Platform for dynamic virtual auditory environment real-time rendering system

ZHANG ChengYun¹ & XIE BoSun^{1,2*}

¹Acoustic Laboratory, Department of Physics, School of Sciences, South China University of Technology, Guangzhou 510640, China;

²State Key Laboratory of Subtropical Building Science, South China University of Technology, Guangzhou 510640, China

Received February 16, 2012; accepted May 10, 2012

This paper reports the recent works and progress on a PC and C++ language-based virtual auditory environment (VAE) system platform. By tracing the temporary location and orientation of listener's head and dynamically simulating the acoustic propagation from sound source to two ears, the system is capable of recreating free-field virtual sources at various directions and distances as well as auditory perception in reflective environment via headphone presentation. Schemes for improving VAE performance, including PCA-based (principal components analysis) near-field virtual source synthesis, simulating six degrees of freedom of head movement, are proposed. Especially, the PCA-based scheme greatly reduces the computational cost of multiple virtual sources synthesis. Test demonstrates that the system exhibits improved performances as compared with some existing systems. It is able to simultaneously render up to 280 virtual sources using conventional scheme, and 4500 virtual sources using the PCA-based scheme. A set of psychoacoustic experiments also validate the performance of the system, and at the same time, provide some preliminary results on the research of binaural hearing. The functions of the VAE system is being extended and the system serves as a flexible and powerful platform for future binaural hearing researches and virtual reality applications.

virtual auditory environment (VAE), dynamic and real-time rendering, head related transfer function (HRTF), principal components analysis (PCA)

Citation: Zhang C Y, Xie B S. Platform for dynamic virtual auditory environment real-time rendering system. *Chin Sci Bull*, 2013, 58: 316–327, doi: 10.1007/s11434-012-5523-2

Virtual auditory or acoustic environment (VAE) recreates auditory perceptions or events as those would happen in real world by controlling acoustic environment artificially. As a high-technology, VAE interdisciplines the fields of acoustics/physics, signal processing, computer science and human perception. It not only serves as an important experimental tool for the human hearing research, but also constitutes a major part of the multimedia and virtual reality technique. VAE has a lot of potential applications in various fields such as communication, room acoustic design, as well as military and aeronaut training [1].

In real environments, a complex sound field consists of direct sound waves from sound sources and reflected/scattered sound waves from boundaries. The temporal and

spatial characteristics of the sound field capture the information of sources and surrounding environment. The presence of a listener will further disturb the sound field, so that the pressures received by two ears are modified by the scattering, diffraction and reflection of the human anatomical structures. It is just these source direction-dependent scattering, diffraction and reflection that encode the temporal and spatial information of sound field into binaural signals at eardrums. In addition, source or listener's head movement alters the transmission from source to two ears and thereby alters the binaural pressures, which brings dynamic auditory information. Auditory (including high-level nerve) system analyzes and processes the binaural signals and related dynamic information, resulting in spatial auditory perception or events, such as source localization and subjective spatial perceptions of acoustic environment.

*Corresponding author (email: phbsxie@scut.edu.cn)

Because binaural signals contain the primary information of a sound field, a VAE can be realized by rendering the synthesized binaural signals that are similar to those in real acoustic environments via headphones or loudspeakers. This is achieved by modeling the static acoustic courses from sound sources to two ears according to prior knowledge of sources, environment and listener. In other words, VAE processing includes sources physical characteristics modeling, direct and reflected transmission (room acoustic) modeling and listener (individualized scattering/diffraction caused by diversity of human anatomical structure) modeling. In addition, the dynamic auditory information, which is vital to resolving front-back or even up-down confusion in localization as well as recreating authentic and convincing auditory perceptions, should be incorporated into the VAE processing. Therefore, a sophisticated VAE should be an interactive, dynamic, individualized (or at least, appropriate mean across population) and real-time rendering system [1].

There have been great researches on dynamic VAEs. A few systems have been set up since the mid of 1990s, including the first generation SCATIS and the second generation IKA-SIM by Ruhr-Universitat Bochum in Germany [2,3], SLAB by NASA in U.S. [4,5], DIVA by Helsinki University of Technology in Finland [6], the system by Boston University in U.S. [7], as well as the loudspeaker-based system by RWTH Aachen University in Germany [8]. The early-developed systems were implemented on special DSPs and recent-developed systems are usually implemented on PC or server platforms along with software written in C/C++ language.

Improving the performance of dynamic VAE is still challenging, although great progress has been made in the past decade. A dynamic VAE requires large computational resource for source, transmission and listener simulation in real-time [9], especially in the case of multiple virtual sources rendering (including direct sources and image sources for reflections). In practice, limited system resource requires some tradeoffs between performance and computational cost, which embodies as restricting the number of virtual sources; modeling the direct sound and the first or at most the second order reflections only with higher-order

reflections ignored; using low-order filters to model the scatter/diffraction effect of listener; modeling the directional dependence of far-field sources only with the distance dependence of near-field sources ignored; taking the horizontal movement of head into account regardless other degrees of freedom; reducing the dynamic performance of the system, etc. All these inevitably degrade the overall performance of the system. With the development of hardware and software techniques, above problems can be alleviated to some extent. However, the key point to the problems still lies on appropriate simplification of signal processing utilizing some physical and auditory rules while retaining overall performance of system. Continual studies have been done to simplify the system and improve the performance [10], but the problems are too complicated to be solved completely within the frames of a few research projects. Therefore, simplification and improvement of VAE systems is still an open problem. Moreover, no report was found on the hardware/software platform for dynamic VAE in China, not to mention subsequent works based on the platform.

Supported by the National Natural Science Foundation of China, we launched a research project on dynamic and real-time VAE in 2007 and have set up the hardware/software system, which will serve as a basic and powerful experimental platform for human hearing and virtual reality researches in China. The signal processing of current system has been improved so as to enhance the overall performance in the case of multiple virtual sources rendering, auditory distance perception controlling and dynamic information processing for head movement in six-degrees of freedom. This paper reports the recent work and progress on our VAE system, including the principle, structure and design of the system, some measurement results on the physical performances of the system, as well as preliminary results of psychoacoustic experiments to validate the perceptual performances of the system.

1 Basic principle and structure of the system

1.1 Structure of the system

As shown in Figure 1, the standard structure of a dynamic

Figure 1 Structure of the dynamic VAE system.

VAE is adopted in our systems, which consists of three parts.

(i) Information input and definition. This part inputs prior information and data for VAE via a user interface. These information and data are classified into three categories, the information of sources, environment and listener. The information of sources includes type of sources stimuli, the number, spatial locations, orientation and directivities (radiation pattern) and level of sources, or predetermined trajectory for a moving source. The information of environment includes room or environment geometry, absorption coefficients of boundaries and air. The information of listener includes the initial spatial location, orientation and individual features of listener.

(ii) Dynamic VAE signal processing. According to prior information and data in part (i), this part simulates both direct and reflecting transmission from sound sources to two ears using certain physical algorithms. A head-tracking device detects the location and orientation of listener's head, based on which modeling parameters of the scattering/diffraction of listener (HRTF-based filters) are constantly updated. Dynamic binaural signals are thus synthesized.

(iii) Reproduction. The resultant binaural signals are reproduced via headphone after headphone-to-ear-canal transmission equalized.

1.2 Free-field virtual source synthesis

Free-field virtual source synthesis is the simplest case. The source location relative to the head center of a listener is specified by spherical coordinate (r, θ, ϕ) to denote distance, azimuth, and elevation, respectively. The elevation ranges from -90° to 90° with 0° and 90° corresponding to the horizontal plane and directly above the head. The azimuth ranges from 0° to 360° with 0° and 90° corresponding to the front and right directions in the horizontal plane, respectively. The head-related transfer functions (HRTFs), which describes the overall scattering/diffraction or acoustic filtering effects of human anatomical structures, are defined as

$$H_\alpha(r, \theta, \phi, f) = \frac{P_\alpha(r, \theta, \phi, f)}{P_0(r, f)}, \quad (1)$$

where $\alpha = \text{L or R}$ denotes either left or right ear; f is the frequency; P_α denotes the pressures at concerned ear caused by a point source at (r, θ, ϕ) ; P_0 denotes the free-field pressure at the location of head center with the head absent. Generally HRTFs vary as a complex value function of frequency, source location, and individual. In the far field with $r \geq 1.0$ m, HRTFs are approximately independent of distance. While in the near-field with $r < 1.0$ m, HRTFs are related to source distance and thus contain distance perception information. The head-related impulse responses (HRIRs) are the time-domain counterparts of HRTFs, and

related to HRTFs by inverse Fourier transform.

The conventional scheme for a free-field virtual source synthesis is implemented by convolving an appropriately delayed and scaled mono input stimulus $e_0(t)$ with a pair of HRIRs (or equally filtering with a pair of HRTFs), as

$$e_\alpha(t) = \frac{1}{r} h_\alpha(r, \theta, \phi, t) * e_0(t - T), \quad (2)$$

where h_α is HRIRs for the concerned ear; t denotes the time; $T = r/c$ is the propagation delay from the source to the listener; c is the sound speed. The scaling or gain factor $1/r$ accounts for the distance attenuation of pressures magnitude for a point source in the free-field.

For multiple virtual sources synthesis at M locations simultaneously, let $e_{0,i}(t)$ with $i = 1, 2, \dots, M$ denote M mono input stimuli, and binaural signals are then obtained by a linear combination of these of each virtual source, as

$$e_\alpha(t) = \sum_{i=1}^M \frac{1}{r_i} h_\alpha(r_i, \theta_i, \phi_i, t) * e_{0,i}(t - T_i), \quad (3)$$

where r_i is the distance of the i th virtual source and $T_i = r_i/c$ corresponding propagation delay.

A HRIR can be usually approximated as a pure delay version of its minimum phase response $h_{\min}$ [11], i.e.

$$h_\alpha(r, \theta, \phi, t) = h_{\min, \alpha}(r, \theta, \phi, t - \tau_\alpha), \quad (4)$$

where the pure delay $\tau_\alpha = \tau_\alpha(r, \theta, \phi)$ is related to the source location. Substituting eq. (4) into eq. (3) yields

$$e_\alpha(t) = \sum_{i=1}^M \frac{1}{r_i} h_{\min, \alpha}(r_i, \theta_i, \phi_i, t) * e_{0,i}(t - T_{\alpha,i}), \quad (5)$$

where $T_{\alpha,i} = T_i + \tau_\alpha(r_i, \theta_i, \phi_i)$.

It can be seen from eq. (3) that $2M$ HRIR-based convolution manipulations (or equally $2M$ HRTF-based filters) are required for synthesizing binaural signals of M virtual sources. If each HRIR convolution is implemented by a N -point FIR filter, excluding the scaling factors accounting for distance attenuation, $2MN$ multiplications and $2(MN-1)$ additional manipulations are required for obtaining each output sample of the binaural signals. Therefore, the computational cost is directly proportional to the virtual source number M and ultimately exceeds available computational resource. Using minimum phase approximation and truncating the resultant HRIRs with an appropriate time window reduces the impulse response length and thereby reduces the computational cost to some extent. However, an immoderately short time window, for example, a time window shorter than 64 points at 44.1 kHz sampling frequency, will cause audible artifacts. To reduce the computational cost effectively, it is a common practice to limit the number of virtual sources (including direct and image sources for reflections), which inevitably degrades the performance of VAE.

The base function decomposition of HRTFs or HRIRs can be applied to simplify multiple virtual sources synthesis [1,12]. The principle component analysis (PCA) is an efficient scheme for HRTFs (or HRIRs) decomposition. In existing work, the PCA was only applied to distance-independent far-field HRIRs or HRTFs. Here the PCA decomposition is extended to distance-dependent near-field HRIRs. Then the minimum-phase HRIRs for location (r, θ, ϕ) are decomposed as a weighted combination of Q time-domain base functions, as

$$h_{\min, \alpha}(r, \theta, \phi, t) = \sum_{q=1}^Q w_{q, \alpha}(r, \theta, \phi) g_q(t) + h_{\min, \text{av}}(t), \quad (6)$$

where $g_q(t)$ and $h_{\min, \text{av}}(t)$ are, respectively, base functions and mean function related to time but independent of source location and concerned ear. The weights $w_{q, \alpha}(r, \theta, \phi)$, one set for each ear, are only source location-dependent. Therefore, weights in eq. (6) represent location dependence of HRIRs. Given a set of measured HRIRs, $g_q(t)$ and $h_{\min, \text{av}}(t)$ are derived via statistical methods. Then $w_{q, \alpha}(r, \theta, \phi)$ are derived from orthogonal projection. The scheme of near-field HRIRs decomposition by PCA is similar to that of far-field HRIRs [13] and will be discussed in detail in another paper.

The simplified scheme for multiple virtual sources synthesis are obtained by substituting eq. (6) into eq. (5):

$$e_{\alpha}(t) = \sum_{i=1}^M \frac{1}{r_i} \left[\sum_{q=1}^Q w_{q, \alpha}(r_i, \theta_i, \phi_i) g_q(t) + h_{\min, \text{av}}(t) \right] * e_{0, i}(t - T_{\alpha, i})$$

$$= \sum_{q=1}^Q g_q(t) * \left[\sum_{i=1}^M w_{q, \alpha}(r_i, \theta_i, \phi_i) \frac{1}{r_i} e_{0, i}(t - T_{\alpha, i}) \right] + h_{\min, \text{av}}(t) * \left[\sum_{i=1}^M \frac{1}{r_i} e_{0, i}(t - T_{\alpha, i}) \right]. \quad (7)$$

Figure 2 is the block diagram of the signal processing designed according to eq. (7). It can be seen that all M input stimuli share the same $(Q+1)$ convolution kernels $g_q(t)$ and $h_{\min, \text{av}}(t)$, or equally the same parallel bank of $(Q+1)$ filters due to source location-independent feature of $g_q(t)$ and $h_{\min, \text{av}}(t)$. While the virtual source locations are controlled by the weights (gains) $w_{q, \alpha}(r, \theta, \phi)$ as well as scaling factors $1/r_i$ and delay $T_{\alpha, i}$. Therefore, the number of convolution manipulations or filters for binaural synthesis is fixed to $2(Q+1)$, regardless the number of virtual sources. This feature makes the PCA-based scheme more efficient than the conventional one in multiple sources synthesis. Actually, if each convolution is implemented by an N -point FIR filter, excluding the scaling factors accounting for distance attenuation, $2(NQ+MQ+N)$ multiplications and $(2N+2M-4)(Q+1)$ addition manipulations are required for obtaining each sample of the binaural signals. Compared with the conventional scheme, the computational cost are reduced when $M > N(Q+1)/(N-Q)$ and the reduction is obvious for very large number M .

Three optional HRIRs database are adopted in our VAE system.

(i) The far-field HRIRs of KEMAR artificial head with

Figure 2 The block diagram of the PCA-based scheme for multiple virtual sources synthesis (one ear only).

DB 61 small pinna measured by MIT Media Laboratory [14]. The database consists of the HRIRs at source distance $r = 1.4$ m and 710 directions. The sampling frequency is 44.1 kHz and the length of each HRIR is 512 points. This is a popular non-individualized HRTF database. We apply it in our work as reference.

(ii) The individualized far-field HRIRs of Chinese subjects measured by our laboratory [15]. The database was measured by the blocked-ear-canal technique. It comprises the data of 52 subjects at source distance $r = 1.5$ m and 493 directions for each subject. The measured elevations and azimuths are shown in Table 1. The sampling frequency is 44.1 kHz and the length of each HRIR is 512 points.

(iii) The near-field HRIRs of KEMAR artificial head measured by our laboratory [16]. The binaural sound pressures were recorded at the end of the canal simulator. The database consists of the HRIRs at nine source distances from 0.2 to 1.0 m with an interval of 0.1 m and 493 different directions for each distance. The directional distribution of measured HRIRs is identical to that in Table 1. The sampling frequency is 44.1 kHz and length of each HRIR is 512 points. Till now, near-field HRIRs of human subjects are unavailable due to the difficulty in measurement, and even the near-field HRIRs data of artificial head are rare. Therefore, only the data of KEMAR are temporarily adopted in our dynamic VAE system. Our laboratory is now carrying out the near-field HRIRs measurement on human subjects. Once the database is constructed, it can be used to replace the current near-field data.

There are two optional schemes or working models for virtual source synthesis in our dynamic VAE system.

(i) In the conventional scheme, the binaural signals are synthesized according to eq. (3) or eq. (5). The 128-point HRIRs, which are obtained by truncating the original 512-point HRIRs with a rectangular window, are used in processing. The length of HRIRs can be further reduced to 64 points by appropriate smoothing and minimum-phase reconstruction of HRTF magnitudes [17].

(ii) In the PCA-based scheme, the binaural signals are synthesized according to eq. (7) and Figure 2. The 128-

point minimum phase HRIRs are first derived from original measured HRIRs. PCA is then applied to the minimum-phase HRIRs, resulting in $Q = 15$ base functions and related weights which account for more than 97.4% energy variation of minimum phase HRIRs (for both individualized far-field HRIRs and KEMAR near-field HRIRs). The ultimate length of each of $(Q+1) = 16$ filters is 128 points.

The synthesized binaural signals are equalized by the inverse of individualized headphone-ear-canal transfer functions prior to reproduction via headphone [18].

1.3 HRIR spatial interpolation and moving virtual source simulation

HRIRs vary as continuous functions of source location. Measurement usually yields HRIRs at discrete locations, however. HRIRs at unmeasured source locations need to be evaluated by spatial interpolation. Directional interpolation is required for far-field HRIRs. Existing directional interpolation schemes are applicable [1]. Here both directional and distance interpolations are required for near-field HRIRs due to their directional and distance-dependency. The bilinear interpolation scheme is applied for directional interpolation, in which HRIRs are measured at the directions of rectangular grid on a spherical surface, and HRIRs at unmeasured directions are approximated by the weighted sum of four nearest measured HRIRs. While the simplest adjacent linear interpolation scheme is used for distance interpolation, in which HRIRs at unmeasured distance are evaluated by a weighted sum of measured ones at two closest adjacent distances.

Directional and distance interpolations are directly applied to each temporal sample of HRIR in conventional HRIRs convolution-based virtual source synthesis, which also costs considerable computational resources. Alternatively, interpolations are only applied to $2Q$ PCA-weights $w_{q,a}(r, \theta, \phi)$ in PCA-based virtual source synthesis given in eq. (7) because the directional dependence of HRIRs are completely captured by the weights. This simplifies the interpolation and thus is another advantage of PCA-based virtual source synthesis, especially for dynamic synthesis discussed in the following.

For a moving virtual source, its spatial trajectory is described by the following parameter equation:

$$r = r(t), \quad \theta = \theta(t), \quad \phi = \phi(t). \quad (8)$$

The conventional scheme for moving virtual source rendering is implemented by constantly interpolating and updating HRIRs as well as the scaling factor $1/r_i$ and propagation delay T_i in eq. (3) according to the temporary location of the virtual source relative to listener's head. This scheme is complex and inclined to causing audible commutation artifacts. In the PCA-based synthesis, a moving virtual source can be simply recreated by continually interpolating and updating the weights $w_{q,a}(r, \theta, \phi)$ as well as scaling factor

Table 1 The directional distribution of measured HRIRs in our Chinese-subject HRTF database

Elevation ϕ	Azimuth interval $\Delta\theta$
−30°	5°
−15°	5°
0°	5°
15°	5°
30°	5°
45°	5°
60°	10°
75°	15°
90°	360°

$1/r_i$ and delay T_{asi} in eq. (7), avoiding noise caused by directly switching HRIRs. In addition, the Doppler frequency shift can also be incorporated into the VAE processing for simulating a fast moving source [19].

1.4 Dynamic signal processing

In a dynamic case, however, even if the spatial location of virtual source is fixed, listener's head movement alters the source location relative to head. Accordingly, the parameters for static virtual source synthesis, such as the scaling factor and the delay, HRIRs in eq. (3) or weights in eq. (7), should be constantly updated according to the source location (r', θ', ϕ') relative to the temporary location and orien-

$$\begin{bmatrix} r' \cos \phi' \sin \theta' \\ r' \cos \phi' \cos \theta' \\ r' \sin \phi' \end{bmatrix} \begin{bmatrix} \cos \alpha \cos \gamma + \sin \alpha \sin \beta \sin \gamma & -\sin \alpha \cos \gamma + \cos \alpha \sin \beta \sin \gamma & \cos \beta \sin \gamma \\ \sin \alpha \cos \beta & \cos \alpha \cos \beta & -\sin \beta \\ -\cos \alpha \sin \gamma + \sin \alpha \sin \beta \cos \gamma & \sin \alpha \sin \gamma + \cos \alpha \sin \beta \cos \gamma & \cos \beta \cos \gamma \end{bmatrix} \begin{bmatrix} r \cos \phi \sin \theta - \Delta x \\ r \cos \phi \cos \theta - \Delta y \\ r \sin \phi - \Delta z \end{bmatrix}. \quad (9)$$

Figure 3 Coordinate system of listener's head.

Once the source location and direction relative to head are evaluated according to eq. (9), the schemes for dynamic virtual source synthesis are similar to those in moving virtual source synthesis. In the conventional scheme, the HRIRs, scaling factor $1/r_i$ and delay in eq. (3) or eq. (5) should be constantly updated according to the time-variant locations and directions of sources relative to the listener. In addition, HRIRs should be interpolated to obtain the data at arbitrary locations. This dynamic processing costs a large computational resource. Moreover, because head location and orientation are detected at discrete times in a practical dynamic VAE system, updating HRIRs means switching them from one location to another. In order to avoid the noise caused by abruptly switching HRIRs, the most common method is to convolute the input stimulus with two pair of adjacent HRIRs simultaneously and then cross-fades between two outputs. But this leads to twice computational cost. Alternatively, in the PCA-based scheme, dynamic virtual sources synthesis is implemented by changing the weights $w_{q,a}(r, \theta, \phi)$, the scaling factor $1/r_i$ and the delay T_{asi}

tation of listener's head.

The head is able to move in six degrees of freedom in space, with three of them being translation and other three being turning. Accordingly, as shown in Figure 3, three Cartesian coordinate parameters (x, y, z) and three angular parameters (α, β, γ) are needed for fully describing the location and orientation of head, respectively. It is supposed that the head is initially located at the origin ($x = 0, y = 0, z = 0$). The x, y and z axes point to the right, front and above, respectively. The translation of the head is described by ($\Delta x, \Delta y, \Delta z$). The turning of head is described by (α, β, γ), which represent the turning around z, x and y axis, respectively. Then the source location and direction relative to the temporary location and orientation of head is given by

in eq. (7). This implementation is simpler and more efficient than the conventional one, meanwhile, without audible artifacts.

The six degrees of freedom of the head movement are detected, and then both conventional and PCA-based schemes are optionally adopted in current dynamic VAE processing.

1.5 The simulation of environmental reflections

Reflections occur in most real environments and are vital for a convincing VAE. Simulation of environmental reflections is an important branch of room acoustics. Various methods for reflection simulation have been proposed [20]. Because reflection simulation is a somewhat independent issue, for brevity, we do not discuss it in detail and adopt existing image-source method for room reflection simulation.

In the simplest rectangular room [20], for example, there are six image sources accounting for the 1st order reflections from the six surface of room, 30 image sources for the 2nd order and $6 \times 5^{(L-1)}$ image sources for the L th order reflections. Therefore there are $3 \times (5^L - 1) / 2$ image sources in total for the reflections up to L th order. The location or coordinate of each image source is evaluated by reflecting against the boundary surface. Of course, to simulating reflections in a room, visibility check is required for each image source because some image sources may be occluded by protruding boundaries.

Once the location of each image source is determined, each reflection can be regarded as if it emits from a corresponding image source and then be processed in the same manner as a direct or free-field virtual source by eq. (3) or eq. (7). Listener's head movement also alters the locations and directions of image sources relative to the listener, so that dynamic processing introduced in Section 1.4 should be

incorporated to each of image source or reflection. However, the number of the image sources grows exponentially with increasing the order of reflections, which means a fast increase of computational cost in conventional virtual source synthesis scheme and ultimately exceeding of the available resource of the system. Hence, the early reflections only up to the first or second order are simulated by image-source method in most of the previous VAE system. The PCA-based virtual sources synthesis adopted in current system can alleviate the computational burden due to its nearly virtual source number-independent property of convolution manipulations.

Frequency-dependent boundary and air absorption can be incorporated into VAE processing using low-order filters designed from reflection or absorption coefficients. And the frequency-dependent directivity of a complex sound source can also be simulated through corresponding directional filters. In the current dynamic VAE system, each early reflection (up to three-order) is simulated individually and later reverberation is simulated by some artificial reverberation algorithms. There are various algorithms for reverberation simulation [21]. Although these algorithms provide rough simulation of late reflections perception, they are applicable to current VAE system.

2 The hardware and software of dynamic VAE system

2.1 The hardware structure

Figure 4 shows the hardware structure of current dynamic VAE system, which consists of a 4-core PC (Intel Q9550 CPU@2.83 GHz, 4 G internal memory), an ASIO sound card (ESI UGM96), a headphone (Sennheiser HD250) and an electromagnetic head tracker (Polhemus FASTRAK).

Two dynamic parameters, scenario update rate and system latency time, are important for the dynamic performance. The scenario update rate is defined as the number of auditory scenario update manipulations per second, which is mainly determined by the performance of head tracker. The system latency time is defined as the time from the listener's movement to corresponding change in the synthesized binaural signals output. It is influenced by the performance of head tracker, sound card, signal processing algorithms,

Figure 4 The hardware structure of VAE system.

data transmission and communication, data buffer, but most contributed by the preceding two items. Therefore, the head tracker and sound card mainly determine the dynamic performance of system and should be chosen carefully.

The Polhemus FASTRAK electromagnetic head tracker used in the current system consists of a system electronics unit, a transmitter and up to four receivers. It is capable of detecting six degrees of freedom of head translation and turning, with a distance precision of 0.08 cm, distance resolution of 0.0005 cm, angular precision of 0.15° and angular resolution of 0.025°. When one receiver is active alone, the update rate is 120 Hz and the delay is 4 ms.

The driving mode for sound card heavily affects its delay. Familiar driving modes include MME, WDM, DirectSound and ASIO, among which ASIO provides the shortest delay. ASIO, which was specified by Steinberg, demands for a special circuit integrated on the sound card. The ESI UGM96 sound card with ASIO driver is adopted in the current system with a buffer size of 128 samples and a word length of 24 bits.

PC is the “headquarter” of the whole system. The head tracker and sound card are initialized via the USB port, and the prior information for source, environment and the listener is defined via the software interface. When the system starts to work, CPU receives the location and orientation data of head via USB port and carries out the dynamic VAE processing. The resultant binaural signals are replayed via the sound card.

2.2 The software structure

The software was written with C++ language on Microsoft Visual Studio.NET 2003. It performs the functions of human-machine interface, receives the data from head tracker, dynamic VAE processing, inputs and outputs the audio stimuli. Figure 5 shows the structure of the software consisting of three threads and five functioning modules.

(i) The human-machine interface module responds for information input and definition of sound sources, environment and listener as well as real-time visual display of virtual source direction and distance, if necessary.

Figure 5 The modules of the software for the dynamic VAE system.

(ii) The head tracker module connects the head tracker to the PC, by which the location and orientation data of listener's head is transmitted to PC.

(iii) The sound card module connects the sound card to the PC, by which the sound card is configured and the processed audio data are transmitted to sound card. It is important to use the ASIO sound card driver. Steinberg provides the free API functions in its official webpage.

(iv) The parameter module calculates the temporary directions and distances of virtual sources with respect to listener according to the data from head tracker and eq. (9). It also calculates the propagation delay and distance attenuation of the virtual source. If necessary, this module checks the visibility of each image source when simulating the reflections in a room.

(v) The signal processing module performs the sources, environment and listener modeling in VAE as well as headphone-to-ear-canal transmission equalization.

A small size of input data block is preferred in view of reducing system latency time. However, the size of data block is restricted by the performance of sound card. Test indicated that the ESI UGM96 ASIO sound card still works normally with a 64-samples buffer. To be on the safe side, a 128-sample buffer size is chosen in current system, which leads to a 2.9 ms delay in output at sampling frequency of 44.1 kHz.

3 The objective performance of the system

The scenario update rate, system latency time, and the maximal number of virtual sources (including direct and image sources for reflections) which the system is able to render simultaneously, are important parameters for dynamic performance of the VAE system. These three parameters were tested in current VAE system.

The scenario update rate is determined by the performance of head tracker. The specification of Polhemus FASTRAK head tracker claims an update rate of 120 Hz when one receiver channel is active alone. This can be easily validated by measurement using the high-precision internal clock of the PC. The result was 8.33 ms, just corresponding to an update rate of 120 Hz. The system latency

time was measured by a simpler method [22]. The final result for mean overall system latency time is 25.4 ms.

The maximal number of virtual sources depends on scheme, hardware and software of the system. The current system exhibits following performance:

(i) For the conventional scheme, it is able to synthesize up to 280 free-field virtual sources simultaneously.

(ii) For the PCA-based scheme, it is able to synthesize up to 4500 free-field virtual sources simultaneously.

(iii) For the PCA-based scheme, it is able to render a direct virtual source and up to four order reflections of a rectangular room. In order to compare with the results of conventional scheme, however, the highest-order of simulated reflections is limited to three, corresponding to 186 image sources in total.

Table 2 compares the performance of current system with some previous systems [2–6], and all of these systems were reproduced via headphone. Although the improvement on performance of current system partly results from the advance on hardware/software, the differences in performance proves the advantage of current system.

4 The psychoacoustic experiments

The goals of psychoacoustic experiments are to validate the perceived performance of system and provide some preliminary results on binaural hearing research. Here, the preliminary results of three psychoacoustic experiments are reported. More results on binaural hearing research based on the VAE platform will be reported soon.

White noise with full-audible bandwidth generated by software with 44.1 kHz sampling frequency and 16 bit resolution was used as input stimulus. All three experiments were carried out in a listening room with reverberation time of 0.15 s and the background noise below 30 dBA. The listener sat at the center of the room.

4.1 The free and far-field virtual source localization experiment

Free and far-field virtual source synthesis with non-individualized HRIRs using conventional scheme is the sim-

Table 2 The performance of some dynamic VAE systems

System name	Latency (ms)	Update rate (Hz)	Number of sources	Filter length (points)
SCATIS	66	60	32	80
IKA-SIM	60	60	40–60	64/96/128
SLAB	24 (excluding tracker latency)	120	24	128 (direct path) 32 (reflection path)
DIVA	110–160	20	10–20	30–35 (FIR) 15–18 (IIR)
Current system	25.4	120	280 (conventional mode) 4500 (PCA-based mode)	128

plest case, which provides a primary validation to the performance of system. The MIT-KEMAR far-field HRIRs were used in synthesis. Virtual source direction and distance localization was made for both static and dynamic synthesis via the conventional scheme.

The 28 intended virtual directions, which were distributed in the right-hemispherical space at four elevations $\phi = -30^\circ, 0^\circ, 30^\circ, 60^\circ$ with seven azimuths $\theta = 0^\circ, 30^\circ, 60^\circ, 90^\circ, 120^\circ, 150^\circ$ and 180° at each elevation, were chosen in the experiment. Six listeners with three male and three female took part in the experiments. Listeners were asked to judge the perceived virtual source location in terms of its direction and distance. Each listener repeatedly judged three times for each intended virtual source location. There were 18 judgments in total for each intended location. There are different ways for a listener to report the perceived direction and distance of a virtual source. A tracker was used in our experiments [23]. The PC picked up and saved the location data instantly once the listener points the tracker to the perceived location of virtual source.

The percentages of front-back and up-down confusion in localization were evaluated. Front-back confusion refers to a virtual source intended for the frontal-hemisphere that was actually perceived in the rear-hemisphere, or vice versa; and up-down confusion refers to a virtual source intended for the upper-hemisphere that was actually perceived in the bottom-hemisphere, or vice versa. The results indicated that the percentages of front-back and up-down confusion are 30.3% and 24.1% for static rendering, 2.5% and 13.7% for dynamic rendering, respectively. Most of the front-back confusions in dynamic synthesis occur in the high elevation plane of $\phi = 60^\circ$.

Some front-back and up-down confusions are observed

in the case of static rendering. Dynamic rendering almost completely resolves the front-back confusion and obviously reduces the percentage of up-down confusion, even though the non-individualized HRIRs (HRTFs) are used. This result provides evidence for comparing the relative contribution between dynamic cue and individualized HRTFs (spectral cue) to front-back and up-down localization. Actually, localization information provided by pressures at two ears and their variation is redundant. Dynamic information provides stable front-back and up-down localization cues so that the hearing relies less on the individualized spectral cues. This agrees with some previous results. Notably, because all three degrees of freedom for head turning are incorporated into current processing, percentages for both front-back and up-down confusions are obviously reduced. Therefore, this experiment extends the results of some previous experiments in which the head rotation around the vertical axes alone was incorporated and thus only the front-back confusion was reduced.

Prior to calculating the statistical results of perceived virtual source direction, the perceived azimuth for front-back confusion case was corrected by reflecting against the lateral plane. A similar procedure was applied to the cases of up-down confusion case. Figure 6 is graphical representation of perceived virtual source direction, which makes the results more visible, as suggested by Leong et al. [24]. The results are demonstrated on the surface of a sphere and viewed from the front, right and rear, respectively. Notation “+” represents the intended virtual source direction. The black points at the center of ellipses are the average perceived direction. The ellipses are the confidence region at the significance level of $\alpha = 0.05$. The ellipses drawn with dash line indicate that the data are highly symmetric around

Figure 6 The graphical representation of the results of directional localization for free and far-field virtual source.

the mean. As seen, dynamic synthesis also reduces localization error in terms of the difference between intended and average perceived virtual source direction.

At each intended virtual source direction, the average perceived virtual source distance across six listeners and three judgments per listener ranges from 0.16 to 0.62 m (close to the head surface) in the static synthesis. Meanwhile the results of the dynamic synthesis range from 0.86 to 1.23 m. Therefore, dynamic synthesis is helpful to create appropriate virtual source distance perception.

Above conclusions can be further proved by applying statistic T test (at a significance level of $\alpha = 0.05$) to the localization results of static and dynamic synthesis, which validate the preliminary performance of the VAE system.

Individualized far-field HRIRs were adopted in the experiment as well, which further reduces the localization error compared with using non-individualized ones. In addition, localization experiment for virtual source with environmental reflections was also carried out. For brevity, however, the results are omitted here and will be discussed in detail in another paper.

4.2 The free and near-field virtual source localization experiment

This experiment provides the results for free and near-field virtual source localization, and at the same time, a comparison between the performance of conventional and PCA-based dynamic virtual source synthesis.

The measured KEMAR near-field HRIRs were adopted. Dynamic virtual sources were synthesized by either conventional or PCA-based scheme. Five different intended virtual source distances at $r = 0.2, 0.4, 0.6, 0.8$, and 1.0 m and four elevations $\phi = -30^\circ, 0^\circ, 30^\circ$ and 60° for each distance were chosen. The virtual source azimuths distributed in the right-hemispherical space, with three azimuths at $\theta = 0^\circ, 90^\circ$ and 180° for elevation $\phi = 60^\circ$, five azimuths at $\theta = 0^\circ, 45^\circ, 90^\circ, 135^\circ$ and 180° for other three elevations. Therefore, there were 720 intended virtual source locations.

Ten listeners with five male and five female took part in the experiments. The procedures for localization were similar to those in Section 4.1. Each listener repeatedly judged four times for each intended virtual source location. There were 40 judgments in total for each location.

The results were processed and illustrated similar to those in Section 4.1. Table 3 shows the average percentages of front-back and up-down confusion in localization for different intended virtual source distances and conventional or PCA-based synthesis. The two schemes for synthesis yield similar results. The results are also similar to those in Section 4.1 except an increase in percentage of front-back and up-down confusion at the close intended virtual source distance of $r = 0.2$ m.

The statistical results of perceived virtual source direction after correction on front-back and up-down confusion are similar to those in Figure 6, except an increase in localization dispersion at close intended virtual source distance of $r = 0.2$ m. Moreover, conventional and PCA-based schemes yield similar localization results. As an example, Figure 7 shows the graphical representation of statistical results of perceived virtual source direction at intended distance of $r = 0.8$ m, for conventional and PCA-based schemes respectively.

The results of distance localization show that the average perceived distance increases with the intended distance for virtual sources intended at $r \leq 0.6$ m and outside the median plane ($\theta = 0^\circ$ or 180°), especially for those intended at the lateral direction in the horizontal plane. When the intended distance exceeds 0.6 m, the average perceived distance is almost invariable against the intended distance. Moreover, the conventional and PCA-based schemes yield similar results. Actually, the source distance-dependent head shadow brings distance localization cue. Head shadow comes into effect only when the source departs from the median plane. As an example, Figure 8 shows the average perceived distance and the standard deviation over ten listeners and four repeated judgments per listeners, for intended virtual directions at ($\theta = 90^\circ, \phi = 0^\circ$). Results for both conventional and PCA-based schemes are given. It is interesting to see that the average perceived virtual source distance is quantitatively inconsistent with intended distance, although the variation tendency of former qualitatively matches with the latter. Moreover, the standard deviation of perceived distance is great, suggesting a large inter-subject variation. This is similar to the previous results [25]. In real environment, auditory distance perception is contributed by multiple cues; the variation of near-field HRTFs with distance is only one of those cues. A more reliable distance

Table 3 Percentage of front-back and up-down confusion in near-field virtual source localization

Intended distance (m)	Percentage of front-back confusion (%)		Percentage of up-down confusion (%)	
	Conventional synthesis	PCA-based synthesis	Conventional synthesis	PCA-based synthesis
0.2	9.0	8.2	20.4	24.8
0.4	2.7	2.0	13.5	14.6
0.6	1.6	2.3	11.5	16.3
0.8	2.3	1.6	10.8	16.0
1.0	2.9	2.3	11.5	16.7

Figure 7 The graphical representation of results of directional localization for free and near-field virtual source ($r = 0.8$ m).

Figure 8 Results of distance localization for free and near-field virtual source at ($\theta = 90^\circ$, $\phi = 0^\circ$).

perception in VAE can be achieved by incorporating other relevant cues, such as environment reflections, into simulation. But the situation is too complex to be discussed here.

Above results proved that near-field HRIRs processing is capable of controlling perceived virtual source distance to some extent in a dynamic VAE. Moreover, the conventional and PCA-based schemes yield equivalent virtual source localization results. The latter can be further validated by applying statistic T test (at a significance level of $\alpha = 0.05$) to the localization data of conventional and PCA-based schemes [26].

4.3 The subjective comparison experiment

The subjective comparison experiment was conducted to evaluate whether the PCA-based scheme for dynamic free and near-field virtual source synthesis results in other audible artifact (such as timbre coloration) in addition to localization compared with the conventional scheme.

A three-interval two-alternative force-choice (3I-2AFC) paradigm was adopted in the experiment, in which binaural

signals synthesized by the conventional scheme were regarded as a reference signal A and those synthesized by the PCA-based scheme were regarded as a target signal B. Each stimulus presentation consisted of three segments; the length of each segment was 10 s. The first segment was always signal A followed by signals A and B at a random order (AAB or ABA). The subject's task was to judge which segment of the second or third perceptually differed from the first segment, according to any audible attributes such as virtual source location and timbre. If the subject was unable to discriminate the differences, he/she should choose an answer randomly. The 90 intended virtual source locations in the experiment were identical to those in Section 4.2. For each intended location, each listener made six judgments with three repetitions for either "AAB" or "ABA". Ten listeners with six male and four female took part in the experiments.

There were 60 judgments in total for each intended virtual source location. Then the ratio p of correct judgments for each intended virtual source location is evaluated. When the number of judgments is sufficiently large and the listener cannot discriminate the difference between reference A and target B, the expected value of p approaches $p_0 = 0.5$. Meanwhile, $p_0 \geq 0.75$ is often used as a criterion for correctly discriminating the difference between reference and target. According to the one-side and two-side hypothesis testing for (0,1) distribution in statistics, at a significance level of $\alpha = 0.05$, $0.373 \leq p \leq 0.627$ means the listeners were unable to discriminate the difference between reference and target signals; $0.627 < p \leq 0.658$ represents an uncertain case; and $p > 0.658$ means the subjects were able to discriminate the difference. The experimental results indicate that listeners were unable to discriminate the difference between reference and target signals at most locations, except for three

uncertain cases at locations ($\theta=135^\circ$, $\phi=-30^\circ$, $r=0.2$ m), ($\theta=135^\circ$, $\phi=30^\circ$, $r=0.4$ m) and ($\theta=0^\circ$, $\phi=60^\circ$, $r=0.8$ m), respectively. Therefore, the PCA-based scheme is perceptually equivalent to the conventional scheme, without causing audible artifact.

5 Conclusions

A PC and C++ language-based platform of the dynamic VAE real-time rendering system is constructed. Conventional and proposed PCA-based schemes are alternatively adopted in dynamic far and near-field virtual source synthesis. Non-individualized or individualized HRIRs are used in processing. Image-source method is applied to simulate early reflections. The system is able to recreate both far and near free-field virtual source at various spatial locations as well as the auditory perception in reflective environments. The system performances are improved in aspects of dynamic multiple virtual sources rendering, near-field virtual source synthesis, and six degrees of freedom of head movement simulation. Test indicates that the scenario update rate is 120 Hz and system latency time is 25.4 ms. The system is able to simultaneously render up to 280 virtual sources using conventional scheme, or 4500 virtual sources using the PCA-based scheme. Hence, the improvements in performance are obvious.

The psychoacoustic experiments validate the perceived performance of the system, and at the same time, provide some preliminary results on binaural hearing research. The results indicate that the dynamic virtual source synthesis incorporating six degrees of freedom of head movement obviously reduces the up-down confusion and even almost eliminates the front-back confusion in localization, which often occur in static virtual source synthesis. Near-field HRIRs processing is capable of controlling perception virtual source distance to some extent. Moreover, conventional and PCA-based schemes yields equivalent perceived results.

The functions of system are currently being extended. The system serves as a platform for future research on binaural hearing and multimedia, as well as one of the core parts of other virtual reality systems.

This work was supported by the National Natural Science Foundation of China (11174087, 10774049) and State Key Laboratory of Subtropical Building Science, South China University of Technology. The authors also acknowledge Dr. Zhong Xiaoli for her help.

- 1 Xie B S. Head Related Transfer Function and Virtual Auditory (in Chinese). Beijing: National Defense Industry Press, 2008
- 2 Blauert J, Lehnert H, Sahrhage J, et al. An interactive virtual-environment generator for psychoacoustic research I: Architecture and implementation. *Acta Acust United Acust*, 2000, 86: 94–102

- 3 Silzle A, Novo P, Strauss H. IKA-SIM: A system to generate auditory virtual environments. In: AES 116th Convention, Berlin, Germany, 2004. 6016
- 4 Wenzel E M, Miller J D, Abel J S. Sound Lab: A real-time, software-based system for the study of spatial hearing. In: AES 108th Convention, Paris, France, 2000. 5140
- 5 Miller J D, Wenzel E M. Recent developments in SLAB: A software-based system for interactive spatial sound synthesis. In: Proceedings of the 2002 International Conference on Auditory Display, Kyoto, Japan, 2002
- 6 Savioja L, Huopaniemi J, Lokki T, et al. Creating interactive virtual acoustic environments. *J Audio Eng Soc*, 1999, 47: 675–705
- 7 Scarpaci J W, Colburn H S, White J A. A system for real-time virtual auditory space. In: Proceedings of the 11th International Conference on Auditory Display, Limerick, Ireland, 2005. 241–246
- 8 Lentz T, Assenmacher I, Vorländer M, et al. Precise near-to-head acoustics with binaural synthesis. *J Virtual Reality Broadcast*, 2006, 3: urn:nbn:de:0009-6-5890
- 9 Sandvad J. Dynamic aspects of auditory virtual environments. In: AES 100th Convention, Copenhagen, Denmark, 1996. 4226
- 10 Begault D R, Wenzel E M, Godfroy M, et al. Applying spatial audio to human interfaces: 25 years of NASA experience. In: AES 40th Conference, Tokyo, Japan, 2010
- 11 Kulkarni A, Isabelle S K, Colburn H S. Sensitivity of human subjects to head-related transfer-function phase spectra. *J Acoust Soc Am*, 1999, 105: 2821–2840
- 12 Larcher V, Jot J M, Guyard J, et al. Study and comparison of efficient methods for 3D audio spatialization based on linear decomposition of HRTF data. In: AES 108th Convention, Paris, France, 2000. 5097
- 13 Kistler D J, Wightman F L. A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction. *J Acoust Soc Am*, 1992, 91: 1637–1647
- 14 Gardner W G, Martin K D. HRTF measurements of a KEMAR. *J Acoust Soc Am*, 1995, 97: 3907–3908
- 15 Xie B S, Zhong X L, Rao D, et al. Head-related transfer function database and analyses. *Sci China Ser G: Phys Mech Astron*, 2007, 50: 267–280
- 16 Yu G Z, Xie B S, Rao D. Characteristics of near-field head-related transfer function for KEMAR. In: AES 40th Conference, Tokyo, Japan, 2010
- 17 Xie B S, Zhang T T. The audibility of spectral detail of head-related transfer functions at high frequency. *Acta Acust United Acust*, 2010, 96: 328–339
- 18 Møller H. Fundamentals of binaural technology. *Appl Acoust*, 1992, 36: 171–218
- 19 Krebber W, Gierlich H W, Genuit K. Auditory virtual environments: Basics and applications for interactive simulations. *Signal Processing*, 2000, 80: 2307–2322
- 20 Lehnert H, Blauert J. Principle of binaural room simulation. *Appl Acoust*, 1992, 36: 259–291
- 21 Gardner W G. Reverberation algorithms. In: Kahrs M, Brandenburg K, eds. *Applications of Signal Processing to Audio and Acoustics*. Boston: Kluwer Academic Publishers, 1998
- 22 Zhang C Y, Xie B S, Rao D, et al. Dynamic parameters measurement of virtual auditory environment real-time rendering system. In: Proceedings of CISP, Yantai, China, 2010. 3852–3855
- 23 Zhang C Y, Xie B S. Analysis and verification on the accuracy of virtual sound source position reported by magnetic tracker (in Chinese). *J South China Univ Tech (Nat Sci Ed)*, 2011, 39: 100–105
- 24 Leong P, Carlile S. Methods for spherical data analysis and visualization. *J Neurosci Meth*, 1998, 80: 191–200
- 25 Zahorik P. Assessing auditory distance perception using virtual acoustics. *J Acoust Soc Am*, 2002, 111: 1832–1846
- 26 Sheng Z, Xie S Q, Pan C Y. *Probability Theory and Mathematical Statistics* (in Chinese). 3rd ed. Beijing: Higher Education Press, 2001