Chapter 13
Acoustic VR System

Abstract An acoustic VR system provides a three-dimensional acoustical sensation of an existing sound field (as at concert halls, stadiums, or disaster sites) and/or an imaginary sound field (as in movies or video games), independent of time and space. This chapter introduces the system configuration and the applications of acoustic VR systems.

13.1 System Configuration

An example of a typical configuration of an acoustic VR system is shown in Fig. 13.1. The system consists of hardware (a PC, a digital audio interface, headphones, earplug-type microphones, and a head-tracker), software for signal processing, and a database (HRTF and pinna shape).

The main function of the acoustic VR system is to reproduce the ear-input signals obtained in an arbitrary sound field through headphones by the signal processing described in Chap. 12. This signal processing (convolution between a sound source signal and HRIRs) is performed on the PC.

Another function is to change the HRTFs to those of another direction in response to the head movement of a listener. In order to capture the direction of the listener's head, a head tracker is used. Rewriting of the HRTFs must be done within the threshold for the detection of system delay, i.e., 80 ms (Yairi et al. 2005).

Various systems have adopted an individualization function for HRTFs to ensure accurate sound image localization. An example system is shown in Fig. 13.2.

The external specifications of the acoustic VR system, i.e., the Sound Image Reproduction system with Individualized-HRTF, graphical User-interface and Successive head-movement tracking (SIRIUS), which was developed in the author's lab, are shown in Table 13.1.

This system runs on a Windows PC. An HRIR database (response length: 512 samples) is stored on the PC. A sound source signal and HRIRs are convolved in real time in order to control the direction and distance of a sound image.

K. Iida, *Head-Related Transfer Function and Acoustic Virtual Reality*,
https://doi.org/10.1007/978-981-13-9745-5_13

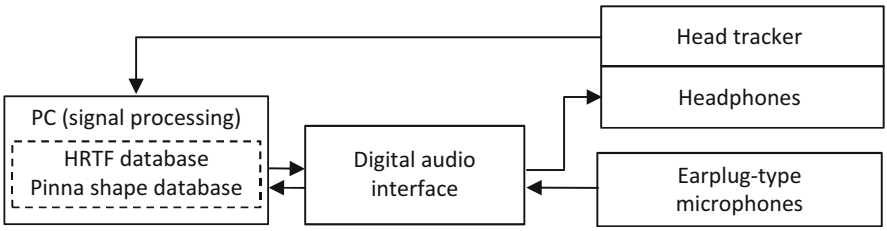

Fig. 13.1 Block diagram of three-dimensional auditory display

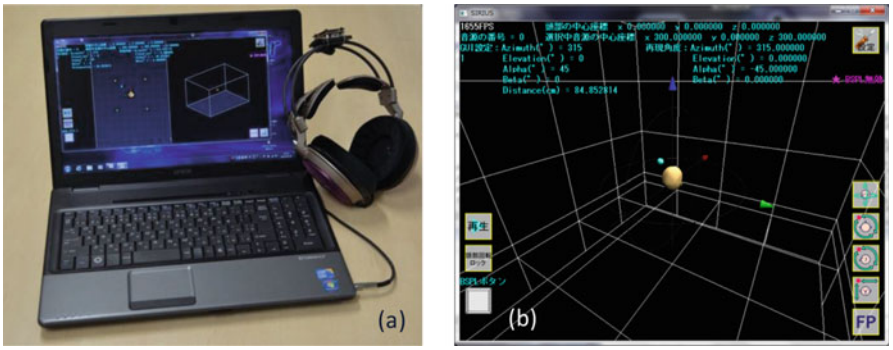

Fig. 13.2 (**a**) Photograph of system and (**b**) GUI of three-dimensional auditory display

Table 13.1 External specifications of the SIRIUS acoustic VR system

Programming language	C++, C#, MATLAB
OS	Windows 10
CPU	Core i3 2.13 GHz
Head-tracker	Acceleration (3 axes) + angular velocity (3 axes)
HRTFs	1) measured HRTFs
	2) measured median plane HRTFs + ITD
	3) parametric median plane HRTFs + ITD
Azimuth (resolution)	0° to 360° (< 1°)
Vertical angle (resolution)	−90° to +90° (< 1°)
Individualization of HRTF	Selection from the minimal parametric HRTF database
Distance of a sound image	Control based on BSPL
Maximum number of sound sources	7
System delay	< 21 ms

By using a head tracker and a three-dimensional position sensor, the direction and position of the listener's head are captured, and these changes are reflected in the signal processing in real time.

The sound image direction can be controlled for the entire sky (azimuth angle: 0° to 360°, vertical angle: −90° to +90°), and the sound image distance can also be

controlled based on the BSPL (Sect. 7.2.3). By performing convolution using the overlap-add method (Sect. 10.3.2), a CPU, the clock frequency of which is 2.13 GHz, can simultaneously process up to seven sound sources. The system delay is approximately 21 ms.

13.2 Signal Processing Flow

The signal processing flow is shown in Fig. 13.3. After starting the program, initialization is performed, and the processing enters the main loop, which performs the following processes:

(1) Acquire information on the direction of the head and the position of the listener from the head tracker and three-dimensional position sensor.
(2) Calculate the relative angle and distance between the sound source and the listener based on the sound source position and the listener position set by the GUI (mouse) and the head direction.

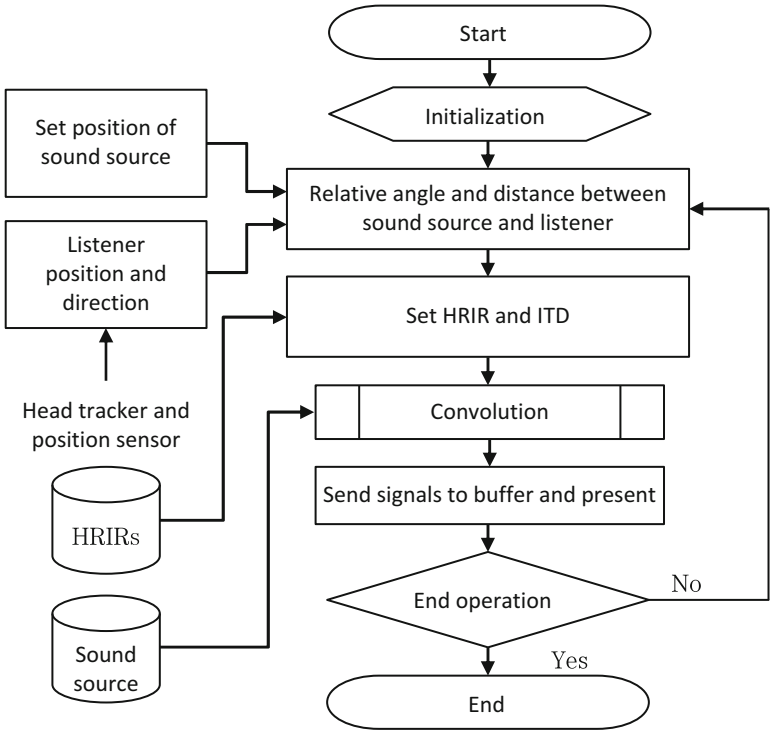

Fig. 13.3 Signal processing flow of SIRIUS

(3) Select the HRIRs corresponding to the vertical angle calculated in step (2) from the database. Then, the ITD corresponding to the lateral angle is added to the HRIRs.
(4) Perform convolution of a sound source signal and HRIRs by the overlap-add method described in Sect. 10.3.2.
(5) Send the signals calculated in step (4) to the buffer. Then, present these signals through headphones.

13.3 Application to Concert Hall Acoustics

Acoustic VR for a concert hall can be achieved by convolution of a sound source and the reflections in addition to a direct sound.

A binaural room impulse response (BRIR, see Appendixes 2 and 4) is generated by convolution of the HRIRs and the room impulse response (RIR) calculated by geometric acoustic simulation beforehand. Convolution of the BRIR and a source signal generates ear-input signals to be heard in the concert hall.

However, since the BRIR has a long response length, convolution in real time is difficult even using the overlap addition method. Therefore, a high-speed convolution algorithm using the frame division method was developed in the author's laboratory.

Figures 13.4 and 13.5 show a conceptual diagram and a flowchart, respectively, of the fast convolution algorithm. The process is as follows.

(1) Cut out the source signal every frame length (512 samples). Let the cut out blocks be denoted as S_1, S_2, . . ., S_n.

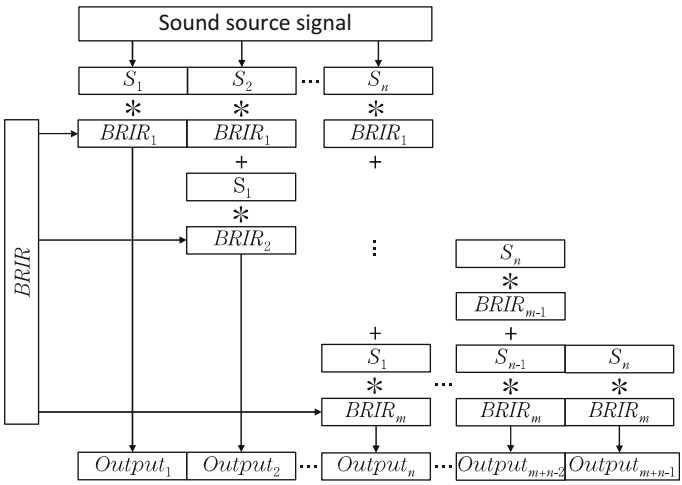

Fig. 13.4 Conceptual diagram of fast convolution algorithm using frame division

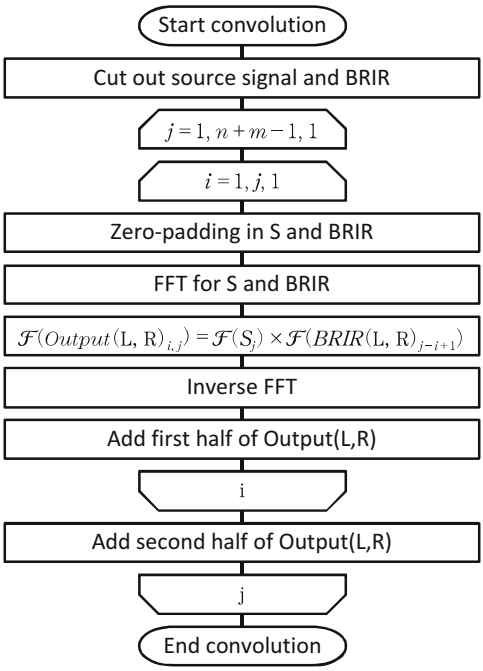

Fig. 13.5 Flow chart of fast convolution algorithm

(2) Cut out BRIR(L, R) for each frame length in the same manner. Let the cut out blocks be denoted as $\mathrm{BRIR(L,\ R)_1}$, $\mathrm{BRIR(L,\ R)_2}$, . . ., $\mathrm{BRIR(L,\ R)_m}$.
(3) In order to obtain the output signal, $\mathrm{Output(L,\ R)_j}$, of the j-th frame, calculation is performed according to Eq. (13.1).

$$Output(\mathrm{L,\ R})_j = \sum_{i=1}^{j} S_i * BRIR(\mathrm{L,\ R})_{j-i+1} \tag{13.1}$$

(4) The overlap-add method is used for the convolution of Eq. (13.1). In order to obtain the output signal, $\mathrm{Output(L,\ R)_j}$, of the j-th frame according to the algorithm of the overlap-add method, the following processing is performed.
   (a) Zeros are added in each block ($\mathrm{S_i}$, $\mathrm{BRIR(L)_{j\text{-}i\text{-}1}}$ and $\mathrm{BRIR(R)_{j\text{-}i\text{-}1}}$) so that the block is twice as long as the frame length.
   (b) An FFT is performed for each block.
   (c) Perform complex multiplication of $\mathrm{S_i}$ and $\mathrm{BRIR(L)_{j\text{-}i\text{-}1}}$, and $\mathrm{S_i}$ and $\mathrm{BRIR(R)_{j\text{-}i\text{-}1}}$, respectively.
   (d) An inverse FFT is performed for each of L and R of the signal obtained in c).
   (e) Add each of L and R of the signal obtained in d) to a temporary array (1024 samples).
   (f) The process from a) to e) is repeated, while i is increased until i = j.

(g) Add the first half (for 512 samples) of the signal obtained in f) and the second half (for 512 samples) of the signal obtained by the calculation of the previous frame (Output(L, $R)_{j-1}$).
(h) The first half (for 512 samples) of the signal obtained in g) is sent to the reproduction buffer. The output signal for reproduction is obtained by repeated calculation using the above processes.

Using such an algorithm, the CPU (core i7, 2.7 GHz, four-core, eight-thread) installed on a Windows 7 laptop PC is operated in six-thread parallel processing using OpenMP. The maximum response length of the BRIR was confirmed to be 92,160 samples (1.92 s), for which real time convolution is achieved.

13.4 Application to a Public Address System

Since an outdoor public address system simultaneously emits sound from loudspeakers installed at multiple points, multiple voices often arrive while overlapping listening points. The time difference of the incident sound is often on the order of hundreds of milliseconds or seconds, and the subsequent incident sound becomes a long-pass echo that reduces the speech intelligibility (see Chap. 8 and Appendix A.3).

The speech intelligibility can be evaluated directly if a speech delivered by an existing or designed outdoor public address system can be auralized. In order to simulate the intelligibility of speech accurately, it is necessary to reproduce not only the time characteristic and frequency characteristic of the incident sound but also the spatial characteristic, i.e., the three-dimensional incidence direction.

An example of simulation of the word intelligibility of an outdoor emergency public address system with a three-dimensional acoustic VR system developed in the author's laboratory is shown below. Four types of sound field, which modeled existing outdoor emergency public address systems in Tokyo, were simulated (Fig. 13.6). The sound source was a quadruple word, in which four-mora Japanese words were connected at four 1.5-mora (281.25-ms) intervals. Background noise was presented from $\pm 45^\circ$ and $\pm 135^\circ$ with a time delay.

Four combinations of HRTFs and headphones, as shown in Table 13.2, were used. Here, the subject's own HRTFs is denoted as "own", and the best-matching HRTF (see Sect. 4.4.1) is denoted as "bm". The HRTFs of the dummy head (B&K, Type 4128C) are denoted as "HATS". Moreover, "FEC" indicates the FEC headphones (AKG, K1000) described in Sect. 12.1, and "OPEN" denotes commercially available open-air headphones (audio-technica, ATH-AD700).

For comparison, loudspeakers were installed in an anechoic chamber, and the four types of sound fields were reproduced as an original sound field (Fig. 13.7).

Table 13.3 shows the simulation results for word intelligibility. The simulated word intelligibility has the same tendency as reproduction using loudspeakers in an anechoic room, i.e., sound field 2 < sound field 3 ≅ sound field 4 < sound field 1. The

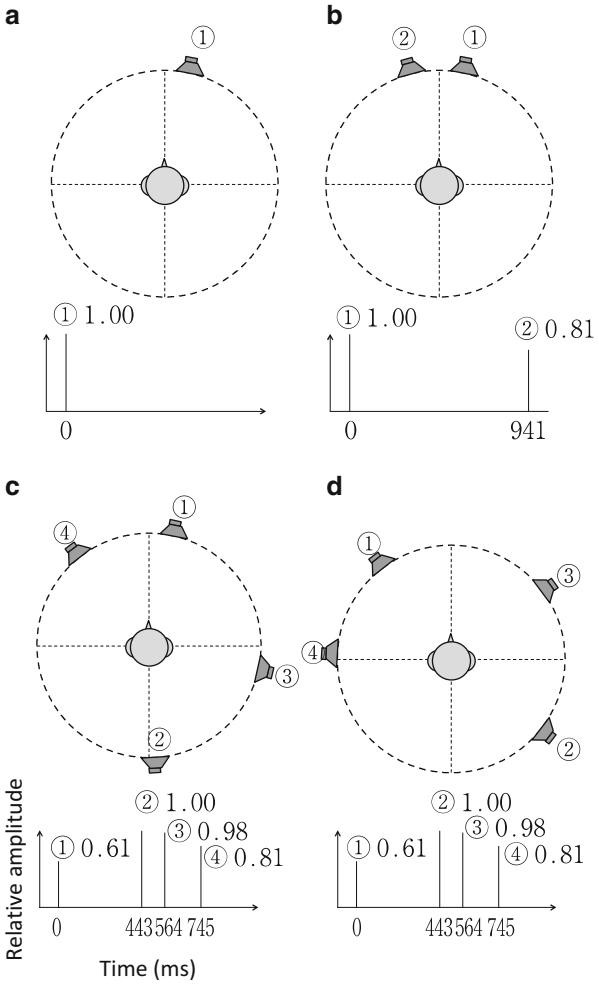

Fig. 13.6 Spatial and temporal structures of four sound fields to be simulated. (**a**) sound field 1, (**b**) sound field 2, (**c**) sound field 3, and (**d**) sound field 4

Table 13.2 Combinations of HRTFs and headphones used for auralization

Method	HRTF	Headphones
1	Subject's own (own)	FEC
2	Subject's own (own)	Open
3	Best-matching (bm)	Open
4	Dummy head (HATS)	Open

results of a chi-square test suggest that own_FEC and bm_OPEN can simulate word intelligibility with an accuracy that does not differ statistically significant from anechoic chamber reproduction.

However, the simulated word intelligibility was slightly higher than that of anechoic room reproduction for all combinations of HRTFs and headphones.

Fig. 13.7 Reproduction of four types of original sound field in anechoic chamber

Table 13.3 Comparison of word intelligibility between the original and the simulated sound field

		Simulated			
Sound field number	Original	own_FEC	own_OPEN	bm_OPEN	HATS_OPEN
1	0.77	0.80	0.83	0.81	0.84∗
2	0.46	0.51	0.54∗	0.47	0.50
3	0.65	0.66	0.70	0.70	0.71
4	0.64	0.65	0.71	0.70	0.74∗∗

∗∗ and ∗ indicate that there exist statistically significant differences between the original sound field and the simulated sound field with significance levels of 1% and 5%, respectively

13.5 Application to Searching for a Sound Source Direction

In addition to three-dimensional reproduction and presentation of sounds, systems have also been developed that use HRTFs to search for sound source directions.

For example, a method of estimating the sound source direction focusing on the interaural phase difference obtained from the input signals to the left and right ears has been proposed (Nakashima et al. 2003, Chisaki et al. 2008). This method divides the ear-input signals into multiple frequency bands and calculates the interaural phase difference in each band. Then, the intersection point of the cones (see Sect. 2.4), which are caused by the phase difference, indicates the sound source direction.

Figure 13.8 shows the estimated directions using this method for the case in which two sound sources (S1 and S2) exist. The vertical angle and azimuth angle of S1 and S2 are (30° and 20°) and (−30° and − 20°), respectively. As shown in Fig. 13.8(a), the directions of the two sound sources can be estimated accurately when the two sound sources are male and female voices. However, as shown in Fig. 13.8(b), the estimation accuracy decreases slightly for the case of a male voice and white noise.

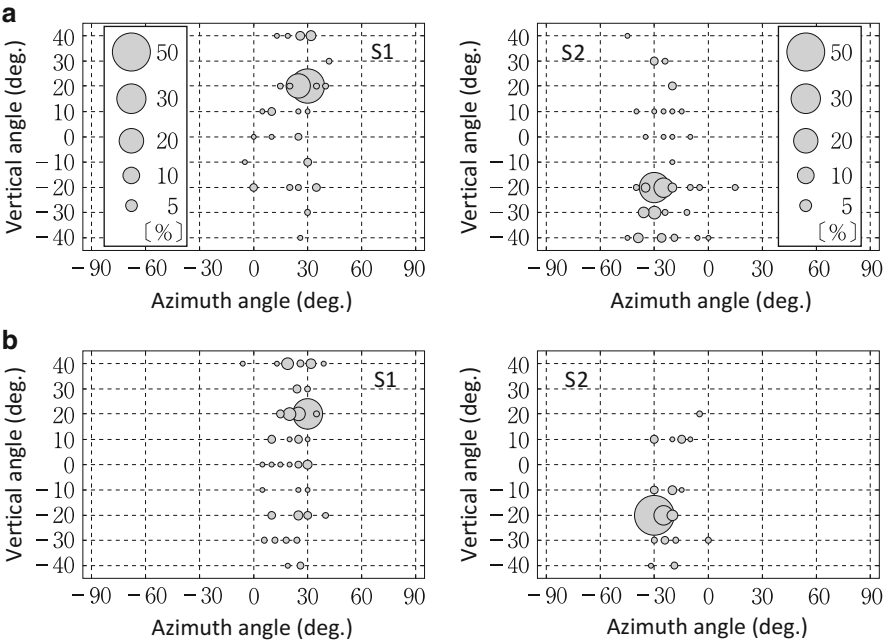

Fig. 13.8 Estimated directions for case in which two sound sources exist (Chisaki et al. 2008) (**a**) S1: Male voice, S2: Female voice (**b**) S1: Male voice, S2: White noise

A method to estimate vertical angle of the sound source by extracting the frequencies of the notches (N1 and N2) of the HRTF from sound signals recorded at both ears of a dummy head or a real head has been proposed (Iida 2010).

This method is based on the idea of using the findings that the frequencies of N1 and N2 depend strongly on the vertical angle of the sound source. This method performs signal processing as follows.

(1) Transform the input signal to the ear ipsilateral to the sound source in the time domain to spectral information by FFT.
(2) Obtain the envelope of the amplitude spectrum of the ear-input signal in order to eliminate the microscopic fluctuations using a moving average.
(3) Extract all local minima of the amplitude spectrum envelope of the input signals to the ear above 4 kHz, and set the local minima as the candidates for N1 and N2.
(4) The most probable vertical angle is estimated by collating with the relationship between the vertical angle and the frequencies of N1 and N2 (Fig. 13.9).

Figure 13.10 shows the estimated vertical angle for a sound source located in the upper median plane in 30° steps. In general, the estimation was accurate regardless of the kind of sound source. However, front-back estimation errors were observed in the cases of 0° for popular music and 30° for a female voice. This error could be related to the fact that the N1 frequency in the front direction is similar to that in the rear direction.

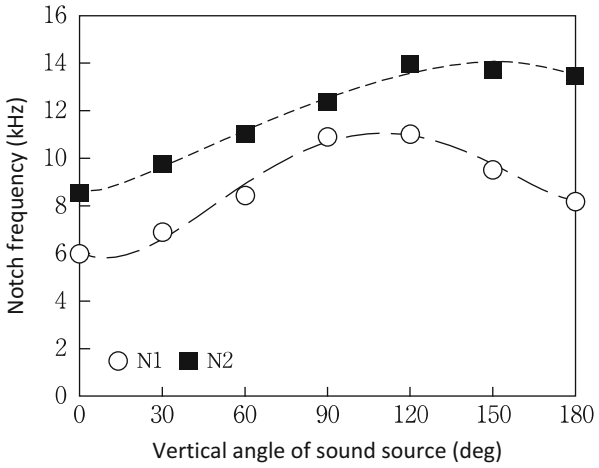

Fig. 13.9 Relationship between vertical angle of sound source and notch frequency (Iida 2010)

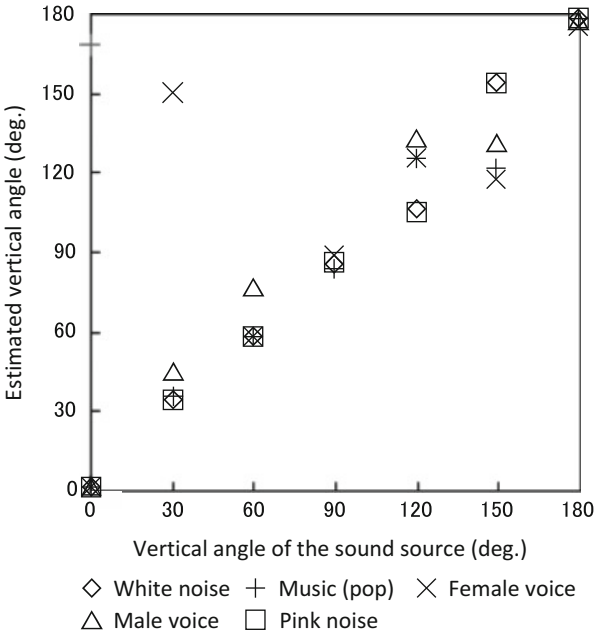

Fig. 13.10 Estimated vertical angle for sound source located in upper median plane (Iida 2010)

Furthermore, this front-back error exhibits a similar tendency to human front-back confusion. The estimated direction was behind (around 170°) for some 1-s-long parts of popular music located at vertical angle of 0° in the median plane, and to the front (around 0°) for other 1-s-long parts (Fig. 13.11). These results indicate instability in front-back estimation, which is a well-known behavior in human sound localization (Wightman and Kistler 1999).

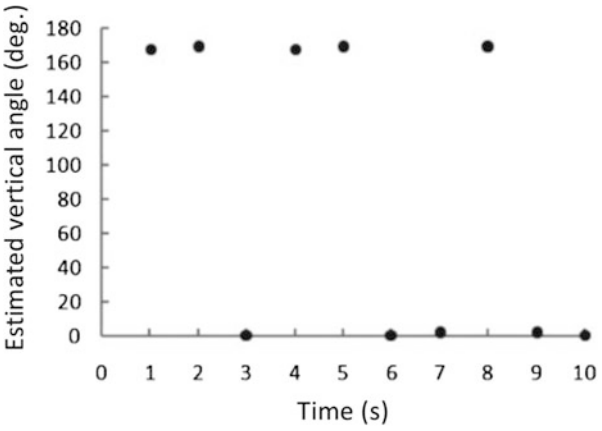

Fig. 13.11 Estimated vertical angle for various 1-s-long-parts of popular music located at vertical angle of 0° in median plane (Iida 2010)

However, as described in Chap. 4, since there are individual differences in the relationship between the vertical angle of a sound source and the N1 and N2 frequencies, it is necessary to obtain this relationship in advance for the dummy head or the real head used for recording.

As described above, although sound source direction estimation was successful in a certain range, there are problems to be solved in the future, such as the decrease of the influence of noise and reverberation. Further research is expected.

References

Chisaki Y, Kawano S, Nagata K, Matsuo K, Nakashima H, Usagawa T (2008) Azimuthal and elevation localization of two sound sources using interaural phase and level differences Acoust. Sci Tech 29:139–148

Iida K (2010) Model for estimating elevation of sound source in the median plane from ear—input signals. Acoust Sci Tech 31:191–194

Nakashima H, Chisaki Y, Usagawa T, Ebata M (2003) Frequency domain binaural model based on interaural phase and level differences. Acoust Sci Tech 24:172–178

Wightman FL, Kistler DJ (1999) Resolution of front-back ambiguity in spatial hearing by listener and source movement. J Acoust Soc Am 105:2841–2853

Yairi S, Iwaya Y, Suzuki Y (2005) Relationship betweenhead movement and total systemdelay of virtual auditory display system. IEICE Technical Report EA:2005–2038. in Japanese