400Hz is regarded as a nasal-like portion, or segment (D), because nasals are periodic and have power concentration in a low frequency region. The rest of the input signal is regarded as non-speech portion, or segment (E). Following the partitioning procedure described above, a received signal is classified into segments shown in Fig. 4.

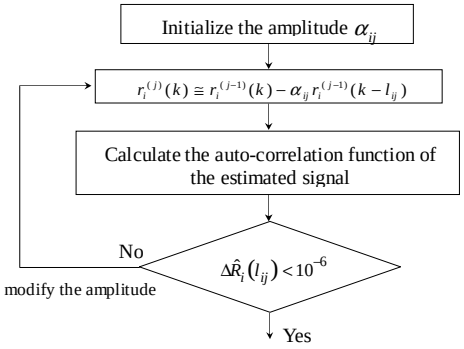

Fig. 2. An algorithm of estimating the path amplitude.

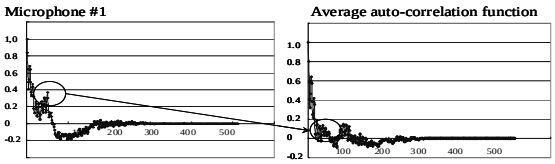

Fig. 3. Explanatory sketch of processing on ACFs.

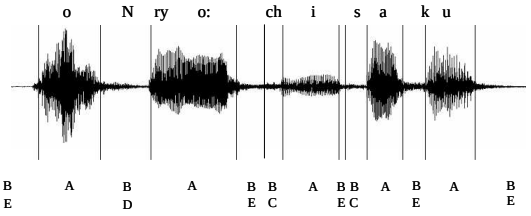

Fig. 4. Segmentation of a received signal.

E. Processing scheme

Figure 5 shows the processing scheme of the proposed method. First, the starting point of speech on each microphone is detected by a double threshold method. Then, R_i, ACF of the signal received by microphone #*i* is calculated. Next, l_{ij}, the delay time of *j*th path is estimated as the time lag τ that gives the maximum value of $\Delta R_i(l_{ij}) = \hat{R}_i^{(j)}(l_{ij}) - \overline{R}_i^{*(j-1)}(l_{ij})$. Then, received signal is segmented into large amplitude portion (A) and small amplitude portion (B), and then, (B) is classified into fricative-like portion (C), nasal-like portion (D) and non-speech portion (E). Finally, the path amplitude α_{ij} is chosen to be the value that minimizes the difference $\Delta R(\alpha_{ij})$. Then the supposed reflection wave is subtracted according to (5). Proceed by $i=i+1$ until $i=I$, then $j=j+1$ until $j=J$.

III. EXPERIMENTS

A. Experimental set-ups

To examine the performance of the proposed method, a living room is used as a reverberant room having reverberation time 440ms measured with white noise. One loudspeaker and three microphones are used in this experiment. Figure 6 shows the configuration of the loudspeaker and microphones in the living room.

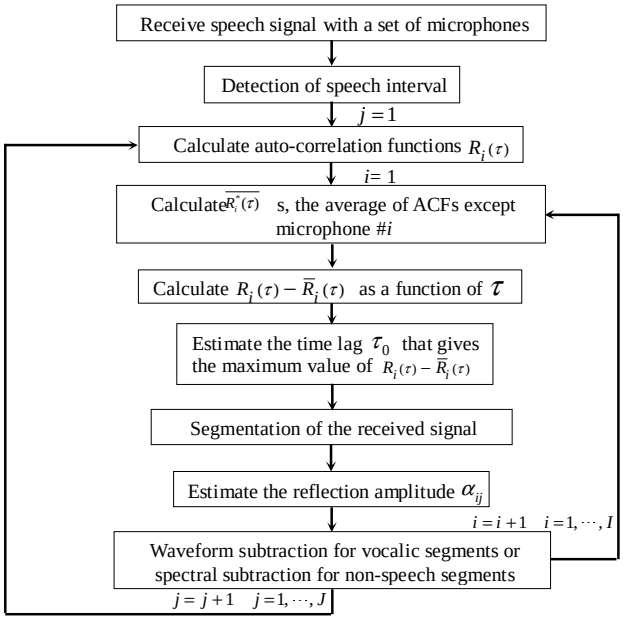

Fig. 5. A processing scheme of removing reflected waves.
(I denotes the number of microphones and J denotes the number of reflected waves to be removed)

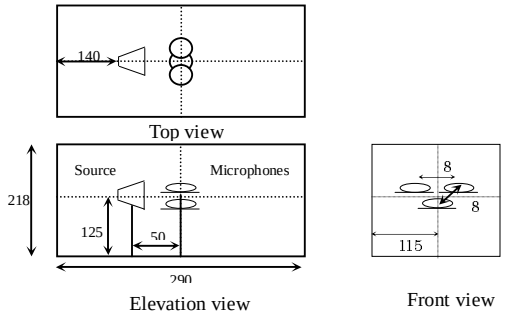

Fig. 6. Positions of a loudspeaker and microphones in a living room.

B. Experimental conditions

Source speech is recorded by a close contact microphone in a sound proof chamber. The speech samples are 255 Japanese words uttered by a female and two males. Contents of the speech data are Japanese commands for controlling TV sets. For example, “terebi oN(TV on),” ”chaN-neru ichi(channel one)” and so on. Table 1 shows specification of data acquisition and the language: the vocabulary size and the number of grammar rules for speech recognition. “Julian” is employed as a decoder [9].

C. Results

We compare the spectrum of the received signal $r_i(k)$ and that of the estimated direct signal $\hat{s}_i(k)$ to evaluate improvements in sound quality and recognition rate.

Figure 7 shows spectrograms of a source, received and estimated signals. As shown in the figures, reflected waves are effectively removed and fricative and nasal sounds well remain unover-suppressed. As the result of applying the proposed method

TABLE I
Specifications of data acquisition and language

Sampling rate	16ksamples/sec
Quantization	16bits

Vocabulary size	99
Grammar rules	13

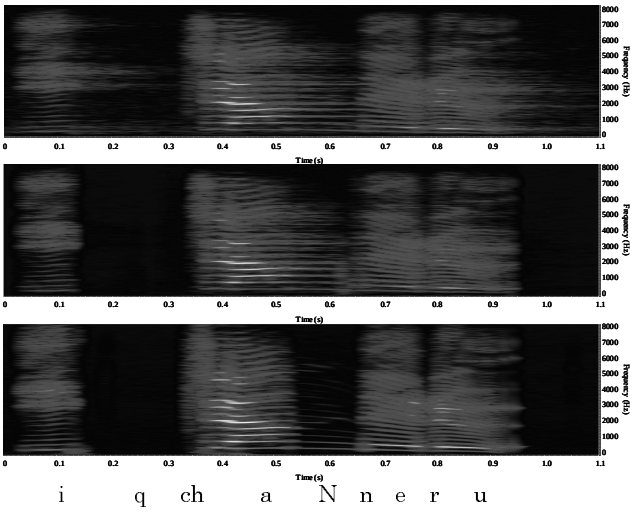

Fig. 7. Comparison of spectrograms.
(Top: received signal in the living room, center: estimated signal, bottom: source signal)

to reverberant signals, the source signal is approximately recovered from the reverberant signals. We can also recognize that reverberation in non-speech segments is sufficiently removed. Listening to the reflection-removed signal, slight perceptual improvement of sound quality was recognized.

Table 2 compares the recognition rates by the previous method and those by the proposed method. The recognition rate of all microphones is improved by applying the proposed method except microphone #1 for M3 and microphone #2 for F1 and M1. Table 3 shows the recognition rate introducing majority decision among the recognition results of three microphones. As the result, the recognition rate is not much improved with the previous method. While, the recognition rate is improved about 8% by applying the proposed method to reverberant signals. The result of sign test between recognition rates of received signals and recovered signals for speaker M2 shows significant difference at 1% level of significance.

IV. DISCUSSIONS

The result of sign test using all 255 data between recognition rates of received and estimated signals shows significant difference at 1% level of significance. So, it can be said that the proposed method well removes reflected waves and improves the recognition rate.

To improve the recognition rate much more, it would be necessary to refine the proposed method. For example, in proposed method, we do not consider the frequency characteristics of reflective surfaces and spectrum subtraction is not applied to fricative-like or nasal-like portions. So, reflected waves are not reduction sufficiently.

We plan to apply the proposed method to preprocessing for

TABLE II
Comparison of recognition rates(%) by three microphones for three speakers

	Microphone #1			Microphone #2			Microphone #3		
	baseline	previous	proposed	baseline	previous	proposed	baseline	previous	proposed
M1	88	90	90	88	90	86	82	88	90
M2	61	76	76	67	82	88	61	78	82
M3	80		75	80		80	76		80
M4	90		92	90		96	86		92
F1	82	84	88	82	80	78	63	78	84

TABLE III
Comparison of recognition rates(%) recalculated employing majority decision among three microphone(* < 0.05, **<0.01)

	baseline	previous	proposed
M1	84	90	90
M2	67	84*	88**
M3	82		82
M4	90		96
F1	76	84	86

ICA (Independent Component Analysis). In reverberant environments, the performance of source signal separation using ICA is degraded by reflected signals, so it is difficult to use ICA in real environments. By using the proposed method, expected effect is improvement of performance of source sound separation by ICA in real environments.

V. CONCLUSION

Proposed is a method to remove reflected waves from a signal received in a reverberant room. The proposed method solves some problems the previous method had. Problems almost solved by the proposed method are: unreliability in delay time estimation, approximation degree of supposed source waves in the processing, and over-suppression for fricative and nasal segments. In the proposed method, the performance of the delay time estimation is improved by majority decision on ACFs. Furthermore, to solve the problem of over-suppression for fricative and nasal segments, received signal is segmented into large amplitude portion (A) and small amplitude portion (B), and then, (B) is classified into fricative-like portion (C), nasal-like portion (D) and non-speech portion (E). As the result of applying the proposed method, the recognition rate is improved by about 8%.

CONVERGENCE OF THE ALGORITHM FOR ESTIMATING THE PATH AMPLITUDE

A. Characteristic Properties of $\Delta R(\alpha_{ij})$

Shown in this section is the convergence property of the algorithm for estimating the path amplitude. The goal of this section is to prove that the solution, or the estimated path amplitude, of equation $\Delta R_i(l_{ij}) = 0$,exists in the interval [-1,1] and the algorithm described in the previous section converges to the solution.

The first step of the algorithm is to estimate $\hat{r}_i(k)$, the expected *k*th sampled value for microphone *i* having been principal reverberations removed. Then, its ACF $\hat{R}_i(l_{ij})$ at delay l_{ij} is calculated, where $\hat{R}_i(l_{ij})$ is normalized by ACF at null delay.

Next, calculated is the difference $\Delta R_i(l_{ij})$ between $\bar{R}_i^*(l_{ij})$, the expected average for $R_i(l_{ij})$, and $\hat{R}_i(l_{ij})$, the estimated ACF of the signal received by microphone #*i*, expressed as follows:

$$\hat{R}_i(l_{ij}) - \bar{R}_i^*(l_{ij}) = \frac{\sum_{k=0}^{N-1}(r_i(k)-\alpha_{ij}r_i(k-l_{ij}))(r_i(k+l_{ij})-\alpha_{ij}r_i(k))}{\sum_{k=0}^{N-1}(r_i(k)-\alpha_{ij}r_i(k-l_{ij}))^2} - \bar{R}_i^*(l_{ij}) \quad (7)$$

$$= \frac{\alpha_{ij}^2\sum_{k=0}^{N-2l_{ij}-1} r_i(k)r_i(k+l_{ij}) - \alpha_{ij}\left(\sum_{k=0}^{N-l_{ij}-1} r_i^2(k) + \sum_{k=0}^{N-2l_{ij}-1} r_i(k)r_i(k+2l_{ij}) + \sum_{k=0}^{l_{ij}-1} r_i(k)r_i(k+N-2l_{ij})\right) + R_i(l_{ij})}{\alpha_{ij}^2\sum_{k=0}^{N-l_{ij}-1} r_i^2(k) - 2\alpha_{ij}\sum_{k=0}^{N-l_{ij}-1} r_i(k)r_i(k+l_{ij}) + R_i(0)} - \bar{R}_i^*(l_{ij}) \quad (8)$$

$$= \left(\frac{\frac{1}{R_i(0)}\left\{\alpha_{ij}^2\sum_{k=0}^{N-2l_{ij}-1} r_i(k)r_i(k+l_{ij}) - \alpha_{ij}\left(\sum_{k=0}^{N-l_{ij}-1} r_i^2(k) + \sum_{k=0}^{N-2l_{ij}-1} r_i(k)r_i(k+2l_{ij}) + \sum_{k=0}^{l_{ij}-1} r_i(k)r_i(k+N-2l_{ij})\right) + R_i(l_{ij})\right\}}{\frac{1}{R_i(0)}\left\{\alpha_{ij}^2\sum_{k=0}^{N-l_{ij}-1} r_i^2(k) - 2\alpha_{ij}\sum_{k=0}^{N-l_{ij}-1} r_i(k)r_i(k+l_{ij}) + R_i(0)\right\}}\right) - \bar{R}_i^*(l_{ij}) \quad (9)$$

where the average ACF $\bar{R}_i^*(l_{ij})$ is independent of α_{ij}, so it can be regarded as a constant, and (9) is normalized by the ACF at null delay to avoid truncation errors, which may occur by calculation as there is the large difference between absolute values *A* and others. Now, each normalized component in (9) is replaced with notation defined as follows:

$$A = \frac{1}{R_i(0)}\sum_{k=0}^{N-l_{ij}-1} r_i^2(k) \quad B = \frac{1}{R_i(0)}\sum_{k=0}^{N-2l_{ij}-1} r_i(k)r_i(k+l_{ij}) \quad C = \frac{1}{R_{i0}}\sum_{k=0}^{N-l_{ij}-1} r_i(k)r_i(k+l_{ij})$$

$$D = \frac{1}{R_i(0)}\sum_{k=0}^{N-2l_{ij}-1} r_i(k)r_i(k+2l_{ij}) \quad E = \frac{1}{R_i(0)}\sum_{k=0}^{l_{ij}-1} r_i(k)r_i(k+N-2l_{ij}) \quad R_i(l_{ij}/0) = \frac{R_i(l_{ij})}{R_i(0)}$$

where these are constants and satisfy

$$-1 \le A, B, C, D, E, R_i(l_{ij}/0)\bar{R}_i^*(l_{ij}) \le 1$$

Substituting these constants into (9), we can obtain a simple form of (9).

$$\Delta R(\alpha_{ij}) = \frac{B\alpha_{ij}^2 - (A+D+E)\alpha_{ij} + R_i(l_{ij}/0)}{A\alpha_{ij}^2 - 2C\alpha_{ij} + 1} - \bar{R}_i^*(l_{ij}) \quad (10)$$

Here we introduce plausible conditions to prove that

there is a solution α_{ij} for $\Delta R(\alpha_{ij})=0$, satisfying $-1\le\alpha_{ij}\le 1$.

$A\ge 0$ as A is the squared sum of the input signal. Comparing the definitions of A and C, we can see that they are product sums over the same interval, where A is a squared sum, while C is not. Based on the property of the ACF, we have a relation between A and C as $-1<-A<C<A<1$. Similarly, we have a relation between A and $D+E$ as $-1<-A<D+E<A<1$, or $0<A+D+E$. These inequalities yield $C^2<A$. Now, the denominator of (10) is a convex function to the downside as quadratic coefficient A is positive. The discriminant of the denominator of (10) $Dis=C^2-A$ is negative because $C^2<A$. So, the denominator of (10) has no solution taking positive values for any α_{ij}. It can be concluded that the function $\Delta R(\alpha_{ij})$ is continuous over all range of the value of α_{ij}.

In the case of $\alpha_{ij}=0$, $\Delta R(0)$ becomes $R_i(l_{ij})-\bar{R}_i^*(l_{ij})$, and is positive because $R_i(l_{ij})$ is larger than $\bar{R}_i^*(l_{ij})$. l_{ij} has been chosen at which the difference between the ACF of microphone #1 and the average ACF is to be a positive maximal value. As α_{ij} goes to $\pm\infty$, $\Delta R(\alpha_{ij})$ asymptotically reaches $B/A-\bar{R}_i^*(l_{ij})$.

B. Features of $\Delta R(\alpha_{ij})$ for between –1 and 1

Assuming that target sources are located apart from walls, l_{ij}, the time delay that yields the maximum difference between the ACF of the signal received by microphone #*i* and the average ACF, is sufficiently larger than unity. This assumption leads to inequalities $A\gg D+E$, $A\gg B$, $A\gg R_i(l_{ij})$ and $A\gg C$. Here, α_{ij} which gives the extremum of the function $\Delta R(\alpha_{ij})$ is calculated as

$$\alpha_{ij}=\frac{B-AR_i(l_{ij})\pm\sqrt{(B-AR_i(l_{ij}))^2-\{A(A+D+E)-2BC\}\{2CR_i(l_{ij})-(A+D+E)\}}}{A(A+D+E)-2BC}$$

Dividing the numerator and denominator of α_{ij} with A, and α_{ij} can be simplified as follows considering that inequalities $A\gg B$ etc. lead to approximate expressions $B/A\cong 0$ etc.

$$\alpha_{ij}\cong\frac{-R_i(l_{ij})\pm\sqrt{R_i^2(l_{ij})+A}}{A}$$

The discriminant is positive because $A\gg R_i(l_{ij})$ and $A\ge 0$, so, $\Delta R(\alpha_{ij})$ has two extremums at

$$\frac{-R_i(l_{ij})-\sqrt{R_i^2(l_{ij})+A}}{A}<0 \text{ and } \frac{-R_i(l_{ij})+\sqrt{R_i^2(l_{ij})+A}}{A}>0$$

The sign of $\Delta R(\alpha_{ij})$ at the extremum is determined by the relative values of α_{ij} which give extremums of the numerator or the denominator of (10). These extremums are given as

$$\text{numerator}=B\{(\alpha_{ij}-\frac{A+D+E}{2B})^2-(\frac{A+D+E}{2B})^2+\frac{R_i(l_{ij})}{B}\}$$

$$\text{denominator}=A\{(\alpha_{ij}-\frac{C}{A})^2-(\frac{C}{A})^2+\frac{1}{A}\}$$

Under current situation, these α_{ij} which give extremums of the denominator and numerator satisfy $|C/A|\cong 0$ and $|(A+D+E)/2B|>1$ respectively. Furthermore, in case of $B>0$, $A+D+E/2B$ becomes larger than unity and C/A. While, in case of $B<0$, $A+D+E/2B$ becomes smaller than negative unity, because of $A+D+E>0$, and C/A. So, in generally the function $\Delta R(\alpha_{ij})$ is plotted as (a):

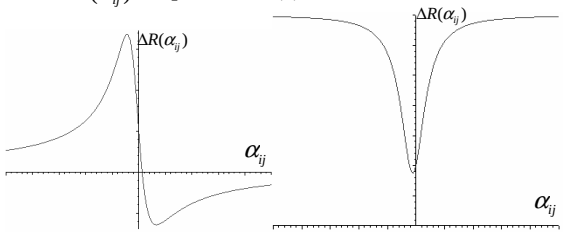

Fig. 8. Function $\Delta R(\alpha_{ij})$ where (a) is in general case and (b) is in special case.

However, in case the extremum of the denominator corresponds to that of the numerator, the function $\Delta R(\alpha_{ij})$ is plotted as (b), where (b) is concave, but the function $\Delta R(\alpha_{ij})$ possibly become convex.

C. Existence of a solution for α_{ij} between –1 and 1

Here, we will confirm that the product $\Delta R(-1)\Delta R(1)$ is negative in order to show that a solution of the equation $\Delta R(\alpha_{ij})=0$ exists within the interval [-1,1], under situation that $\Delta R(\alpha_{ij})$ is continuous for $-\infty\le\alpha_{ij}\le\infty$. The product $\Delta R(-1)\Delta R(1)$ yields

$$\Delta R(-1)\Delta R(1)=\left(\frac{B+(A+D+E)+R_i(l_{ij})}{A+2C+1}-\bar{R}_i^*(l_{ij})\right)\left(\frac{B-(A+D+E)+R_i(l_{ij})}{A-2C+1}-\bar{R}_i^*(l_{ij})\right)$$

Dividing denominators and numerators of the fraction part of the above equation by A, we get the result expressed as

$$\Delta R(-1)\Delta R(1)=\left(\frac{\frac{B}{A}+(1+\frac{D}{A}+\frac{E}{A})+\frac{R_i(l_{ij})}{A}}{1+\frac{2C}{A}+\frac{1}{A}}-\bar{R}_i^*(l_{ij})\right)\left(\frac{\frac{B}{A}-(1+\frac{D}{A}+\frac{E}{A})+\frac{R_i(l_{ij})}{A}}{1-\frac{2C}{A}+\frac{1}{A}}-\bar{R}_i^*(l_{ij})\right)$$
$$\cong\left(0.5-\bar{R}_i^*(l_{ij})\right)\left(-0.5-\bar{R}_i^*(l_{ij})\right)$$

where fractions B/A, C/A, $(D+E)/A$ and $R_i(l_{ij})/A$ are approximately null and $1/A$ is approximately unity. Let us consider that the product $\Delta R(-1)\Delta R(1)$ is dependent on the average ACF $\bar{R}_i^*(l_{ij})$. The product $\Delta R(-1)\Delta R(1)$ is the quadratic function of the average ACF $\bar{R}_i^*(l_{ij})$. Furthermore, the quadratic coefficient is positive, so this function is concave. Hence, if the average ACF $\bar{R}_i^*(l_{ij})$ satisfies $-0.5<\bar{R}_i^*(l_{ij})<0.5$, the product $\Delta R(-1)\Delta R(1)$ is negative.

On the other side, it is obvious that $A\gg R_i(l_{ij})$ and

$\overline{R}_i^*(l_{ij}) - R_i(l/0) < 0$. So, the average ACF $\overline{R}_i^*(l_{ij})$ is approximately null at $\tau = l_{ij}$ and then it satisfies i $-0.5 < \overline{R}_i^*(l_{ij}) < 0.5$. It can be concluded that a solution of the equation $\Delta R(\alpha_{ij}) = 0$ exists within the interval [-1,1].

D. Convergence of the proposed algorism

Let us verify that the estimation algorithm converges to the solution of the equation $\Delta R(\alpha_{ij}) = 0$ that exists within the interval [-1,1]. Here, we consider the initial value for α_{ij}. It is clear that the extremum of the function $\Delta R(\alpha_{ij})$ exist only one within the interval [-1,1]. So, the algorism can be converged easily by using Newton-Raophson method and choosing the initial value as zero.

REFERENCES

[1] Miyoshi, M., and Kaneda, Y., "Inverse filtering of room acoustics," *IEEE Trans. ASSP*, Vol. 36, No. 2, pp. 145-152, 1988.

[2] Nakajima, H., Tsuru, H., and Tohyama, M., "Reverberation suppression by time frequency analysis," *Proc. of ASJ*, 1-4-5, pp. 483-484, Sep., 1997.

[3] Unoki, M., Sakata, K., and Akagi, M., "A speech dereverberation model based on the Modulation Transfer Function," *Proc. of ASJ*, 1-5-2, pp. 481-482,Sep. 2003.

[4] 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.

[5] 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, Sep., 2002.

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

[7] Ohta, K., and Yanagida, M., "Removing reflected waves using delay time detected by majority decision on auto-correlation functions," *Proc. of FIT2004*, vol. 2, pp. 365-366, 2004.

[8] Nakajima, H., Tsuru, H., and Tohyama, M., "Reverberation suppression under frequency-dependent reverberation time condition," *Proc. of ASJ*, 1-Q-4, pp. 567-568, Mar, 1998.

[9] http://julius.sourceforge.jp/