PAPER PUBLISHED IN IEEE SIGNAL PROCESSING MAGAZINE SPECIAL ISSUE ON SIGNAL PROCESSING TECHNIQUES FOR ASSISTED LISTENING, vol. 32, no. 2, pp. 100-113, Mar. 2015.
DOI: 10.1109/MSP.2014.2372062, http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=7038257

Natural Sound Rendering for Headphones: Integration of Signal Processing Techniques

Kaushik Sunder, Jianjun He, *Student Member, IEEE,* Ee-Leng Tan, and Woon-Seng Gan, *Senior Member, IEEE*

With the strong growth of assistive and personal listening devices, natural sound rendering over headphones is becoming a necessity for prolonged listening in multimedia and virtual reality applications. The aim of natural sound rendering is to recreate the sound scenes with the spatial and timbral quality as natural as possible, so as to achieve a truly immersive listening experience. However, rendering natural sound over headphones encounters many challenges. This tutorial paper presents signal processing techniques to tackle these challenges to assist human listening.

I. INTRODUCTION

Sound is an inherent part of our everyday lives for information, communication and interaction. Sound improves the situational awareness by providing feedback for actions and situations that are out of the view of the listener. An advantage of sound is that multiple sound sources can be perceived from any location around the head in the three dimensional (3D) space [1]. The role of natural 3D sound, or spatial sound, in high stress applications, like flight navigation and communication systems, is indisputable [1]. Naturally rendered sound has also been proven to be beneficial in personal route guidance for visually impaired people and in medical therapy for patients [1]. Last but not least, the ever growing market of consumer electronics calls for natural sound rendering for digital media, such as movies, games, and augmented, virtual reality applications like teleconferencing.

In most of these applications, listening is seldom from the physical sound sources but instead from playback devices, such as headphones or loudspeakers. Headphones, by virtue of their

convenience and portability, are typically chosen as the preferred playback device, especially for personal listening. Therefore, to assist headphone listening, it is critical for the sound to be rendered in a way that listeners can perceive it as natural as possible. In this context, natural sound rendering essentially refers to rendering of the original sound scene using headphones to create an immersive listening experience and the sensation of "being there" at the venue of the acoustic event. To achieve natural sound rendering, the virtual sound rendered should exactly emulate all the spatial cues of the original sound scene, as well as the individual spectral characteristics of the listener's ears. In this paper, we mainly consider the most widely used channel-based audio as the input signals for the natural sound rendering system, though some of the signal processing techniques discussed could also be used in other audio formats, such as object-based format and ambisonics [2], [3].

In recent years, the design criteria for commercial headphones have undergone significant development. At Harman, Olive *et al.* investigated the best target responses for designing headphones based on the listener's preference for the most natural sound [4]. Creating realistic surround sound in headphones has become a common pursuit of many headphone technologies from Dolby, DTS, etc. Furthermore, personalized listening experience and incorporating the information of listening environment has also been the trends in headphone industry. These trends in headphones share one common objective: ***To render natural sound in headphones.***

II. CHALLENGES

The listening process in humans can generally be considered as a source-medium-receiver model, as stated by Begault [1]. This model is used in this paper to highlight the differences between natural listening in real environment and listening over headphones. In natural listening, we listen to the physical sound sources in a particular acoustic space, with the sound waves undergoing diffraction, interference with different parts of our morphology (torso, head and pinna) before reaching the eardrum. This information of sound wave propagation can be encapsulated in spatial

digital filters termed as head-related transfer functions (HRTFs) [1]. Listeners also get valuable interaural cues for sound localization with head movements. However, headphone listening is inherently different from natural listening as the sources we are listening to are no longer physical sound sources but are recorded and edited sound materials. These differences between natural and headphone listening lead to various challenges in rendering natural sound over headphones, which can be broadly classified into the following three categories:

1) **From the perspective of source,** the sound scenes rendered for headphone listening should comprise not only the individual sound sources but also the features of the sound environment. Listeners usually perceive these sound sources to be directional, i.e., coming from certain directions. Moreover, in most of the digital media content, the sound environment is usually perceived by the listener to be diffuse (partially). This perceptual difference between the sound sources and the sound environment requires them to be considered separately in natural sound rendering [2]. Though there are other formats that can represent the sound scenes (e.g., object-based, ambisonics), the convention for today's digital media is still primarily channel-based format. Hence, the focus of this paper lies in the rendering of channel-based audio, where sound source and environment signals are mixed in each channel [2]. In channel-based signals, where only the sound mixtures are available (assuming one mixture in every channel), it is necessary to extract the source signals and environment signals, which can be quite challenging. Furthermore, most of the traditional recordings are processed, and mixed for optimal playback over loudspeakers, rather than headphones. Direct playback of such recordings over headphones results in an unnatural listening experience, which is mainly due to the loss of crosstalk, and localization issues.

2) **From the perspective of medium**, headphone listening does not satisfy free-air listening conditions as in natural listening. Since the headphone transfer function (HPTF) is not flat, equalization of the headphone is necessary. However, this equalization is tedious and challenging

as the headphone response is highly dependent on the individual anthropometrical features and also varies with repositioning.

3) **From the perspective of receiver**, the omission of listener's individualized filtering with the outer ear in headphone listening often leads to coloration and localization inaccuracies. These individualized characteristics of the listener are lost when the sound content is recorded or synthesized non-individually, i.e., the subject in the listening is different from the subject in the recording or synthesis. Furthermore, the sound in headphone listening is not adapted to the listener's head movements, which departs from a natural listening experience.

III. SIGNAL PROCESSING TECHNIQUES

To tackle the above challenges and enhance natural sound rendering over headphones, digital signal processing techniques are commonly used. In Fig. 1, we summarize the differences between natural listening and headphone listening, and introduce the corresponding signal processing techniques to tackle these challenges, which are:

1) Virtualization: to match the desired playback for the digital media content;
2) Sound scene decomposition using blind source separation (BSS) and primary-ambient extraction (PAE): to optimally facilitate the separate rendering of sound sources and sound environment;
3) Individualization of HRTF: to compensate for the lost or altered individual filtering of the sound in headphone listening;
4) Equalization: to preserve the original timbral quality of the source and alleviate the adverse effect of the inherent headphone response;
5) Head tracking: to adapt to the dynamic head movements of the listener.

The remainder of this paper is structured as follows. Virtualization and head tracking, due to their high interactions, are explained together in Section IV, followed by the decomposition of sound scenes in Section V. Sections VI and VII describe individualization and equalization,

respectively. These signal processing techniques are integrated and evaluated using subjective tests in Sections VIII and IX, respectively. Finally, the conclusions and future trends are presented in Section X.

IV. VIRTUALIZATION

In digital media, sound is typically mixed for loudspeaker rather than headphone playback. The spatial sound to be rendered naturally over headphones should emulate the natural propagation of the acoustic waves emanated from the loudspeaker to the eardrum of the listener. To emulate stereo or surround sound loudspeaker rendering over headphones, virtualization techniques based on HRTF corresponding to the loudspeaker positions are commonly used. Given these acoustic transfer functions (i.e., HRTFs), the virtualization technique is applicable to any multichannel loudspeaker setup, be it stereo, 5.1, 7.1, 22.2, or even loudspeaker arrays in wave-field synthesis. As shown in Fig. 2(a), for every desired loudspeaker position, the signal in the *m*th channel $x_m(n)$ is filtered with the corresponding HRTF $h_{xmL}(n)$, $h_{xmR}(n)$, and summed before being routed to the left and right ears [1], [5], respectively, as:

$$\begin{aligned} y_L(n) &= \sum_{m=1}^{M} h_{xmL}(n) * x_m(n), \\ y_R(n) &= \sum_{m=1}^{M} h_{xmR}(n) * x_m(n), \end{aligned} \tag{1}$$

where * denotes convolution and *M* is the total number of channels. When the HRTFs are directly applied to multichannel loudspeaker signals, the rendered sound scenes in headphone playback suffer from inaccurate virtual source directions, lack of depth, and reduced image width [5], [6].

To solve these problems in virtualization of multichannel loudspeaker signals and achieve a faithful reproduction of the sound scenes, the HRTFs should be applied to the individual source signals that are usually extracted (using BSS, PAE) from the loudspeaker signals (i.e., mixtures). In

this virtualization as shown in Fig. 2(b), the sources are rendered directly using the HRTFs of the corresponding source directions $h_{skL}(n)$, $h_{skR}(n)$:

$$\begin{aligned} y_L(n) &= \sum_{k=1}^{K} h_{skL}(n) * s_k(n) + a_L(n), \\ y_R(n) &= \sum_{k=1}^{K} h_{skR}(n) * s_k(n) + a_R(n), \end{aligned} \tag{2}$$

where K is the total number of sources, $s_k(n)$ is the kth source in the multichannel signal, and the environment signals $a_L(n)$, $a_R(n)$ are the rendered signals representing the sound environment perceived at two ears. To render the acoustics of the environment, the environment signals can be either synthesized according to the sound environment [7] or extracted from the mixtures. Techniques like decorrelation [5], [8] and artificial reverberation [9] are commonly employed to render the environment signals in order to create a more diffuse and natural sound environment.

Furthermore, adding the reverberation of sources (or the loudspeaker signals in virtualization of multichannel loudspeaker signals) can also improve the realism of the reproduced sound scene [10]. Therefore, in virtualization, it is quite common to use binaural room impulse response (BRIR) [1], [5] that encapsulates HRTFs and reverberation. On this note, selecting the correct amount of early reflections as well as late reverberation is critical to recreate a faithful sound environment [1]. In general, the BRIR that matches the sound environment of the scene or BRIR of a mixing studio are considered to be more suitable [4]. As discussed in Section II, natural sound rendering requires the accurate reproduction of both the sound sources and the sound environment. Compared to the virtualization of multichannel loudspeaker signals (Fig. 2(a)), the latter technique of virtualizing the source and environment signals (Fig. 2(b)) is more desirable as it is closer to natural listening [6], [8], [9]. These virtualization techniques can also be incorporated into spatial audio coding systems, such as binaural cue coding [11], spatial audio scene coding [5], and directional audio coding [3].

In virtualization, the directions of the sources (or the loudspeakers in virtualization of multichannel loudspeaker signals as in Fig. 2(a)) have to be calibrated according to the head movements (as in natural listening). To fulfill this need, the HRTFs/BRIRs in the virtualization are updated on the fly based on these head movements that are often tracked by a sensor (e.g., accelerometer, gyroscope, camera, etc.). The latency between the head tracking and sound rendering should be such that the localization accuracy is not affected [12]. Such a head tracking system when incorporated in the virtualization process can provide useful dynamic cues to resolve the localization conflicts [1] and enhance natural sound rendering [10], [12]. It shall be noted that head tracking is more critical for the directional sources but less important for the diffuse signals like environment signals and late reverberation [12]. This is because the perception of diffuse signals is less affected by head movements.

Recreating the perception of distance of the sources close to natural listening is another critical aspect in virtualization for natural sound rendering. However, the challenges in simulating accurate distance perception are aplenty. The ability of human beings to accurately estimate the distance has long been known to be poorer compared to their direction localization ability even in the physical listening space [1]. Virtual listening over headphones further hinders the distance perception as it leads to inside-the-head localization (IHL) of sound [1]. IHL of sound is caused by several factors, such as the use of non-individualized HRTFs, absence of equalization, lack of reverberation, impedance mismatch due to the presence of headphones [1], [13]. Presence of individualized HRTFs, equalization and reverberation can improve the externalization of sound but does not ensure accurate distance perception [1]. Direct to reverberation energy ratio is found to be the most critical cue for absolute distance perception, even though the intensity, loudness, and binaural cues can provide relative cues for distance perception [1]. Since reverberation is an essential cue for both distance perception and perception of a real environment context, a veridical simulation of the reverberation is highly imperative for natural sound rendering [1]. However, accurate simulation of distance perception is challenging since reverberation entirely depends on the room characteristics.

The correct amount of reverberation to be added to simulate distance perception in a particular room can be obtained only by carrying out acoustical measurements.

V. SOUND SCENE DECOMPOSITION USING BSS AND PAE

To achieve natural sound rendering in headphones, two important constituents of the sound scenes are required in the virtualization, namely, the individual sound sources and characteristics of the sound environment. However, this information is usually not directly available to the end user. One has to work with the existing digital media content that is available, i.e., the mastered mix distributed in channel-based formats (e.g., stereo, 5.1). Therefore, to facilitate natural sound rendering, it is necessary to extract the sound sources and/or sound environment from their mixtures. In this section, we discuss two types of techniques applied in sound scene decomposition, namely, BSS and PAE.

A. DECOMPOSITION USING BSS

Extracting the sound sources from the mixtures, often referred to as BSS, has been extensively studied in the last few decades. The basic mixing model in BSS can be considered as anechoic mixing, where the sources $s_k(n)$ in each mixture $x_m(n)$ have different gains g_{mk} and delays τ_{mk}. Hence, the anechoic mixing is formulated as follows:

$$x_m(n) = \sum_{k=1}^{K} g_{mk} s_k(n - \tau_{mk}) + e_m(n), \quad \forall m \in \{1, 2, \ldots, M\}, \tag{3}$$

where $e_m(n)$ is the noise in each mixture, which is usually neglected for most cases. Note that estimating the number of sources is quite challenging and it is usually assumed to be known in advance [14]. This formulation can be simplified to represent instantaneous mixing by ignoring the delays, or can be extended to reverberant mixing by including multiple paths between each source and mixture. An overview of the typical techniques applied in BSS is listed in Table I.

Based on the statistical independence and non-Gaussianity of the sources, independent component analysis (ICA) algorithms have been the most widely used techniques in BSS to separate the sources from mixtures in the determined case, where the numbers of mixtures and sources are equal [14]. In the over-determined case, where there are more mixtures than sources, ICA is combined with principal component analysis (PCA) to reduce the dimension of the mixtures, or combined with least-squares (LS) to minimize the overall mean-square error (MSE) [14]. In practice, the under-determined case is the most common, where there are fewer mixtures than sources. For the under-determined BSS, sparse representations of the sources are usually employed to increase the likelihood of sources to be disjoint [15]. The most challenging under-determined BSS is when the number of mixtures is two or lesser, i.e., in stereo and mono signals.

Stereo signals (i.e., $M = 2$), being one of the most widely used audio format, have been the focus in BSS. Many of these BSS techniques can be considered as time-frequency masking and usually assume one dominant source in one time-frequency bin of the stereo signal [16]. In these time-frequency masking based approaches, a histogram for all possible directions of the sources is constructed, based on the range of the bin-wise amplitude and phase differences between the two channels. The directions, which appear as peaks in the histogram, are selected as source directions. These selected source directions are then used to classify the time-frequency bins, and to construct the mask. For every time-frequency bin (n,l), the kth source at mth channel $\hat{S}_{mk}(n,l)$ is estimated as:

$$\hat{S}_{mk}(n,l) = \Psi_{mk}(n,l) X_m(n,l), \tag{4}$$

where the mask and the mth mixture are represented by $\Psi_{mk}(n,l)$ and $X_m(n,l)$, respectively.

In the case of single-channel (or mono) signals, the separation is even more challenging since there is no inter-channel information. Hence, there is a need to look into the inherent physical or perceptual properties of the sound sources. Non-negative matrix factorization (NMF) based approaches are extensively studied and applied in single-channel BSS in recent years. The key idea

of NMF is to formulate an atom-based representation of the sound scene [17], where the atoms have repetitive and non-destructive spectral structures. NMF usually expresses the magnitude (or power) spectrogram of the mixture as a product of the atoms and time varying non-negative weights in an unsupervised manner. These atoms, after being multiplied with their corresponding weights, can be considered as potential components of sources [18]. Another technique applied in single-channel BSS is the computational auditory scene analysis (CASA) that simulates the segregation and grouping mechanism of human auditory system [19] on the model-based representation (monaural case) of the auditory scenes. An important aspect worth considering is the directions of the extracted sources, which can usually come as a by-product in multichannel BSS. In single-channel BSS, this information of source directions has to be provided separately.

B. DECOMPOSITION USING PAE

In most sound scenes, the mixture comprises not only the dry sources but also the reverberation and ambient sound, which are contributed by the acoustics of the surrounding space. Therefore, the mixing model of the sources in BSS usually does not match with the actual sound scenes. In this paper, we refer to the dominant sources as primary (or direct) components, while the signals contributed by the sound environment as ambient (or diffuse) components. The primary and ambient components are perceived to be directional and diffuse, respectively. Different rendering methods should be applied to the primary and ambient components [6], [7] due to their perceptual differences. Therefore, rendering of natural sound scenes requires the decomposition of the mixtures into primary and ambient components [6], [7], [9]. Since stereo is still the most widely used format for digital media content, our discussion on the decomposition using primary-ambient extraction is focused on stereo signals ($M = 2$).

In PAE, we often follow some intuitive signal models as discussed in [3], [5], [7], [8], [20]. In the mth channel, the mixture $x_m(n)$ is assumed to be the sum of the primary component $p_m(n)$ and ambient component $a_m(n)$, i.e., $x_m(n) = p_m(n) + a_m(n)$. The discrimination of directional primary

components and diffuse ambient components is mainly based on their inter-channel correlations, where the primary and ambient components in the two channels are assumed to be correlated and uncorrelated, respectively. In the basic mixing model for PAE, the primary components are assumed to be amplitude panned, while the ambient components are of approximately equal levels in all channels.

Based on these assumptions, various approaches are proposed in PAE for stereo signals. Similar to BSS, time-frequency masking approaches are introduced to extract ambient components $\hat{A}_m(n,l)$ [7], [20] and these approaches can be generalized as

$$\hat{A}_m(n,l) = X_m(n,l)\Psi_A(n,l), \tag{5}$$

where $0 \le \Psi_A(n,l) \le 1$ is the real-valued ambient mask at time-frequency bin (n,l). Time-frequency bins having high inter-channel correlation are considered to be primary components (or mostly primary components in the soft masking case), whereas low correlation bins are more likely to be ambient components.

Several linear estimation based PAE approaches were also introduced [21], which exploits the differences between the two channels of the stereo signal to perform the primary-ambient extraction, including PCA based approaches [20] and LS based approaches. In these approaches, the extracted primary components $\hat{p}_0(n)$, $\hat{p}_1(n)$ and ambient components $\hat{a}_0(n)$, $\hat{a}_1(n)$ are expressed as weighted sums of the mixtures:

$$\begin{bmatrix} \hat{p}_0(n) \\ \hat{p}_1(n) \\ \hat{a}_0(n) \\ \hat{a}_1(n) \end{bmatrix} = \begin{bmatrix} w_{\mathrm{P0,0}} & w_{\mathrm{P0,1}} \\ w_{\mathrm{P1,0}} & w_{\mathrm{P1,1}} \\ w_{\mathrm{A0,0}} & w_{\mathrm{A0,1}} \\ w_{\mathrm{A1,0}} & w_{\mathrm{A1,1}} \end{bmatrix} \begin{bmatrix} x_0(n) \\ x_1(n) \end{bmatrix}. \tag{6}$$

The solutions for the weights in (6) are derived based on different performance-related criteria [21]. More specifically, PCA extracts the primary components having maximum variance, and extracts the ambient components having minimum variance with the constraint that the primary and

ambient components are uncorrelated, while LS extracts these components having minimum MSE. Based on the study in [21], it is recommended that PCA based approaches should be used for signals that contains dominant primary components (e.g., gaming), while LS based approaches are preferred for signals that contain a balanced mix of primary and ambient components (e.g., movies). In addition, to deal with more complex types of input signals that do not fit into the basic mixing model, other techniques have also been introduced, such as, time shifting to compensate for time differences [22] and adaptive frequency bin partitioning for multiple sources in primary components [23]. Furthermore, though it is possible to extend the framework of PAE from stereo signals to multichannel signals, e.g., [24], more comprehensive studies on PAE for multichannel signals are required.

C. A COMPARISON BETWEEN BSS AND PAE

Both BSS and PAE are extensively applied in sound scene decomposition, and a comparison between these approaches is summarized in Table II. The common objective of BSS and PAE is to extract useful information (mainly the sound sources and their directions) about the original sound scene from the mixtures, and to use this information to facilitate natural sound rendering. On this note, there are three common characteristics in BSS and PAE. First, only the mixtures are available and usually no other prior information is given. Second, the extraction of the specific components from the mixtures is based on certain signal models. Third, both techniques require objective and subjective evaluation.

As discussed earlier, the applications of different signal models in BSS and PAE lead to different techniques. In BSS, the mixtures are considered as the sums of multiple sources, and the independence among the sources is one of the most important characteristics. In contrast, the mixing model in PAE is based on human perception of directional sources (primary components) and diffuse sound environment (ambient components). The perceptual difference between primary and ambient components is due to the directivity of these components which can be characterized

by their correlations. The applications that adopted BSS and PAE also have distinct differences. BSS is commonly used in speech and music applications, where the clarity of the sources is usually more important than the effect of the environment. On the other hand, PAE is more suited for the reproduction of movie and gaming sound content, where the ambient components also contribute significantly to the naturalness and immersiveness of the sound scenes. Subjective experiments revealed that BSS and PAE based headphone rendering can improve the externalization and enlarge the sound stage with minimal coloration [6].

Despite the recent advances in BSS and PAE, the challenges due to the complexity and uncertainty of the sound scenes still remain to be resolved. One common challenge in both BSS and PAE is the increasing number of audio sources in the sound scenes, while only a limited number of mixtures (i.e., channels) are available. In certain time-frequency representations, the sparse solutions in BSS and PAE would require the sources to be sparse and disjoint [15]. Considering the diversity of audio signals, finding a robust representation for different types of audio signals is extremely difficult. The recorded or post-processed source signals might even be filtered due to physical or equivalently simulated propagation and reflections. Moreover, the audio signals coming from adverse environmental conditions (including reverberation, and strong ambient sound) usually degrade the performance of the decomposition. These difficulties can be addressed by studying the features of the resulting signals and by obtaining more prior information on the sources, the sound environment, the mixing process [18], and combining with visual information of the scene.

VI. INDIVIDUALIZATION OF HRTF

Binaural technology is the most promising solution for delivering spatial audio in headphones, as it is the closest to natural listening. Unlike conventional microphone recordings, which are meant for loudspeaker playback, the binaural signals are recorded or synthesized at the ears of the listener. In a binaural audio system, the spatial encoding (i.e., HRTFs) should encapsulate all the spectral

features due to the interaction of the acoustic wave with the listener's morphology (torso, head, and pinna). The pinna, which is also considered as the acoustic fingerprint, embeds the most idiosyncratic spectral features into HRTFs, which are essential for accurate perception of the sound (Fig. 3(a)). Thus, the HRTF features of the individuals are extremely unique, as shown in Fig. 3(c). Often the HRTFs used for virtualization are non-individualized HRTFs, typically measured on a dummy head, since they are easily accessible.

However, the use of non-individualized HRTFs leads to several artefacts like IHL, elevation confusions, and front-back, up-down reversals. Additionally, subjects display poor angular resolution and sometimes find it difficult to pinpoint the exact location of the auditory image in the case of using non-individualized HRTFs. Thus, individualization of the HRTFs (Fig. 3(b)) plays a critical role to create an immersive experience closest to the natural listening experience. There are various individualization techniques to obtain the individualized HRTFs from acoustical measurements, anthropometric features of the listener, customizing generic HRTFs with perceptual feedback or frontal projection of sound, as summarized in Table III.

Acoustical measurements: The most straightforward individualization technique is to actually measure the individualized HRTFs for every listener at different sound positions [25], [26]. This is the most ideal solution but it is extremely tedious and involves highly precise measurements. These measurements also require the subjects to remain motionless for long periods, which may cause fatigue to the subjects. Zotkin *et al.* developed a fast HRTF measurement system using the technique of reciprocity, where a micro-speaker is placed into the ear and several microphones are placed around the listener [13]. Other researchers developed a continuous 3D azimuth acquisition system to measure the HRTFs using a multichannel adaptive filtering technique [27]. However, all these techniques to acoustically measure the individual HRTFs require a large amount of resources and expensive setups.

Anthropometric data: Individualized HRTFs can also be modelled as weighted sums of basis functions, which can be performed either in the frequency or spatial domain. The basis functions

are usually common to all individuals and the individualization information is often conveyed by the weights. The HRTFs are essentially expressed as weighted sums of a set of eigen vectors, which can be derived from PCA or ICA [26], [13]. The individual weights are derived from the anthropometric parameters that are captured by optical descriptors, which can be derived from direct measurements, pictures or a 3D mesh of the morphology [13]. The solution to the problem of diffraction of an acoustic wave with the listener's body results in individual HRTFs. This solution may be obtained by analytical or numerical methods, such as the boundary element method (BEM) or the finite element method (FEM) [13], [26]. Other methods used include multiple linear regressions [26], multiway array analysis [28], and artificial neural networks [26]. The inputs to these methods can be a simple geometrical primitive [29] (e.g., a sphere, cylinder or an ellipsoid), a 3D mesh obtained from MRI or laser scanner or a set of 2D images [13]. An important advantage of these techniques is that the relative effects of a particular morphological element (e.g., torso, head, and pinna) and their variation with size, location and shape can be independently investigated [13]. Another technique used a simple customization technique, where a HRTF is selected by matching certain anthropometric parameters [30]. One of the major challenges today to numerically model the HRTF is the very high resolution of imaging techniques required for accurate prediction of HRTFs at high frequencies. The required resolution of the mesh imaging depends on the shortest wavelength, which is around 17mm at 20 kHz [13]. Moreover, obtaining these optical descriptors demands for the use of extremely expensive laser, MRI scanners, and also requires highly skilled qualified users.

Perceptual feedback: Several attempts have been carried out to personalize HRTF from a generic HRTF database using perceptual feedback. Subjects select the HRTFs through listening tests, where they choose the HRTFs based on the correct perception of frontal sources and reduced front-back reversals [13]. Listeners can also adapt to the non-individualized HRTF by modifying the HRTFs to suit his or her perception. Middlebrooks observed that the peaks and notches of HRTFs are frequency shifted for different individuals and that the extent of the shift is related to the

size of pinna [31]. Listeners often tune the spectrum until they achieve a good and natural spatialization [13]. Other techniques involve active sensory tuning [26], and tuning the PCA weights [32] to individualize the HRTFs. These perceptual based methods are much simpler in terms of the required resources, and effort compared to the individualization methods using acoustical measurements or anthropometric data. However, these listening sessions can sometimes be quite long and result in listener fatigue.

Frontal projection playback: More recently, a study by Sunder *et al.* [33] customized the non-individualized HRTFs using a frontal projection headphone. Unlike side projection of sound in conventional headphones, a frontal projection headphone projects the sound from the front to emulate the playback from a physical set of loudspeakers. By projecting the sound from the front, the idiosyncratic frontal pinna spectral cues of the listener are captured inherently during the playback [33]. It is found that the idiosyncratic high frequency pinna cues captured in the frontal projection headphones response match well with the frontal HRTF cues, giving it a better frontal perception (as shown in Fig. 4). The authors reported that the front-back reversals reduced by almost 50% [33] using the frontal projection headphone, thus improving the veracity of the 3D audio. The advantage of this technique is that it does not require any measurements, training or the anthropometric data of the listener. However, the frontal projection individualization technique has been limited to only the horizontal plane and also requires a special kind of headphone equalization (Type-2).

As discussed in Section IV, head tracking is important in the virtualization process. It was found that head tracking, when used with non-individualized HRTFs, can improve the localization [10]. However, head tracking primarily helps in reducing the front-back confusions and has minimal effect in reducing the elevation localization errors, IHL [10], and coloration caused by non-individualized HRTFs. Since individualization of HRTFs can alleviate some of these limitations, it is suggested that head tracking be used with individualized rendering.

To sum up, there is a noticeable trend to achieve more and more accurate individualization with lesser data, complexity and effort. However, the effect of individualization of HRTFs can be hindered by the presence of the headphone. Hence, the headphone has to be compensated to ensure that the spectrum at the eardrum has only the individualized HRTF features. Additionally, equalization of the binaural recording itself may be necessary in certain applications (e.g., musical recordings). The challenges and methods of equalization for both binaural and stereo recordings are explained in the next section.

VII. EQUALIZATION

Headphones are not acoustically transparent as they not only color the sound that is played from the headphone but also affect the free-air characteristics at the ear. Typically the HPTF comprises of the headphones transducer response and the acoustic coupling between the headphones and the listener's ears. To compensate for the headphone response, the HPTF is first measured at the same point where the recording was carried out at the blocked ear canal or at the eardrum [35]. The binaural recording is then de-convolved with the HPTF to eliminate the effect of the recording microphones and the headphone. This type of direct equalization is also known as the "non-decoupled" mode of equalization (Table IV) [36]. This method is often used when the HPTF is measured with the same measurement setup as the recording and particularly works well when the HPTF measurement and recording are carried out on the same dummy head.

It is observed that, in the absence of headphone equalization, the front-back reversals are increased and the elevation localization is distorted [1], [26], [13]. Thus, headphone equalization is critical to create a convincing perception of virtual sound sources. However, headphone equalization is challenging since the HPTF depends on individual morphology (headphone-ear coupling). Researchers have also reported that the use of non-individualized equalization can reduce the externalization and the effect can be as critical as the use of non-individualized HRTFs [13]. Thus, equalization using individual HPTFs is strongly recommended. Another difficulty in

carrying out accurate headphone equalization is the variability of the HPTFs with repositioning. The effect of repositioning of headphones is lower at low frequencies but displays high standard deviations up to 10 dB at high frequencies [37]. Kulkarni *et al.* [37] observed that equalization based on a single measurement may become worse than no equalization at all. The positional dependency has no specific solution and its effect can only be reduced by taking the average of a number of trials as a representative HPTF [37]. Thus, to create a convincing immersive sound environment, use of individualized HRTFs and individualized equalization is entailed, which may not be viable all the time. To reduce the dependency on individualized equalization, Sunder *et al.* [33] designed a Type-2 equalization technique for the playback through frontal projection headphone, which is independent of the headphone-ear coupling. Unlike the conventional equalization technique, Type-2 equalization compensates only for the distortion due to the emitter, thereby preserving the individual pinna cues due to frontal projection.

The other type of equalization is the "decoupled" equalization technique and it is the most commonly used method of equalization for rendering music. In this technique, the binaural recording (BIR or HRTFs) as well as the headphone are equalized using a reference sound field (e.g., free-field, diffuse-field, etc.) [36]. If the reference sound field (REF) of the recording environment is well known and reproduced reliably, this method of equalization can result in a very natural perception of sound similar to the non-decoupled equalization technique. This method of equalization is mainly carried out to make the binaural recordings compatible with stereophonic (conventional microphone) recordings in terms of timbral quality.

If the recording is binaural, then a reference field equalized binaural recording (BIR/REF) achieves a sound quality equivalent to a conventional microphone recording. When the equalized recording is played from a reference field equalized headphone (HPTF/REF), the perceived timbre of the spatial sound would be as natural as the original binaural recording. Individualized binaural recordings are thus necessary in order to experience the true immersiveness of sound without any timbral coloration and spatial degradation. Note that for rendering conventional stereo recorded

music, it is sufficient to carry out just the headphone equalization using an appropriate reference field. Some of the commonly used reference fields are:

1) **Free-field (FF) equalization**: With the aim to replicate the ear signals produced by frontal loudspeakers, the target response of FF equalization is the HRTF of frontal incidence. Hammershoi *et al.* proposed an FF equalization curve, which has additional high frequency energy above 3 kHz to approximate listening to stereo loudspeakers in the free-field [4]. A FF equalized headphone can reproduce a frontal sound with natural sound quality but colors the sound that originates from other directions. Moreover, it is important to note that there are large inter-individual variations in the FF equalization filters [38].

2) **Diffuse-field (DF) equalization**: In this case, the target response for equalization is the diffuse-field response, i.e., the average of the HRTFs of all measured directions in horizontal plane. The inter-individual variations are reduced drastically due to the averaging effect [38]. Thus, the DF target response can be achieved universally over a great number of individuals. Møller [35] identified certain headphones which are already DF equalized and recommended such type of headphones for stereo listening.

3) **Other target responses**: A typical listening room is not completely diffuse but it can be considered somewhere between a free-field and a diffuse-field. Møller [38] illustrated other alternative target responses which are partially diffuse by applying unequal weighting to different directions within ± 45 degrees azimuth and elevation. Other researchers also modified the DF equalization filters with the help of certain parametric filters and found that the subjects generally preferred the target response with a 3 kHz peak lower in amplitude than in the diffuse-field response for both music and speech [4]. Recent experiments [4], [38] showed that listeners prefer other alternative target responses more than the conventional FF and DF equalizations. Examples of these preferred target curves include RR_G and RR1_G proposed by Olive *et al*. [4] based on the impulse response of the loudspeaker system in the Harman Reference rooms.

Ideally, the best reference field that preserves the true quality of the recording would be the field where the recording is carried out. Furthermore, the choice of headphones can also greatly affect the transparency of the binaural rendering even with the correct headphone equalization. The external ear is un-hindered in the natural listening conditions, where the sound pressures at the ear are governed by free-air characteristics. With headphones placed over the ear, the pressure characteristics of the sound arriving at the eardrum are greatly affected compared to the free-air characteristics due to the interaction between the external ear and the headphone enclosure. The closer the coupling characteristic of the headphones with that of the free-air, the more accurate and transparent is the reproduced sound. Møller [35] defined the effect of the headphone for a binaural recording at the blocked ear canal in terms of the electrical transmission gain, *G*:

$$G = \left(\frac{1}{\mathrm{MPTF} \cdot \mathrm{HPTF}} \right) \cdot \mathrm{PDR}, \tag{7}$$

where MPTF is the transfer function of the recording microphone, and PDR is the pressure division ratio. PDR is defined as the ratio of the equivalent thevenin impedances when the ear is in free-air to the case when the headphone is placed on the ear, and is given as [35]:

$$\mathrm{PDR} = \frac{Z_{\mathrm{earcanal}} + Z_{\mathrm{headphones}}}{Z_{\mathrm{earcanal}} + Z_{\mathrm{radiation}}}, \tag{8}$$

where Z_{earcanal} and $Z_{\mathrm{headphones}}$ are the input impedances of the ear canal and the impedance of the headphone, respectively; $Z_{\mathrm{radiation}}$ is the free-air radiation impedance as seen from the ear canal. The PDR reduces to unity when the pressures in the free-air and with headphones become equal. Such headphones are defined as FEC (free-air equivalent coupling) headphones, which are also sometimes termed as “open headphones” [35]. The “open headphones” is different from the commercially available “open-back headphones”. Most of the commercially available headphones have less than ideal FEC characteristics [35]. It is important to note that the FEC condition for the headphone is necessary only for binaural recordings made at the blocked ear canal, which is also the most common technique for individualized binaural recording [35]. In such a case, headphone

equalization alone is sufficient to achieve auralization transparency. To summarize, equalization (both recording and playback) and individualization play a critical role in the natural rendering of sound of any formats (binaural or stereo) over headphones.

VIII. INTEGRATION OF NATURAL SOUND RENDERING TECHNIQUES

An integration of these signal processing techniques for natural sound rendering reviewed in this paper is depicted in Fig. 5. The original sound sources along with their environmental information are represented as a sound mixture after the mixing process. The sound scenes from the mix are then decomposed into primary components (sources) and/or ambient components (environment) using BSS and/or PAE. The extracted primary components, which are basically directional sound sources as perceived by the listener, can be rendered using (individualized) HRTFs [1]. Ambient components are rendered in a manner so as to recreate a natural sound environment. Modelling the acoustics of the natural sound environment by adding the correct amount of early reflections and reverberation also helps in enhancing the perception of the sound environment as well as veridical distance, which is critical for natural listening. Moreover, a suitable individualization technique has to be applied to the directional sources such that the rendered sound scenes played over headphones are maximally tailored for the individual listener. Meanwhile, use of a robust equalization technique can significantly reduce the adverse coloration of the source. Finally, the influence of the head movements on the rendered sound can be taken into account by incorporating head tracking in virtualization.

In general, natural sound rendering requires both the spatial and timbral quality of the reproduced sound to be realistic. For digital media content that contains plenty of spatial cues (e.g., movies, games), all the five techniques reviewed are important in creating a sense of immersiveness. For other content, where the timbral quality is of utmost importance (e.g., music recordings), a subset of the techniques (e.g., individualization, equalization) are sufficient in natural sound rendering.

IX. SUBJECTIVE EXPERIMENTS

Subjective experiments were carried out to validate the reviewed natural sound rendering system by comparing it with the conventional stereo playback system. A total of 18 subjects (15 males and 3 females), who were all between 20-30 years old, participated in this listening experiment. None of the subjects reported any hearing loss. The test was conducted in a semi-anechoic listening room at NTU, Singapore. The two systems of headphone listening tested in this experiment were:

(i) Conventional stereo system. The materials are directly played back over headphones without any processing.

(ii) Natural sound rendering system. The signal processing techniques introduced in the paper were applied to the audio content. In this study, we chose PAE as the sound scene decomposition method since our primary interest lies in movie and gaming audio content that contains the individual sound sources and the sound environment [21]. Individualization is carried out by frontal projection headphone pinna cues during playback and does not require any individual acoustical experiments, anthropometric data or training [33]. To fully exploit the frontal projection in the natural sound rendering, we have developed a new four-emitter headphone [39] that houses a frontal emitter and a conventional side emitter in each ear cup of the headphone [33]. In the virtualization process, the frontal emitters are used to render the directional sources, while all the emitters (both frontal and side) are used to render the sound environment. Type-2 EQ is applied to the frontal emitters for source rendering [33], and diffuse-field EQ is used to render environment signals over all the emitters. Head tracking has not been incorporated in this system.

The stimuli used in this experiment were binaural (motorcycle in a storm and bee at a waterfall), movie (Brave, Prometheus), and gaming tracks (Battlefield 3), which contain plenty of spatial cues. Each track was played back using the two headphone playback systems tested here. The tracks corresponding to the two systems were named “A” and “B” and played back in a random order. The

listening tests were conducted in a double-blind manner, where both the experimenter and the subjects were unaware of the order of the stimuli. In this experiment, four audio quality measures were considered to evaluate the performance of the two systems. Their descriptions are given below:

1. *Sense of direction*: how clear or distinct are the perceived directions of the sound objects?
2. *Externalization*: how clear is the stimulus perceived outside the head?
3. *Ambience*: how clear and natural is the ambience of the sound environment perceived?
4. *Timbral quality*: how realistic is the timbral quality of the sound?

Subjects were asked to give the scores for the four measures for each of the two tracks "A" and "B". The scores were based on a 0-100 scale where subjects rated 0-20 (Bad), 21-40 (Poor), 41-60 (Fair), 61-80 (Good), and 81-100 (Excellent). Finally, the subjects were also required to indicate their overall preference for the two tracks by selecting one of the following three choices: "Prefer A", "Not sure", or "Prefer B". To carry out this experiment, a GUI was created which randomized the order of the stimuli and automatically stored the responses of the subjects in a file.

The responses of the subjects were analyzed for both sound rendering systems. Fig. 6 shows the overall comparison between the two systems in terms of the mean opinion score (MOS), scatter plot and the overall preference of the subjects. In Fig. 6(a), MOS of the four measures for the two systems were computed across all the 18 subjects and 5 stimuli. While the MOS for the conventional stereo system for all the measures were around 60, the natural sound rendering system performed much better with MOS of over 70. An analysis of variance (ANOVA) was conducted to generalize these results to the whole population of listeners. The *p*-values were found to be very small ($<< 0.01$) for all the measures, indicating that the improved performance of the natural sound rendering system over the conventional stereo system is statistically significant. The scatter plot in Fig. 6(b) implies that most of the subjects gave a higher score for the natural sound rendering system for all the four measures. The overall preference of the subjects across all the five tracks is

shown in Fig. 6(c). The pie chart suggests that 61% of the subjects preferred the natural sound rendering, while only 33% preferred the conventional stereo rendering.

To sum up the subjective test results, we found that the natural sound rendering system using the various signal processing techniques explained in this paper enhances the listening experience compared to a conventional stereo system. Additionally, the presence of head tracking in the system will only improve the natural sound rendering as observed in several studies [10].

X. CONCLUSIONS AND FUTURE TRENDS

With the advent of low cost, low power, small form factor, and high speed multi-core embedded processor, we can now implement the above signal processing techniques in real-time and embed processors into the headphone design. However, various implementation issues regarding the computation cost of sound scene decomposition, HRTF/BRIR filtering in virtualization, and equalization as well as the latency in head tracking should be carefully considered. One example of such a natural sound rendering system is the four-emitter 3D audio headphone [39] developed at the DSP Lab in NTU. This system has been psychophysically validated and found to perform much better than the conventional stereo headphone playback system.

Besides the five types of techniques discussed in this paper, there have been other efforts to enhance the natural experience of headphone listening. To enable the natural pass through of the sound from outside world without coloration, headphones can be designed with suitable acoustically transparent materials. When this is not effective, microphones integrated into headphones and associated signal processing techniques, such as equalization, and active noise control are employed. The headphones with built-in microphones open a new dimension to augment the listening experience with the physical world.

The future of headphones for assistive listening applications would be the one where listeners cannot differentiate between the virtual acoustic space created from headphone playback and the

real acoustic space. This would require the combined effort from the whole audio community from the headphone manufacturers, sound engineers to audio scientists. More information about the content production has to be distributed from the content developers to the end user to enhance the extraction process. Moreover, obtaining and exploiting every individual's anthropometrical features or hearing profiles is crucial for a natural listening experience. Finally, with more sensors, such as GPS, gyroscopes, and microphones that can be integrated into headphones, future headphones can be more location-aware, content-aware listener-aware, and hence become more intelligent and assistive.

ACKNOWLEDGEMENTS

This work is supported by the Singapore National Research Foundation Proof-of-Concept program under grant NRF 2011 NRF-POC001-033. We thank the guest editors and reviewers for their constructive comments and suggestions.

AUTHORS

Kaushik Sunder (KAUSHIK1@e.ntu.edu.sg)received his B.Tech degree in electrical and electronics engineering from the National Institute of Technology Karnataka, Surathkal, India, in 2011. He is currently pursuing his Ph.D. degree in electrical and electronics engineering at the Nanyang Technological University, Singapore. His research interest includes spatial audio, psychoacoustics, and music signal processing.

Jianjun He (JHE007@e.ntu.edu.sg)received his BEng degree in Automation from Nanjing University of Posts and Telecommunications, P. R. China in 2011 and is currently pursuing his Ph.D. degree in Electrical and Electronic Engineering in Nanyang Technological University, Singapore. His research interests include: audio and acoustic signal processing, 3D audio, psychoacoustics, active noise control, and emerging audio and speech applications.

Ee-Leng Tan (ETanEL@ntu.edu.sg)received his BEng (1st Class Hons) and PhD degrees in Electrical and Electronic Engineering from Nanyang Technological University in 2003 and 2012, respectively. Currently, he is with NTU as a research fellow. His research interests include image/audio processing and real-time digital signal processing.

Woon-Seng Gan (ewsgan@ntu.edu.sg) received his BEng (1st Class Hons) and PhD degrees, both in Electrical and Electronic Engineering from the University of Strathclyde, UK in 1989 and 1993 respectively. He is currently an Associate Professor and the Head of Information Engineering Division, School of Electrical and Electronic Engineering in Nanyang Technological University. His research interests span a wide and related areas of adaptive signal processing, active noise control, and directional sound system.

REFERENCES

[1] D. R. Begault, *3-D sound for virtual reality and multimedia*: AP Professional, 2000.
[2] S. Spors, H. Wierstorf, A. Raake, F. Melchior, M. Frank, and F. Zotter, "Spatial sound with loudspeakers and its perception: A review of the current state," *Proc. IEEE,* vol. 101, no. 9, pp. 1920-1938, Sep. 2013.
[3] V. Pulkki, "Spatial sound reproduction with directional audio coding," *J. Audio Eng. Soc.,* vol. 55, no.6, pp. 503-516, Jun. 2007.
[4] S. Olive, T. Welti, and E. McMullin, "Listener Preferences for Different Headphone Target Response Curves," in *Proc. 134th Audio Engineering Society Convention*, Rome, Italy, May 2013.
[5] M. M. Goodwin and J. M. Jot, "Binaural 3-D audio rendering based on spatial audio scene coding," in *Proc. 123rd Audio Engineering Society Convention*, New York, Oct. 2007.
[6] J. Breebaart and E. Schuijers, "Phantom materialization: a novel method to enhance stereo audio reproduction on headphones," *IEEE Trans. Audio, Speech, and Language Processing,* vol. 16, no.8, pp. 1503-1511, Nov. 2008.
[7] C. Avendano and J. M. Jot, "A frequency-domain approach to multichannel upmix," *J. Audio Eng. Soc.*, vol. 52, no.7/8, pp. 740-749, Jul. 2004.
[8] C. Faller, "Multiple-loudspeaker playback of stereo signals," *J. Audio Eng. Soc.*, vol. 54, no.11, pp. 1051-1064, Nov. 2006.
[9] F. Menzer and C. Faller, "Stereo-to-binaural conversion using interaural coherence matching," in *Proc. 128th Audio Engineering Society Convention*, London, UK, May 2010.
[10] D. R. Begault, E. M. Wenzel, and M. R. Anderson, "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source," *J. Audio Eng. Soc.*, vol. 49, no. 10, pp. 904-916, Oct. 2001.
[11] C. Faller and F. Baumgarte, "Binaural cue coding—part II: schemes and applications," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 520–531, Nov. 2003.
[12] V. R. Algazi and R. O. Duda, "Headphone-based spatial sound," *Signal Processing Magazine, IEEE,* vol. 28, no. 1, pp. 33-42, Jan. 2011.
[13] R. Nicol, *Binaural Technology*: AES, 2010.
[14] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*. New York: John Wiley & Sons, 2004.

[15] M. D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M. E. Davies, "Sparse representations in audio and music: from coding to source separation," *Proc. IEEE,* vol. 98, no. 6, pp. 995-1005, Jun. 2010.
[16] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Processing,* vol. 52, no.7, pp. 1830-1847, Jul. 2004.
[17] T. Virtanen, "Sound source separation in monaural music signals," PhD Thesis, Tampere University of Technology, 2006.
[18] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, "From blind to guided audio source separation," *IEEE Signal Processing Magazine,* vol. 31, no. 3, pp. 107-115, 2014.
[19] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications.* NJ: Wiley-IEEE Press, 2006.
[20] J. Merimaa, M. M. Goodwin, and J. M. Jot, "Correlation-based ambience extraction from stereo recordings," in *Proc. 123rd Audio Engineering Society Convention,* New York, Oct. 2007.
[21] J. He, E. L. Tan, and W. S. Gan, "Linear estimation based primary-ambient extraction for stereo audio signals," *IEEE/ACM Trans. Audio, Speech, and Language Processing,* vol. 22, no.2, pp. 505-517, 2014.
[22] J. He, E. L. Tan, and W. S. Gan, "Time-shifted principal component analysis based cue extraction for stereo audio signals," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'13)*, Canada, May 2013, pp. 266-270.
[23] J. He, E. L. Tan, and W. S. Gan, "A study on the frequency-domain primary-ambient extraction for stereo audio signals," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'14)*, Florence, Italy, 2014, pp. 2892-2896.
[24] J. Thompson, B. Smith, A. Warner, and J. M. Jot, "Direct-diffuse decomposition of multichannel signals using a system of pair-wise correlations," in *Proc. 133rd Audio Eng. Soc. Conv.*, San Francisco, 2012.
[25] H. Møller, M. F. Sørensen, D. Hammershøi, and C. B. Jensen, "Head-related transfer functions of human subjects," *J. Audio Eng. Soc.*, vol. 43, no. 5, pp. 300-321, May 1995.
[26] S. Xu, Z. Li, and G. Salvendy, "Individualization of head-related transfer function for three-dimensional virtual auditory display: a review," in *Virtual Reality*, ed: Springer, 2007, pp. 397-407.
[27] G. Enzner, "3D-continuous-azimuth acquisition of head-related impulse responses using multi-channel adaptive filtering," *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2009, pp. 325-328.
[28] M. Rothbucher, M. Durkovic, H. Shen, and K. Diepold, "HRTF customization using multiway array analysis," in *Proc. 18th European Signal Processing Conference (EUSIPCO'10)*, Aalborg, August 2010, pp. 229-233.
[29] R. O. Duda, V. R. Algazi, and D. M. Thompson, "The use of head-and-torso models for improved spatial sound synthesis," in *Proc. 113th Audio Engineering Society Convention*, Los Angeles, Oct. 2002.
[30] D. N. Zotkin, J. Hwang, R. Duraiswaini, and L. S. Davis, "HRTF personalization using anthropometric measurements," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'03)*, New York, Oct. 2003, pp. 157-160.
[31] J. C. Middlebrooks, "Individual differences in external-ear transfer functions reduced by scaling in frequency," *J. Acoust. Soc. Amer.*, vol. 106, no. 3, pp. 1480-1492, Sep. 1999.
[32] K. J. Fink and L. Ray, "Tuning principal component weights to individualize HRTFs," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'12)*, Kyoto, Mar. 2012, pp. 389-392.
[33] K. Sunder, E. L. Tan, and W. S. Gan, "Individualization of binaural synthesis using frontal projection headphones," *J. Audio Eng. Soc.*, vol. 61, no. 12, pp. 989-1000, Dec. 2013.
[34] A. Bondu, S. Busson, V. Lemaire, and R. Nicol, "Looking for a relevant similarity criterion for HRTF clustering: a comparative study," in *Proc. 120th Audio Engineering Society Convention*, Paris, France, May 2006.
[35] H. Møller, D. Hammershoi, C. B. Jensen, and M. F. Sorensen, "Transfer characteristics of headphones measured on human ears," *J. Audio Eng. Soc.*, vol. 43, no. 4, pp. 203-217, Apr. 1995.
[36] V. Larcher, J. M. Jot, and G. Vandernoot, "Equalization methods in binaural technology," in *Proc. 105th Audio Engineering Society Convention*, SanFrancisco, Sep. 1998.
[37] A. Kulkarni and H. S. Colburn, "Variability in the characterization of the headphone transfer-function," *J. Acoust. Soc. Amer.*, vol. 107, no. 2, pp. 1071-1074, Feb. 2000.
[38] H. Møller, C. B. Jensen, D. Hammershøi, and M. F. Sørensen, "Design criteria for headphones," *J. Audio Eng. Soc.*, vol. 43, no. 4, pp. 218-232, Apr. 1995.
[39] W. S. Gan and E. L. Tan, "Listening device and accompanying signal processing method," US Patent 2014/0153765 A1, 2014.

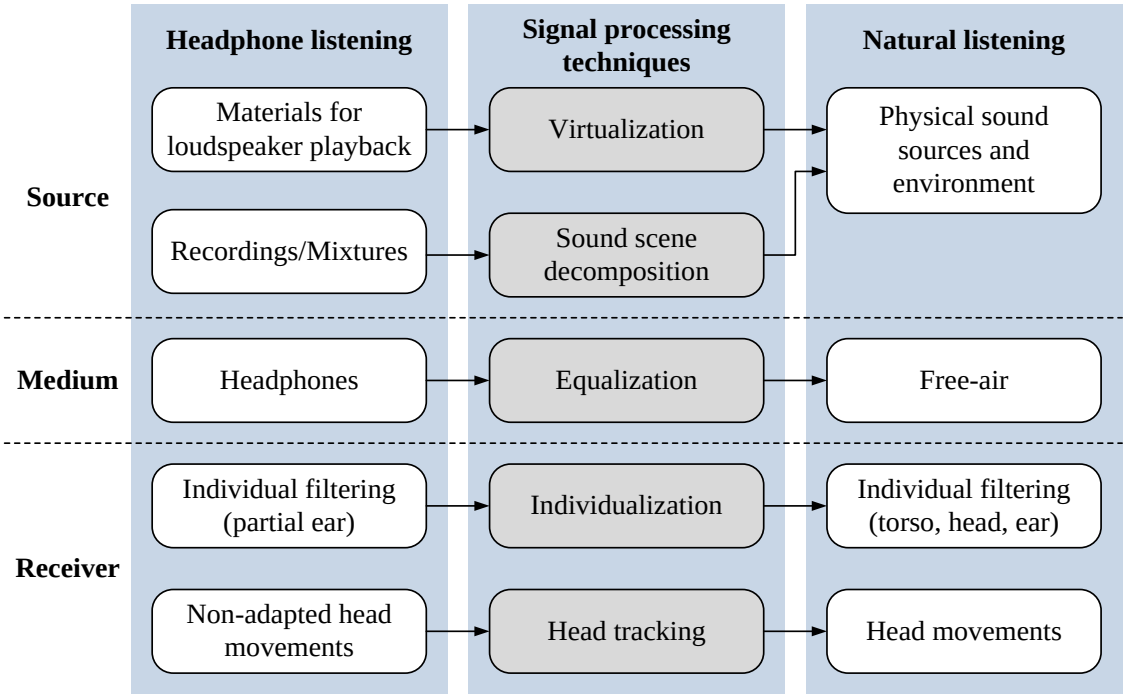
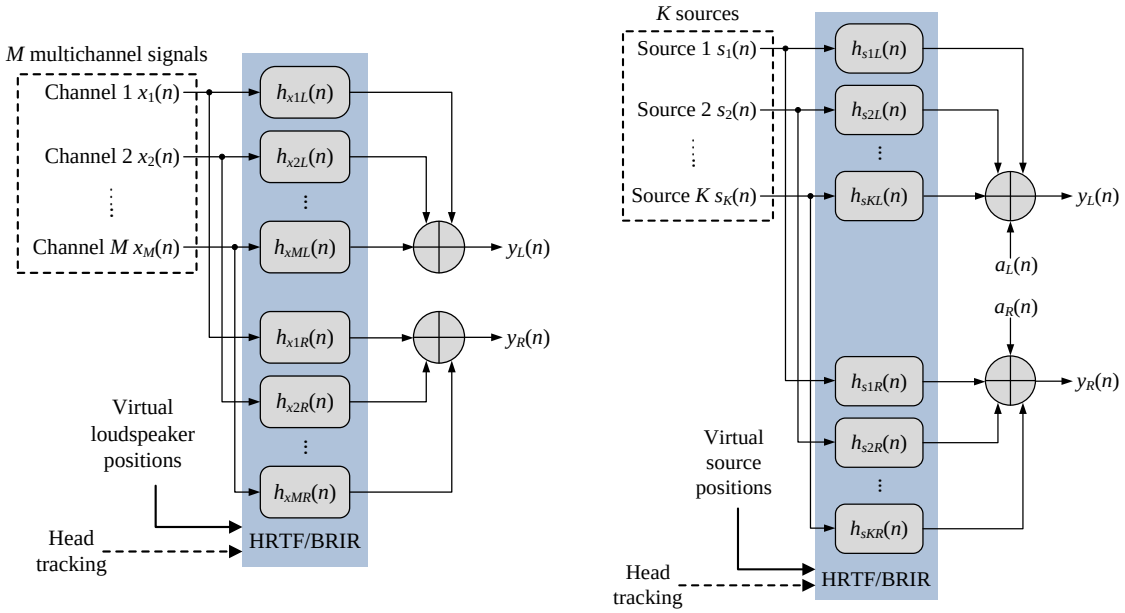


Fig. 1. A summary of the differences between natural listening and headphone listening and the corresponding signal processing techniques to solve these challenges for natural sound rendering. The main challenges and their corresponding signal processing techniques in each category (source, medium, and receiver) are highlighted and their interactions (not shown here) are further discussed in the paper.



(a) (b)

Fig. 2. Virtualization of (a) multichannel loudspeaker signals $x_m(n)$ (adapted from [5]), and (b) multiple sources $s_k(n)$ and environment signals $a_L(n), a_R(n)$. $y_L(n), y_R(n)$ is the signal sent to the left and right ear, respectively. Note that head tracking can be used to update the selected directions of HRTFs/BRIRs.

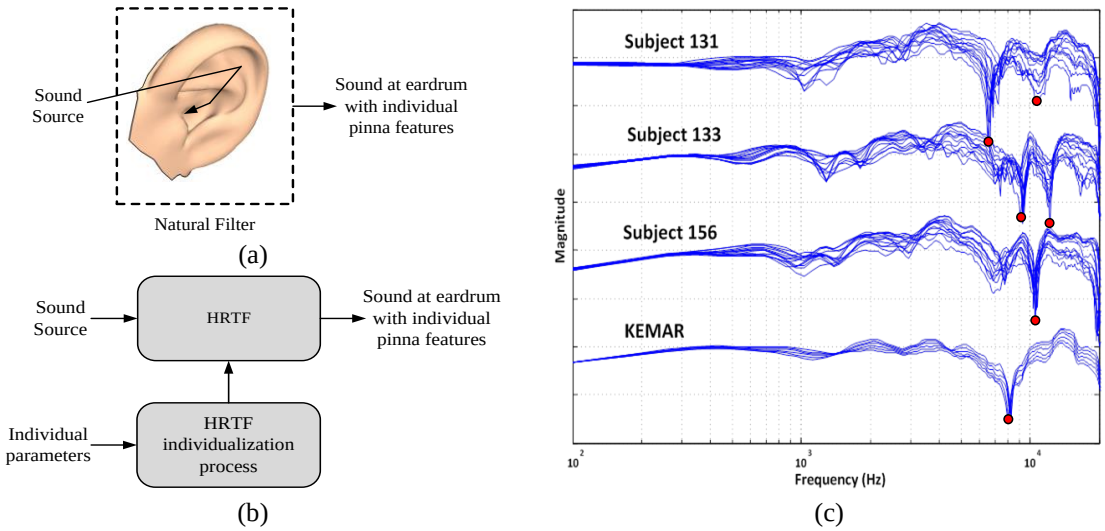


Fig. 3. (a) Human ears act as a natural filter in physical listening. (b) The natural HRTF filter is modelled by a digital filter using various individualization techniques. (c) Note the vast variation of the HRTF spectrum at high frequencies of the various subjects taken from CIPIC database and the MIT KEMAR dummy head database [26]. This is due to the idiosyncratic nature of the pinna.

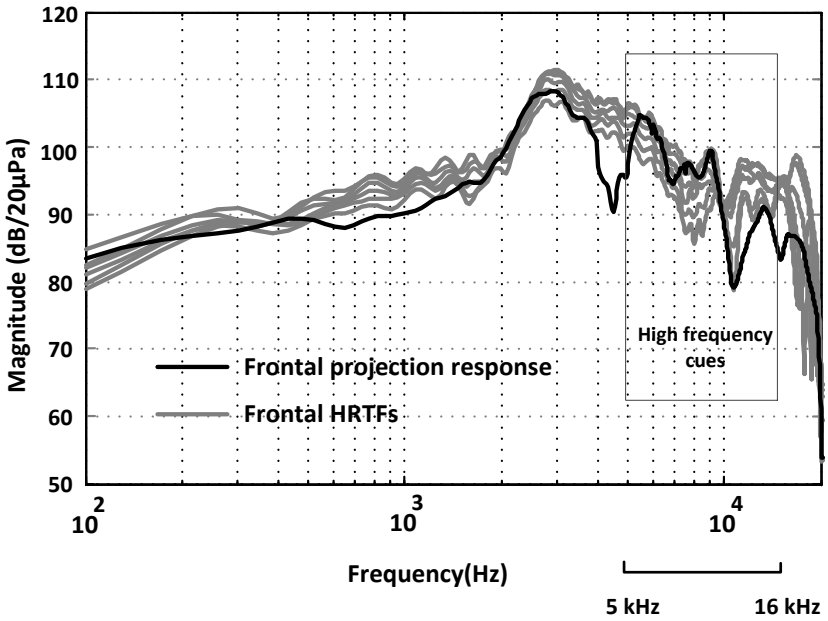


Fig. 4. Comparison of the frontal projection headphone response and the frontal directional HRTFs measured on a dummy head. Figure extracted from Sunder *et al.* [33].

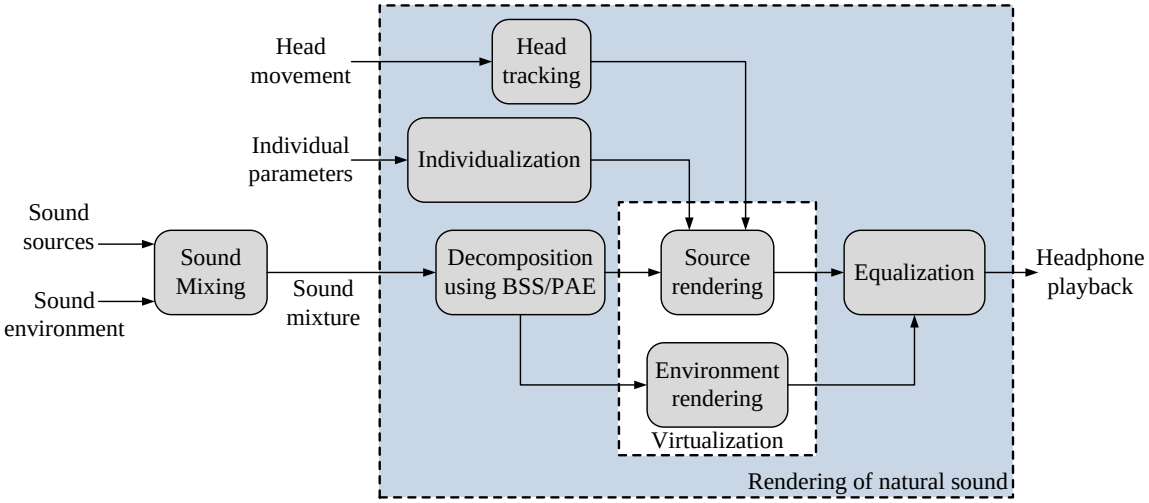


Fig. 5. Natural sound rendering system for headphones: an integration of all the signal processing techniques reviewed in this paper.

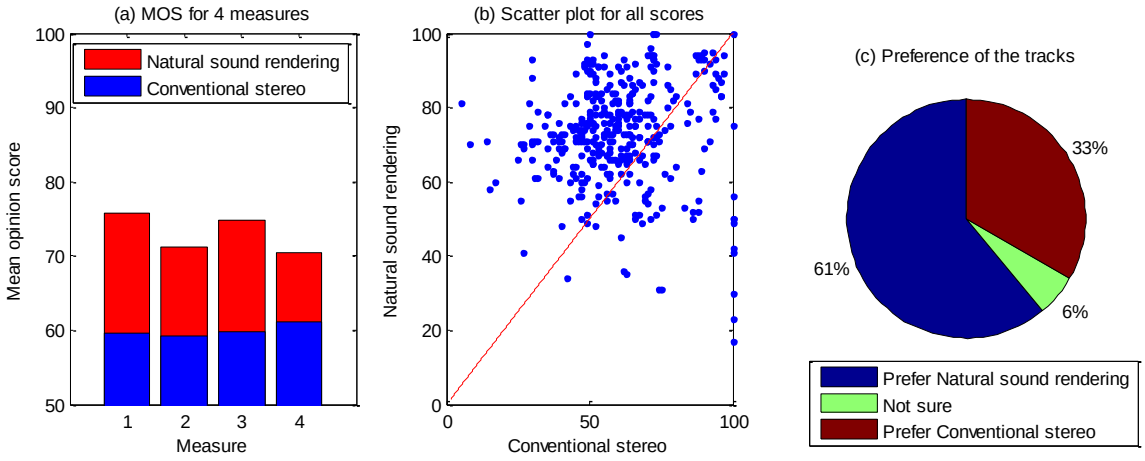


Fig. 6. Results of the subjective experiments: (a) MOS, (b) scatter plot, and (c) overall preference.

TABLE I
OVERVIEW OF TYPICAL TECHNIQUES IN BSS

Objective: To extract K ($K > 2$) sources from M mixtures		
Case		**Typical techniques**
Determined: $K = M$		ICA [14]
Over-determined: $K < M$		ICA with PCA or LS [14]
Under-determined: $K > M$	$M > 2$	ICA with sparse solutions [14], [15]
	$M = 2$	Time-frequency masking [16]
	$M = 1$	NMF [17], [18]; CASA [19]

TABLE II
COMPARISON BETWEEN BSS AND PAE IN SOUND SCENE DECOMPOSITION

	BSS	**PAE**
Objective	To obtain useful information about the original sound scene from given mixtures, and facilitate natural sound rendering.	
Common characteristics	• Usually no prior information, only mixtures; • Based on certain signal models; • Require objective as well as subjective evaluation.	
Basic mixing model	Sums of multiple sources (independent, non-Gaussian, etc.)	Primary components (highly correlated)+ Ambient components (uncorrelated)
Techniques	ICA [14], sparse solutions [15], time-frequency masking [16], NMF [17], [18], CASA [19], etc.	PCA [20], LS [8],[21], time-frequency masking [7],[20], time/phase-shifting [22], [23], etc.
Typical applications	Speech, music	Movie, gaming
Related applications	Speech enhancement, noise reduction, speech recognition, music classification	Sound reproduction, sound localization, coding
Limitations	• Small number of sources • Sparseness/disjoint • No/simple environment	• Small number of sources • Sparseness/disjoint • Low ambient power • Primary ambient components uncorrelated

TABLE III
COMPARISON OF THE VARIOUS HRTF INDIVIDUALIZATION TECHNIQUES

How to obtain individual features	Techniques	Pros	Cons	Performance and remarks
Acoustical Measurements	Individual measurements [25], IRCAM France, CIPIC, Uni. of Maryland, Tohoku Uni, Nagoya Uni, Austrian Academy of Sciences [26]	Ideal, accurate	Requires high precision; tedious; impractical for every listener	Reference for individualization techniques
Anthropometric data	Optical Descriptors : 3D mesh, 2D pictures [13]	Based on acoustic principles; studies the effects of independent elements of the morphology	Need a large database; Tedious; Requires high resolution imaging; Expensive equipments; Qualified users	Uses the correlation between individual HRTF and anthropometric data
	Analytical or Numerical Solutions: PCA + multiple linear regression [26] Finite element method, boundary element method [26], [13], Multiway array analysis [28], Artificial neural network [26]			
	Structural model of HRTFs [13], HRTF database matching [30]			
Listening/ Training	Selection from non-individualized HRTF [13], Frequency scaling [31]	Easy to implement; directly relates to perception	Takes time; requires regular training; causes fatigue	Obtains the best HRTFs perceptually
	Tune magnitude spectrum [13], Active Sensory Tuning [26], PCA weight tuning [32]			
	Select cepstrum parameters [34]			
Playback Mode	Frontal projection headphone [33]	No additional measurement, listening training	New structure; not applicable to normal headphones; Type-2 equalization	Automatic customization, reduced front-back confusions
Non-individualized HRTF	Generalized HRTF [1]	Easy to implement	Not accurate; Poor localization	Not an individualization technique

TABLE IV
EQUALIZATION TECHNIQUES FOR DIFFERENT PLAYBACK MODES (BINAURAL, STEREOPHONY)

Mode of Equalization	Aim	Types of Equalization and Target Response	Characteristics
Non-decoupled (Binaural)	Spectrum at eardrum is the individual HRTF features	Conventional equalization (flat target response)	• For conventional headphones. The spectrum at the eardrum has individual features (if individualized HRTF is used) • Dependent on the individual's unique pinna features
		Type-2 equalization [33]	• For frontal projection headphones. The spectrum at eardrum automatically models the individual pinna spectral cues • Removes only the distortion due to the headphone emitter • Independent of the idiosyncratic features of the ear
Decoupled (Binaural, stereophony)	Emulate the most natural reproduction closer to the perception in a reference field	Free-field equalization (FF) [38]	• Target response is the free-field response corresponding to the frontal incidence
		Diffuse-field equalization (DF) [38]	• Target response is the diffuse-field response • Lesser inter-individual variability
		Diffuse-field target response based on Møller [38]	• Target response based on average of HRTFs between ± 45 degrees azimuth and elevation with unequal weighting
		Diffuse-field target response based on Lorho [4]	• Reduced a 3 kHz peak from about 12 dB to 3dB of diffuse-field response
		RR_G and RR1_G [4]	• RR_G: Based on the impulse response of Harman Reference Listening Room • RR1_G has lesser bass and treble