Complex Cepstrum Based Single Channel Speech Dereverberation

Shen Xizhong
Mechanical & Electrical department
Shanghai Institute of Technology
Shanghai, China
Shenxizhong2004@msn.com

Meng Guang
National Key Laboratory for Mechanical System and Vibration
Shanghai Jiao Tong University
Shanghai, China

***Abstract*—A new method of blind dereverberation of single channel is proposed to solve the existing problem about the study of speech signal. Firstly, segmentation of speech signal with variant length is proposed on the sampled speech signal to obtain relatively accurate estimation of the subsequent cepstrum, and the complex cepstrum is applied to the valid segments. Then, the pre-estimation of the room impulse response is obtained, and so is the inverse-filter solution. Finally, single channel blind deconvolution is used by taking the inverse-filter as the initial parameter of the algorithm. Simulations and experiments demonstrate its validity and robusty.**

***Index Terms*—Dereverberation; Complex cepstrum; Room Impulse response; Speech; Segmentation**

I. INTRODUCTION

In the real world, reverberation is one of the primary factors that degrade the quality of speech signals when captured by a distant microphone. It makes sounds unintelligible, and prevents any system from adequately extracting any speech features. This problem becomes more severe as the reverberation time becomes longer. For example, when the reverberation time is longer than 0.5 sec, the performance of an automatic speech recognition system does not improve sufficiently even when the recognizer is trained on reverberant signals captured in the same environment [1].

In general, one sampled speech signal includes two categories, one is the direct path wave directly from the source to the microphone, and the other is the resulted echoes, also known as reflected wave, which are reflected from material such as wall, etc., to the receiver. Besides, the signal is often disturbed by environment noise. Acoustically, the effect of reflection is named reverberation. The reverberation is needed by some architecture such as musical hall, church, etc., to make better sound effect for the human ear; however it is annoying for some receivers such as speech recognition system, or speech-to-text appliances. Therefore, it is a very important subject how to decrease the reverberation. Blind dereverberation is aimed to perform dereverberation in the case of the unknown channel and the unknown source, and it is currently one of the main research directions in the field of speech signal processing.

In this paper, we further study the room acoustic reverberation in terms of the method proposed by Bees, etc [2], and propose a method to estimate the room acoustic impulse response (RIR) by complex cepstrum mainly on adaptive segmentation technique. The RIR pre-estimation is obtained by this method, and so is the inverse-filter solution. Blind deconvolution(BD) is used by taking the inverse-filter as the initial parameter of the single channel blind deconvolution algorithm(SCBD) proposed by Douglas etc [3]. Simulations and experiments show its validity and robusty.

II. THE PRINCIPLE OF REVERBERATION

A. Definition of reverberation

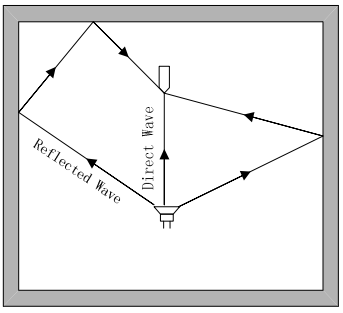

Figure 1. Reverberation of Room

In order to explain the principle of reverberation, a receiver (microphone) $x(t)$ is built up near a source signal (speaker) $s(t)$ in a room at certain distance, which includes the direct wave $s(t)$ and the echoes reflected from different faces in the room. The echoes are delayed in time, and attenuated in amplitude. The signal received by the microphone is described in mathematical model as

$$x(t) = s(t) + \sum_{k=1}^{M} \alpha_k s(t - t_k) + n(t), \qquad (1)$$

where $0 < t_1 < t_2 < \cdots < t_M$, and $|\alpha_k| < 1$. Set RIR as

$$h(t)=\delta(t)+\sum_{k=1}^{M}\alpha_k\delta(t-t_k). \quad (2)$$

Then we get from(1),

$$x(t)=s(t)*h(t) \quad (3)$$

$s(t)$ is distorted severely as its length is far greater than the delayed width $t_k, k=1,\cdots,M$. The observation is generated by the echoes as described in (3), and the dereverberation of speech signal is aimed to separate the source $s(t)$ and RIR $h(t)$, which is also named as homomorphic filter problem. Homomorphic signal processing mainly solves the problem of homomorphic filter, and its task is to accomplish the separation of convolution of two signals by transforming the convolutional relation to summative relation, that is, deconvolution[4].

B. Analysis of reverberation

We firstly consider only one echo[4], that is,

$$h(t)=\delta(t)+\alpha_1\delta(t-t_1). \quad (4)$$

Then, $$x(t)=s(t)+\alpha_1 s(t-t_1). \quad (5)$$

The z-transform of (3) is

$$X(z)=S(z)H(z) \quad (6)$$

where $H(z)=1+\alpha_1 z^{-t_1}$. By log operation of the two sides of eq.(6), we get

$$\hat{X}(z)=\ln X(z)=\hat{S}(z)+\hat{H}(z). \quad (7)$$

To make inverse z-transform of (7), we get

$$\hat{x}(t)=\hat{s}(t)+\hat{h}(t) \quad (8)$$

where $\hat{x}(t)$, $\hat{s}(t)$ and $\hat{h}(t)$ are complex cepstra of $x(t)$, $s(t)$ and $h(t)$. It is easily proved that [4] when $|\alpha_1|<1$,

$$\hat{h}(t)=\sum_{k=1}^{+\infty}(-1)^{k+1}\frac{\alpha_1^k}{k}\delta(t-kt_1). \quad (9)$$

Here, $\hat{h}(t)$ is impulse series with every t_1 distance space, and its every term in (9) exponential attenuation with the increase of k , as depicted in Figure 2. Sampling point is expressed in the abscissa, and cepstrum is expressed in the ordinate. The figure is taken with $t_1=24$ in (9). The wall reflection coefficient is set as $\alpha_1=0.8$. When $t=6t_1$, the term in (9) is attenuated to 0.2621, and it is trivial enough to be ignored in subsequent calculation under noise environment. If we can design a time trap filter $w(t)$ with invariant frequency to remove the effect of $\hat{h}(t)$, the original speech signal is obtained by complex inverse cepstrum[2].

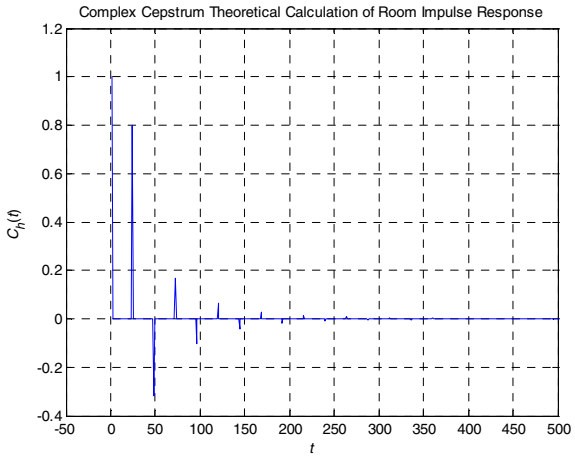

Figure 2. Theory calculation of cepstrum of room impulse response (The wall reflection coefficient is $\alpha_1=0.8$.)

The complex cepstrum of the room impulse response is in the form of peaks depicted in Figure 2, and most of its components are far away from zero cepstrum time in the cepstrum region. Attention is deserved that the complex cepstrum of speeh is relatively closer to the original point than the RIR points. Therefore, we can use peak-picking technique to pick the needed signal [2].

III. THE PRINCIPLE OF DEREVERBERATION

We estimate room impulse response by means of complex cepstrum[2] and peak-picking technique. Because deconvolution in time field corresponds to the subtraction in cepstral field, that is,

$$\hat{w}(t)=-\hat{h}(t), \quad (10)$$

the time filter is obtained and then we take it as the initial solution of blind deconvolution to get relatively more accurate source $s(t)$. The concrete calculating processes is depicted in Figure 3. The detail problems are analyzed and solved in this section as follows.

978-1-4244-3521-0/09/$25.00 ©2009 IEEE

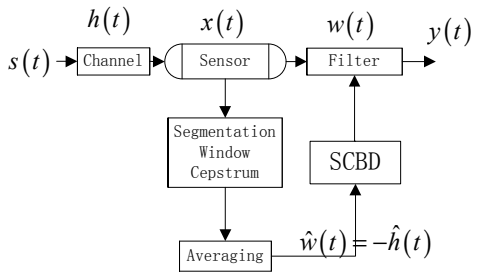

Figure 3. Block diagram of dereverberation system

A. *Initial Solution: Complex Cepstrum*

The RIR is obtained by segmenting the speech signal, adding exponential window, averaging [2], in which the window processing is to make the signal characterized with minimum phase. The computed complex cepstrum is averaged over several segments, we get from(8),

$$\bar{\hat{x}}(t) = \bar{\hat{s}}(t) + \bar{\hat{h}}(t). \tag{11}$$

Speech pitch is not accurately constant over most speech segments, and the large cepstral peak $\bar{\hat{s}}(t)$ at the period of the pitch becomes smeared and reduced by averaging. Otherwise, the cepstrum remains constant due to $h(t)$, that is, $\bar{\hat{h}}(t) = \hat{h}(t)$. This allows cepstral components corresponding to the impulse response located around the pitch period to be identified after averaging by peak-picking [4]. Therefore, when $\bar{\hat{s}}(t)$ is reduced to be ignored so that

$$\bar{\hat{x}}(t) \approx \bar{\hat{h}}(t). \tag{12}$$

Then by(10), we get dereverberant filter $w(t)$. However, we cannot make $\bar{\hat{s}}(t)$ infinitesimal on account of the limit samples and unstability of acoustical signal especially with long time series.

Fixed segmenting was proposed to be applied to acoustic dereverberation which began only after a period of speech silence, and exponential windowing was applied to reduce the segmentation errors further in the time domain [2]. However, truncation error is made by fixed length in the proposed algorithm. To further overcome the error, we apply variant length to segment the sampled speech, and make averaging over the segments. The length is self-adapted by the silent state of the speech signal; in the way that one segment is beginning from the end of one silence and ending at the start of next silence. We can find the way in Figure 4. Exponential window is also added to improve the performance in the sense that the error is further reduced.

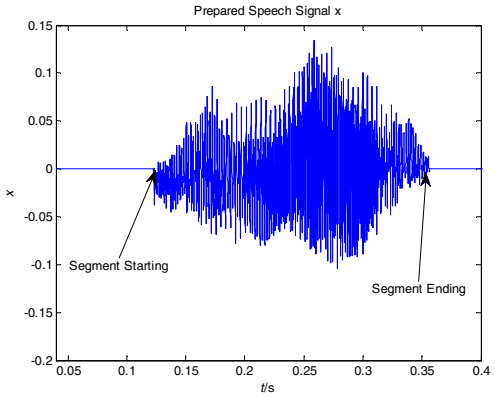

Figure 4. Segmentation of speech signal

We take synthetic averaging on the cepstrum of the different segments, which is quite different from the traditional moving averaging[2]. Here, synthetic averaging means that the common parts of each cepstrum are averaged, but the overflowing parts due to the different length are given up. We take the length of the common parts as 4096, 2048, 1024, etc. Simulated room impulse response is shown in Figure 5(Top figure), and it is used to generate the reverberant speech signal. Variant length segment is applied to estimating the impulse response, depicted in Figure 5(middle figure), and fixed length segment is used in Figure 5(below figure). We set the coefficient of exponential window $\gamma = 0.99996$. The estimated peaks of the room impulse response are close to the simulated one, and the method with variant length is better than the method with fixed one.

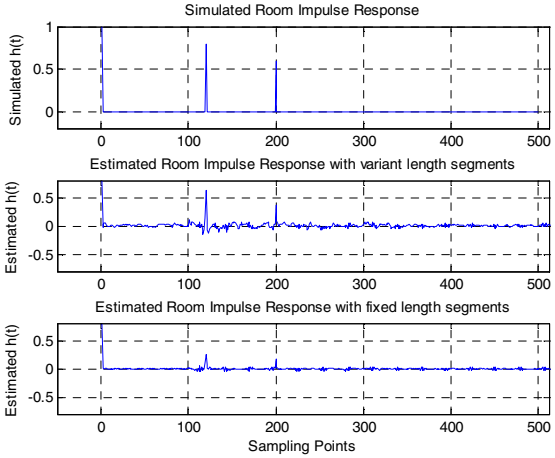

Figure 5. Simulation experiment of room impulse response

B. *More Accurate Solution: SCBD*

Peak-picking technique is used in Figure 5(b) and (c) to select the cepstrum peaks $\bar{\hat{h}}(t)$ of the main echoes. Here, we take the cepstrum data responding to $t > 12.5$ ms, and set the other cepstrum zero as a matter of experience. At last, dereverberation filter $w(t)$ is obtained by (10). Therefore, error in design is existing due to the zero cepstrum setting, and we use blind deconvolution to complement it.

Douglas et al proposed an algorithm of single channel blind deconvolution, and it is claimed to estimate source signal under the unkown source and the unknown impulse response. The filter is updated as follow,

$$\mathbf{w}(t+1)=\mathbf{w}(t)+\mu(t)\left[\mathbf{w}(t)-f\left(y(t)\right)\mathbf{z}^{\mathrm{T}}(t)\right] \tag{13}$$

where $\mathbf{w}(t)=\left[w_0(t) \quad w_1(t) \quad \cdots \quad w_L(t)\right]$ is blind deconvolutional filter with length L; $\mathbf{x}(t)=\left[x(t) \quad x(t-1) \quad \cdots \quad x(t-L)\right]^{\mathrm{T}}$ is input observation signal; $y(t)=\mathbf{w}(t)\mathbf{x}(t)$ is the output of the filter at instant time t; $f\left(y(t)\right)=y(t)^3$; and $\mathbf{z}(t)=\mathbf{R}(t)\mathbf{x}(t)$.

Here, $\mathbf{R}(t)=\begin{bmatrix} r_0(t) & r_1(t) & \cdots & r_L(t) \\ r_{-1}(t) & r_0(t) & \cdots & r_{L-1}(t) \\ \vdots & \vdots & \ddots & \vdots \\ r_{-L}(t) & r_{-L+1}(t) & \cdots & r_0(t) \end{bmatrix}$ is coefficient autocorrelation matrix; and $r_l(t)=\sum_{p=0}^{L-|l|} w_p(t)\,w_{p+|l|}(t), -L\le l\le +L$ is coefficient autocorrelation function; T is matrix transposition; and $\mu(t)$ is the learning rate.

The blind deconvolutional algorithm of single channel aims to obtain steady signal and channel, and the speech signal is approximately regarded as steady. Thus, the algorithm proposed by Douglas et al can be applied to blind dereverberation. However, center-tap method is usually taken as the initial solution of blind deconvolution. It needs many iterations of the algorithm, and the convergence is very slow, or even divergent when the learning rate is not adequate. Therefore, we take the filter (10) as the initial value $\mathbf{w}(0)$ of the algorithm (13) and the number of iteration is reduced greatly to obtain the same performance as the general center-tap method.

We study the speech signal, and only select the valid data to update the algorithm (13) to further reduce the iterations. Here, the data is valid when its variance is greater than the variance of environment noise, that is, we set the threshold value as 0.2 of maximum variance.

IV. SIMULATION

In this section, we perform two examples, one is to estimate the artificial room acoustic reverberation, and the other is to perform dereverberation of practical speech signal.

Firstly, we use a phase of "clean" speech signal recorded in an anechoic chamber with 60s length, and 41.666kHz sampling rate. Here, "clean" means that the recorded signal has tiny or no reverberation. We use an artificial impulse response as depicted in Figure 5(a). to simulate the room case, and then use the algorithm in Figure 3 to dereverberate. We set $\gamma=0.99996$, zero in the range [0 100] of cepstrum, and $\mu(t)=10e^{-0.01t}+10^{-6}$ in SCBD algorithm. Figure 6 shows the inter-symbol interference(ISI), which defined as

$$\mathrm{ISI}(k)=\frac{\sum_{l=0}^{M} c_l^2(k)}{\max\limits_{0\le j\le M} c_j^2(k)}-1, \tag{14}$$

where $c(t)=w(t)*h(t)$. The top view in Figure 6 shows ISI with cepstrum initialization; and the bottom one ISI with center-tap initialization. The effect of dereverberation is further improved in the sense that ISI is more robustly converged to zero in top view than in bottom view. The same effect is also obtained in the other simulated signals recorded in the anechoic chamber.

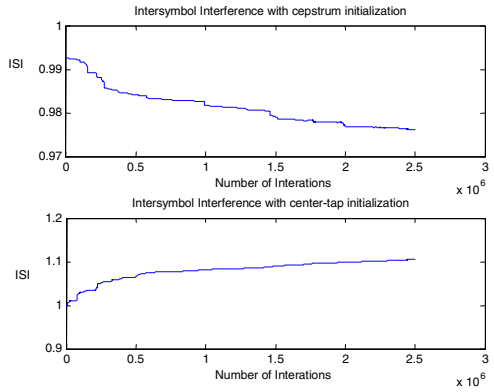

Figure 6. Intersymbol Interference

Secondly, we select a room of size 2000x1500x2500mm^3 to record speech signals with 500mm distance between the microphone and the speaker. Several Chinese speech signals are saved with sampling rate 8kHz, and length 80s. Figure 7 shows one of the results of the real speech signal experiments. There are two figures showing estimated room impulse response, where top view shows it just with cepstrum, and bottom view with cepstrum and SCBD. We set $\gamma=0.99996$, divide the signal into 1302 segments by silent state, and average over 4096 length. Then, the filter is obtained by $w(t)=-\hat{\bar{p}}(t)$, and use SCBD to perform blind deconvolution. As a result, we get a clean speech, which sounds very clear.

978-1-4244-3521-0/09/$25.00 ©2009 IEEE

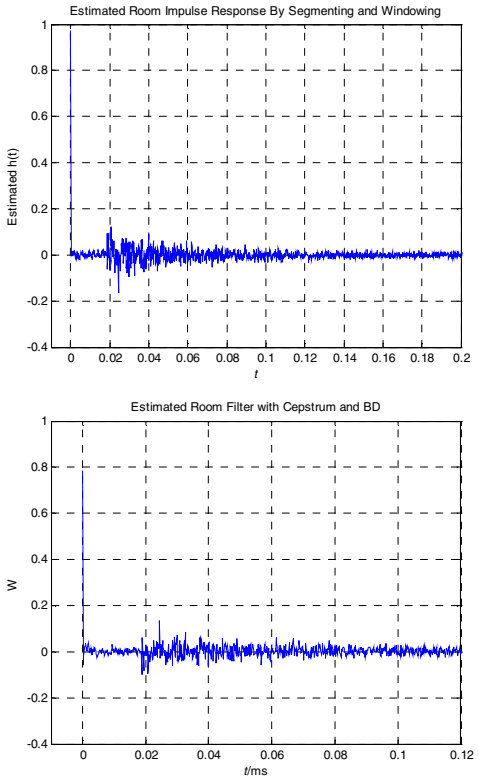

Figure 7. Real speech signal experiments

V. CONCLUTION

We have studied the case of room acoustic reverberation, and propose a new method of blind dereverberation of single channel. Here, two main ideas are applied in this paper; one is adaptive segmentation of speech signal, and the other is the combination of complex cepstrum and single channel blind deconvolution. Simulations and experiments are shown its validity and robusty.

ACKNOWLEDGMENT

This work has been supported by NSFC with No. 10732060, and also Shanghai Education with No. ZX2006-01.

REFERENCES

[1] Nakatani T; Miyoshi M and K Kinoshita. Single-Microphone Blind Dereverberation, in Book: J Benesty; S Makino; J Chen. Speech Enhancement. Springer Berlin Heidelberg. 2005, pp: 247-270.

[2] Bees, D.; Blostein, M.; Kabal, P. Reverberant speech enhancement using cepstral processing. Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on. 14-17 April 1991 Page(s):977 - 980 vol.2.

[3] Douglas, S.C.; Sawada, H.; Makino, S.; Natural gradient multichannel blind deconvolution and speech separation using causal FIR filters. IEEE Transactions on Speech and Audio Processing, Volume 13, Issue 1, Jan. 2005 Page(s):92 – 104

[4] Oppenheim A V, Schafer R W. Digital Signal Processing [M]. Prentice Hall Inc., 1975.

[5] Nakatani T; Miyoshi M. Blind dereverberation of single channel speech signal based on harmonic structure. NTT Commun. Sci. Labs., NTT Corp., Kyoto, Japan; Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on. 6-10 April 2003. Volume: 1: I- 92-5 vol.1