Improved Blind Dereverberation Performance by Using Spatial Information

Marc Delcroix \textsuperscript{a, b}, Takafumi Hikichi \textsuperscript{a} and Masato Miyoshi \textsuperscript{ab}

\textsuperscript{a}NTT Communication Science Laboratories, NTT Corporation
2-4, Hikaridai, Seika-cho “Keihanna Science City”, Soraku-gun, Kyoto 619-0237 Japan
\textsuperscript{b}Graduate School of Information Science and Technology, Hokkaido University,
Kita 14, Nishi 9, Kita-ku, Sapporo, 060-0814 Japan
{marc.delcroix, hikichi, miyo}@cslab.kecl.ntt.co.jp

Abstract

In this paper we consider the numerical problems faced by a blind dereverberation algorithm based on a multi-channel linear prediction. One hypothesis frequently incorporated in multi-microphone dereverberation algorithms is that channels do not share common zeros. However, it is known that real room transfer functions have a large number of zeros close to the unit circle on the z-plane, and thus many zeros are expected to be very close to each other. Consequently if few microphones are used, the channels would present numerically overlapping zeros and dereverberation algorithms would perform poorly. We study this phenomenon using the previously reported Linear-predictive Multi-input Equalization (LIME) algorithm. Spatial information can be used to deal with the problem of overlapping zeros. We describe the improved dereverberation performance that we achieve by increasing the number of microphones.

1. Introduction

In this paper, we consider a single-input multiple-output system where a speech signal from one source is captured by \( P \) microphones, as shown in Figure 1. The signal received by the i-th microphone, \( u_i(n) \), can be modeled by the input signal convolved with the room impulse response between the source and microphone \( h_i(n) \) with some additive noise \( \nu_i(n) \), \( i = 1, \ldots, P \).

\[ u_i(n) = h_i(n) * s(n) + \nu_i(n) = \sum_{k=0}^{M-1} h_i(k)s(n-k) + \nu_i(n). \]

where \( M \) is the duration of the room impulse responses. The blind dereverberation problem consists in recovering target signal \( s(n) \) from the observed signals \( u_i(n) \), \( i = 1, \ldots, P \). As we are concentrating on the dereverberation problem, hereafter we consider the noise free case, i.e. \( \nu_i(n) = 0 \) for \( i = 1, \ldots, P \).

Solving the blind dereverberation problem is important for applications such as hands-free teleconferencing and automatic speech recognition (ASR). Indeed, reverberations affect speech characteristics and current ASR systems perform poorly in reverberant conditions [1]. The dereverberation problem can be viewed as the deconvolution of room acoustics. Much research has been undertaken on the topic and in particular, multi-channel deconvolution methods appear interesting because theoretically perfect deconvolution can be achieved [1]. However, traditional methods [2] assume that input signals are independent and identically distributed (i.i.d.). This hypothesis does not hold for speech-like signals. Consequently, when applying traditional deconvolution methods to speech, the speech generating process is somehow deconvolved and the target signal is excessively whitened. In [3] and [4] we presented a two-channel dereverberation algorithm called Linear-predictive Multi-input Equalization (LIME) with a view to solving the whitening problem of traditional methods. Simulations showed that LIME could achieve almost perfect dereverberation for short duration impulse responses. However, we faced computer accuracy problems when trying to deal with longer room impulse responses. The problems are attributed to the presence of numerically overlapping zeros among the channels. It is known [5] that real Room Transfer Functions (RTF) have a large number of zeros close to the unit circle on the z-plane. Many zeros are thus expected to be very close to each other. Consequently, for a small number of microphones, the channels would present numerically overlapping zeros and the dereverberation algorithm would perform poorly. In this paper, we show how the use of spatial information, obtained by increasing the number of microphones, enables us to deal with long duration impulse responses. In the second section we recall the principles of the LIME algorithm in the general case of \( P \) microphones. We then present the results of simulations showing the algorithm behavior when we increase the number of microphones from 2 to 6. Finally we employ LIME for the dereverberation of a measured impulse response with a reverberation time of 0.5 seconds.

2. LIME algorithm

2.1. Notation and hypotheses

We construct the following hypotheses:

- First, we assume that input signal \( s(n) \) is modeled by an autoregressive (AR) process applied to white noise \( e(n) \). The Z-transform of the AR process is \( 1/a(z) \) where \( a(z) \) is the AR polynomial. We would like to define \( a(z) \) using \( N+1 \) coefficients as:

\[ a(z) = 1 - \{a_1z^{-1} + \ldots + a_Nz^{-N}\}, \quad (2) \]

however, the actual order of \( a(z) \) may be smaller than \( N \). Using a matrix formulation we can write [6]:

\[ s(n) = C^T s(n-1) + e(n), \quad (3) \]
where $C$ is the companion matrix defined as:

$$
C = \begin{pmatrix}
a_1 & 1 & 0 & \cdots & 0 \\
a_2 & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & \ddots & \cdots & \vdots \\
a_n & 0 & 0 & \cdots & 1 \\
\end{pmatrix},
$$

and $s(n) = [s(n), s(n-1), \ldots, s(n-N)]^T$. $e(n) = [e(n), 0, \ldots, 0]^T$.

- We model the Room Transfer Functions (RTF) by polynomials that we assume to be time invariant. We also assume that the RTF do not share common zeros. The Z-transform of the $i$-th RTF is given by

$$
H_i(z) = \sum_{k=0}^{m-1} h_i(k)z^{-k}
$$

We use the matrix formulation to rewrite Equation (1) as:

$$
u_i(n) = H_i^T s(n),
$$

where $H_i$ is an $N \times L$ convolution matrix expressed as

$$
H_i = \begin{pmatrix}
h_i(0) & 0 & \cdots & 0 \\
h_i(1) & h_i(0) & \cdots & \vdots \\
\vdots & \vdots & \ddots & \vdots \\
h_i(M-1) & h_i(M-1) & h_i(0) & \cdots \\
\vdots & \vdots & \vdots & \cdots & 0 \\
\end{pmatrix},
$$

$N = L + M - 1$, $h_i([h_i(0), \ldots, h_i(M-1)])^T$, $i = 1, \ldots, P$, $u_i(n) = [u_i(n), \ldots, u_i(n-(L-1))]^T$, $L$ is the number of samples of $u_i(n)$.

We further define $u(n) = [u_i^T(n), \ldots, u_P^T(n)]^T$, and $H = [H_1, \ldots, H_P]$ full row rank matrix of size $N \times PL$, thus

$$u = H^T s.
$$

The blind dereverberation problem consists in recovering a target input signal $s(n)$ from the $P$ observed microphone signals, $u_i(n)$ $i = 1, \ldots, P$. First, we propose the calculation of prediction filters that cancel out the effect of the RTF. However as those filters also whiten the signal, in the second step of our method, we estimate the AR process in order to recover the target signal precisely.

2.2. Linear prediction

According to the linear prediction theory [7], we can define the prediction error as:

$$
\hat{e}(n) = u_1(n) - u_1^T (n-1) w
$$

where $w$ is a prediction filter set of length $PL$.

$$w = \begin{pmatrix}
w_1^T \\
w_2^T \\
\vdots \\
w_L^T
\end{pmatrix},
$$

$$w_i = [w_i(0), \ldots, w_i(L-1)]^T, i = 1, \ldots, P.$$

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

$$w = \left(H^T E\{s(n-1)s^T(n-1)\}H\right)^+ H^T E\{s(n-1)s^T(n)\}u_1
$$

where $A^+$ is the Moore-Penrose generalized inverse of matrix $A$ [8], and $E\{\}$ is an expectation operator. If we replace the column vector $u_1$ with matrix $H$, we can define matrix $Q$ as:

$$Q \triangleq \left(H^T E\{s(n-1)s^T(n-1)\}H\right)^+
$$

From Equation (3) we can write:

$$E\{s(n-1)s^T(n)\} = E\{s(n-1)s^T(n-1)\}.
$$

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

$$Q = \left(H^T X^T XH\right)^+ H^T X^T XCH
$$

We define the first column of $Q$ gives us the prediction filter set,

$$w = H^T (HH^T)^{-1} C u_1.
$$

The prediction error is thus:

$$\hat{e}(n) = s^T(n)h_1 - s^T(n-1)Hw
$$

By definition, the first column of $Q$ can be obtained from the characteristic polynomial of matrix $A$. From (18) we deduce that the estimated AR polynomial, $\hat{a}(z)$, is obtained by a relatively long averaging process. Consequently, the estimated AR polynomial, obtained from the characteristic polynomial of matrix $Q$, corresponds to a so-called average AR polynomial [4].
2.4. Calculation of Matrix $\mathbf{Q}$

The algorithm is “blind” because the dereverberation is achieved without prior knowledge of the RTF. 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 relation (6) Equation (11) becomes:

$$\mathbf{Q} = \left( E\{\mathbf{u}(n-1)\mathbf{u}^\top(n-1)\} + E\{\mathbf{u}(n-1)\mathbf{u}^\top(n)\} \right).$$

Equation (19) is used in practice to calculate $\mathbf{Q}$.

3. Algorithm

We can summarize the dereverberation algorithm as follows:

1. First we calculate matrix $\mathbf{Q}$ with signals received at the microphones using Equation (19).
2. The first column of matrix $\mathbf{Q}$ gives us the prediction filter set, $\mathbf{w}$.
3. The prediction error is calculated using formula (8).
4. The estimated AR parameters are obtained from the characteristic polynomial of $\mathbf{Q}$.
5. The input signal is recovered by filtering the prediction error with the estimated AR parameters.

4. Simulation

4.1. Simulation conditions

We conducted two types of simulations. With the first type, we simulated a room sound field with the image method [9]. We can set the value of the reflection coefficient of the room to control the order of the RTF. This enables us to study the behavior of LIME as a function of RTF order. With the second type of simulation the RTF were measured in a real room. The second simulation shows the applicability of the algorithm to real RTF.

For both simulations, the disposition of the room and microphones was the same and is shown in Figure 3. The microphone signals were simulated by convolving RTF with speech taken from the ATR database [10]. The sampling frequency was 8 kHz. To show how LIME could improve the performance of ASR systems, we evaluate the results using the Spectral Distortion (SD) defined as:

$$SD = \frac{2}{F} \sum_{f=0}^{F-1} \left( 20 \log \left| P(f) \right| - 20 \log \left| \hat{P}(f) \right| \right)^2,$$

where:

- $P(f)$ is the short time Fourier transform of target signal $s(n)$ using a Hamming window,
- $\hat{P}(f)$ is the short time Fourier transform of the estimated target signal $\hat{s}(n)$ using a Hamming window,
- $F$ is the number of frequency bins.

The SD was calculated for time frames of 30 ms, and the results were averaged for all the frames.

4.2. Impulse responses generated by the image method

We generated RTF with duration of 0.01 to 0.45 seconds. We used from 2 to 6 microphones and 2 seconds of training data to obtain the following results. Figure 4 plots the SD as a function of the RTF duration for 2, 4 and 6 microphones. With 2 microphones, the spectral distortion becomes large for RTF longer than 0.03 seconds. We face accuracy problems due to numerically overlapping zeros. We observed that the prediction error was precisely obtained. The prediction filters are thus correctly calculated and the distortion comes from the estimated AR process. Overlapping zeros are wrongly considered to be part of the AR process. Consequently, the estimated AR process is degraded. From these results, we conclude that using only 2 microphones, LIME fails to estimate the AR parameters for long RTF and thus the spectral distortion is large. If we increase the number of microphones, the spectral distortion can be considerably reduced even for long RTF. Indeed, as the number of channels increases, the diversity among the channels also increases and fewer overlapping zeros are expected. The AR process can thus be estimated more precisely. In Figure 5 we show the estimated AR process obtained using 4 and 6 microphones for RTF of 0.45 seconds. With 4 microphones, the AR process is estimated with few distortions. The SD of the recovered signal is 7.6 dB. When using 6 microphones the AR process is estimated precisely and the SD of the recovered signal is only 1.9 dB.

4.3. Measured impulse response

We tested our method with impulse responses measured in a real room with a reverberation time of around 500 ms ($M = 4000$ Hz).
taps). We used 4 microphones and 2.5 seconds of training data to calculate the prediction filters and AR process. The prediction filter length was set at \( L = 1344 \) taps and consequently \( N = 5344 \) taps. Figure 6 plots the reverberation energy density curve for the original impulse response and equalized impulse response. When we use LIME, the reverberation is almost completely suppressed with an attenuation of around 40 dB. The

SD of the microphone signal is 13.9 dB and this can be reduced to 3.3 dB by using our method. The remaining distortions are mainly localized in the lower frequency band below 80 Hz and to 3.3 dB by using our method. The remaining distortions are mainly localized in the lower frequency band below 80 Hz and this can be reduced using 4 microphones (dashed line).

Figure 6: Reverberation energy density curve of the original impulse response (solid line) and dereverberated impulse response using 4 microphones (dashed line).

6. References