A New HB Weighted Time Delay Estimation Method in Reverberation Environment

Liyan Zhang[†] [††], *Fuliang Yin*[†]

[†]*School of Electronic and Information Engineering,Dalian University of Technology, Dalian 116023 China*
[††] *School of Electronic Engineering, Dalian JiaoTong University, Dalian 116028, China*

Summary
In the tele-conference, the speech recognition system is mainly affected by the reverberation and not affected by the noise. The model for receiving signals adopted by most time delay estimation methods are additive noise model, which are not suitable for the system. Thus this paper illustrates one HB weight time delay estimation method in the reverberation environment based on onset signals. This method uses the onset signals without reverberation to estimate the cross power spectrum of the signals, then compensates the power spectrum of the noise, and at last adopts HB method to realize the accurate time delay estimation. The simulation result shows the validity of the method.

Key words:
onset signal, reverberation, time delay estimation, microphone array

1. Introduction

The international research on the using of microphone array technology for speech signals process begins from the 80s in the 20th century. These years, because of the wide range application of speech enhancement technology, which is based on the microphone array, on the tele-conference[1], vehicle communication[2], hearing aids[3] and robots control system[4], a series of speech enhancement methods based on microphone array were puts forward, such as SVD-based beamforming speech enhancement algorithm[4], a subband speech enhancement algorithm[5] and so on. For these speech enhancement methods, all the signals received are time delay compensated, so the function of the time delay estimation will affect the speech enhancement system to some extent. Most of the time delay algorithms[6-8] don't take into account the existence of reverberation, but in fact the reverberation does exist. For example, in the tele-conference system, the reverberation affects the system function badly. In 1997, the first tele-conference system[9], presented by Picture Tel Co., cannot be used because of the worse robust on reverberation and noise. So how to estimate accurately the time delay under reverberation environment needs to be solved urgently.

Manuscript received July 5, 2008.
Manuscript revised July 20, 2008.

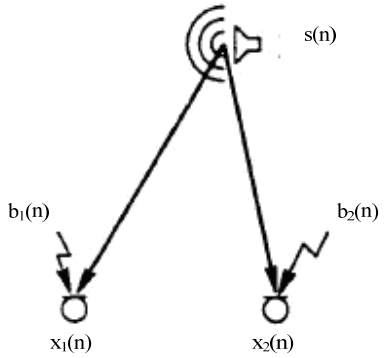
In these ten years, many time delay methods are studied. In 1976, Knapp and Cater put forward the generalized correlation method for estimation of time delay[6], which makes the study of time delay to the climax. The generalized correlation method[7] for estimation of time delay is very simple, with less calculation and better estimation precision, applied widely in the microphone array speech signals process. But this method doesn't take into accountant the influence of the reverberation, so is not suitable for the system of this paper. Adaptive eigenvalue decomposition algorithm[10] is different with correlation method for estimation of time delay, and this algorithm begins from reverberation model, then to estimate the time delay through adaptive approach shock response. So this method can still have good function under big reverberation environment. But this method needs to estimate the correlated matrix of noise and carry out matrix inversion algorithm, which demands more calculation and cannot be used practically. This paper puts forward a new time delay estimation method, based on generalized correlation time delay estimation and combined with the real characteristics of speech signals, suitable for reverberation environment onset signals and HB methods.

The second and third part of this paper illustrates separately the acoustics model and Room impulse response, and the fourth part introduces HB weighted time delay estimation method. The fifth part of this paper illustrates in details the HB weighted time delay methods based on onset signals, and the sixth part shows the computer simulation result, then the last part is the conclusion of the paper.

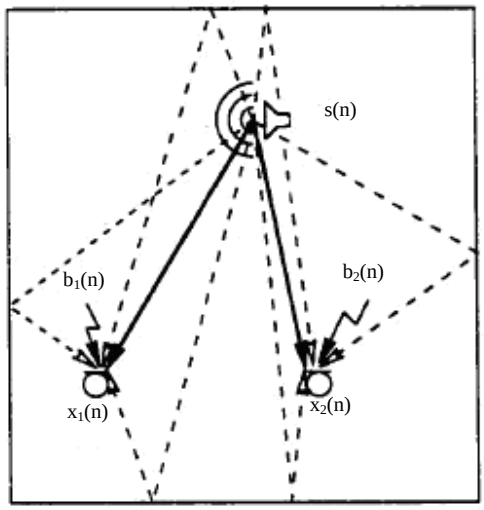
2. Acoustics model

Microphone array realizes the speech signals process by using many microphones to collect signals, and the acoustic model canned be divided into ideal acoustic model and real acoustic model[11], as shown in Fig. 1.

Fig.1(a) denotes that the signals received by microphone array can be described with ideal model without



(a) ideal acoustic model



(b)real acoustic model

Fig.1 Signal model

consideration of reverberation, and the vector form is shown as follows:

$$\mathbf{x}(n) = \boldsymbol{\alpha} \sqcup \mathbf{s}(n-\tau) + \mathbf{B}(n). \quad (1)$$

$\mathbf{x}(n)$ is the column vector of signals received by microphone array, $\boldsymbol{\alpha}$ is the column vector of each channel transmission attenuation. $\mathbf{B}(n)$ is the column vector of additive noise. Vector signal $\mathbf{s}(n)$ is the time-lapse with delay vector τ, and the vector τ is correlated with the transmission delay decided by array system geometry size. Operator "$\sqcup$" denotes the product among elements.

Fig.1 (b) denotes under the real environment, taking the reverberation into account, the signals received by microphone array can be descriped with real model, and the vector form is as follows:

$$\mathbf{x}(n) = \mathbf{h}(n) * \mathbf{s}(n) + \mathbf{B}(n). \quad (2)$$

$\mathbf{h}(n)$ is the column vector of room transmission function, operator "$*$" denotes convolution operation.

3. Room impulse response

Image model[12] is offen used to simulate the room reverberation during the small tele-conference system. Supposed that the wall is slick, reflects direct, and will loose part of the wave energy during each reflecttion. The iterative reflection is equivalent with a series of intension attenuation mirror source. The room shock response can be gained by discovering all the mirror source intention and location distribution. The calculation formula of room shock response come from Image model is as follows:

$$h(t,\mathbf{x},\mathbf{x}') = \sum_{\mathbf{p}=0}^{1}\sum_{\mathbf{r}=-\infty}^{\infty} \beta_{x1}^{|n-i|}\beta_{x2}^{|n|}\beta_{y1}^{|l-j|}\beta_{y2}^{|l|}\beta_{z1}^{|m-k|}\beta_{z2}^{|m|} \times \frac{\delta(t-\left|\mathbf{R}_p+\mathbf{R}_r\right|/c)}{4\pi\left|\mathbf{R}_p+\mathbf{R}_r\right|}. \quad (3)$$

Here, $\beta_{x1},\beta_{x2},\beta_{y1},\beta_{y2},\beta_{z1},\beta_{z2}$ is the reflection coefficients of each plane in the room, x and x' is the speech source coordinate and microphone coordinate, c denotes velocity of sound, L,W and H is separatel the length, width and height of the room. Let x coordinate is $\mathbf{p}=\begin{bmatrix} i & j & k \end{bmatrix}$, and x' coordinate is $\mathbf{r}=\begin{bmatrix} n & l & m \end{bmatrix}$, $R_r=\begin{bmatrix} 2nL & 2lW & 2mH \end{bmatrix}$, $R_p=\begin{bmatrix} x-x'+2ix' & y-y'+2jy' & z-z'+2kz' \end{bmatrix}$。

If supposed the reflection coefficient of each wall is the same and equal to γ, the Eq. (3) can be shown as follows:

$$h(t,\mathbf{x},\mathbf{x}') = \sum_{\mathbf{p}=0}^{1}\sum_{\mathbf{r}=-\infty}^{\infty} \gamma^{|n-i|}\gamma^{|n|}\gamma^{|l-j|}\gamma^{|l|}\gamma^{|m-k|}\gamma^{|m|} \times \frac{\delta(t-\left|\mathbf{R}_p+\mathbf{R}_r\right|/c)}{4\pi\left|\mathbf{R}_p+\mathbf{R}_r\right|}. \quad (4)$$

If the room size is 5m x 6m x 3m, and the value of the reflection coefficient γ is 0.74s, the reverberation time is about 200ms, then the room shock response curve is shown as Fig. 2.

4. HB weighted time delay estimation method

4.1 Signal model

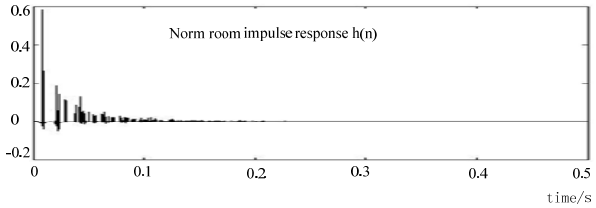


Fig.2 Norm room impulse response

Supposed the signals model of one pair microphones is

$$x_1(t) = s(t) + b_1(t). \tag{5}$$

$$x_2(t) = s(t-\tau) + b_2(t). \tag{6}$$

$s(t)$ is the speech signals, τ is the transmission time delay, and hereby doesn't consdier the difference of the transmission range attenuation, $b_1(t)$ 、 $b_2(t)$ is the additive background noise , suppose $s(t)$、$b_1(t)$、$b_2(t)$ is not correlated with each other。

The generalized correalation function of two microphones $R_{12}(\tau)$ can be shown as :

$$R_{12}(\tau) = \int_0^{\pi} W_{12}(\omega) X_1(\omega) X_2^*(\omega) e^{-j\omega\tau} d\omega. \tag{7}$$

Here, $X_1(\omega)$ and $X_2(\omega)$ is the Fonrier transform of $x_1(n)$ and $x_2(n)$, $W_{12}(\omega)$ is the generalized correalation function. To different noise and reflection situation, different weighted function $W_{12}(\omega)$ can be used, which can make $R_{12}(\tau)$ has more aculeated peak value. The peak value of $R_{12}(\tau)$ is the time delay of two microphones.

The HB weighted function [8] of time delay estimation defined as :

$$W_{12}(\omega) = \frac{|G_{x_1 x_2}(\omega)|}{G_{x_1 x_1}(\omega) G_{x_2 x_2}(\omega)}. \tag{8}$$

The HB weighted time delay estimation method doesn't consider the influence of reverberation, so the key of the problem is how to make the cross power spectrum estimation not affected by reverberation.

5. The proposed method

For the method in this paper, we adopts speech onset signals without reverberation to estimate the time delay. Becasue the scope of the signals without reverberation is small, the algorithm must estimate the cross power spectrum of the background noise, then get rid of the contribution of noise in the cross power spectrum of signals received. In fact, there is various noises and the noises is correalated with each other partly. In this paper, we suppose the noise is short time stable with minimum average energy. First, calculate the minimum signals frame energy during the current long enough time period (for example 5s), and if for one long period (for example 0.5s), the signals energy distribution almost keeps the same and is equal to the energy of the minimal signal frame, then consider these frams as the noise, which can be used to estimate the cross power spectrum of current background noise. At last smooth last time estimation result and current estimation result, get the final estimation of noise cross power spectrum. $P_{old}(w)$ denotes last time estimated noise cross power spectrum, $P_{current}(w)$ denotes current estimated noise cross power, then the final estimated noise cross power spectrum $P_{new}(w)$ is

$$P_{new}(w) = \alpha P_{current}(w) + (1-\alpha) P_{old}(w), 0 < \alpha \le 1. \tag{9}$$

α is the smooth coefficient, normally endowed with a small value, but if the change of of the minimal frame energy is bigger, which shows the noise is not stable, then α shall be endowed with 1。

The collection of the onset signals without reverberation is based on EA model (Echo-Avoidance)[13-14] , which is based on human hearing system preference effect. The basic idea is to inspect the onset signals without reverberation before time delay estimation. Onset signals are part of the sound, locates behind mute speech phase. Compared with background noise and significative signals thresh hold values, the amplitude of onset signals is bigger. In addition, the effective onset signals length can be decided by the time difference between the through signals and reflected signals.

Suppose τ_{fe} is the time delay of the first arrived reflected signals, and each τ_{fe} passed, choose suitable window function to analyze signals with short-time window Fonrier transform,and the result is $X_i(j,k)$, k is the frequency tag, j is the frame tag, i is the microphone tag. Because the maximum reverberation

amplitude $X_{echo}(j,k)$ attenuates in the index form, it can be forecasted as follows :

$$X_{echo}(j,k) = \max\{\lambda^n \mid X(j-n,k)\mid\}, 0<\lambda<1, \quad n=1,2,\cdots . \tag{10}$$

Here, $\lambda = e^{-\tau_{fe}/\tau}$ is the reverberation attenuation speed. Then can get the same result by using recursive sequence calculation as follows:

$$X_{echo}(j,k) = \max\{\lambda X_{echo}(j-1,k),\ \lambda \mid X(j-1,k)\mid\}. \tag{11}$$

Then the signals gained from below formula can be considered as signals without reverberation

$$\frac{|X(j,k)|}{X_{echo}(j,k)} > th. \tag{12}$$

Here, th is the threshhold value ascertained already, thus the cross power spectrum of the signals can be estimated is

$$G_{x_1x_2}(k) = \sum_{\substack{j \\ echo-free(j,k)}} X_1(j,k)X_2^*(j,k). \tag{13}$$

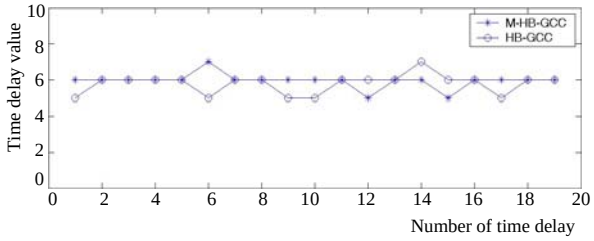
Through the cross power spectrum got from the above formula, HB weighted method can get the unsensitive time delay estimation method for reverberation.

The advantage of HB weighted method based on onset signals is that the time distinguishing rate is higher and the calculation is less, even if the reverberation is strong, the method can also estimate the time delay right, and has better robust to background noise and reverberation.

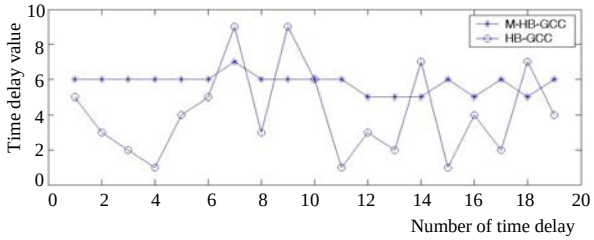
6. Experiment result

In our experiment, the sampling frequency of the speech signals is 16KHz, and each frame is 512 dots, the size of the room is 5m x 6m x 3m, smooth coefficient α is 0.06, time delay for first arrived reflected signal τ_{fe} is 0.8, the programme is written with Matlab6p1 language.

Fig. 3 shows the results of two time delay estimation under different reverberation situation, the row axis denotes the time delay estimation times, and the column axis denotes the value of the time delay estimation, HB-GCC denotes traditional generalized correlation HB weighted time delay estimation method, M-HB-GCC denotes generalized correlation HB weighted time delay estimation method based on onset signals recommended by this paper, i.e improved generalized correlation HB weighted time delay estimation method. From Fig. 3(a) we can see that under the environment without reverberation, two methods canl both get better time delay estimation result. From Fig. 3(b) we can see that when the reverberation time is 200ms, the traditional HB-GCC method cannot get right time delay estimation, but the improved method in this paper still has very good performance.



(a) When the reverberation time is 0ms



(b) When the reverberation time is 200ms

Fig. 3 The time delay estimation results of two methods

Table 1 and Table 2 show the comparing results between the time delay estimation average value and variance gained from two methods mentioned above, seperatly. From table 1 and table 2, we can see that two methods can both get very good performance without reveberation, but traditional HB-GCC method is more simple. Under reverberation environment, the traditional HB-GCC method cannot estimate the time delay accurately, but the method proposed still has very good performance.

7. Conclusion

This paper illustrates a new HB time delay estimation method based on onset signals under the reverberation environment. This method estimates the signals cross power spectrum by using the onset signals gained from the reverberation index attenuation model and fast Fonrier transform, then compensates the cross power spectrum of the noise, finally adopts HB method to estimate time delay. The experiment result shows the validity of this method, and compared with the method that doesn't consider reververation, the result has apparent improvement.

Table 1: the TDE average value by using the traditional HB-GCC method and the proposed method (M-HB-GCC)

Without reverberation (γ is 0)		Reverberation time is 200 ms (γ is 0.74s)	
HB-GCC	M-HB-GCC	HB-GCC	M-HB-GCC
5.9474	5.7895	5.5789	4.1053

Table 2: the TDE variance by using the traditional HB-GCC method and the proposed method (M-HB-GCC)

Without reverberation (γ is 0)		Reverberation time is 200 ms (γ is 0.74s)	
HB-GCC	M-HB-GCC	HB-GCC	M-HB-GCC
0.4047	0.5353	0.5353	2.5363

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant No.60372082 and No. 60172073, Trans-Century Training Program Foundation for the Talents by the Ministry of Education of China.

References

[1] T. Nishiura, R. Gruhn and S. Nakamura, "Collaborative steering of microphone array and video camera toward multi-lingual tele-conference through speech-to-speech translation," IEEE Workshop on Automatic Speech Recognition and Understanding, Trento, Italy, pp.119-122, 2001.

[2] J. Meyer and K. U. Simmer, "Multi-channel speech enhancement in a car environment using wiener filtering and spectral subtraction," IEEE Trans. on Acoust. Speech, and Signal Processing, vol.21, pp.1167-1170,1997.

[3] B. Widrow and F. L. Lou, "Microphone array for hearing aids: an overview," Speech Communication, Jan. vol.39, pp. 139-146, 2003.

[4] Liyan Zhang, Fuliang Yin and Daiwen Hou, "SVD-based beamforming speech enhancement algorithm," Proc. of the International Conference on Electronic Measurement & Instruments, Beijing, China, vol.3, pp. 493-497, 2005.

[5] Dongxia Wang and Fuliang Yin, "A subband adaptive learning algorithm for microphone array based speech enhancement." Lecture Notes in Computer Science, Berlin: Springer, pp. 592-597, 2005.

[6] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," IEEE Trans. ASSP, vol.24, pp.320-327, 1976.

[7] Xiaohong Ma, Xiaoyan Lu and Fuliang Yin, "Time delay estimation by using modified crosspower spectrum phase technique." J. Electronics & Infor. Technology, vol.26, pp. 53-59, 2004 (in Chinese).

[8] J. C. Hassab and R. E. Boucher, "Optimum estimation of time delay by a generalized correlation." IEEE Trans. on ASSP, vol. 27, pp. 373-380, 1979.

[9] PictureTel. A dynamic location camere. http://www.picturetel.com/pdf/MTPicturetel80.pdf.

[10] Y. Huang, J. Benesty and G. W. Elko, "Adaptive eigenvalue decomposition algorithm for real time acoustic source localization system." IEEE International conference on acoustic, speech and signal processing, vol.2, pp.937-940, 1999.

[11] Y. Huang, J. Benesty, and G. W. Elko, "Microphone arrays for video camera steering," in Acoustical Signal Processing for Telecommunication, Kluwer Academic Publishers, Chapter 11, pp.239-257, 2000.

[12] L. Ziomaek, Fundamentals of Acoustic Field Theory and Space-Time Signal Processing. Boca Raton, FL: CRC Press, pp.210-248, 1995.

[13]J.Huang, N.Ohnishi and N.Sugie, "sound localization in reverberant environment based on the model of the precedence effect," IEEE trans. on Instr. and measurement. vol.46, pp.842-846, 1997.

[14] Huang Jie, "A model-based sound localization system and its application to robot navigation." Robotics and autonomous systems, vol.27, pp. 109-209, 1999.

Liyan Zhang received the B.S. and M.S. degrees from Harbin Engineering Univ. in 1998 and 2000, respectively. After graduation She is working as a teacher in Dalian Jiaotong Univ. and now is working for her Dr. degree at Dalian Univ. Technology. Her research interest includes acoustic signal processing, array signal processing. She is a member of CIC China.

Fuliang Yin received the B.S. and M. S. degree from Dalian Univ. of Technology in 1984 and 1987, respectively. From 1991 to 1994, he was an Associate Professor at Dalian Univ. of Technology. And he has been a Professor since 1995 and the dean of School of Electronic and Informatin Engineering Since 2000. His research interest includes the theories and application of the digital signal processing, the acoustic/array signal processing and the pattern recognition and digital communication, He is a member of CIC, CIE, CESE, and STCME China.