Simplification and subjective evaluation of filters for virtual sound using loudspeakers

XIE Bo-sun[1], ZHANG Lin-shan[1], GUAN Shan-qun[2], ZHANG Cheng-yun[3]

(1. Acoustics Laboratory, School of Physics, South China University of Technology, Guangzhou, 510641, China; 2. Telecommunication College, Beijing University of Posts and Telecommunications, Beijing, 100876, China; 3. Guangzhou University, School of Information and Mechanical Electronics Engineering, Guangzhou, 510006, China)

Abstract: The filters of virtual reproduction for 5.1 ch surround sound are simplified, and the method of reducing the impulse response of filters by poles and zeros cancellation is proposed. The results of subjective evaluation experiment show that, at the sampling frequency of 48 kHz, filters with 128 points (2.7 ms) impulse response produce satisfying performances. Filters with 64 points (1.3 ms) impulse response produce a little degraded performance, but in practical they are still usable. Therefore the impulse response of proposed filters are shorter than that of traditional design (5-10 ms), and the resulted filters can be implemented directly by FIR structure in time domain.

Key words: virtual sound; surround sound; head-related transfer function (HRTF); filter

采用扬声器的虚拟声滤波器简化及其主观评价

谢菠荪[1]，张林山[1]，管善群[2]，张承云[3]

(1. 华南理工大学物理科学与技术学院声学研究所，广州 510641；2. 北京邮电大学电信工程学院，北京 100876；3. 广州大学信息与机电工程学院，广州 510006)

摘要：对用扬声器的 5.1 通路环绕声虚拟重放系统的滤波器进行了简化，提出通过零极点的抵消可减少滤波器的脉冲响应长度。主观评价实验表明，在 48 kHz 的采样频率下，采用 128 点(2.7 ms)脉冲响应长度的滤波器进行处理可得到理想的主观听觉效果。采用 64 点(1.3 ms)脉冲响应长度的滤波器会导致主观听觉效果略为下降，但实际中仍然可用。因而简化滤波器的脉冲响应长度较传统方法的 5 到 10 ms 的量级要短，采用简单的 FIR 滤波器即可实现。

关键词：虚拟声；环绕声；头相关传输函数；滤波器

中图分类号：O42　　文献标识码：A　　文章编号：1000-3630(2006)-06-0547-08

1 INTRODUCTION

By filtering the signal with HRTFs (head-related transfer functions), virtual sounds recreate the same sound pressure at two ears as that of true sources and bring the listener with the effect of spatial auditory. Virtual sounds have the advantage of simple in structure and more natural in the reproduced spatial auditory, hence they are used widely in virtual reality, the research on psychoacoustics and domestic sound reproduction.

There are two classes of virtual sound, i.e., virtual sound using headphone and virtual sound us-

Received Aug. 31, 2006; Revised Oct. 30, 2006

* This work is supported by National Natural Science Foundation of China (1037-4031)

XIE Bosun (1960-), Professor, Research areas: Electroacoustics and psychoacoustics. Email: phbsxie@scut.edu.cn Phone: 020-87113191 ing loudspeakers. Accordingly, the signal processing algorithms of them are somewhat different. Usually, signal processing for a virtual sound is implemented by various finite or infinite impulse response (FIR or IIR) filters. There have been a lot of works on the design and simplification of fil ters for virtual sound[1,2]. However, for virtual sound using loudspeakers, due to the recursive structure of cross-talk canceling, the impulse response of the filters is in the order of 5 to 10 ms, or about 256 to 1024 points at 48 kHz sampling rate[3]. Hence it is difficult to implement filters with such a length of impulse response in real time by FIR structures in time domain. Implementing FIR filters in frequency domain by FFT can improve the efficiency, but introduce delay on the output signals. In practice, the filters are often implemented by IIR structure. But an IIR structure is complicated, moreover, it may cause the problem of stability.

A multichannel sound requires more channels so that it is complicated. For example, 5.1 ch (chan nel) surround sound has been recommended by ITU as the standard of multichannel sound[4] and (or will be) used in DVD, DTV and home theatre. There are five independent channels with full frequency range in 5.1 ch system, including left L, right R, center C, left surround LS, right surround RS, plus a low frequency effect channel LFE. One of the important applications of virtual sound is the virtual reproduction for multichannel sound (called virtual surround sound commercially). By using the method of virtual sound source and from a pair of headphone or front loudspeakers, virtual reproducing systems create virtual loudspeakers of multichannel sound and reproduce the spatial auditory similar to that of multichannel sound, resulting in simplifying of multichannel sound. However, in virtual reproduction, multichannel signals are required to be processed in real time, therefore the simplification of filters is significant.

In the following, combined with the application to virtual reproduction for 5.1 ch surround sound, the simplification of signal processing for virtual sound using loudspeakers are analyzed, and the resulted filter are evaluated by subjective experiment.

2 PRINCIPLE OF VIRTUAL REPRODUCTION FOR MULTICHANNEL SOUND

Coordinate in horizontal plane is chosen as 0° θ<360°, θ=0° is the front, and θ=90° is the right. If the transfer functions (HRTFs) for a single source in azimuth θ to two ears are denoted by $H_L(\theta,\omega)$ and $H_R(\theta,\omega)$, then the pressure at two ears can be written as (where P_0 is a constant):

$$P_L=H_L(\theta,\omega)P_0 \qquad P_R=H_R(\theta,\omega)P_0 \tag{1}$$

As shown in Fig.1, for virtual sound using loudspeakers, the four transfer functions from the left and right loudspeakers to two ears are H_{LL}, H_{LR}, H_{RL} and H_{RR} respectively. By left-right symmetry, we can assume that $H_{LL}=H_{RR}=\alpha$ and $H_{LR}=H_{RL}=\beta$ It can be proved that if a mono signal E_0 is filtered by Eq.(2) and then reproduced by a pair of loudspeakers[5], the pressures at two ears are identical to (or proportional to) Eq.(1), resulting in a virtual sound image at azimuth θ:

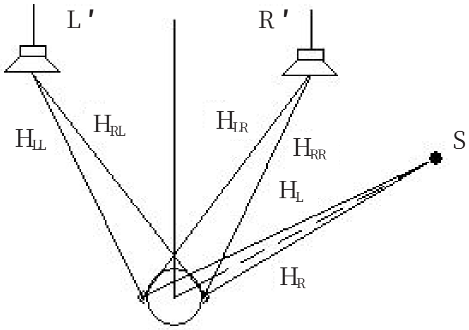

Fig.1 Virtual sound using loudspeakers

$$L'=A(\theta,\omega)E_0 \qquad R'=B(\theta,\omega)E_0 \tag{2}$$

$$A(\theta,\omega)=\frac{\alpha H_L-\beta H_R}{\alpha^2-\beta^2}, \quad B(\theta,\omega)=\frac{-\beta H_L+\alpha H_R}{\alpha^2-\beta^2} \tag{3}$$

However, the filters given by Eq.(2) and Eq.(3) are recursive, and the poles of transfer functions $A(\theta,\omega)$, $B(\theta,\omega)$ depends on the zeros of denominator$(\alpha^2-\beta)$. After taking the transform z=exp $(j\omega)$, as seen, it is the poles of transfer functions (especially the poles near the unit circle z=exp$(j\omega)$

in the Z plane) that cause a long impulse response and make the signal processing difficult. Moreover, the poles (as well as zeros) in Eq.(3) change the frequency spectrum of signals, consequently, result in distortion in frequency domain and subjective timbre change in reproduction. This is a defect of virtual sound.

Recently, an algorithm of timbre equalization has been incorporated into virtual sound using loudspeakers[6]. If $A(\theta,\omega)$ and $B(\theta,\omega)$ in Eq.(2) are replaced by $A'(\theta,\omega)$ and $B'(\theta,\omega)$ given in Eq.(4), a virtual sound image at azimuth θ can also be recreated:

$$A'(\theta,\omega)=\frac{\alpha H_L-\beta H_R}{\sqrt{|\alpha H_L-\beta H_R|^2+|-\beta H_L+\alpha H_R|^2}}\frac{|\alpha^2-\beta^2|}{\alpha^2-\beta^2}$$

$$B'(\theta,\omega)=\frac{-\beta H_L+\alpha H_R}{\sqrt{|\alpha H_L-\beta H_R|^2+|-\beta H_L+\alpha H_R|^2}}\frac{|\alpha^2-\beta^2|}{\alpha^2-\beta^2} \quad (4)$$

It is easy to prove that the signals given by Eq.(2) and Eq.(4) meet the following equation:

$$|L'|^2+|R'|^2=E_0^2 \quad (5)$$

The total output power in virtual reproduction equals to that of original mono signal, therefore the algorithm of timbre equalization is also called the algorithm of power equalization.

It has also been proved that[7], the poles and zeros (near the unit circle) in filters given by Eq.(4) are cancelled effectively. Then the timbre change in reproduction is reduced. If pinna-less HRTFs measured by the blocked-ear-cannel method is used, the notches (zeros) in the transfer functions given by Eq.(4) can be eliminated, resulting in a further reduction in timbre change in reproduction. Results of psychoacoustic experiment have shown that there are no significantly different between the localization results for HRTFs without pinnae and that for HRTFs with pinnae and ear cannel[8]. Furthermore, cancellation of poles and zeros in transfer functions shortens the impulse response and results in a simplification on signal processing. This is the starting point of our work.

The principle of virtual sound image can be applied to virtual reproduction for multichannel sound. For a normal 5.1 ch surround sound, five independent signals are fed to five loudspeakers respectively. The azimuths of five loudspeakers are:

$$\theta_L=330° \quad \theta_R=30° \quad \theta_C=0° \quad \theta_{LS}=250° \quad \theta_{RS}=110° \quad (6)$$

In the virtual reproducing system for 5.1 ch surround sound proposed by us recently[6], five dependent signals are processed and then fed to a pair of symmetrical front loudspeakers. Assuming that the real loudspeakers are arranged in the direction of θ_0 and 360°-θ_0 respectively, the transfer function from loudspeakers to two ears are shown in Fig.1, and $H_{LL}=H_{RR}=\alpha$, $H_{LR}=H_{RL}=\beta$ The signals fed to two real loudspeakers are:

$$\begin{aligned}L'=&A'(\theta_L,\omega)L+A'(\theta_R,\omega)R+0.707C+\\&A'(\theta_{LS},\omega)LS+A'(\theta_{RS},\omega)RS+0.707LFE \quad (7)\\R'=&B'(\theta_L,\omega)L+B'(\theta_R,\omega)R+0.707C+\\&B'(\theta_{LS},\omega)LS+B'(\theta_{RS},\omega)RS+0.707LFE\end{aligned}$$

where $A'(\theta,\omega)$ and $B'(\theta,\omega)$ are given by Eq.(4).

As shown, original signal C is attenuated by-3 dB and fed to left and right loudspeakers simultaneously. From Eq.(7), if signals L=R=LS=RS=0, then $L'=R'=0.707C$, an image in the front of $\theta=0°$ will be recreated.

For original signal R, if the other signals C=L=LS=RS=0, signals fed to loudspeakers are $L'=A'(\theta_R,\omega)R$ and $R'=B'(\theta_R,\omega)R$. Compared with Eq.(2) and Eq.(4), it can be seen that a sound image at $\theta_R=30°$ is recreate by R signal. In other words, signal R is reproduced by a virtual loudspeaker in $\theta_R=30°$.

Similarly, signals L, LS and RS are reproduced by virtual loudspeakers in θ_L, θ_{LS} and θ_{RS} respectively. As a whole, if five original signals are processed by Eq.(7) and reproduced by a pair of front loudspeakers, the effect is equivalent to that reproduced by five virtual loudspeakers, resulting in a perceived effect similar to that of original 5.1 ch surround sound.

There are three remarks:

(1) In the proposed system, a pair of real loudspeakers are arranged in 15° and 345° respectively. The span of loudspeaker pair is 30°, which is less

than traditional standard of 60°. On one hand, a narrowed span of loudspeakers can improve the stability of sound image and enlarge the listening area[6][9]. On the other hand, in some practical use, for example, TV set and multimedia computer, the span of loudspeakers pair is inherently less than 60°. Therefore a narrow span of loudspeakers is preferable.

(2) In virtual sounds, the sound image range recreated by a pair of front loudspeakers is not better than 90° or 270°. Trying to recreate sound image in the back of horizontal plane is unpractical, and the image intended for the back will appear in the mirror direction of front. As a kind of com promise, two virtual surround loudspeakers should be set in the direction of 90° or 270°. This is also allowed in a practical 5.1 ch surround sound reproduction.

(3) Algorithm of timbre equalization incorporated into Eq.(7) can also cancel the poles and zeros which is near the unit circle.

3 SIMPLIFICATION OF THE SIGNAL PROCESSING

In Eq.(7), there are eight multiplying opera tions in frequency domain. Accordingly, eight digital filters are needed. However, from the leftright symmetry[5], we have $A'(\theta_L, \omega)=B'(\theta_R, \omega)$, $A'(\theta_R, \omega)=B'(\theta_L, \omega)$, $A'(\theta_{LS}, \omega)=B'(\theta_{RS}, \omega)$, $A'(\theta_{RS}, \omega)=B'(\theta_{LS}, \omega)$, then Eq.(7) is equivalent to the following equation:

$$\begin{bmatrix} L' \\ R' \end{bmatrix}=0.707\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\left\{\begin{bmatrix} 1 \\ 0 \end{bmatrix}(C+LFE)+\begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Delta_1 \end{bmatrix}\cdot\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} L \\ R \end{bmatrix}+\begin{bmatrix} \Sigma_2 & 0 \\ 0 & \Delta_2 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} LS \\ RS \end{bmatrix}\right\} \quad (8)$$

where

$$\begin{aligned} \Sigma_1&=0.707[A'(\theta_L, \omega)+A'(\theta_R, \omega)] \\ \Delta_1&=0.707[A'(\theta_L, \omega)-A'(\theta_R, \omega)] \\ \Sigma_2&=0.707[A'(\theta_{LS}, \omega)+A'(\theta_{RS}, \omega)] \\ \Delta_2&=0.707[A'(\theta_{LS}, \omega)-A'(\theta_{RS}, \omega)] \end{aligned} \quad (9)$$

In Eq.(8), signals L and R or LS and RS are firstly multiplied with a 2×2 MS matrix, then multiplied with Σ_1 and Δ_1 or Σ_2 and Δ_2 in frequency domain respectively; after mixing with signal C and LFE, the signals are multiplied with another 2×2 MS matrix and a scale factor of 0.707, resulting in signals L′ and R′.

In Eq.(8), there are four multiplying operations in frequency domain, then four filters are needed. Hence signal processing algorithm is primarily simplified. Since the characters of filters are completely determined by four transfer functions Σ_1, Δ_1, Σ_2 and Δ_2, a further simplification of signal processing algorithm relies on the simplification of filters design.

Eq.(8) can also be converted into time domain:

$$\begin{bmatrix} l' \\ r' \end{bmatrix}=0.707\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\left\{\begin{bmatrix} 1 \\ 0 \end{bmatrix}(c+lfe)+\begin{bmatrix} \sigma_1 & 0 \\ 0 & \delta_1 \end{bmatrix}*\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} l \\ r \end{bmatrix}+\begin{bmatrix} \sigma_2 & 0 \\ 0 & \delta_2 \end{bmatrix}*\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} ls \\ rs \end{bmatrix}\right\} \quad (10)$$

where "*" represents convolution, and all the symbols in Eq.(10) are connected with corresponding symbols in Eq.(9) by inverse Fourier transform.

The HRIRs (time domain representation of HRTFs) measured from a KEMAR artifical head without pinna by the blocked-ear-cannel method are used. The procedure of measurement is given in Ref[10]. The original HRIRs are at the sampling frequency of 44.1 kHz, 16 bit quantization, and 512 points in length. In order to fit the use in DVD, the sampling frequency of original HRIRs are converted to 48 kHz by interpolation. The transfer functions Σ_1, Δ_1, Σ_2 and Δ_2 with 1024 frequency points are calculated from Eq.(9) by padding zeros at the end of HRIRs. After an inverse Fourier transform, the resulted four impulse responses at sampling frequency of 48 kHz and with the length of 1024 points, namely $\sigma_1(n)$, $\delta_1(n)$, $\sigma_2(n)$, $\delta_2(n)$, n=0,1…1023, are given. These four impulse responses are used as references of the filter design.

Take $\sigma_1(n)$ as example. The impulse response with 1024 points is shown in Fig.2. As shown, the main part of the impulse response lasts about 50 to 60 samples. The amplitude of the beginning and ending part of the impulse is very small. As mentioned above, timbre equalization incorporated into Eq.(4) cancels the poles and zeros, and thereby shortens the impulse response of filters. Therefore a rectangular window can be used to truncate $\sigma_1(n)$. The rectangular window is:

$$W(n)=\begin{cases}1 & N_1\leq n\leq N_2\\ 0 & \text{others}\end{cases} \qquad (11)$$

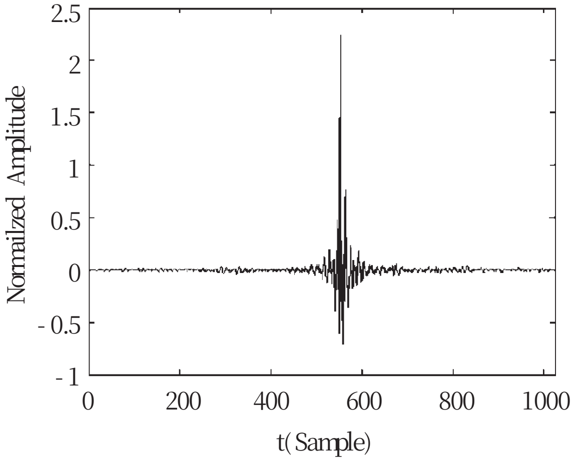

Fig. 2 The impulse response $\sigma_1(n)$ with 1024 points

The relative error in energy caused by truncation is calculated by:

$$\mathrm{Err}=10\lg\frac{\sum_{n=0}^{N_1-1}\sigma_1^2(n)+\sum_{n=N_2+1}^{1023}\sigma_1^2(n)}{\sum_{n=0}^{1023}\sigma_1^2(n)} \quad (\mathrm{dB}) \quad (12)$$

Two rectangular windows with different width are used:

(1) 128 points rectangular window ($N_1=501$, $N_2=628$):

(2) 64 points rectangular window ($N_1=527$, $N_2=590$):

Similarly, the other three impulse response $\delta_1(n)$, $\sigma_2(n)$ and $\delta_2(n)$ can also be truncated by rectangular window. The errors caused by truncation are shown in Table 1. It can be seen that relative errors is less than - 18.3 dB for 128 points rectangular window, and less than - 13.9 dB for 64 points rectangular window.

Table 1 Error caused by truncation

Err (dB)	$\sigma_1(n)$	$\delta_1(n)$	$\sigma_2(n)$	$\delta_2(n)$
128 point	- 18.3	- 21.5	- 18.3	- 19.1
64 point	- 14.8	- 19.5	- 13.9	- 15.7

Truncating the impulse response by windows reduces the resolution in frequency. Fig.3 shows the magnitude spectra of $\sigma_1(n)$ with length 1024, 128 and 64 points respectively (for $\sigma_1(n)$ less than 1024 points, before calculating magnitude spectra by FFT, some zeros have been padded at the end of the impulse). As seen, truncating the impulse response can also smooth the magnitude spectra. Fortunately, the frequency resolution of human auditory is in herently poor at high frequencies, and the magnitude spectra of 1024 points $\sigma_1(n)$ is flat at low frequencies, therefore smoothing of the magnitude spectra results in insignificant effect in auditory. Analyses on other three filters give similar conclusions.

In conclusion, the length of impulse response of filters can be truncated to 128 or 64 points. Filters with such a length of impulse response can be implemented directly by FIR structure in time domain, and FFT calculation is not needed. Therefore the signal processing algorithm is simplified. of course, the final result should be evaluated by subjective experiments.

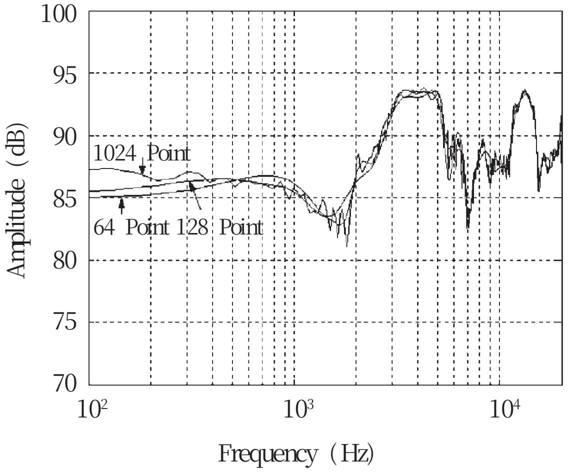

Fig.3 The magnitude spectra of $\sigma_1(n)$

4 SUBJECTIVE EVALUATION EXPERIMENT

A subjective experiment is carried out to eval uate the performance of simplifying the filters design. The key to evaluate the virtual reproduction of 5.1 ch surround sound is to evaluate the effect of virtual loudspeakers, i.e., to evaluate the perceived direction and timbre of virtual loudspeakers. Thus, in Eq.(10) only one of the original input signals l, r, ls or rs is remained, and others are set to zeros. Accordingly, the perceived direction and timbre of each virtual loudspeaker are evaluated. Since original signals c and lfe are attenuated by -3 dB and fed to left and right loudspeakers simultaneously, the result is identical to that of stereophonic down mixing and is familiar to all. Hence it is unnecessary to evaluate the effect of signal c repeatedly. Furthermore, due to left-right symmetry, only the r and rs signals and corresponding virtual loudspeakers are evaluated.

The experiment is carried out in a listening room with reverberation time of 0.15 s. A pair of monitor loudspeaker systems (Genelic 1032A) are arranged in a circle with radius 2.0 m, at the azimuth of 15° and 345° respectively. Listener is seated at the center of the circle, with his (her) ear is at the level of loudspeakers. Mono signal is used as the original input signals of r or rs, then processed by software according to Eq.(10), the resulted l′ and r′ signals are passed though a sound card (Echo Layla 24), preamplifier, and then fed to the loudspeakers.

Nine kinds of audio stimuli are used in the experiment, including 1/3 oct noise with center frequency of 125 Hz, 250 Hz, 500 Hz, 1 kHz, 2 kHz, 4 kHz and 8 kHz respectively, Chinese speech (male), music (a section of orchestra, J. Strauss, The blue Danube). The pressure level at the center of the circle in reproduction is about 70 to 75 dBA.

In order to evaluate the effect of simplified filters, a 3AFC (three-alternative forced-choice) listening experiment is adopted. For a given virtual loudspeaker (for example, θ_R=30°) and given type of stimulus (for example, music), the signal processed by filters with 1024 points impulse response is used as a reference. While the signal processed by filters with 64 points impulse response is used as a comparison. There are three parts on each trail of evaluated signals. The first part is always the reference, the second and third parts are the reference and the comparison, with two different orders, i.e., "reference-reference-comparison", or "reference-comparison-reference".

Firstly, only the reference signal is reproduced and the listener determines the direction of virtual loudspeaker (sound image). Then three parts of signal on each trail are reproduced one after the other. According to the subjective difference (on

Table 2 Statistical results for 3AFC experiments

		30° Virtual loudspeaker		90° Virtual loudspeaker	
		Mean perceived direction and standard deviation	Rate of correction in 3AFC	Mean perceived direction and standard deviation	Rate of correction in 3AFC
Stimuli	Speech	29.3 (1.0°)	0.47	68.1 (14.0°)	0.91
	Music	28.8 (1.9°)	0.50	72.1 (16.4°)	0.50
	125 Hz	32.9 (6.1°)	0.38	91.2 (5.8°)	0.50
	250 Hz	30.1 (4.4°)	0.50	82.5 (7.6°)	0.81
	500 Hz	29.3 (1.8°)	0.47	81.2 (13.8°)	0.72
	1 kHz	28.9 (3.4°)	0.66	86.9 (4.6°)	0.53
	2 kHz	-----	0.63	-----	0.56
	4 kHz	-----	0.53	-----	0.53
	8 kHz	-----	0.56	-----	0.50

image direction and timbre etc.) among three parts, listener tells the comparison signal from the second or the third part of the trail. If listener is unable to distinguish the comparison signal, he (she) should select the answer randomly. Each kind of signal with two different orders are reproduced twice in random order, therefore there are four judgments for each listener. Eight listeners take part in the experiments. For given virtual loudspeakers and stimulus, there are 32 judgments in total.

For R and RS virtual loudspeakers, after all nine stimuli are evaluated, the results are analyzed by statistical methods. The mean and standard deviation of perceived directions of virtual loudspeakers, as well as the correct rates of 3AFC experiment are calculated. Table 2 shows the results. For the stimuli of 1/3 oct noise with center frequency 2 kHz, 4 kHz and 8 kHz, the perceived directions of virtual loudspeakers are very sensitive to the position of head, therefore the results are omitted in Table 2. In fact, since wavelength of sound is short at high frequencies, a slight movement of head results in an obvious change in the pressures at two ears (especially in phase). This is a common defect of all virtual sound systems.

From Table 2, it can be seen:

(1) For right virtual loudspeaker intended at $\theta_R=30^\circ$, the mean perceived directions for six stimuli are near 30°, the difference is slight. Except for the stimulus of 125 Hz, the standard deviations of perceived direction are also small. A somewhat large in the standard deviations of perceived direction for 125 Hz may be due to the poor localization ability of human auditory at low frequencies.

(2) For right surround virtual loudspeaker intended at $\theta_{RS}=90^\circ$, and for six stimuli, the mean perceived directions are somewhat different from the intended directions. Especially for speech and music, the mean perceived directions are about 70°. Compared with intended direction, the perceived directions move a little bit towards the front. Moreover, the standard deviations of perceived direction are large. Two reasons are responsible for these. Firstly, speech and music include some spectral components with frequency above 2 kHz, accordingly at high frequencies the perceived directions of virtual loudspeaker are very sensitive to the position of head. Then, the nonindividual HRTFs (from KEMAR artifical head) are used in the signal processing. An unmatched listener's head size with KEMAR is likely to cause the directional distortion in lateral virtual loudspeaker[11]. Individual HRTFs can improve the performance.

(3) According to the U test in statistics, at 0.05 level, if the correct rate in a 3AFC experiment is larger than 0.65, the reference and comparison signals are significantly different. For right virtual loudspeaker in $\theta_R=30^\circ$ and for eight of nine stimuli, the correct rates are less than 0.65. While for stimulus of 1 kHz, the correct rate is 0.66. However, all the listeners report that they are unable to distinguish the comparison signal and they select the answer randomly. Therefore a correct rate of 0.66 is due to the fluctuation in statistics.

(4) For virtual loudspeaker in $\theta_{RS}=90^\circ$ and for stimuli of 250 Hz, 500 Hz as well as speech, the correct rates of 3AFC experiment are large than 0.65. Especially, for speech, the correct rate is as high as 0.91. In fact, considerable amount of energy of speech are distributed within the frequency range of 250 Hz to 500 Hz. Therefore, in this frequency range, the reference and comparison signals are significantly different in auditory. Moreover, listeners report that the perceived directions of virtual loudspeakers move forward (in a order of 5°) for comparison signals.

As shown, if the impulse responses of filters for virtual loudspeakers are reduced to 64 points, for virtual loudspeaker in $\theta_{RS}=90^\circ$ and for stimuli of 250 Hz, 500 Hz as well as speech, the subjective performances are degraded. For comparison, a complementary 3AFC experiment for virtual loudspeaker in $\theta_{RS}=90^\circ$ and for stimuli of 250 Hz, 500 Hz as well as speech is carried out. The experimental condi tion is identical to above except that the comparison signals are processed by filters with 128 points impulse response. Statistical results show that the correct rates of 3AFC experiment are 0.59, 0.50 and 0.59 respectively, all of them are less than 0.65.

Therefore, the subjective performances of filters with 128 points impulse response are identical to those of filters with 1024 points impulse response. There is no difference in auditory between them Although filters with 64 points impulse response cause degradation for virtual loudspeaker in $\theta_{RS}=90°$ and for part of the stimuli (cause the virtual loudspeaker move forward), they are still usable.

5 CONCLUSION

Theoretical and experimental results show that, in virtual reproduction for 5.1 ch surround sound using loudspeakers, algorithm of timbre equalization can cancel the poles and zeros near the unit circle in signal processing functions, thereby reduce the impulse response of filters and simplify the signal processing. The results of subjective evaluation experiment show that, at the sampling frequency of 48 kHz, the performances of filters with impulse response 128 points (about 2.7 ms) are identical to that of filters with impulse response 1024 points (about 21.3 ms). The subjective performances of filters with impulse response 64 points (about 1.3 ms) are somewhat degraded, but in practice the filters are still usable. Therefore the filters used for virtual sound proposed in this paper are simpler than that of traditional design (with impulse response 5-10 ms), and can be implemented directly by FIR structure in time domain. This is convenient for practical applications.

It should be pointed out that virtual reproduction for 5.1 ch surround sound are taken as an example in the discussion of this paper. However, methods developed in this paper can be extended to the virtual reproduction for other multichannel sound, and more generally, to all virtual sound using loudspeakers.

References

[1] Huopaniemi J, Zacharov N. Objective and subjective evaluation of head-related transfer function filter design [J]. J.Audio.Eng.Soc., 1999, 47(4): 218-239.

[2] Huopaniemi J, Karjalainen M, Review of digital filter design and implementation methods for 3D sound[A]. AES 102th Convention[C]. Preprint 4461, Munich, Germany, 1997.

[3] Jot J M. Digital signal processing issues in the context of binaural and transaural stereophony[A]. AES 98th Convention[C]. Preprint 3980, Paris, France, 1995.

[4] ITU-R Rec. BS 775-1, Multichannel stereophonic sound system with and without accompanying picture, Doc 10/63, Geneva, Switzerland, International Telecommunication Union, 1994.

[5] Bauck J, Cooper D H. Generalization transaural stereo and applications[J]. J Audio Eng Soc, 1996, 44(9): 683-705.

[6] Xie Bosun. Virtual reproducing system for 5.1 channel surround sound[J]. Chinese Journal of Acoustics, 2005, 24(1): 76-88.

[7] HE Pu, XIE Bosun, RAO Dan. Subjective and objective analysis of timber equalized algorithms for virtual sound reproduction with loudspeakers[J]. Applied Acoustics (in Chinese), 2006, 25(1): 4-12.
何璞, 谢菠荪, 饶丹. 虚拟声音色均衡信号处理方法的主客观分析[J]. **应用声学**, 2006, 25(1): 4-12.

[8] Hepu. Research on the timbre equalization of virtual sound[D]. South China University of Technology, Guangzhou, 2005.
何璞. 虚拟声音色均衡的研究[D]. **华南理工大学**, **广州**, 2005.

[9] Kirkeby O. The "Stereo Dipole": A virtual source imaging system using two closely spaced loudspeakers[J]. J. Audio Eng. Soc., 1998, 46(5): 387-395.

[10] ZHONG Xiaoli, XIE Bosun. Overall influence of clothing and pinnae on shoulder reflection and HRTF[J]. Technical Acoustics. 2006, 25(2): 113-118.
钟小丽, 谢菠荪. 衣服、耳廓对肩部反射及头相关传输函数的综合影响[J]. **声学技术**, 2006, 25(2): 113-118.

[11] XIE Bosun. Effect of head size on virtual sound image localization[J]. Applied Acoustics (in Chinese), 2002, 21(5): 1-7.
谢菠荪. 头部尺寸对虚拟声像定位的影响[J]. **应用声学**, 2002, 21(5): 1-7.