Proposed is a method for removing reflected waves from a mixed wave consisting of a direct signal and reflected waves. The method is a kind of waveform subtraction referring to auto-correlation functions (ACFs) of multi-channel speech signals. A reflected wave is assumed to have 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, while fricative-like and nasal-like segments are left as they are, and conventional spectral subtraction is applied to the rest of the input speech. The rate of waveform subtraction, or the path amplitude of a 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 path in concern. The proposed method can be realized without a priori knowledge about room characteristics and the target speech. Speech recognition rate for the signals picked up with 3 microphones in a reverberant environment is improved about 9% employing the proposed method. Convergence of the algorithm for estimating the path amplitude is shown in Appendix.

1. 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 falls 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 the recognition process. The proposed method is an approach classified to the former.

Inverse filtering of the source-microphone transfer function is widely 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 obtainable 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 require measuring transfer functions [3], but a source signal and transfer characteristics are modeled by MTF for recovering the power envelope from the reverberant speech. Although the method requires a procedure for estimating the reverberation time and path amplitude, no definite method for determining them has been developed yet.

Takiguchi et al. proposed an adaptive recognition method as an alternative which doesn’t require transfer functions [4]. The method, however, 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]. Their method constitutes an inverse filter based on a large number of reverberant speech data. The method, however, takes a lot of time to develop the accurate inverse filter. Hence, it is difficult to put their method into practical use.

We have proposed a method to remove reflected waves by subtracting supposed reflective components on ACFs without measuring the transfer function [7]. The method, however, cannot sufficiently improve recognition rate because of the following three reasons: (1) delay time estimation is unreliable, (2) the received wave is used in place of the source wave in the iterative procedure, and (3) over-suppression occurs on fricative and nasal segments are over-suppressed.

The above-mentioned problems of our previous method will be almost cleared by the method proposed here. Accuracy of delay time estimation is improved by introducing majority decision on ACFs. Partitioning and classifying the non-vocalic segments into three categories, i.e., fricative, nasal and non-speech segments, solve the problem of over-suppression.

The current paper describes details of the method and the effects of the method on recognition rate.

2. Description of the proposed method

2.1 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 sources. The signal $r_i(t)$ received by microphone $#i$ consists of many waves from the source including the direct wave and reflected waves 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 the microphone. The signal \( r_i(t) \) received by microphone \#i can be expressed by the following equation.

\[
r_i(t) = s(t) * h_i(t)
\]

where \(*\) denotes convolution, \( h_i(t) = \sum_{j=0}^{\infty} h_{ij}(t) \) 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 \( \alpha_j \) (between -1 and 1) and a certain amount of delay time \( l_{ij} \) that is, the reflected signal received by microphone \#i from the source via \( j \)th path is expressed as \( \alpha_j s(t - l_{ij}) \). Here, we assume that major reflection waves consist of those of single reflection, and waves of multiple reflection decay much more compared with single reflection waves. That is, the effect of reflection waves is expected to be almost reduced by removing single reflection waves. Based on this assumption, an equation for estimating the direct wave \( \alpha_{io} s(t - l_{io}) \) is expressed as:

\[
\alpha_{io} s(t - l_{io}) = r_i(t) - \sum_{j=1}^{J} \alpha_j s(t - l_{ij})
\]

where \( J \) denotes the effective number of reflected waves.

Now, \( s(t - l_{io}) \) is replaced by its approximate wave because we don’t know \( s(t - l_{io}) \). We use the received signal with delay \( l_{io} \), \( r(t - l_{io}) \) instead of \( s(t - l_{io}) \).

\[
\alpha_{io} s(t - l_{io}) \cong r_i(t) - \sum_{j=1}^{J} \alpha_j r(t - l_{ij})
\]

In discrete form we have

\[
\alpha_{io} s(k - l_{io}) \cong r_i(k) - \sum_{j=1}^{J} \alpha_j r(k - l_{ij})
\]

where \( k \) denotes \( k \)th sampling point and \( j \) denotes the path ID for microphone \#i. Here, we simply use \( l_{io} \) to represent time delay counted by the sampling interval. The frequency characteristics of reflective surfaces are assumed to be flat [8]. Then, in iterative form we have

\[
r_i^{(j)}(k) \cong r_i^{(j-1)}(k - \alpha_j r_i^{(j-1)}(k - l_{ij})) \quad j = 1 \ldots J
\]

\[
\hat{s}_i(k) \cong r_i^{(j)}(k)
\]

where \( \alpha_{io} = 1 \) and \( l_{io} = 0 \).

2.2 Estimation of the delay time based on AFCs

In case a speech signal is picked up with a distant microphone in a real environment, reflected signals jointly constitute the received signal. So, 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 microphones 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 the direct wave 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 presumed 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 three microphones.

Assume that we want to remove a principal reflection from signal \( r_1(t) \) received by microphone \#1. First, the average ACF \( \bar{R}_1 \) is calculated by averaging ACFs of the signals received by microphones other than \#1, i.e. microphones \#2 and \#3 for this case. \( \bar{R}_1 \) is regarded to approximate \( R_s \), the ACF of the source signal. Furthermore, \( \bar{R}_1^* \) is used later as the reference for estimating \( \alpha_{io} \), the path amplitude of the reflected wave along the \( j \)th path. Next, the difference between \( R_1 \), ACF of signal \( r_1 \), and \( \bar{R}_1^* \) is calculated. The amount of delay due to reflection would be dependent on the relative position of microphones and walls. So, the time lag at which \( R_1 - \bar{R}_1^* \) gets large is thought to be the time difference between the \( j \)th reflection path and the direct path. The delay time \( l_{ij} \) of \( j \)th path to microphone \#1 is estimated by detecting the positive maximal value of the difference \( R_1 - \bar{R}_1^* \).

2.3 Estimation of path amplitude \( \alpha_y \)

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 \( \alpha_y \). The delay estimated as described in subsection B and the average ACF are used to estimate the path amplitude. \( \bar{R}_1^* \) is assumed to approximate \( R_s \), so it is used as the reference in the
recursive procedure, which will be explained soon, to estimate the path amplitude by minimizing the difference between $R_i$ and $\overline{R}_i$ at time delay $l_j$. Figure 2 shows the algorithm for estimating the path amplitude.

First, the initial value for path amplitude $\alpha_j$ is set to be null. Then, the ACF of the estimated signal, $\hat{r}_s^\alpha(t)$, is calculated based on $r_s^\alpha$ expressed by (5).

Next, $\Delta R_i(l_j) = \hat{r}_s^\alpha(l_j) - \overline{r}_s^\alpha(l_j)$ is calculated, where $\hat{r}_s(l_j)$ and $\overline{r}_s(l_j)$ are the values of $\hat{r}_s$ and $\overline{r}_s$, respectively, at the delay time $l_j$. If the difference $\Delta R_i(\alpha_j)$ is less than $10^{-6}$, then take the current $\alpha_j$ to be the path amplitude for $j$th path, then it goes to the next step. Otherwise, the path amplitude is increased further. Figure 3 explains the reducing scheme on ACFs. Analysis of this estimation algorithm is shown in Appendix.

2.4 Segmentation of the received signal

In our previous method [7], supposed reflection waves are subtracted from the received wave in the time domain. The method, however, has difficulties that over-subtraction occurs in fricative and nasal portions because of relatively low power compared with other phonemes. Over reduction may occur by waveform subtraction for removing reflected waves in case received wave $r(k)$ is used instead of source wave $s(k)$ even if both delay time and path amplitude are properly estimated. So, recognition rate for speech signals picked up in reverberant circumstances cannot be satisfactory in case our previous method is employed.

The proposed method detects fricative-like and nasal-like portions in input signals and leaves them as they are to avoid over-subtraction. Fricatives and nasals are detectable as they show power concentration in the high and low frequency regions, respectively. To subtract reflected waves properly, the input signal is segmented into speech or non-speech portions. Temporal waveform subtraction should be employed only for speech portion except fricative-like or nasal-like portions, and conventional spectral subtraction is preferably employed 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 a type of segments (class “A”) whose short-term powers are larger than the threshold employed to detect utterance initials, and another type of segments (class “B”) whose short-term powers are smaller than that. If the power spectrum of a segment of class “B” shows the maximum value at frequency beyond 4kHz (where the sampling rate is 16kHz/s), the segment is classified into class “C”, a fricative-like portion. Then, a segment having its spectral peak in a low frequency region below 1kHz with fundamental frequency less than 400Hz is regarded as a segment of class “D”, a nasal-like portion, because nasals are periodic and have power concentration in low frequency regions. The rest of the input signal is regarded as non-speech portion, or class “E”. Following the partitioning procedure described above, a received signal is partitioned into segments and each segment is classified into one of classes A,C,D or E as shown in Fig. 4.

2.5 The processing scheme

Figure 5 shows the processing scheme of the proposed method. First, utterance initials of speech signals are detected by a double threshold method. Then, $R_i$, the ACF of the signal received by microphone #i is calculated. Next, $l_j$, the delay time of $j$th path is estimated as the time lag $\tau$ that gives the maximum value of $\Delta R_i(l_j) = \hat{r}_s^\alpha(l_j) - \overline{r}_s^\alpha(l_j)$. Then, received signal is partitioned into segments of either large amplitude portion “A” or small amplitude portion “B”, and then, “B” is further classified into fricative-like portion “C”, nasal-like portion “D” or non-speech portion “E”. Finally, the path amplitude $\alpha_j$ is obtained as the value that minimizes the difference $\Delta R(\alpha_j)$. Then the supposed reflection wave is
3. Experiments

3.1 Experimental set-ups

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

3.2 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 commands in Japanese for controlling TV sets. For example, “terebi oN(TV on),” “chaN-neru ichi (channel one)”, “nyuusu(news)” and so on. The vocabulary size is 99 and the number of grammar rules is 13. Speech sounds are sampled at 16ksamples/sec with 16bit accuracy. “Julian” is employed as the decoder [9].

3.3 Results

The spectrum of the received signal $r_f(k)$ and that of the estimated direct signal $\hat{s}_d(k)$ are compared 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-like sounds well remain not over-subtracted. As the result of applying the proposed method 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 signals, slight improvement in sound quality was perceived.

Table 1 shows the recognition rate introducing majority decision among the recognition results of three microphones. Though the recognition rate is not much improved compared with the previous method, it is improved about 9% 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% hazard rate.
4. Discussions

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

The current experiment was carried out on the condition that a speaker does not move. So, it is problematic whether a system can manage movements of speakers or not. The proposed method, however, estimates a reflection wave one by one. So, it is expected to be robust against movement of speakers.

To improve the recognition rate further, it would be necessary to refine the method. For example, in the proposed method, we do not take the frequency characteristics of reflective surfaces into consideration, and furthermore, spectral subtraction is not applied to fricative-like or nasal-like portions. There must be further reflected waves as the system deals with only one-time reflection.

We plan to apply the proposed method to preprocessing of ICA (Independent Component Analysis). In reverberant environments, the performance of source signal separation by ICA is degraded by reflected signals, so it is difficult to use ICA in real environments. Employing the proposed method as preprocessing for ICA, performance improvement of source sound separation is expected even in real environments.

5. Conclusions

Proposed is a method to remove reflected waves from a signal received in a reverberant room. The proposed method solves some problems that our previous method has. Problems almost solved by the proposed method are: unreliability in delay time estimation, approximation degree of supposed source waves, and furthermore, spectral subtraction is not applied to fricative-like or nasal-like portions. There must be further reflected waves as the system deals with only one-time reflection.

We plan to apply the proposed method to preprocessing of ICA (Independent Component Analysis). In reverberant environments, the performance of source signal separation by ICA is degraded by reflected signals, so it is difficult to use ICA in real environments. Employing the proposed method as preprocessing for ICA, performance improvement of source sound separation is expected even in real environments.

References


APPENDIX: Convergence of the algorithm for estimating the path amplitude

A. Characteristic Properties of $\Delta R(\alpha_k)$

Shown in this section is the convergence property of the algorithm for estimating the path amplitude. The purpose of this section is to prove that the solution, or the estimated path amplitude, of equation $\Delta R(l_j) = 0$ exists in the interval [-1,1] and the algorithm described in 2.3 converges to the solution.

The first step of the algorithm is to estimate $\hat{r}_i(k)$, the expected $80th$ sampled value for microphone $\hat{i}$ whose principal reverberations are to be removed. Then, its ACF $\hat{R}(l_j)$ at delay $l_j$ is calculated, where $\hat{R}(l_j)$ is normalized by ACF at null delay.

Next, calculated is the difference $\Delta R(l_j)$ between $\hat{R}(l_j)$, the estimated ACF of the signal received by microphone $\hat{i}$, and $\mathbf{R}(l_j)$, the expected average for $R(l_j)$, and is expressed as follows:

$$\hat{R}(l_j) - \mathbf{R}(l_j) = \sum_{i=1}^{N} \sum_{k=0}^{N-1} \hat{r}_i(k) - \alpha_i \hat{r}_i(k-l_j) - \alpha_i \hat{r}_i(k)$$

where the average ACF $\mathbf{R}(l_j)$ is independent of $\alpha_i$, 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 a large difference between absolute value A and others. Each normalized component in (9) is replaced with notation defined as follows:

\[ A = \frac{1}{R_0} \sum_{k=0}^{N-1} r(k), \quad B = \frac{1}{R_0} \sum_{k=0}^{N-1} r(k) \tau(k + l_0), \quad C = \frac{1}{R_0} \sum_{k=0}^{N-1} \tau(k) r(k + l_0), \quad D = \frac{1}{R_0} \sum_{k=0}^{N-1} r(k) \tau(k + 2l_0), \quad E = \frac{1}{R_0} \sum_{k=0}^{N-1} \tau(k) r(k - l) \]

where these are constants and satisfy the following inequations.

\[-1 \leq A, B, C, D, E, R^\prime R^\prime \leq 1\]

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

\[
\Delta R(\alpha_\gamma) = \frac{B\alpha_\gamma^2 - (A + D + E)\alpha_\gamma + R^\prime R^\prime}{A\alpha_\gamma^2 - 2C\alpha_\gamma + 1} - R^\prime \leq 1
\]

(10)

Here we introduce plausible conditions to prove that there exists a solution \( \alpha_\gamma \) for \( \Delta R(\alpha_\gamma) = 0 \) within interval \([-1,1]\).

\[ A \geq 0 \]

Because 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 concave as quadratic coefficient A is positive. The discriminant of the denominator of (10) is negative because \( C^2 < A \). So, the denominator of (10) has no solution for \( \alpha_\gamma \). It can be concluded that the function \( \Delta R(\alpha_\gamma) \) is continuous over all region for \( \alpha_\gamma \).

In the case of \( \alpha_\gamma = 0 \), \( \Delta R(0) \) becomes \( R^\prime R^\prime - R^\prime \), and is positive because \( R^\prime R^\prime \) is larger than \( R^\prime \). \( l_0 \) has been chosen at which the difference between the ACF of microphone #i and the average ACF is positive maximal. As \( \alpha_\gamma \) goes to \( \pm \infty \), \( \Delta R(\alpha_\gamma) \) asymptotically reaches \( R^\prime - R^\prime \).

### B. The shape of \( \Delta R(\alpha_\gamma) \) for \(-1 \leq \alpha_\gamma \leq 1\)

Assuming that target sources are located apart from walls, \( l_0 \), 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 \( A \geq D + E, \quad \alpha_\gamma \in R^\prime, \quad A \leq C \). Here, \( \alpha_\gamma \) that gives the extremum of the function \( \Delta R(\alpha_\gamma) \) is calculated as

\[ \alpha_\gamma = \frac{B - \Delta R(\alpha_\gamma) \pm \sqrt{(B - \Delta R(\alpha_\gamma))^2 - (A + D + E - 2BC)(2CR^\prime - (A + D + E))}}{2(A + D + E - 2BC)} \]

Considering relative values for \( \alpha_\gamma \) and continuity of \( \Delta R(\alpha_\gamma) \), we can see that \( \Delta R(\alpha_\gamma) \) takes the shape shown in Fig.8(a) for general case and (b) in special case though (b) can be convex or concave.

### C. Existence of a solution for \( \alpha_\gamma \) 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 \( \Delta R(\alpha_\gamma) = 0 \) exists within the interval \([-1,1] \), under situation that \( \Delta R(\alpha_\gamma) \) is continuous for \( \alpha_\gamma \leq \alpha_\gamma < \alpha_\gamma \). The product \( \Delta R(-1)\Delta R(1) \) yields

\[ \Delta R(-1) \Delta R(1) = \frac{B(\alpha_\gamma^2 + D + E)R^\prime R^\prime - R^\prime}{A + 2C + 1} = \frac{B(\alpha_\gamma^2 + D + E)R^\prime R^\prime - R^\prime}{A - 2C - 1} \]

where fractions \( \frac{\alpha_\gamma^2 + D + E}{A} \) and \( \frac{\alpha_\gamma^2 + D + E}{C} \) are approximately null and \( \frac{\alpha_\gamma^2 + D + E}{A - 2C - 1} \) is approximately unity. Let us consider that the product \( \Delta R(-1)\Delta R(1) \) is dependent on the average ACF \( R^\prime(\tau) \). The product \( \Delta R(-1)\Delta R(1) \) is the quadratic function of the average ACF \( R^\prime(\tau) \). Furthermore, the quadratic coefficient is positive, so this function is concave. Hence, if the average ACF \( R^\prime(\tau) \) satisfies \(-0.5 < R^\prime(\tau) < 0.5 \), the product \( \Delta R(-1)\Delta R(1) \) is negative.

On the other hand, it is obvious that \( A \geq R^\prime(\tau) \) and \( R^\prime(\tau) - R^\prime(\tau) < 0 \). So, the average ACF \( R^\prime(\tau) \) is approximately null at \( \tau = l_0 \) and then it satisfies \(-0.5 < R^\prime(\tau) < 0.5 \). It can be concluded that a solution of the equation \( \Delta R(\alpha_\gamma) = 0 \) exists within the interval \([-1,1]\).