Natural listening over headphones in augmented reality using adaptive filtering techniques

Ranjan, Rishabh; Gan, Woon-Seng

2015

Ranjan, R., & Gan, W.-S. (2015). Natural Listening over Headphones in Augmented Reality Using Adaptive Filtering Techniques. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(11), 1988-2002.

<https://hdl.handle.net/10356/81366>

<https://doi.org/10.1109/TASLP.2015.2460459>

© 2015 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. The published version is available at: [<http://dx.doi.org/10.1109/TASLP.2015.2460459>].

Downloaded on 26 Apr 2023 10:50:17 SGT

Natural Listening over Headphones in Augmented Reality using Adaptive Filtering Techniques

Rishabh Ranjan, *Student Member, IEEE*, and Woon-Seng Gan, *Senior Member, IEEE*,

Abstract—Augmented reality (AR), which composes of virtual and real world environments, is becoming one of the major topics of research interest due to the advent of wearable devices. Today, AR is commonly used as assistive display to enhance the perception of reality in education, gaming, navigation, sports, entertainment, simulators, etc. However, most of the past works have mainly concentrated on the visual aspects of AR. Auditory events are one of the essential components in human perceptions in daily life but the augmented reality solutions have been lacking in this regard till now compared to visual aspects. Therefore, there is a need of natural listening in AR systems to give a holistic experience to the user. A new headphones configuration is presented in this work with two pairs of binaural microphones attached to headphones (one internal and one external microphone on each side). This paper focuses on enabling natural listening using open headphones employing adaptive filtering techniques to equalize the headset such that virtual sources are perceived as close as possible to sounds emanating from the physical sources. This would also require a superposition of virtual sources with the physical sound sources, as well as ambience. Modified versions of the filtered-x normalized least mean square algorithm (FxNLMS) are proposed in the paper to converge faster to the optimum solution as compared to the conventional FxNLMS. Measurements are carried out with open structure type headphones to evaluate their performance. Subjective test was conducted using individualized binaural room impulse responses (BRIRs) to evaluate the perceptual similarity between real and virtual sounds.

Index Terms—augmented reality, natural listening, head related transfer function, spatial audio, adaptive filtering

I. INTRODUCTION

AUGMENTED reality (AR) is changing the way we live in real world by adding a virtual layer to our sense of sight, smell, sound, taste, and touch along with real-world senses to give us an enriched experience. With the advent of wearable devices, such as Google Glass and Oculus Rift, sensors like microphones and cameras that capture our surroundings with global positioning system, AR technology provides sensory dimensions to the user to navigate more effectively in the real and virtual world. AR system is defined with three main characteristics, namely, superimposition of virtual objects onto physical world, ability to interact in real-time, and projection in three dimensional space [1]. AR devices are currently used in several application areas like assistive navigation for the disabled, augmented reality gaming [2], medical [3], audio-video conferencing [4], [5], binaural hearing aids, and audio guides in museum or tourist places [6]. So far, visual

information is predominantly used in AR-enabled devices to provide additional guidance or information to the user. Spatial sound is also being incorporated into AR devices to provide auditory cues of virtual and real objects in the listener's space via headphones [4]. These spatial cues can be used to alert listener of imminent danger/obstacle in a certain direction; add to the realism in gaming; and give a feeling of being there in the augmented environment. The ultimate goal of deploying spatial sound via headphones in AR devices is to create the impression that virtual sounds are coming from the physical world. At the same time, virtual sources should merge with the real sources in a transparent manner, enabling awareness to the real sources.

There have been several works in recent years in an attempt to playback spatial audio in AR based headphones, as well as existing commercial headphones. Haptic audio (sound that is felt rather than heard) is applied in headphones to enhance the user experience [7]. Bone conduction headset enabling hear-through augmented reality gives comparable performance to speaker array based system [8]. An augmented reality audio (ARA) headset has been introduced in [9] using an in-ear headphones with binaural microphones to assist the listener with pseudo-acoustic scenes. In another work, the same authors further developed an ARA mixer [10], [11] to be used with headset for equalizing and mixing the virtual objects with real environments. The main problem addressed in the ARA headset is the blockage of natural sounds coming from outside and reaching ear drum due to the in-earphone structure. Thus, their goal is to reproduce natural sounds unaltered with the help of binaural microphones to capture, process, and playback so as to make the ARA headset acoustically transparent. However, the direct sound leakage cannot be completely avoided and earphone repositioning might also affect the reproduced sound quality. Schobben and Aarts [12] proposed a headphone based 3-D sound reproduction system with binaural microphones positioned inside the ear cup near ear opening using active noise control (ANC) based calibration method. Filtered-x least mean square (FxLMS) adaptive filtering algorithm is used to achieve sound reproduction close to the 5.1 multichannel loudspeaker setup. The key problem being solved here is the large localization errors for most listeners due to non-individualized equalization of headphones. ANC is used to calibrate the system for every individual to identify loudspeakers' transfer function at listeners' ears before playing 5.1 multichannel virtual auditory scenes through headphones. Therefore, the primary challenge in AR based headphones is to reproduce sound as close to natural possible so that augmented audio environment presented are well

The authors are with the Digital Signal Processing Lab, School of Electrical and Electronics Engineering, Nanyang Technological University, Singapore, 639798 (e-mail: rishabh001@ntu.edu.sg, ewsgan@ntu.edu.sg).

Fig. 1. Bruel & Kjaer dummy head with different type of headphones used for the HRTF measurements using the binaural microphones located at ear drum.

externalized with no front-back confusions. Most importantly, virtual audio objects/scenes are seamlessly augmented in the real environment.

In this work, we present a natural augmented reality (NAR) headset with two pairs of binaural microphones to achieve natural listening experience using online adaptive filtering. An open type headphone structure is chosen over closed in-ear type headphones mainly because of two reasons: (1) its open-cup design allows external sound to pass through without much attenuation, resulting in a more natural listening experience; (2) open ear canal resonance, which is more natural compared to the blocked ear canal resonance of the in-ear headphones. With the use of sensing microphones installed in the headphone structure and by applying real-time adaptive training, virtual sources are reproduced as close as possible to real sources and thus, adding realism to the augmented reality environment (ARE). Modified versions of the filtered-x normalized least mean square (FxNLMS) algorithm are proposed in order to improve the slower convergence rate and steady state performance of the conventional FxNLMS. One of the main objectives is to ensure virtual sound objects become part of the real auditory space as augmented space. Therefore, the proposed approach is extended for the case when both real and virtual sources are to be mixed together, such that signal due to external sources does not interfere with the convergence process of the FxNLMS. The main advantage of applying FxNLMS technique here is that it adapts to the individualized head-related transfer functions (HRTF), while compensating for the individual headphone transfer function (HPTF), as it alters the desired spectrum at listeners' ears. Thus, adaptive process ensures NAR headset is individualized to a listener and virtual sources are reproduced alike real sources. Using dummy head measurements, it is found that the proposed approach is able to closely match the natural sound reproduction with faster convergence rate. The proposed method is also found to be equally effective in the presence of external sounds. Subjective study based on individualized binaural impulse responses (BRIRs) is conducted to validate the proposed approach and assess whether listeners can distinguish between real and virtual sounds.

This paper is structured as follows. Section II outlines the theoretical background on the binaural hearing and synthesis, focusing on the spatial cues necessary for accurate localization of sources. Section III evaluates the effects of different types of headphones on the direct sound spectrum. Section IV introduces natural listening techniques using the proposed

NAR headset. Adaptive equalization methods for reproducing virtual sources, with and without the presence of external real signals are presented in the sub-sections IV.D and IV.E, followed by the subjective test results in Section V. Finally, Section VI concludes the paper highlighting key results of this work.

II. THEORETICAL BACKGROUND ON BINAURAL HEARING AND SYNTHESIS

With the help of just two ears, we are able to acquire all the auditory information about incoming sounds, such as distance, direction based on the time and level difference between the sounds received at the two ears. Thus, if the two ears' signals can be reproduced exactly the same as in the direct listening case, a perfect replica of true auditory scene can be synthesized and subsequently, natural listening is attained. HRTF plays a significant role in localizing the sound image accurately. HRTF filters the source signal to account for the propagation of sound from the source to the listeners' ears in a free-field listening environment. HRTF comprises of three main binaural cues, namely, inter-aural time differences (ITD), inter-aural level differences (ILD) and spectral cues (SC) based on the reflections and refractions from torso, head, and pinnae [13], [14]. Among these, ITD and ILD are the main cues for localizing the source in the azimuthal plane or so-called lateral angle and are independent from the source spectrum [15]. ITD is the dominant cue in low frequency below 1.5 kHz, where head shadowing is weak as sound wavelength is more than the distance between the two ears. Above 1.5 kHz, ILD prevails over ITD and helps in localizing the sound more accurately. Spectral cues due to the pinnae reflections is the dominant cue for frontal and elevation perception above 3 kHz [16]. However, interaction with the torso also adds to the elevation cues. Cues due to the pinnae reflections are unique owing to the different pinnae structures in each individual and therefore, pose a special problem in binaural synthesis using headphones. Hence, human auditory perception is strongly dependent on the anthropometry of individual ear and head, especially on the unique pinnae shape of every individual. Using non-individualized HRTFs may result in front-back confusions and in-head localization (IHL). Due to the impracticality of measuring individual HRTFs in anechoic room, generic HRTFs are widely used in binaural synthesis, and studies [17] have shown to partially improve the performance.

Fig. 2. Effects of four types of headphones on the direct sound source spectrum at ear drum of the dummy head

In addition, dynamic cues due to the head motion and visual cues help in minimizing the front-back confusions and large localization errors [18]. Since most of the HRTF measurements are carried out in anechoic chamber as free-field, they are far from natural when played back over the headphones. In this context, room reflections are very important for source localization of distant sources in a reverberant room. Use of artificial reverberation can thus, help in synthesizing externalized sound image in headphone listening. Furthermore, ILD cues in low frequencies are important for sources in near-field [19]. It has been observed in recent research that motion cues dominate pinnae cues in resolving the front-back reversals as well as enhancing externalization [20]. Dynamic cues can be generated with the help of head-tracking device attached to a headphone by continuously adapting to the head movements. In the next section, we investigate the effect of different headphones on direct sound spectrum in a natural listening environment. We focus on the use of open type headphones over closed-end headphones. It is generally found that “more” open headphones, which do not obstruct external sounds much and allow them to be perceived naturally, possess much less spectral errors as well as insignificant localization errors as compared to closed-back headphones [21].

III. HEADPHONES EFFECT ON DIRECT SOUND SPECTRUM

In an ARE, a user wearing a NAR headset must not feel isolated from the surroundings. The choice of headphones is crucial in designing an NAR headset as it should allow external sounds to pass unblocked and reach listeners’ ear drum in a natural manner. Different types of commercial headphones have been used to evaluate their effects on direct sound source spectrum. Fig. 1 (b)-(e) show four types of headphones, namely, open circumaural (CA), open supra-aural (SA), open personal field speaker (PFS) and closed circumaural (CA), which are worn on the dummy head for measurement. The open PFS headphones are completely open with external drivers facing towards the pinna from the frontal direction, as shown in Fig. 1 (c).

Effects of the aforementioned headphones on the direct sound source spectrum for different azimuths are shown in Fig. 2 along with the direct sound spectrum as reference (Ref) HRTF measured without the headphones. Measurements were conducted in an anechoic room using the Bruel & Kjaer 4128D head and torso simulator. Exponential sine sweep signal with sampling frequency of 44.1 kHz was played from sound source placed at 1.4 m away from the center of dummy head and recorded at the binaural microphones located at ear drum of the dummy head. All of the HRTFs are 1/3rd octave smoothed to decrease perceptually redundant fluctuations, especially in high frequencies [22]. As shown in Fig. 2, headphones act

Fig. 3. (a) Proposed NAR headset structure (top) and prototype using open CA headphones with two microphones (below) (b) Headphone modified transfer functions (HMTF) (c) Transfer functions measurement set up

as passive low pass filters and that is why the difference between the headphone modified spectrum and direct sound spectrum is observed only in high frequencies above 1.5 kHz, except for the closed CA headphones, which attenuates up to 10-15 dB is observed below 1.5 kHz. One common observation from the HRTF plots in Fig. 2 is that the closed CA headphones attenuate direct sound the most as compared to other open headphones, while open PFS headphones is the most acoustically transparent headphone for all azimuths. For the closed CA headphones, attenuation up to approximately 30 dB is observed for most of the azimuths. This leads to significant coloring of the direct sound source spectrum. The open SA headphones and open CA headphones possess average headphone attenuation of roughly up to 10-15 dB in high frequencies. Other important aspect that needs to be observed is the high frequency pinnae specific notches, which is particularly essential for the frontal localization [23], [24] as well as elevation [25]. These notches for the open CA and open PFS headphones are consistent with the reference HRTF, but mismatch/absence of the notch positions is observed for the open SA and closed CA headphones. For the open SA headphones, it might be due to the fact that headphones are resting on the ear, suppressing the reflections due to pinna. For the closed CA headphones, strong passive isolation by the headphones structure possibly leads to reduction/disappearance of notches. To summarize, closed headphones are not suitable for AR based application due to its strong isolation property and coloration of the sound spectrum. Similar observations were also reported in [21] on the influence of headphones to localization of loudspeaker source. In particular, it was found that the localization accuracy degraded only slightly due to wearing of headphones as compared to open-ear listening. It was also found that listeners used head rotation more frequently as additional cue to assist in localization when headphones were worn. However, large ILD errors due to high frequency loss will result in audible coloration, as well as dulling of the sound. Spectral and ILD distortions were less pronounced for headphones with “more” open design. Therefore, it is suggested that care must be taken while selecting headphones for ARE. In practice, absolute acoustic transparency cannot be achieved, but headphones characteristics can be modified through signal processing techniques and/or passive techniques to achieve realistic impression of physical sound sources and environments. The following sections outline the adaptive sig-

nal processing techniques to achieve natural listening through headphones.

IV. NATURAL LISTENING VIA NAR HEADSET BASED ON ADAPTIVE FILTERING

Natural listening using headphones require sounds to be reproduced as natural as possible. For AR based scenarios, we would need both real and virtual sounds to be perceived in the same way such that one cannot distinguish between the two. In addition, a realistic fusion of virtual sound objects with the real auditory environment is also desired for ARE. We thus, divide our analysis into three possible practical scenarios, which will be presented in the subsequent sub-sections:

- Case I:** Only real source present
- Case II:** Only virtual source present
- Case III:** Virtual source in the presence of real source/surroundings

In the following sub-sections, we present the proposed NAR headset structure followed by adaptive signal processing techniques to achieve natural listening in ARE. Our main focus in this work is on the adaptive algorithms to create virtual auditory events that are engrossed with the real environment, giving an immersive experience to the listener. Case I scenario with only real external sources, may not need any additional processing if using an open type headphones. Hence, the focus of this paper is mainly on Case II and Case III, where virtual sources are needed to be reproduced naturally to the listener, without and with the presence of the real sound sources, respectively.

A. Proposed Headset Structure

The proposed NAR headset structure and the prototype constructed using AKG K702 open CA studio headphones is shown in Fig. 3 (a). The open CA headphones is preferred over the other two types of open headphones for ease of microphones placement in our prototype. There are two microphones attached on each side of the headphones ear cup. As shown in the figure, internal microphone, denoted by m_{int} , is positioned very near to ear opening, whereas external microphone, denoted by m_{ext} , is positioned just outside the headphones ear cup. The main purpose of internal microphone (also known as the error microphone), m_{int} is to adapt to the desired virtual sound field measured at listeners' ears.

Fig. 4. Measured modified transfer functions ($H_{int}(z)$ and $H_{ext}(z)$) for two azimuths (Top: Ipsilateral ear; Bottom: Contralateral ear)

External microphone (or reference microphone), m_{ext} is used to capture the external sounds. Besides these, both pairs of microphones are also used for off-line measurements of the transfer functions modified due to the presence of headphones. These transfer functions are used in the binaural reproduction of virtual sources and to be stored in our own personalized HRTF database for different listening environments. In the next sub-section, the characteristics of the headphone modified transfer functions (HMTFs) measured at the two microphone positions are discussed.

B. HMTF measurements and observations

Fig. 3 shows the measurement setup for HMTFs at the two microphone positions using the NAR headset prototype. Four miniature AKG C417 microphones are used in the measurements having mostly flat response. The two HMTFs denoted as $H_{int}(z)$ and $H_{ext}(z)$, (modified due to the passive headphones structure) are measured on the dummy head using the two pair of microphones (See Fig. 3 (b)). $H_{int}(z)$ represents the transfer function similar to HRTF to account for the sound propagation from source to ear entrance but modified by the presence of headphones, while $H_{ext}(z)$ accounts for the sound propagation from source to the just outside the ear cup.

It should be noted that since $H_{ext}(z)$ is measured just outside the ear cup (~2 cm away from the pinna), its spectrum/impulse response will contain all the individual related characteristics and environment without pinna and shell reflections. Spectrums of the two HMTFs for two azimuths are shown in Fig. 4. In addition, a reference HRTF measured without headphones at the ear canal entrance is also shown for comparison with the two HMTFs. One noticeable observation is that spectrum of $H_{ext}(z)$ is closer to the direct sound spectrum than that of $H_{int}(z)$ with little or no high frequency loss. Furthermore, there are no pinnae specific frontal notches (especially for the frontal azimuths) observed in $H_{ext}(z)$ as compared to the reference HRTF as well as $H_{int}(z)$. Therefore, the spectrum of $H_{ext}(z)$ is much smoother with only smaller peaks and notches due to the absence of

Fig. 5. Case I: Only real source present scenario and corresponding signal flow block diagram

reflections within the shell and the pinnae. In contrast, $H_{int}(z)$ have sharper peaks and notches compared to $H_{ext}(z)$. This prior information in $H_{ext}(z)$ (environment, torso, head related characteristics) can help in estimating the signal accurately at listeners' ears. As will be shown later in the paper, $H_{ext}(z)$ is very useful in improving the performance of adaptive equalizer methods presented. In addition, external signals received at m_{ext} are also used to further estimate the real signals at m_{int} adaptively. We now present the analysis for three practical scenarios in the following three sub-sections.

C. Case I: Only real source present – No additional processing required

In this scenario, only real sound sources are present, which is what we experience in day-to-day listening. But in this case, it is required to hear the sounds coming from the environment and external sources, while wearing the NAR headset. This scenario is depicted in Fig. 5 using an open ear cup headphone along with its corresponding signal flow block diagram representation. Natural sound from real source, $r(n)$ propagates through air and reaches the listeners' ear after passing through the ear cup. Thus, $h_{int}(n)$ (corresponding impulse response of HMTF, $H_{int}(z)$) accounts for the natural sound propagation in air from source to the listeners' ear. Alternatively, external sound source propagation can also be seen as real source signal, $r_{ext}(n)$ just outside the ear cup (m_{ext}) passed through a passive headphone-effect filter with impulse response $h_{he}(n)$, accounting for the ear cup effect and pinnae reflections, before reaching listeners' ear. Hence, $H_{he}(z)$ is a transfer function from m_{ext} to m_{int} :

$$H_{he}(z) = \frac{H_{int}(z)}{H_{ext}(z)}. \quad (1)$$

For an acoustically transparent headphone, (1) does not contain the headphone-effect but is just the free-field transfer function between the two microphone positions without headphones. As discussed in section III, open headphones are more suitable for this mode resembling natural listening and “relatively acoustically transparent” as compared to closed headphones. Therefore, no additional processing is required if there is not much attenuation of direct sound assuming headphones with more open design are used in the construction of NAR headset having less pronounced spectral loss and ITD errors. However, with less open headphones, the augmented content should be modified (using active and/or passive techniques) in order to make it closer to acoustical

Fig. 6. Case II: Only Virtual source present scenario and corresponding signal flow block diagram

transparent. Schobben and Aarts showed in [12] that high frequency attenuation in open headphones can be partially compensated by replacing the headphone ear pads by an acoustically transparent foam type material. Härmä *et al.* [9] developed a generic equalization method to compensate for the closed in-ear type headphones isolation by capturing the external sound using an external microphone and playing back through earphone after filtering through an equalization filter. Similar to [9], with the help of external microphone in our NAR headset, high frequencies (mainly above 6 kHz) can be boosted to compensate for the headphones isolation. Since in our study, it was observed that headphones isolation characteristics varied with azimuths, difference between the signals levels at internal and external microphones can be used to determine the gain increment. Realization of such equalization for headphone effects compensation is beyond the scope of this work. It is also the subject of further research on detailed perceptual analysis of the headphones effect in long-term listening and listeners can be expected to become accustomed to a somewhat modified direct sound spectrum based on past studies [26]. In this work, we primarily focus on the adaptive equalization methods for personalized virtual sound source reproduction over headphones presented in the next two subsections.

D. Case II: Only virtual source present – Adaptive equalization for headphones

To enable natural listening via headphones for virtual sounds, it is required to reproduce exact replica of real signals at listener's ears as in natural listening for external sources. To achieve this, desired impulse response, $h_{int}^v(n)$ (superscript v represents virtual sound reproduction), which must be measured for each and every individual in a given listening environment, are required to create an illusion that sound is coming from a physical source. In addition, $h_{int}^v(n)$ must be equalized to compensate for the individual HPTF, an electro acoustical transfer function measured at the listeners' ears as impulse response $h_{hp}(n)$. HPTFs are also unique for every individual and modify the intended sound to be reproduced in an undesired manner. In a recent study on the effects of headphone compensation in binaural synthesis by Brinkman *et al.* in [27], it has been found that only individualized headphone compensation is able to completely eliminate the audible high frequency ringing effects as against non-individual and generic headphone compensation. In short, both individualized desired transfer function ($H_{int}^v(z)$) and

Fig. 7. Conventional FxNLMS Block Diagram for virtual source reproduction

individualized headphone equalization are necessary for the NAR headset to accurately replicate the physical sound spectrum virtually.

Fig. 6 shows the scenario with only virtual source present along with the corresponding signal flow block diagram. A monaural signal can be placed anywhere in the virtual auditory environment by convolving with the desired impulse response based on the intended position (direction, distance) of virtual source. With the NAR headset, the virtual source signal is first convolved with an equalization filter, $w(n)$, to compute secondary source signal, $u(n)$ and subsequently, passed through the inherent HPTF filter, $h_{hp}(n)$, before reaching listeners' ear. $w(n)$ is estimated as convolution of $h_{int}^v(n)$ and inverse filter of $h_{hp}(n)$, such that the virtually synthesized signal, $x_{int}(n)$ approaches the desired signal, $d(n)$. In this sub-section, we are mainly focusing on the individual binaural headphone compensation assuming that individualized set of HRTFs measured in the listener environment are available in the database. The proposed NAR headset has an advantage over most of the current headphones in the market. This is due to individualized headphone equalization, which is possible because of the two internal microphones attached as $h_{hp}(n)$ can be measured and compensated for every individual. Usually, the headphone equalization requires inversion of the HPTF, which need not necessarily exist and regularization techniques are used to avoid large boosts [28]. But regularization can also convert a causal minimum-phase inverse filter into one with non-minimum phase characteristics, which can create audible distortions like pre-ringing effects [29]. Another widely used alternative and generally the most effective approach is to use adaptive algorithm like, filtered-x least mean square algorithm (FxLMS), where an estimate of $h_{hp}(n)$ is placed in reference signal path for weight update to ensure convergence and stability [30]. This type of adaptive process is termed as adaptive equalization, since equalization filter, $w(n)$ is adapted to any time-varying changes in the individual HPTF due to headphone repositioning or even change of listener. Therefore, fast convergence and minimum steady state mean square error (SS-MSE) of the adaptive process is very crucial for the performance of NAR headset. In addition, minimum spectral distortion (SD) is required to ensure similar spectral variation between the desired and estimated secondary path transfer function, while preserving the pinnae cues crucial for localization. Fast convergence will also ensure that virtual signal captured by the error microphone (m_{int}) converges to the desired signal as quickly as possible.

Fig. 8. Block Diagram of modified FxNLMS for virtual source reproduction

FxLMS usually suffers from slow convergence and can be improved by its normalized version (FxNLMS). Fig. 7 shows the block diagram of conventional FxNLMS for virtual source reproduction. In the case of adaptive equalization presented in this paper, signals are electrically subtracted unlike the FxNLMS algorithm used in conventional ANC applications of acoustic duct and ANC headset, where primary signal is acoustically summed at the error microphone. The FxNLMS algorithm is expressed below:

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mu \frac{\mathbf{x}'(n)}{\|\mathbf{x}'(n)\|^2} e(n), \quad (2)$$

where $\mathbf{w}(n)$ is the coefficient vector of $w(n)$ with length L , and $\mathbf{x}'(n) = [x'(n) \ x'(n-1) \ \dots \ x'(n-L+1)]^T$ is the set of current and past $(L-1)$ samples of filtered reference signal $x'(n)$ at time n . $\|\cdot\|^2$ represents the norm-2 of the vector. Optimum value of the equalization filter $w(n)$ is achieved when the expectation of the squared error, $e(n)$ approaches zero and can be found as:

$$W^o(z) = \frac{H_{int}^v(z)}{H_{hp}(z)}. \quad (3)$$

It should be noted that the required number of filter taps for the equalization filter can be large because of the fact that desired signal, $d(n)$, may contain the additional delay due to the distance and room acoustics stored in the desired impulse response, $h_{int}^v(n)$. As larger filter taps will improve the accuracy of the adaptive process by closely following the desired signal, it also slows down the convergence rate at the same time. Since fast convergence being one of the stringent requirements for our system performance with the SS-MSE, we propose a modified version of FxNLMS, as shown in Fig. 8. The secondary path of the adaptive process is modified by including an additional filter, $h_{ext}^v(n)$, and a forward delay (Δ) is also introduced in the primary path to take into account for the overall delay (A/D, D/A, processing) of the secondary path. As discussed in the subsection IV-B, transfer function measured just outside the ear cup, $H_{ext}^v(z)$, contains all the spatial information of the human torso, head, as well as environments without the pinnae and headphone shell reflections. By using this prior-information in estimating the desired signal, the adaptive process is simplified with shorter adaptive filter length and subsequently, faster convergence. As compared to the conventional FxNLMS approach, virtual signal is first pre-filtered with $h_{ext}^v(n)$ before passing to the equalization filter $w(n)$. Using this approach, we are trying to

Fig. 9. Comparison between Conventional FxNLMS and Modified FxNLMS for 40° azimuth (Top: Ipsilateral ear; Bottom: Contralateral ear)

emulate the natural listening process by using an estimate of virtual signal at m_{ext} , i.e., $x_{ext}(n)$, to reproduce the replica of real sound alike at listeners' ear, as shown in Fig. 8. Equalization filter weights will be optimum when square of the residual error is minimized:

$$e(n) = d(n) - x_{int}(n) = 0, \quad (4)$$

where $d(n)$ is defined as:

$$d(n) = h_{int}^v(n) * x(n - \Delta). \quad (5)$$

Substituting (5) into (4) and transforming into Z domain:

$$X(z)H_{int}^v(z)z^{-\Delta} - X(z)H_{ext}^v(z)W^o(z)H_{hp}(z) = 0. \quad (6)$$

By simplifying (6), the optimum equalization filter can be expressed as:

$$W^o(z) = \frac{H_{int}^v(z)z^{-\Delta}}{H_{ext}^v(z)H_{hp}(z)} = \frac{H_{he}^v(z)z^{-\Delta}}{H_{hp}(z)}, \quad (7)$$

where $H_{he}^v(z)$ is the headphone effect transfer function for virtual sound reproduction denoted by the superscript, v . Therefore, the difference between the optimal solution of conventional FxNLMS in (3) and (7) is due to the filter $h_{ext}^v(n)$ and the forward delay in primary path. Delay in the primary path must be at least equal to the secondary path delay for a feed-forward adaptive filter to converge [30]. Weight update equation for the modified FxNLMS approach is expressed similarly as in (2), except that the filtered reference signal is now defined as follows:

$$x'(n) = \hat{h}_{hp}(n) * x_{ext}(n), \quad (8)$$

where $\hat{h}_{hp}(n)$ is an estimate of the secondary path transfer function (HPTF), which is estimated offline by playing a test sequence through the headset and recording the response at internal microphone. As in ANC, FxNLMS algorithm converges with the limit of 90° phase error between $\hat{h}_{hp}(n)$ and $h_{hp}(n)$ [30].

Fig. 10. Spectral distortion comparisons for both the approaches over 0 to 180° azimuths (Top: Ipsilateral ear; Bottom: Contralateral ear)

1) *Case II Results: Conventional FxNLMS Vs Modified FxNLMS*: In this sub-section, we compare the performance of the modified FxNLMS with the conventional FxNLMS method. A white noise sequence of 1 second duration is used to estimate the adaptive filter in both methods. Length of impulse responses, $h_{int}^v(n)$ and $h_{ext}^v(n)$, are set at 1024 taps, and 256 taps are used for $h_{hp}(n)$. Longer filter length for the desired responses is required to account for the distance and reverberations. The equalization filter lengths are set at 1024 taps and 256 taps for the conventional FxNLMS and the modified FxNLMS, respectively. A step size of 0.1 is chosen for both the algorithms.

Fig. 9 compares the performance of the two approaches. Three main performance criteria used in this paper are the faster convergence rate, minimum SS-MSE, and minimum SD. As shown in Fig. 9, conventional FxNLMS suffers from slow convergence rate as expected. SS-MSE for both the approaches do not differ much from each other, as can be seen in Fig. 9 (a) and (b). Besides, we can also objectively quantify the spectral error between reference and estimated transfer functions using a widely used objective metric [31], [32] i.e., spectral distortion (SD) score:

$$SD = \sqrt{\frac{1}{K} \sum_{k=1}^K \left(20 \log \frac{|H(f_k)|}{|\hat{H}(f_k)|} \right)^2} [dB], \quad (9)$$

where $H(f)$ is the magnitude response of reference primary path transfer function, $\hat{H}(f)$ is the secondary path estimated transfer function, and K is the total number of frequency samples in the observed range (100 Hz – 16 kHz). Secondary path transfer functions for conventional FxNLMS and modified FxNLMS are expressed, respectively as:

$$S_{conv}(z) = W_{conv}(z)H_{hp}(z), \quad \text{and} \quad (10)$$

$$S_{mod}(z) = H_{ext}^v(z)W_{mod}(z)H_{hp}(z). \quad (11)$$

Fig. 10 shows the spectral distortion scores for low frequency (below 1.5 kHz) and high frequency (above 1.5 kHz).

Fig. 11. Block diagram of hybrid FxNLMS using conventional FxNLMS and modified FxNLMS algorithm

To clearly demonstrate the difference between the two approaches, simulation is stopped after 0.2 second and SD scores using (9) are computed at this instance. It is clearly observed that mean spectral distortion for the conventional FxNLMS is much higher than the modified FxNLMS, especially at low frequencies. Even at higher frequencies, modified FxNLMS has higher accuracy than the conventional FxNLMS for most of the azimuths except for some source positions. It should also be noted that spectral distortion is considerably greater for ipsilateral ear than the contralateral ear for both approaches. This might be due to the pinna effects being more pronounced at the ipsilateral ear. Although, the modified FxNLMS has faster convergence rate as well as better accuracy for most azimuths, larger spectral deviations in higher frequencies can significantly affect the NAR headset performance and may result in higher SS-MSE for some of the azimuths. Based on above observations, a hybrid adaptive equalizer (HAE) is also proposed by combining both the above approaches to obtain an optimum steady state performance of the adaptive algorithm for all azimuths.

2) *Hybrid Adaptive Equalizer (Hybrid FxNLMS)*: The conventional FxNLMS suffers from slow convergence and generally requires large filter order for equalization filter to converge to the optimum solution. The modified FxNLMS uses an additional pre-filter in secondary path, which contains most of the spatial information of the primary path. This ensures faster convergence of the adaptive process. High spectral distortions have been observed for the conventional FxNLMS in low frequency, while modified FxNLMS has relatively larger errors in high frequency regions for some source positions. The proposed hybrid FxNLMS uses simple combination of both the conventional and modified FxNLMS structures discussed above, which can result in significant steady state performance improvements as well as fast convergence most of the times [33]. The block diagram for the hybrid FxNLMS is shown in Fig. 11. The secondary source signal $u(n)$ is generated using outputs of both the conventional FxNLMS equalization filter $w_1(n)$ and the modified FxNLMS equalization filter $w_2(n)$. As shown in Fig. 11, the equivalent equalization filter $w(n)$ has two reference inputs: $x(n)$ as the virtual signal, and $x_{ext}(n)$ is the virtual signal estimated at the reference microphone (m_{ext}). Filtered versions of the two reference signals $x'(n)$ and $x'_{ext}(n)$ are used to adapt the filter coefficients $w_1(n)$ and $w_2(n)$, respectively.

The secondary signal $u(n)$ is computed by the equivalent

Fig. 12. Hybrid adaptive equalizer performance for 40° azimuth (Top: Ipsilateral ear; Bottom: Contralateral ear)

equalization filter $w(n)$, which consists of the two adaptive filters' length of L_1 and L_2 , respectively, for $w_1(n)$ and $w_2(n)$ as:

$$u(n) = \mathbf{w}_1^T(n)\mathbf{x}(n) + \mathbf{w}_2^T(n)\mathbf{x}_{ext}(n), \quad (12)$$

where

$$\mathbf{x}(n) = [x(n) \ x(n-1) \ \dots \ x(n-L_1+1)]^T, \quad (13)$$

and

$$\mathbf{x}_{ext}(n) = [x_{ext}(n) \ x_{ext}(n-1) \ \dots \ x_{ext}(n-L_2+1)]^T, \quad (14)$$

The hybrid FxNLMS algorithm for the weight update of the two filters is expressed as:

$$\mathbf{w}_1(n+1) = \mathbf{w}_1(n) + \mu \frac{\mathbf{x}'(n)}{\|\mathbf{x}'(n)\|^2} e(n), \quad (15)$$

and

$$\mathbf{w}_2(n+1) = \mathbf{w}_2(n) + \mu \frac{\mathbf{x}'_{ext}(n)}{\|\mathbf{x}'_{ext}(n)\|^2} e(n), \quad (16)$$

Weight update equation (15) corresponds to the conventional FxNLMS with only difference in the calculation of residual error signal (delayed version) as defined by (4) and (5), while weight update for $w_2(n)$ is same as the modified FxNLMS. The main purpose of the hybrid approach is to take advantage of both the adaptive processes so as to minimize the residual error. Ideal solution for $w(n)$ is derived using (4) and (5) as:

$$X(z)H_{int}^v(z)z^{-\Delta} - U(z)H_{hp}(z) = 0. \quad (17)$$

Taking Z transform of (12) and substituting into (17) :

$$X(z)H_{int}^v(z)z^{-\Delta} - X(z)W^o(z)H_{hp}(z) = 0. \quad (18)$$

$W(z)$ is an equivalent equalization filter representation for the HAE and expressed as:

Fig. 13. Spectral distortion score comparisons: Hybrid FxNLMS versus others (Top: Ipsilateral ear; Bottom: Contralateral ear)

$$W(z) = W_1(z) + H_{ext}^v(z)W_2(z). \quad (19)$$

Thus from (18), the optimal solution of equivalent adaptive filter, $W^o(z)$ is similar to that of conventional FxNLMS with an additional delay term:

$$W^o(z) = \frac{H_{int}^v(z)z^{-\Delta}}{H_{hp}(z)}. \quad (20)$$

Therefore, the optimal solution of the hybrid FxNLMS can be written as linear combination of optimal solutions $W_1^o(z)$ and $W_2^o(z)$ for the conventional and modified FxNLMS approach, respectively:

$$W^o(z) = \alpha W_1^o(z) + \beta H_{ext}^v(z)W_2^o(z). \quad (21)$$

such that,

$$\alpha + \beta = 1; \quad 0 \leq \alpha, \beta \leq 1 \quad (22)$$

The values of α and β are inherently determined by the adaptive algorithm such that the residual error is minimized. Next, we will discuss the performance of the presented HAE and compare the results with the conventional FxNLMS and modified FxNLMS.

3) Case II Results: Hybrid FxNLMS Vs Others: In this subsection, we compare the performance of the proposed hybrid FxNLMS with the conventional and modified FxNLMS algorithms. Same number of taps for the two filters, $W_1^o(z)$ and $W_2^o(z)$ are used, as in subsubsection IV-D1. Fig. 12(a) shows the residual error for the hybrid FxNLMS. Comparing the results with that of Fig. 9, the hybrid FxNLMS performs much better as compared to conventional and modified FxNLMS with optimum MS-SSE. Moreover, its convergence rate is also much faster than the conventional FxNLMS but slightly slower than the modified FxNLMS. Spectral distortion scores versus time plots for three headphone placements (HP1-3) are shown in Fig. 12(b). For the first two headphone placements, headphone was slightly adjusted, while in the third placement headphone was lifted and placed back on the ears. Clearly, the proposed hybrid approach is also robust to headphone placements while its SD converges to the optimal solution

Fig. 14. Spectral distortion score for Hybrid FxNLMS for elevated sources

in all cases. It was also found that adaptive equalization performs better than fixed headphone equalization with 7-8 dB higher error reduction. Spectral distortion scores for the three approaches are shown in Fig. 13 computed at two time instants of 0.2 second and 1 second. Since conventional FxNLMS has the slowest convergence rate, the residual error cannot be completely converged after 0.2 second and results in larger spectral distortion. On the other hand, hybrid approach has the least spectral distortion, as shown in the Fig. 13. For longer noise sequence of 1 second, when steady state performance is reached, relatively larger spectral differences is observed between conventional and modified FxNLMS approach, whereas the hybrid FxNLMS has the best overall performance, as shown in the right side plots of Fig. 13. Mean steady state error attenuation for the hybrid FxNLMS across all azimuths was found to be around 25 dB and 28 dB for the ipsilateral and contralateral ears, respectively. Finally, we show the SD scores for elevated sources in Fig. 14. Target impulse responses for the adaptive equalization were measured at 0°, 15°, 30° and 45° elevated positions for 4 different azimuth positions (0°, 40°, 80° and 120°). As shown, mean SD score for the ipsilateral ear for all the angles is within 5 dB, while it is less than 2 dB for the contralateral ear. Spectral distortion is clearly more pronounced for the ipsilateral ear, especially when source is one side of the dummy head and directly facing the ipsilateral ear (See Fig. 14 (b) for 80° azimuth). Due to the fast converging speed of the hybrid FxNLMS, estimated responses are closely tallied to the desired response for most of the source positions with optimum SS-MSE and SD. The high frequency pinnae cues, which are primary cues for the sources in the front as well as elevation, have also been preserved in the virtually synthesized responses.

E. Case III: Both virtual and real source present (Augmented reality): HAE with adaptive estimation

In this section, we present the most general case for augmented reality with virtual and real sounds being intermixed. As explained earlier, augmented reality requires virtual sources to be coherently fused with the real source and surroundings such as to create an illusion of virtual sources being perceived as one of the real sound sources. Fig. 15 shows this scenario along with the corresponding signal flow block diagram. In an ideal case with virtually no headphones present, virtual source after passing through the target response is acoustically added with real signals reaching the listeners' ear, as shown in Fig. 15. But the HPTF colors the intended sound spectrum

Fig. 15. Case III: Both virtual and real source present scenario and corresponding signal flow block diagram

Fig. 16. Residual error plots for hybrid FxNLMS with and without real source. Virtual source is positioned at 0° azimuth, while real sound is coming from 40° azimuth and added to virtually reproduced signal at m_{int} . (Top: Ipsilateral ear; Bottom: Contralateral ear)

and thus, virtual signal must be equalized before playing back through the headphones. We presented the HAE for virtual source reproduction in the previous section, and ensuring that there is no difference between the real and virtual source signals. In this scenario with NAR headset, real signal is also captured by the internal error microphone (m_{int}) simultaneously with the synthesized virtual signal, as shown in the block diagram of Fig. 15. In addition to the external sounds, leakage signal, $l(n)$ from inside of headset is also captured by the external microphone, with $h_{le}(n)$ as the headphone leakage impulse response measured at m_{ext} . Leakage signal is assumed to be negligible with more than 20 dB attenuation to playback level at ear opening. However, the real signal $r_{int}(n)$ can hamper the adaptive equalization for $w(n)$ by acting as interference to the system. Fig. 16 shows the plots for residual error of the hybrid FxNLMS with and without real source present. Two uncorrelated random noise sequences are used in simulation for the virtual and real source. Clearly, due to the presence of real signals, the hybrid FxNLMS cannot reach the optimum solution and subsequently, resulting in roughly 14 dB lesser reduction in steady state than the case with no real source. Therefore, the effect of real signal must be removed from the adaptive process; otherwise it might result in large steady state error depending on the energy and nature of external signals.

In augmented reality, both real and virtual sounds are equally crucial for an immersive experience and therefore, either of them must not interfere with each other to reproduce a natural superposition. In this respect, the acquired real signal

Fig. 17. Block diagram of hybrid adaptive equalizer with online adaptive estimation of $r_{int}(n)$

$r_{int}(n)$ must be removed from the adaptive process of $w(n)$. There are two ways to compute an estimate of the signal $r_{int}(n)$ using the real signal received at m_{ext} , i.e. $r_{ext}(n)$:

- 1) With the help of pre-computed $h_{he}(n)$: As explained in subsection IV-C, $h_{he}(n)$ represents the headphone-effect transfer function from m_{ext} to m_{int} , an exact estimate of $r_{int}(n)$ is computed from $r_{ext}(n)$ as:

$$r_{int}(n) = r_{ext}(n) * h_{he}(n). \quad (23)$$

But in practice, the precise location of external sound is not known and hence, a filter averaged over entire azimuths has to be used instead of the exact $h_{he}(n)$ as:

$$r_{int}(n) = r_{ext}(n) * h_{he,avg}(n). \quad (24)$$

The headphone-effect transfer function $h_{he}(n)$ is computed as off-line adaptive estimation using the FxNLMS algorithm.

- 2) Using an online adaptive process to estimate $r_{int}(n)$ from $r_{ext}(n)$: It has been observed that $h_{he}(n)$ varies considerably with head movements. Therefore, online adaptive estimation of $h_{he}(n)$ can give better estimate of $r_{int}(n)$ instead of using an average filter $h_{he,avg}(n)$.

We further, extend the hybrid FxNLMS with online adaptive estimation of real signal, as shown in Fig. 17. The adaptive equalization filter $w(n)$ is the equivalent representation of the HAE given by (19). As discussed in the previous section, equalization filter $w(n)$ comprises of two adaptive filters $w_1(n)$ and $w_2(n)$ corresponding to the conventional FxNLMS and modified FxNLMS algorithms, respectively. As shown in Fig. 17, $w_r(n)$ is adapted to generate an estimate of $r_{int}(n)$, $\hat{r}_{int}(n)$ and added to $d(n)$ from which the acoustically superimposed signal $y_{int}(n)$ is subtracted. After $\hat{r}_{int}(n)$ has converged, we obtain the residual error signal as

$$e'(n) = \{d(n) + \hat{r}_{int}(n)\} - y_{int}(n), \quad (25)$$

where $y_{int}(n)$ is defined as

$$y_{int}(n) = r_{int}(n) + x_{int}(n). \quad (26)$$

Substituting (26) into (25) and re-arranging,

$$e'(n) = \{d(n) - x_{int}(n)\} + \{-r_{int}(n) + \hat{r}_{int}(n)\}. \quad (27)$$

Hence, the residual error signal consists of two separate error signals. The first term in RHS of (27) is the error signal for hybrid adaptive process defined by $e(n)$ in (4), while the

Fig. 18. Results for hybrid FxNLMS with and without adaptive estimation. Simulation set up is kept same as in Fig. 16. (Top: Ipsilateral ear; Bottom: Contralateral ear)

second term is the negative error signal due to the online adaptive estimation of $\hat{r}_{int}(n)$. The optimum solution of adaptive estimation process is derived when $\hat{r}_{int}(n) = r_{int}(n)$, or

$$r_{ext}(n) * w_r(n) = r_{int}(n). \quad (28)$$

Taking the z-transform of (28), the optimum control filter $w_r(n)$ is expressed as

$$W_r^o(z) = \frac{R_{int}(z)}{R_{ext}(z)} = \frac{H_{int}(z)}{H_{ext}(z)} = H_{he}(z). \quad (29)$$

Thus, optimum control filter is simply the headphone-effect impulse response. Weight update equation for the two control filters in HAE is defined as in (15) and (16), whereas weight update equation for the control filter $w_r(n)$ is defined using the LMS algorithm as

$$\mathbf{w}_r(n+1) = \mathbf{w}_r(n) - \mu_r \mathbf{r}_{ext}(n) e'(n). \quad (30)$$

Note that negative sign in the weight update equation (30) is due to the way error signal is defined in (27). Fig. 18 shows the performance comparison of the HAE with and without adaptive estimation. Clearly, with the proposed adaptive estimation, the performance of the hybrid FxNLMS is very close to the one without any real source present, as observed in Fig. 16(a) and Fig. 18(a). With off-line estimation i.e., using an average filter, the steady state error increases especially for the ipsilateral ear. However, the approach with offline estimation still performs much better than the one without any estimation (See Fig. 16(b) and Fig. 18(b)). A perceptual validation of the hybrid FxNLMS is carried out via subjective study, which is explained next in the next section.

V. LISTENING TEST

The goal of the NAR headset is to reproduce augmented reality contents such that users cannot distinguish whether the sounds are coming from physical sources/environments or from the NAR headset. A listening test was conducted to subjectively validate the proposed HAE approach using individualized BRIRs. Three main research questions were asked in following listening tests:

Fig. 19. Listening Test Set up (▨ : Elevated speaker; ▩ : Azimuth speaker)

- **Naturalness:** Does virtual sound perceive natural?
- **Sound similarity:** Does virtual sound perceive similar to the real source i.e., sound coming from physical speakers?
- **Source position similarity:** How close is the virtual source position in 3D space as compared to real source?

The set up used for the listening test is shown in Fig. 19. Listener wearing the NAR headset prototype is surrounded by 7 Genelec 1030A loudspeakers. Five of the speakers are in azimuth plane (3 in the front and 2 in the rear), while two speakers are elevated at 30° in the front hemisphere. All the loudspeakers were positioned at a distance of 1.2 m away from the center of listener's head. Two MOTU Ultralite soundcards were used to interface with the 7 loudspeakers, 2 channels of AKG K702 headphones and 4 AKG C417 microphones. Three different listening sets were carried out as follows:

- **SET 1:** Perceptual similarity test between speaker and headphone playback of a male speech signal.
- **SET 2:** Perceptual similarity test between real and virtual mixing of two male speech signals.
- **SET 3:** Perceptual similarity test between real and virtual superposition of a speech signal with ambient sound.

Individualized BRIRs were measured for each of the seven speakers at both the binaural microphones' positions (m_{int} and m_{ext}) attached through the NAR headset, as shown in Fig. 19. Head tracker was mounted on the NAR headset to help subjects to maintain still head position during measurement process. Subjects were asked to repeat the measurement if they moved their head more than 5° in any of the three degrees of freedom. Individual HPTFs were compensated for with the measured individualized BRIRs. White noise sequence was used to train the adaptive filters offline using the proposed hybrid FxNLMS presented in this work. In each of the sets, listeners were presented with a pair of stimuli, one of them is played over physical speakers, while other can be either played over headphone as virtual source or physical speaker serving as hidden anchor.

In SET 1, a 4 second male speech signal was used to evaluate the similarity between real and virtual playback. Speech signal is played through all the seven loudspeakers for real playback, while same speech signal is convolved with the headphone equalized filters for virtual playback for the left and right ears using (12). Hybrid FxNLMS for case II presented in Fig. IV-D2 was used to obtain the equalized filters. The virtually synthesized secondary source signal, $u(n)$ for both

Fig. 20. Source confusion % for the three listening sets

the ears was subsequently played back over headphones, as shown in Fig. 19. Two additional pairs were included in this set as hidden anchors with both the stimuli played over speakers, resulting in a total of 9 test pairs.

In SET 2, a scenario is created, where two persons are having a conversation. Thus, two male speech signals (each around 3.5 seconds long) were played back from two different directions one after another, thereby merging the two signals. 3 pairs of loudspeaker configurations were chosen for the playback, namely, front left-front right (A-B), rear left-rear right (C-D), and elevated left-elevated right (E1-E2). For real playback, the two speech signals are played through each of the 3 loudspeaker pairs. For virtual playback, first speech was played through speaker, while second speech was played over headphones, and vice-versa. Thus, two virtually synthesized tracks were computed for each set of three pairs, in which one of them is played over speaker and the other is rendered virtually over headphones, while keeping the order of speech signals fixed. Thus, a total of 8 virtual signals were used in SET 2 including two hidden anchor pairs of both real sounds.

In SET 3, a male speech signal is superimposed onto ambient sounds of length around 6 seconds. In this scenario, two configurations are chosen for the superposition of speech signal. In the first configuration, ambient signal is played over two frontal loudspeakers A and B, while all four surrounding loudspeakers in horizontal plane (A, B, C and D) were chosen for the ambient signal playback in second configuration. Speech signal was played from front loudspeaker position, F for both real and virtual playback. For real playback, both the speech and ambient sound track is played through the loudspeakers. For virtual playback, the proposed hybrid FxNLMS with adaptive estimation of real ambient signals presented for Case III in subsection IV-E, was used to obtain the headphone equalized filters. The pre-recorded ambient signals ($r_{int}(n)$ and $r_{ext}(n)$) were used to remove the effect of real signals from the hybrid adaptive equalizer. The speech signal was then convolved with the equalized filter, and played back over the headphones simultaneously with the real ambient signals playing from the surrounding loudspeakers. An additional pair was constructed for each configuration with equalized filters computed using the hybrid FxNLMS in Case II with no real source present. The main objective here is to evaluate whether the adaptive equalization in the presence of external sounds i.e., Case III performs as good as with no external source

Fig. 21. Box plot showing subjective scores for sound similarity and source position similarity

present i.e., Case II. Thus, there were total of 5 pairs used in this listening set including one hidden anchor.

A. Listening test results

The listening test was conducted in a small quiet room with reverberation time of around 80 milliseconds. In all the listening sets, BRIRs were truncated to 50 milliseconds so as to include all the early reflections and most part of the late reverberations of the listening room. Order of the pairs in each listening set was randomized. Listeners were asked three questions for each randomly assigned pair. First question asked to subjects was to identify which of the two sounds are real i.e., coming from physical speaker. They were given the option of either choosing one of the two sounds or “both” if they perceive both sounds as natural. Similar subjective rating was used in [34] to do the pairwise comparison of two audio samples. Secondly, they were asked to rate the similarity of the two sounds on a scale of 0-10 from “completely different” to “same”. The main purpose here is to quantify the difference between real and virtual sounds, if any. Finally, they were also specifically asked to rate the proximity of the two sounds in 3D space on a scale of 0-10 from “very different” to “same”. There were a total of 22 pairs of audio tracks used in the listening test (9 pairs in SET 1, 8 pairs in SET 2 and 5 pairs in SET 3). The subjective ratings of the last two questions were decided based on some of the past works on A/B pairwise test to study the perceptual similarity of audio signals [35], [36]. These tests were mainly conducted for evaluation of blind source separation or different audio coding algorithms. A total of 18 subjects participated in the listening test comprising of 3 females and 15 male subjects. Two subjects were discarded as the similarity ratings for the hidden anchors with identical stimuli were given score less than 8. We will now discuss the listening test results based on the three research questions we want to answer in this study.

Naturalness: Naturalness of the virtual source is evaluated based on response of the first question, where subjects were

asked to select the real source among the two sounds or both if they perceived both sounds to be natural. Source confusion (i.e., virtual source is being confused with real source) is used as a measure of naturalness of the virtual source compared to sound coming from speakers. Source confusion can occur in two of the three possible scenarios: (1) subjects chose virtual source as real instead of real source, and (2) subjects perceived both virtual and real sounds as natural and marked “both” as the response. Thus, if virtual sound was reproduced very close to real sound, it was expected to have a very high percentage of responses for the “both” option. Fig. 20 shows the source confusion for the three listening sets estimated in percentage as sum of the two scenarios. As shown, for all the three listening sets, in more than 75% of cases, subjects identified both sounds as real implying virtual sound perceived natural. On the other hand percentage responses for the first scenario was very low, where subjects might have found real source colored and they chose virtual sound as real instead. Overall, very high source confusion of around 90 % is observed for all the listening sets, where subjects marked virtual sound as real. It was also interesting to note that source confusion increases further with increased number of external sources as for SET 2 and SET 3. One way ANOVA test was conducted to test the significance of reference-test pairs in each set across subjects. It was found that there were no significant variations among loudspeaker positions in SET 1 [$F(6, 105) = 0.68, p = 0.67$], speaker pairs in SET 2 [$F(5, 90) = 1.27, p = 0.29$] and different configurations in SET 3 [$F(3, 60) = 0.69, p = 0.56$].

Sound similarity: Fig. 21(a) shows boxplots for subjective ratings of the perceptual similarity between real and virtual sounds for all the three sets. Center line in the box represents the median value, while edges of the box are 25 and 75 percentiles responses. Top and bottom lines represent the extreme subject responses, while outliers are shown in red. In SET 1, most of the subjects found virtual sound highly similar to the real sound source with median of subjective ratings lying in the range 8-10 for all the loudspeaker posi-

TABLE I
MEAN SUBJECTIVE SCORES ALONG WITH THEIR 99 PERCENTILE
INTERVALS FOR THE THREE LISTENING SETS

Attribute	SET 1	SET 2	SET 3
Sound Similarity	8.44 (7.89 - 9.02)	9.19 (8.59 - 9.79)	9.13 (8.61 - 9.65)
Source Position Similarity	8.26 (7.33-9.19)	9.09 (8.32-9.87)	9.28 (8.77-9.78)

tions. However, fewer subjects could easily perceive difference between the two sounds. Using the one way ANOVA test, differences in mean scores for different loudspeaker positions across all subjects were found to be insignificant [$F(6, 105) = 0.98, p = 0.44$]. Similar to the source confusion, sound similarity further increased with increased number of sources and even in the presence of ambient sounds. Almost all the subjects rated the two sounds as highly similar with mean subjective ratings of 9.19 and 9.13 for SET 2 and SET 3, respectively. Different loudspeaker pairs in SET 2 were also found to have insignificant variations in their mean scores [$F(5, 90) = 0.82, p = 0.54$]. In addition, no significant differences were found between the two adaptive equalization methods used in SET 3 with [$F(1, 30) = 0.05, p = 0.83$] and [$F(1, 30) = 0.21, p = 0.65$] for the 2 and 4 ambient channels, respectively.

Source position similarity: Subjects were asked to compare the position of the two sounds and rate them based on their proximity in 3D space in terms of direction, distance and height. Fig. 21(b) shows the boxplot for the subjective ratings of source position similarity. Rating of “*very close*” indicates that the two sounds presented are very close to each other in 3D space, while rating of “*very different*” meant one of the sources may be located in completely different position possibly due to front-back confusions or even in-head localizations. This can be observed for SET 1 in Fig. 21(b) with couple of subjects giving “*different*” score. In general, source position similarities were observed very high with mean rating of 8.26, i.e., the two sounds are perceived very close to each other in 3D space. Similar to the previous two attributes, source position similarity increases further with increased number of sources and mean subjective ratings were found to be 9.1 and 9.3 for SET 2 and SET 3, respectively, implying close proximity of the two sounds. However, few outliers were also observed with rating of “*different*” for SET 2. One way ANOVA results for the effect of reference-test pairs in each set across subjects showed that no significant variations were observed among loudspeaker positions in SET 1 [$F(6, 105) = 0.89, p = 0.51$] and speaker pairs in SET 2 [$F(5, 90) = 1.46, p = 0.21$]. Furthermore, adaptive equalization for Case II and III have no significant differences in their ratings with [$F(1, 30) = 0.4, p = 0.53$] and [$F(1, 30) = 0.85, p = 0.36$], respectively for 2 and 4 ambient channels, which indicates that proposed adaptive equalizer performs equally well, even in the presence of external sounds.

Table I summarizes the mean subjective ratings with their 99 % confidence interval for sound similarity and source position similarity. Clearly, virtual sources were found to be

highly similar (barely distinguishable), as well as very close to the real sources in 3D space using the NAR headset. It was also interesting to find correlation among the three sound source attributes studied above. High similarity between the two sounds also meant that they are very close to each other in 3D space and vice-versa for most subjects, but very close proximity in space doesn’t always mean they are highly similar as reported by few subjects. In addition, naturalness of the virtual sound does not necessarily imply that the two sounds are highly similar or are very close to each other in 3D space. Nevertheless, high sound similarity and source position similarity indeed resulted in virtual source being identified as real.

VI. CONCLUSIONS

In this paper, we presented a new approach in reproducing natural listening in augmented reality headsets based on adaptive filtering techniques. The proposed NAR headset structure consists of an open ear cup with pairs of internal and external microphones. Based on the study of different headphones isolation characteristics, it was found that headphones with open design are more suitable for AR related applications, as they allow direct external sound to reach listeners’ ear without much attenuation. However, for closed-end headphones or less open headphones, additional processing should be applied to compensate for the headphone isolation using the same sensing units. Based on the amplitude/spectral difference between the two microphone signals, a pair of compensation filters can be applied to make the headsets acoustically transparent. This has been identified as an extension of the current prototype. For virtual source reproduction via binaural synthesis, individual headphone equalization is applied using adaptive algorithms to compensate for the HPTFs. Modified FxNLMS is proposed with additional spatial filter introduced in the secondary path to improve the convergence rate. However, it is observed that the modified FxNLMS is not able to entirely adapt to the desired response in high frequencies for some of the source positions, while conventional FxNLMS suffers from spectral distortion in low frequencies. Hence, we proposed a hybrid FxNLMS to combine the two approaches for optimal performance. Using simulation results, it was found that the hybrid FxNLMS is superior to both approaches with mean square steady state error reduction of more than 25 dB for most of the source positions tested. This implies that virtual sound is reproduced perceptually similar as in direct natural listening. Hybrid FxNLMS is further extended with adaptive estimation of external sounds, as they might interfere with the convergence process. Therefore, with the help of hybrid adaptive equalizer, NAR headset can be individually equalized even in the presence of noisy environments. Listening test was conducted to evaluate perceptual similarities between physical speaker playback and virtual headphone playback. Very high source confusion % was observed, which indicates that virtual source sounds natural. Subjects could not differentiate between real and virtual sounds and their positions in 3D space were also in very close vicinity. Moreover, perceptual similarity between real and virtual sounds further increased in an augmented scenario with both real and virtual sources present.

A real-time NAR headset incorporating the proposed adaptive algorithm is currently being developed. Extensive subjective test will be carried out to further validate the robustness and practicality of the NAR headset and how it can be applied to new immersive communication applications.

REFERENCES

- [1] R. T. Azuma, "A survey of augmented reality," *Presence*, vol. 6, pp. 355-385, 1997.
- [2] T. Nilsen, S. Linton, and J. Looser, "Motivations for augmented reality gaming," *Proceedings of FUSE*, vol. 4, pp. 86-93, 2004.
- [3] T. Sielhorst, M. Feuerstein, and N. Navab, "Advanced medical displays: A literature review of augmented reality," *Journal of Display Technology*, vol. 4, pp. 451-467, 2008.
- [4] T. Lokki, H. Nironen, S. Vesa, L. Savioja, A. Härmä, and M. Karjalainen, "Application scenarios of wearable and mobile augmented reality audio," in *Audio Engineering Society Convention 116*, 2004.
- [5] M. Billinghurst and H. Kato, "Collaborative augmented reality," *Communications of the ACM*, vol. 45, pp. 64-70, 2002.
- [6] T. Miyashita, P. Meier, T. Tachikawa, S. Orlic, T. Eble, V. Scholz, et al., "An augmented reality museum guide," in *Proceedings of the 7th IEEE/ACM International Symposium on Mixed and Augmented Reality*, 2008, pp. 103-106.
- [7] H. Ishii and B. Ullmer, "Tangible bits: towards seamless interfaces between people, bits and atoms," in *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*, 1997, pp. 234-241.
- [8] R. W. Lindeman, H. Noma, and P. G. de Barros, "Hear-through and mic-through augmented reality: Using bone conduction to display spatialized audio," in *Proceedings of the 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*, 2007, pp. 1-4.
- [9] A. Härmä, J. Jakka, M. Tikander, M. Karjalainen, T. Lokki, J. Hiipakka, et al., "Augmented reality audio for mobile and wearable appliances," *Journal of the Audio Engineering Society*, vol. 52, pp. 618-639, 2004.
- [10] M. Tikander, M. Karjalainen, and V. Riikonen, "An augmented reality audio headset," in *Proc. of the 11th Int. Conf. on Digital Audio Effects (DAFx-08)*, Espoo, Finland, 2008.
- [11] J. Rämö and V. Välimäki, "Digital Augmented Reality Audio Headset," *Journal of Electrical and Computer Engineering*, vol. 2012, Article ID 457374, 13 pages, 2012. doi:10.1155/2012/457374.
- [12] D. W. Schobben and R. M. Aarts, "Personalized multi-channel headphone sound reproduction based on active noise cancellation," *Acta acustica united with acustica*, vol. 91, pp. 440-450, 2005.
- [13] H. Møller, "Fundamentals of binaural technology," *Applied acoustics*, vol. 36, pp. 171-218, 1992.
- [14] R. Nicol, "Binaural Technology," in *AES Monograph*, ed. 2010.
- [15] V. R. Algazi and R. O. Duda, "Headphone-based spatial sound," *Signal Processing Magazine, IEEE*, vol. 28, pp. 33-42, 2011.
- [16] J. C. Middlebrooks, "Narrow-band sound localization related to external ear acoustics," *The Journal of the Acoustical Society of America*, vol. 92, pp. 2607-2624, 1992.
- [17] J. Blauert, *Spatial Hearing: The psychophysics of human sound localization*: The M.I.T. Press, 1997.
- [18] J. Blauert, *The technology of binaural listening*: Springer, 2013.
- [19] D. Brungart, "Near-field virtual audio displays," *Presence*, vol. 11, pp. 93-106, 2002.
- [20] F. L. Wightman and D. J. Kistler, "Resolution of front-back ambiguity in spatial hearing by listener and source movement," *The Journal of the Acoustical Society of America*, vol. 105, pp. 2841-2853, 1999.
- [21] D. Satongar, C. Pike, Y. W. Lam, and T. Tew, "On the Influence of Headphones on Localization of Loudspeaker Sources," in *Audio Engineering Society Convention 135*, 2013.
- [22] A. Kulkarni and H. S. Colburn, "Role of spectral detail in sound-source localization," *Nature*, vol. 396, pp. 747-749, 1998.
- [23] J. Hebrank and D. Wright, "Spectral cues used in the localization of sound sources on the median plane," *The Journal of the Acoustical Society of America*, vol. 56, pp. 1829-1834, 2005.
- [24] K. Sunder, E.-L. Tan, and W.-S. Gan, "Individualization of Binaural Synthesis Using Frontal Projection Headphones," *Journal of the Audio Engineering Society*, vol. 61, pp. 989-1000, 2013.
- [25] H. Han, "Measuring a dummy head in search of pinna cues," *Journal of the Audio Engineering Society*, vol. 42, pp. 15-37, 1994.
- [26] J. Rämö, "Evaluation of an augmented reality audio headset and mixer," Dissertation, Helsinki University of Technology, 2009.
- [27] F. Brinkmann and A. Lindau, "On the effect of individual headphone compensation in binaural synthesis," *Fortschritte der Akustik: Tagungsband d. 36. DAGA*, pp. 1055-1056, 2010.
- [28] M. Bouchard, S. G. Norcross, and G. A. Soudoudre, "Inverse filtering design using a minimal-phase target function from regularization," in *Audio Engineering Society Convention 121*, 2006.
- [29] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *The Journal of the Acoustical Society of America*, vol. 66, pp. 165-169, 1979.
- [30] S. M. Kuo and D. Morgan, *Active noise control systems: algorithms and DSP implementations*: John Wiley & Sons, Inc., 1995.
- [31] T. Nishino, N. Inoue, K. Takeda, and F. Itakura, "Estimation of HRTFs on the horizontal plane using physical features," *Applied Acoustics*, vol. 68, pp. 897-908, 2007.
- [32] T. Qu, Z. Xiao, M. Gong, Y. Huang, X. Li, and X. Wu, "Distance dependent head-related transfer function database of KEMAR," in *Audio, Language and Image Processing, 2008. ICALIP 2008. International Conference on*, 2008, pp. 466-470.
- [33] R. Ranjan and W.-S. Gan, "Applying Active Noise Control Technique for Augmented Reality Headphones," in *Internoise 2014*, Melbourne.
- [34] B. De Man and J. D. Reiss, "A pairwise and multiple stimuli approach to perceptual evaluation of microphone types," in *Audio Engineering Society Convention 134*, 2013.
- [35] B. Fox, A. Sabin, B. Pardo, and A. Zopf, "Modeling perceptual similarity of audio signals for blind source separation evaluation," in *Independent Component Analysis and Signal Separation*, ed: Springer, 2007, pp. 454-461.
- [36] E. Parizet, N. Hamzaoui, and G. Sabatié, "Comparison of some listening test methods: a case study," *Acta Acustica united with Acustica*, vol. 91, pp. 356-364, 2005.

Rishabh Ranjan received his B.Tech. degree in Electronics and Communication engineering from International Institute of Information Technology (IIIT), Hyderabad, India in 2010. He is currently pursuing his Ph.D. degree in electrical and electronic engineering at Nanyang Technological University (NTU), Singapore. His research interests include 3D audio, wave field synthesis, adaptive signal processing, GPU computing, and embedded systems.

Woon-Seng Gan (M'93-SM'00) 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 in the 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 spatial audio. He has published more than 250 international refereed journals and conferences, and has granted seven Singapore/US patents. He had co-authored three books on Digital Signal Processors: Architectures, Implementations, and Applications (Prentice Hall, 2005), Embedded Signal Processing with the Micro Signal Architecture, (Wiley-IEEE, 2007), and Subband Adaptive Filtering: Theory and Implementation (John Wiley, 2009). He is currently a Fellow of the Audio Engineering Society (AES), a Fellow of the Institute of Engineering and Technology (IET), a Senior Member of the IEEE, and a Professional Engineer of Singapore. He is also an Associate Technical Editor of the Journal of Audio Engineering Society (JAES); Associate Editor of the IEEE Transactions on Audio, Speech, and Language Processing (ASLP); Editorial member of the Asia Pacific Signal and Information Processing Association (APSIPA) Transactions on Signal and Information Processing; and Associate Editor of the EURASIP Journal on Audio, Speech and Music Processing. He is currently a member of the Board of Governor of APSIPA.