Gupta, Rishabh; He, Jianjun; Ranjan, Rishabh; Gan, Woon Seng; Klein, Florian; Schneiderwind, Christian; Neidhardt, Annika; Brandenburg, Karlheinz; Valimaki, Vesa

Augmented/Mixed Reality Audio for Hearables: Sensing, control, and rendering

Published in:
IEEE Signal Processing Magazine

DOI:
10.1109/MSP.2021.3110108

Published: 01/05/2022

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Gupta, R., He, J., Ranjan, R., Gan, W. S., Klein, F., Schneiderwind, C., Neidhardt, A., Brandenburg, K., & Valimaki, V. (2022). Augmented/Mixed Reality Audio for Hearables: Sensing, control, and rendering. *IEEE Signal Processing Magazine*, *39*(3), 63-89. https://doi.org/10.1109/MSP.2021.3110108

Augmented/Mixed Reality Audio for Hearables: Sensing, Control, and Rendering

Rishabh Gupta, Jianjun He, Rishabh Ranjan, Woon-Seng Gan, *Senior Member, IEEE,* Florian Klein, Christian Schneiderwind, Annika Neidhardt, Karlheinz Brandenburg, *Fellow, IEEE* and Vesa Välimäki, *Fellow, IEEE*

AUGMENTED OR MIXED REALITY (AR/MR) is emerging as one of the key technologies in the future of computing. Audio cues are critical for maintaining a high degree of realism, social connection, and spatial awareness for various AR/MR applications such as teaching and training, gaming, remote work, education, and virtual social gatherings. Motivated by a wide variety of AR/MR listening experiences delivered over hearables, this feature article systematically reviews the integration of fundamental and advanced signal processing techniques for AR/MR audio to equip the researchers and engineers in the signal processing community for the next wave of AR/MR.

I. Introduction: AR/MR audio experience over hearables

Alternate reality technologies aim to provide a feeling of presence to humans through engaging multi-modal content. We define AR/MR as an alternate reality experience achieved by the seamless fusion of the reproduced virtual content with the real-world stimuli that can be modified as desired. This definition expands on AR's previous usage, where the experiences provided were limited to the overlay of virtual content in the real world. It can also include other alternate reality technologies, such as virtual reality (VR), where users can get transported to a virtual environment. The immersive AR/MR technologies have demonstrated substantial benefits in various applications such as education and training, tourism, and remote working. We have seen significant progress in rendering AR/MR devices' different modalities in the past few decades. In particular, the new lifestyle of reduced physical connection triggered by the COVID-19 pandemic has made such needs even more critical than ever before.

This paper focuses on sound (or audio), an inherent part of our everyday lives for communication, social interactions, and situational awareness. Unlike vision, where the field of view is limited, natural listening always spans a 360° range. The pervasive spatial perception of sound can be critical in a wide variety of high-stress situations, such as warnings for approaching vehicles or public announcements in case of emergencies where the visual cues may not be enough. Even in day-to-day cases, such as conversations among a group of people, we rely on audio cues to direct our attention towards a particular speaker, referred to as the cocktail party effect.

Audio devices today can be broadly classified into two major categories: speakers and headphones. Speakers are widely used to playback audio content for multimedia applications, where a fixed configuration of speakers can provide the

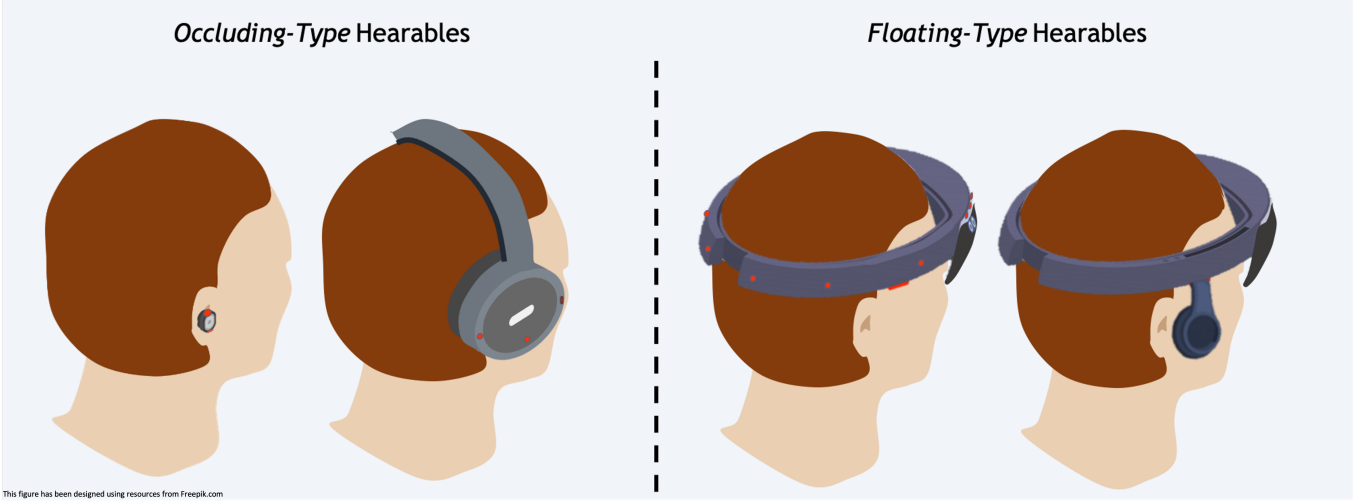

Fig. 1. Two main categories of hearable devices are described in this paper: *occluding-type* hearable which occlude the sound waves entering the ear canal, and *floating-type* hearables that do not touch the ear directly.

desired audio experience. However, large speaker arrays are required to deliver high fidelity audio experience, which is usually limited to a fixed space volume. In AR/MR audio, where the listener can move from any given location to another (not necessarily in the same room), delivering high fidelity AR/MR audio over speakers in all acoustic environments becomes infeasible.

Unlike speakers, headphones are wearable, protect the listener's privacy, and provide a better personal experience by delivering AR/MR audio for any location. Hearables have become a popular umbrella term to denote headphones with essential hardware (e.g., transducers and sensors) and software (e.g., algorithms and applications) capabilities to offer various AR/MR audio experiences [1], [2]. Based on how the hearable devices couple with the ears, we consider two main categories of devices, namely *occluding-type* and *floating-type* hearables as depicted in Fig. 1, though other variants can be derived. The *occluding-type* hearable is in close contact of the ear and occludes the sound waves entering the ear canal directly, which typically includes intra-concha headphones that sit at the entrance of the ear canal, in-ear headphones that is inserted into the ear canal, over-ear (or circumaural) and on-ear (or supra-aural) headphones that fully or partially cover the pinna. The other type of hearable, referred to as *floating-type* in this paper, is a headphone speaker or micro speaker that does not touch the ear directly and usually employs a frame structure on-ear or head to support a head-mounted display (HMD). The size of the hearables also affects the possible features, performance, and experience. In general, with increasing size, the hearable device typically becomes heavier, less portable, and less comfortable. However, it could integrate more sensors, more microphones distributed spatially, and a larger speaker driver for better sound quality, larger battery, and more powerful processors. Other additional devices, such as smartphones, can also assist the design and implementation of AR/MR audio systems.

Fig. 2 illustrates selected milestones in the AR/MR audio and hearables development. One of the first commercial examples for AR/MR audio is Bose QC1, an Active Noise Control (ANC) headphone. In 2004, a study by Härmä et al. proposed a framework for delivering AR/MR audio experiences [3] with binaural microphones integrated with headphones. Tikander et al. [4] investigated the long-term usability of an AR/MR audio device while the users were

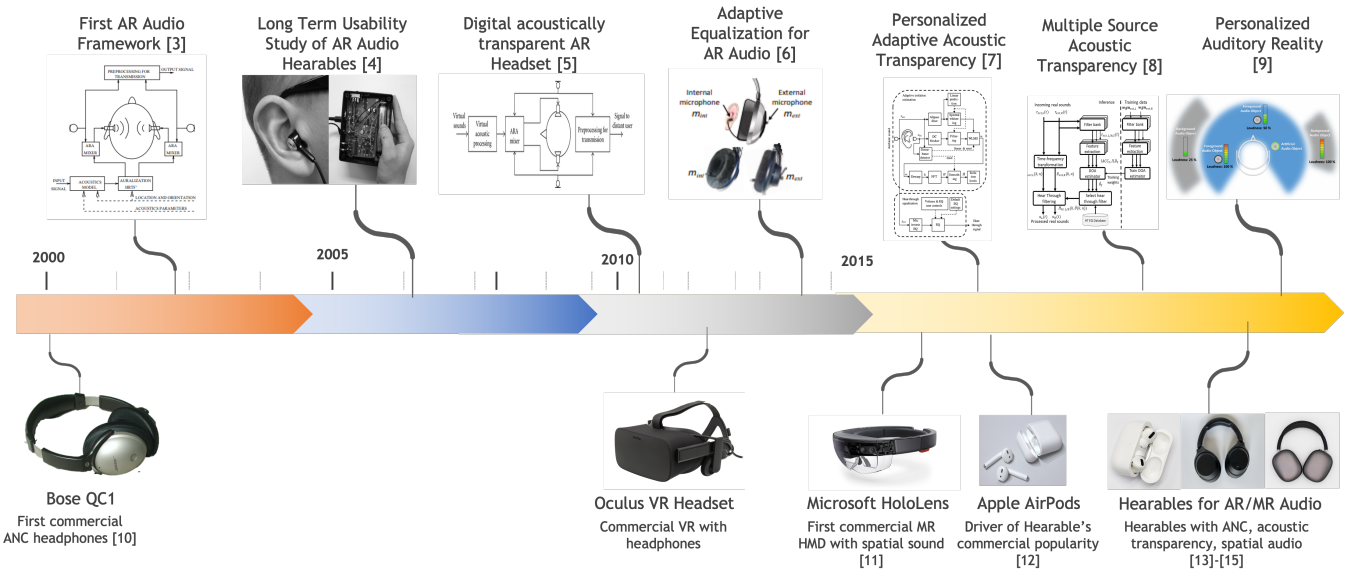

Fig. 2. Selected milestones of AR/MR audio research and commercial examples over the last two decades.[3]–[15]

engaged in several day-to-day activities such as walking, eating, chewing. Rämö and Välimäki have implemented a digital solution for an acoustically transparent AR/MR headset [5], which allows real sound to pass through unaltered. Ranjan and Gan have described an AR/MR audio headset [6], which, unlike fixed filter approaches in the past, performs adaptive equalization for rendering virtual sounds. The AR/MR audio research led to the launch of commercial products equipped with virtual spatial sound, with Oculus VR in 2012 and Microsoft HoloLens in 2016. Since its launch in 2016, Apple AirPods has become the most prominent commercial popularity driver for hearables, which have evolved rapidly to include various features, such as ANC, acoustical transparency, and spatial audio. In recent years, the focus of AR/MR audio research has shifted towards delivering personalized AR/MR audio experiences. Liski et al. [7] proposed an adaptive method to personalize the acoustic transparency mode. Gupta et al. investigated advanced acoustic transparency techniques to improve the performance in terms of the multiple-source acoustic environment and personalization [8]. Brandenburg et al. proposed a personalized auditory reality framework that incorporates real and virtual sound processing [9].

Hearables have evolved over the past few years and become a popular platform to deliver various AR/MR audio experiences. Most hearables, except *floating-type* hearables, attenuate real sound due to the physical presence of the device. In several daily situations, the listener may need to listen to critical sounds in the real environment, such as public announcements, warning alarms, conversations, and traffic passing by. In such situations, microphones can capture real sound, and the hearable can process and play it back instantly to allow the user to create a desired augmented listening experience. The real sound playback combines with the binaural rendering of virtual content to be indistinguishable from real sound, thus, creating a seamless AR/MR audio experience. ANC, acoustic transparency, and spatial audio are popular features either presented in various commercial hearables in the market today or highly desired for future hearables by consumers to enable engaging AR/MR audio experiences.

The AR/MR experiences include listening to music in noisy environments, gaming, and teleconferencing. In a noisy

environment, such as inside a bus in heavy traffic, it can be difficult to enjoy the music due to auditory masking effects. Hearables can reduce the noise levels and playback spatial audio to provide an augmented musical experience. An AR/MR gaming experience can be more immersive and engaging by allowing users to modify the real sounds and playback virtual audio scene that adapts to user movements in the real environment. Teleconferencing is a pivotal technology to facilitate remote communication and build social connections among people situated anywhere globally. In an AR/MR immersive teleconferencing experience, all attendees interact in the same shared environment, which can comprise either real or virtual sounds or a mix of real and virtual sounds. While attendees move freely in their real environment, virtual sound playback makes them feel that they are moving in the shared environment. The real sound modification can allow attendees to selectively hear a person/group without any unwanted ambient noise or other interfering talkers. The following sections describe the signal processing techniques and their integration to create such experiences.

II. Overview: basics and signal processing framework

This section discusses the basics of binaural audio and an overview of signal processing blocks needed to realize AR/MR audio experiences.

A. Basics of Binaural Audio

Binaural listening refers to the natural process in which humans use two ears to listen to sounds in the real world. To explain binaural listening, we refer to the source-medium-receiver model [16]. The real acoustic world consists of sound sources that may originate from different positions in three dimensions (3D) space, defined as the azimuth, elevation, and distance. Typically, sound waves from each source propagate through the medium of air after reflections from all surfaces to reach the human's eardrum at both the left and right ears. Binaural Room Impulse Response (BRIR) denotes this process's acoustic response, which broadly consists of two major components.

The room environment characterizes the first component of propagation, including the direct path between the source position and listener position and the reflections from different surfaces in the environment, defined as Room Impulse Response (RIR).

The frequency-domain representation of RIR is called the Room Transfer Function (RTF). Reflections and scattering around the listener, mainly from the torso, head, and outer ear (pinna in particular), characterize the second propagation component, denoted as Head-Related Impulse Response (HRIR), or its frequency domain representation Head-Related Transfer Function (HRTF) [16]. Multiple sources undergo their respective sound wave propagation paths in the environment and around the listener before summing at the eardrum through wave superposition.

After reaching the listener's eardrum, the received acoustic vibrations are transmitted through the middle ear, converted into the nerve impulses by the cochlear in the inner ear, and finally processed in the brain to create the perception of the sound source loudness, timbre, spatial location, and room environment. Loudness perception varies in the frequency domain in a non-linear manner [17]. For instance, equal Sound Pressure Levels (SPL, measured in

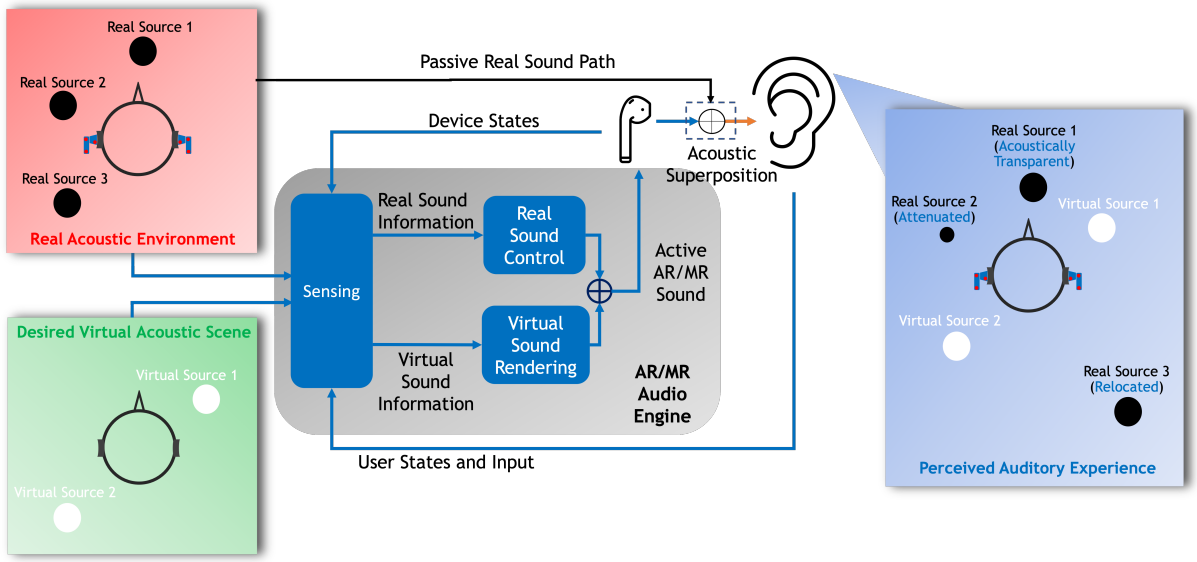

Fig. 3. Signal processing blocks for AR/MR audio. Three major blocks namely, sensing, real sound control and virtual sound rendering, along with their inputs and outputs are shown, along with an AR/MR audio use-case example.

decibel dB) at the eardrum can yield different perceived loudness levels at different frequencies characterized by equal-loudness contour. Perceptually, loudness doubles with every 10 dB increase in SPL [17]. Sound source localization is another critical component of binaural listening, where humans use various static and dynamic audio cues to localize sound in 3D. Major static cues to localize the sound directions in terms of azimuth and elevation are: 1) Interaural Time Difference (ITD) that accounts for the sound wave traveling time difference between the two ears; 2) Interaural Level Difference (ILD) caused primarily by the obstruction of sound waves due to the presence of the head (i.e., head shadowing effect); and 3) Spectral Cues (SC) as a result of the sound reflections mainly from the pinna [16]. HRTFs encode these three major static cues. In addition to static cues, a listener's natural head movement gives rise to a specific pattern of changes in static cues, which can further assist in source localization. Further, humans can perceive approximate distances of distinct sound sources by using some static cues, such as the prior knowledge of the sound characteristics, sound levels (loudness), and the energy ratio between the direct sound and reverberation (i.e., direct-to-reverberant ratio DRR). An increase in ILD for nearby sources and dynamic cues from head and body movements also provide distance cues. Interaction of sound sources can produce changes in perception. One example is the auditory masking phenomenon, where one sound source fully or partially dominates another sound source's perception when overlap exists in time, and frequency [17].

B. Signal Processing Framework for AR/MR Audio

To realize AR/MR audio, we need to augment the user's perception of both the real and virtual sounds using a hearable device with a dedicated AR/MR audio engine. Fig. 3 shows the signal processing blocks and relevant inputs and outputs of the AR/MR audio engine for delivering the desired AR/MR audio experience. The AR/MR sound heard by the user is an acoustic superposition of active and passive sound paths [1], [3]. Passive sound is the real acoustic sound that leaks through the hearable device when worn by the user (except *floating-type* hearables). Depending on

the form factor of the hearable, the passive real sound leakage reaching our eardrum might be modified, as compared to natural listening with open ears. A tighter coupling between the hearable and the ear leads to reduced leaked real sound signal and a larger attenuation. This passive attenuation varies with frequency, usually acting as a low pass filter, with more substantial attenuation at higher frequencies [5]. The AR/MR audio algorithms must compensate for real sound leakage to create a desired real sound experience.

Active AR/MR sound output is the sound generated by the AR/MR audio engine and played back through the hearable speakers. Four major types of inputs are generally involved in generating the desired active AR/MR sound signal, namely, real sound information from the user's local real acoustic environment; virtual sound information about the desired virtual acoustic scene; information about the user; and device states. The real sound information from the local acoustic environment, typically captured by microphones embedded in hearables, includes characteristics of real sound sources, such as frequency content, spatial position, timbre characteristics, and reverberation. Virtual sound information usually includes the sound rendering formats such as object-based, channel-based, and scene-based formats, the corresponding sound signals, associated metadata, and the desired playback mode required by the virtual sound rendering block to deliver desired virtual sonic experiences. The other two inputs are the user and hearable device's current responses and states. Hearables embed microphones and speakers to capture the real sound and playback the active AR/MR output signal, respectively. The responses of external sound sources and hearable speakers at the user's eardrum and hearable microphones must be characterized and taken into consideration. Moreover, the user's state information, including tracking the user's physical movements and biological states, is vital for creating a natural and personalized AR/MR audio experience. Section III provides a detailed description of the definition and estimation of these responses.

We describe the AR/MR audio engine in terms of three processing blocks: sensing block, real sound control block, and virtual sound rendering block. The sensing block aims to understand the user's acoustic environment and user/device responses and states based on input from various sensors. The sensing block relays the captured information to real sound control and virtual sound rendering blocks to synthesize the desired active AR/MR sound signals. The goal of the real sound control block for AR/MR audio is to modify and transform the existing acoustic sound environment into a specific auditory experience to match the user's preference, usually relative to the natural listening experience with open ears. Lastly, the virtual sound rendering block generates audio signals for recreating desired virtual sound experiences.

III. Sensing: Analysis of user/device response, acoustic environment, and user/device states

Sensing block is an indispensable component of hearable devices. It continuously senses the existing dynamic user acoustic environment, user activity, and device characteristics required to create different AR/MR audio listening experiences. Fig. 4 shows an overview block diagram of sensing block's modules that can be included in hearables. They are:

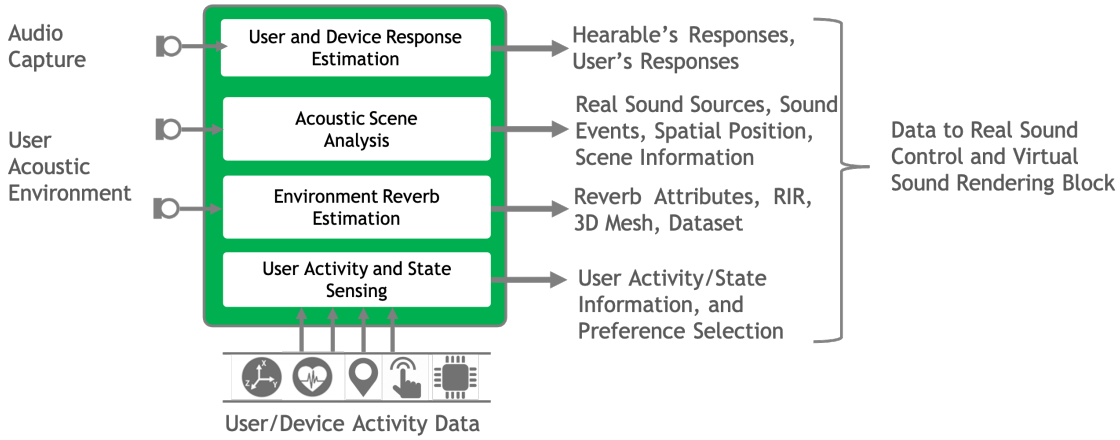

Fig. 4. An overview block diagram of sensing. The different sensors capture user, device and acoustic environment information, and pass the information to three sensing modules to extract useful metadata and signals for further processing by real sound control and virtual sound rendering.

1) **User and device response estimation** to characterize the coupling of hearables with user's ear, microphone responses on hearables, and user's responses
2) **Acoustic scene analysis** to process acoustical information captured from the user acoustic scene to analyze and extract distinct sound sources and accompanying metadata like sound event, spatial position, and scene information.
3) **Environment reverb estimation** to estimate environment reverb characteristics necessary for reverb control of real sound and virtual sound rendering
4) **User activity and device state sensing** to track user's activity, such as physical movements, gestures and preferences, and device state from device sensors

The above processing modules generate multiple outputs in the form of audio signals, extracted metadata from the real sound field (such as the type and position of sources) using acoustic scene analysis, estimating environment reverb (such as RIR estimation, reverb attributes extraction, 3D mesh acquisition), estimated hearables response, as well as personalized HRTFs, current user/device state/activity, and even physiological information. The real sound control and virtual sound rendering blocks use sensing block outputs to create the various AR/MR audio experiences.

A. User and device response estimation

This section describes the definitions and estimation of four responses essential for processing real and virtual sounds. These responses are HRTFs, Hearable Microphone Transfer Function (HeMTF), Hearable Speaker Transfer Function (HeSTF), and Hearable Passive Transfer Function (HePTF). All the transfer functions, except HRTFs, are measured with the hearable placed on the user's ear and, thus, inherently depends on the hearable's form-factor and coupling with the ear. Fig. 5 illustrates their definitions and examples of magnitude responses.

a) HRTF: represents a linear-time-invariant (LTI) system between a static acoustic sound source in free-field and the measurement microphone, ideally situated at the listener's eardrum, which can be defined as [18]:

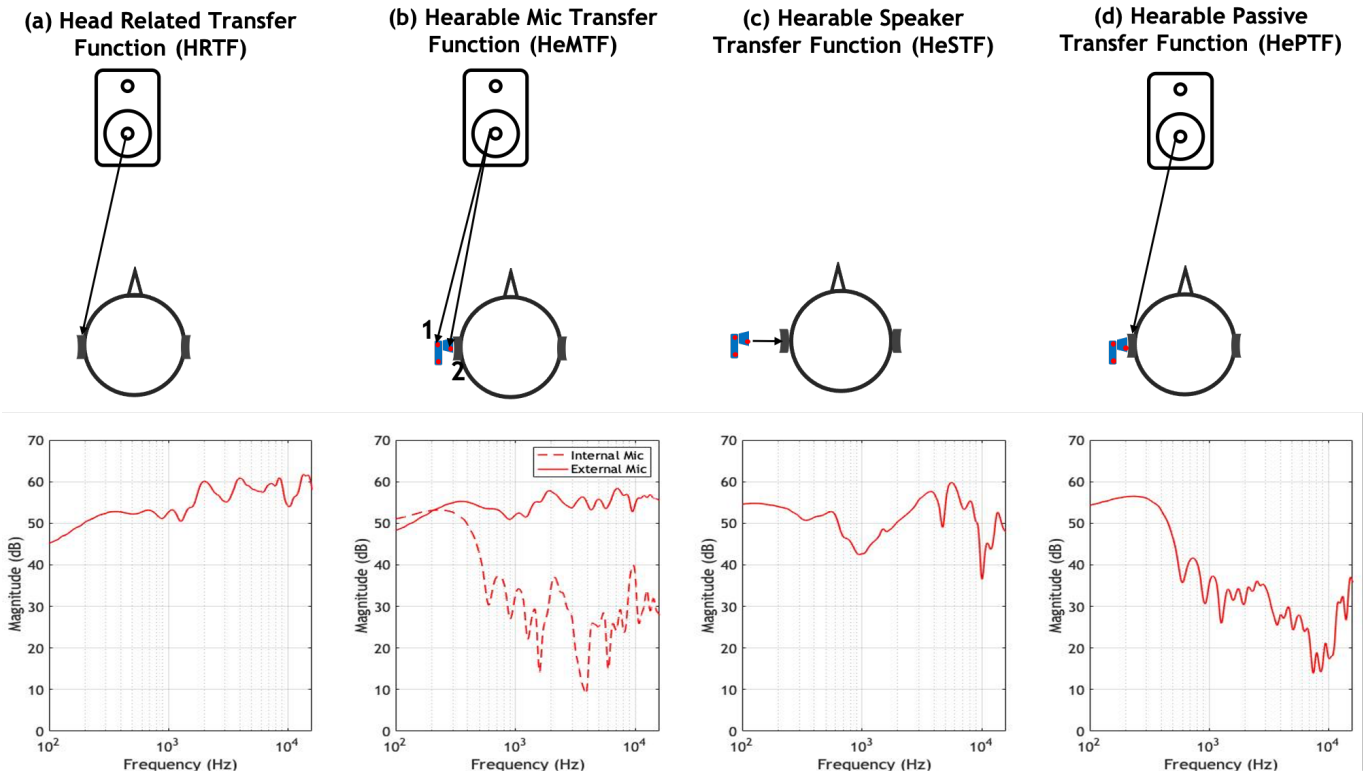

Fig. 5. Transfer functions for AR/MR audio. Example of magnitude response measured: (a) HRTF at the eardrum of the dummy head with open ear; (b) HeMTF at the external and internal microphones of hearables; (c) HeSTF from hearables speaker to eardrum of the dummy head; (d) HePTF at the eardrum of the dummy head with hearables worn on the ear

$$HRTF_i(a, \theta, \phi, r, f) = \frac{P_i(a, \theta, \phi, r, f)}{P_0(r, f)}, \tag{1}$$

where P_i denotes the sound pressure in the frequency domain at the measurement microphone location (with i representing either left (L) or right ear (R)). P_0 denotes the reference sound pressure in the frequency domain at the center position of the listener's head (without the head present). Ideally, the measurement microphone for P_i must be placed at the eardrum, but it has been shown in past studies that measurement conducted at the blocked ear canal can also preserve the major spatial cues [18]. Moreover, a, θ, ϕ, r, f denote the dependence on the listener's anthropometric features (ear, head, and torso), the azimuth angle, elevation angle, distance (between the sound source and the listener), and frequency, respectively. The idiosyncrasy of anthropometric features causes unique sound reflections and diffraction for each individual. For instance, pinna reflections and diffractions result in distinct peaks and notches in HRTF (at 8 kHz in the HRTF curve, as shown in Fig. 5). The most accurate way of obtaining HRTFs is by conducting measurements for each listener [18]. Typically, these measurements are realized by playing a broadband excitation signal from an external speaker at a particular source location and estimating HRTFs from the recordings at microphones placed at the user's ear canal entrance [19]. However, this measurement process is cumbersome and time-consuming. It requires many pieces of equipment, including large speaker arrays and miniature microphones, to capture HRTFs at a large number (usually hundreds or more) of spatial positions. Typically, non-individualized HRTFs selected from an existing database of HRTFs measured on human subjects or dummy heads with averaged anthropometric features of the

human population can reduce this effort [18]. Recently, fast HRTF measurement approaches based on unconstrained head movements have also been used to capture HRTF at high spatial resolution using a single external speaker, with accuracy similar to static measurements used in traditional approaches. These approaches drastically reduce the measurement duration and require lesser equipment than the traditional approaches [19].

Analytical and numerical methods that account for the scattering of sound waves by the listener's head, torso, and pinna can also be used to compute HRTFs. To obtain HRTFs using analytic methods, the spherical head and snowman model (with given head width) approximate the human head shape [18]. Numerical methods like Boundary Element Method (BEM), on the other hand, require multiple 3D models with detailed head and pinna geometries [20], [21]. Laser scanners and cameras are required to capture 3D models for accurately estimating HRTFs. However, numerical methods often require high computation and suffer from poor accuracy, especially for resolving the pinna's finer details [20]. Parametric approaches for HRTF estimation have been proposed to overcome the issues with numerical methods. Parametric approaches involve selecting individualized HRTFs from a large database of HRTFs, typically measured and stored for many subjects. These approaches select best-matching HRTFs based on a few critical anthropometric measurements (such as head width and pinna features) that cause perceptually noticeable spectral differences or alter localization cues (such as ITD or ILD) in HRTFs [18]. Another branch of methods is perceptual-driven, requiring the listener to select the preferred HRTFs from the database [22]. The perceptual-driven selection methods are limited in accuracy by the size and variance of the HRTF dataset, where a more diverse dataset facilitates a more accurate selection.

b) HeMTF: refers to the transfer function from the sound source to the microphones situated on the hearable (microphone index denoted by m). Using two microphones as an example as shown in Fig. 5, HeMTF includes the response of microphone(s) situated on the outer surface of the hearable (called the external microphone, with $m = 1$), and the microphone(s) situated in the cavity formed by the hearable and ear canal (called internal microphone, with $m = 2$). HeMTF can be defined as:

$$HeMTF_{i,m}(a, \theta, \phi, r, f) = \frac{\tilde{P}_{i,m}(a, \theta, \phi, r, f)}{P_0(r, f)}. \tag{2}$$

where $\tilde{P}_{i,m}$ denotes the sound pressure at the hearable's microphone m for each ear measured with the hearable placed on a dummy head or human subject. The reference sound pressure P_0 is the same as in the case of HRTF. Similar to HRTF, HeMTF is also idiosyncratic and ideally must be measured for each listener and source position. However, in addition to being idiosyncratic, HeMTF also varies with the hearable's form factor and placement on the user's ear, which also relates to the microphone position on hearables. Despite the lack of dedicated studies, methods for HRTF estimation can generally be adopted to derive HeMTF.

c) HeSTF: To playback the synthesized sounds from hearables, a transducer or speaker is required. When the user wears the hearable, the hearable's coupling with head and ear generates an idiosyncratic response of hearables speaker at the eardrum, termed as the HeSTF (for headphones evaluated in past studies, this has been termed as

Headphone Transfer Function HpTF) [23]. HeSTF comprises the transducer response and the transfer function from the transducer to the eardrum. HeSTF can be defined as the ratio of sound pressure measured at the eardrum with an excitation signal played back through the hearable for each channel (left/right).

Since HeSTF depends on an individual's ear shape and hearable fitting, conducting the measurement in-situ presents a practical challenge. While some past studies have used non-individualized HeSTF, Pralong and Carlile [24] have compared HeSTF for different listeners and determined that the use of non-individualized HeSTF could cause considerable distortions in the 4-10 kHz range. The distortions above 4 kHz are mainly due to the difference in acoustic scattering by the outer ear, which is highly idiosyncratic.

The methods for HeSTF estimation include online adaptive algorithms [7], and statistical approaches such as mean, median, or upper variance limit [25] of HeSTF databases measured on dummy heads and human subjects.

d) HePTF: represents the passive response of the hearables when worn by the user. The expression for HePTF can be written in a similar way as HeMTF:

$$HePTF_i(a, \theta, \phi, r, f) = \frac{\tilde{P}_i(a, \theta, \phi, r, f)}{P_0(r, f)}, \tag{3}$$

where $\tilde{P}_i$ denotes the transfer function from the source to the eardrum, which is similar to $\tilde{P}_{i.m}$ defined in case of HeMTF, but now the measurement microphone is located at user's eardrum instead of being located on the hearable. Fig. 5 shows the magnitude response of HePTF. Typically, it has a low-pass characteristic due to the filtering effect by the hearable's physical construction, which varies with the amount of coupling of the hearable with the ear. Like HRTF and HeMTF, HePTF depends on the user's anthropometric features, the source location and varies across frequency. The most accurate way to determine HePTF is to conduct an individualized in-situ measurement. However, conducting such a measurement is infeasible due to several factors. In addition to the hearable's physical construction effect, one of the major factors for in-situ variations is the idiosyncrasies in the hearable's placement on the user's ear. Therefore, past studies have tried to estimate the effect of hearable's placement using non-individualized measurements. Some previous studies have proposed lumped parameter models [26] for modeling hearable's physical construction and effects due to changes in hearable's placement, using measurements conducted on dummy heads. Recently, a study modeled the hearable's placement with the user's ear in-situ, using HeMTF and adaptive filtering [7]. Other challenges for in-situ HePTF measurement include difficulties in the placement of measurement microphones at the user's eardrum and the requirement of individualized HePTF measured at hundreds of different spatial positions for each user. Future studies could estimate HePTF by combining the knowledge about hearable placement and physical construction with individualized HeMTF for each source position.

B. Acoustic scene analysis

Acoustic scene analysis is the key to creating various listening experiences by analyzing the dynamic user acoustic environment captured by microphones in real-time. For example, a person walking on a busy street with an overhead train passing by. The user may want to reduce the train noise and stay aware of the nearby traffic. In this case, we

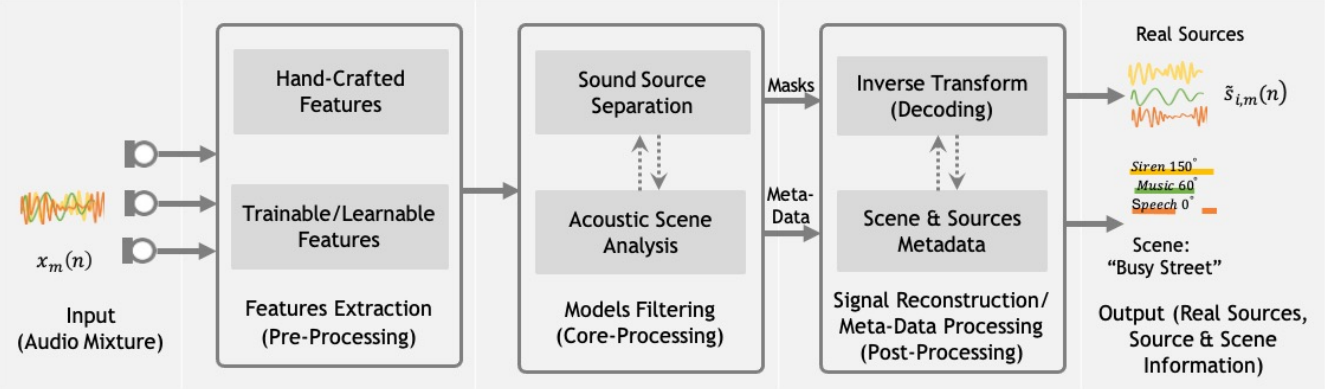

Fig. 6. An example overview architecture of acoustic scene analysis block, which takes audio mixture captured from hearables microphones as input, followed by feature-extraction, processes the extracted features to compute source masks or analyze the acoustic scene, and performs post-processing to reconstruct separated source signals and extract associated meta information like source type, spatial direction, and scene information as outputs.

first need to identify the acoustic scene with two major sound events: train and traffic sounds. Event detection must be accompanied by separation and localization of the sound events, as the separated sound source and their spatial information enable individual manipulation of each real sound source. Therefore, the core functionalities of sensing include sound source separation, sound source localization, sound events classification, and acoustic scene identification. Other sensing functionalities, such as scene captioning [27], speech recognition [28], and language translation [29], are important but beyond the scope of this article.

Fig. 6 shows an example overview architecture for acoustic scene analysis. The primary input is the audio mixture of various real sounds traversing through the user's acoustic environment and captured by the hearable's microphones. Multi-channel audio mixture can be modelled by mixture signal $x_m(n)$ as sum of multiple individual sound source images, $\tilde{s}_{k,m}(n)$ corresponding to source k captured at microphone m, and diffuse/background noise sound $v_m(n)$:

$$x_m(n) = \sum_k \tilde{s}_{k,m}(n) + v_m(n), \quad \text{where} \quad \tilde{s}_{k,m}(n) = s_k(n) * HeMIR_m(\theta_k, \phi_k, r_k, n). \tag{4}$$

Source image $\tilde{s}_{k,m}(n)$ is represented as a convolution of source signal $s_k(n)$ and Hearable Microphone Impulse Response (time-domain representation of HeMTF) $HeMIR_m$ between source k with 3D spatial position (azimuth: θ_k, elevation: ϕ_k, distance: r_k) relative to the center of head and microphone m. Captured time-domain input signal $x_m(n)$ is then passed through a pre-processing block to extract meaningful features. These features include Time-Frequency domain features (Short Time Fourier Transform, or STFT), magnitude spectrograms characterizing human listening (Mel filtering, cochleagram), phase-related features, like phase spectrogram, interaural features (e.g., ITD, ILD) for binaural signals, Time Difference Of Arrival (TDOA) for microphone array, time-domain features like time-segments, energy, and entropy. Features can also be implicitly learned using a supervised learning process based on acoustic data and ground truth labels. Extracted features are fed to the model's filtering block, which performs acoustic scene analysis using specific models for generating individual sound source masks for source separation and meta-data from acoustic scenes and events. Acoustic scene analysis and source separation models may also interact with each other or can be

jointly processed to ensure both individual sources and meta-data are matched. A post-processing block is applied to reconstruct the individual sources. The meta-data is processed to detect and classify events along with their spatial position matched with separated source signals and scene information.

Some of the key requirements for acoustic scene analysis are robustness under diverse acoustic events, dynamically changing reverberant and noisy environment; and low latency to allow real sound control and virtual sound rendering blocks to process the sensing information in real-time. In the past, Digital Signal Processing (DSP) methods and traditional Machine Learning (ML) methods, such as regression, and Support Vector Machines (SVM), have been used in acoustic scene analysis. The key difference between DSP and traditional ML methods is that DSP methods are usually based on explicit hand-crafted signal models. In contrast, traditional ML methods learn an implicit model from training data. DSP-based models usually make strong assumptions about signals (source characteristics) and environmental characteristics, which often perform sub-optimally when exposed to the real environment. While traditional ML models show good performance in several applications, domain expertise is still required to define features and smaller network components. Thus, traditional ML can result in shallow learning, which suffers from sub-optimal performance when evaluated in complex and challenging acoustic environment conditions. Lately, deep learning (DL) methods have been extensively employed in audio [30], [31].

Deep learning is represented by a Deep Neural Network (DNN), which consists of multiple hidden layers (two or more) and input and output layers arranged with linkages of multiple neurons [32]. They provide enhanced capabilities to learn distinct features and temporal variations within the acoustic scene, resulting in robust performance compared to conventional methods. Several complex DNN networks are constructed by stacking one or more layers of building blocks, such as Fully-Connected (FC), convolutional (Conv) layer, recurrent layer (including their variants like dilated convolutional, residual, and attention). An FC-based network is agnostic to input data structure and applies to a broad range of audio applications. A Conv layer is more effective than an FC layer. It is superior in exploiting local spatial features in the input data by connecting each neuron only to neighboring ones, particularly in audio spectrograms or a time series. Dilated convolutional layers have gaps, which skips neurons depending on the dilation factor and can help extract global features in the data without increasing the computations. The recurrent layer is another example of DNN building blocks, where neurons connect in a feed-forward manner, store past information in their memory, and add to current data. Recurrent Neural Network (RNN) stacks multiple recurrent layers, instrumental in exploiting the temporal sequence of input data. A Convolutional Neural Network (CNN) employs stacked Conv layers to learn relevant time-frequency feature representations automatically. For more details about DL for audio, the reader should refer to [33]. However, note that DL models are as good as the training data that derived the models, and performance efficacy naturally decreases when there are many sound events, reverberant and noisy environments. It becomes increasingly hard to train a neural network with a limited amount of data, high-dimensional outputs, and many hidden layers. Various techniques have been employed to improve the model's performance, such as data augmentation methods (e.g., multi-condition training and mix-up) and normalization methods to make the models more generalizable over different un-seen test scenarios. With limited training data, DL models might not perform well, and traditional ML or

DSP-based methods are still preferred. Past studies have proposed various DSP, traditional ML, and DL methods for sensing and some relevant techniques are explained briefly in the following sections.

1) Sound source separation: In the context of hearables, source separation is formulated as solving (4) and find all individual source images (which are filtered versions of sound sources at hearables microphones) for left and right ears. Time-Frequency (TF) masking is a popular DSP technique that converts the mixture into TF domain $\mathbf{X}_i$ (where i representing either left or right ear) as a feature. The STFT representation is multiplied by a TF mask $\mathbf{M}$ for each TF bin in the element-wise manner denoted by $\odot$ to derive the estimated source image $\hat{\hat{\mathbf{S}}}_{k,i}$ spectrum as:

$$\hat{\hat{\mathbf{S}}}_{k,i} = \mathbf{X}_i \odot \mathbf{M}_k. \tag{5}$$

Finally, application of the inverse transform recovers the individual source signals. This method aims to estimate the TF mask, a weighting function with a value between 0 to 1 for each TF bin. TF mask is estimated using a set of rules based on temporal, spectral, and spatial features of the clean sources. For example, TF masks can either be an ideal binary mask based on a hard threshold using Signal-to-Noise Ratio (SNR) or a soft mask using Wiener filtering. Some of the commonly-used assumptions made in TF masking methods are sparsity and disjointedness of sources in the mixture. Beamforming is another popular approach used for the enhancement of desired sources or sources at desired directions in several audio devices, including hearables. Beamforming uses spatial filtering to extract the desired signal and reject the interfering signals according to their spatial locations. Minimum Variance Distortionless Response (MVDR) beamformer is a widely used beamforming technique, especially for speech enhancement. For hearables, binaural speech enhancement techniques, such as binaural MVDR, can be used to preserve the spatial auditory cues (such as ITD and ILD) of the acoustic scene [34]. The majority of these traditional techniques suffers from performance degradation to separate sound sources from mixtures in challenging reverberant and noisy environment either because of estimated TF mask being not representative of real sources or shallow models used in some ML models. Besides, most of the earlier research focused on speech (i.e., speech enhancement) and/or music separation with techniques exploiting specific temporal and/or spectral characteristics of the speech/music mixtures. However, the real sound mixtures in AR/MR audio also comprise temporally and spatially varying sounds, which makes the sound source separation task more challenging in real environment.

DNNs have been applied to source separation problems and yield significant performance improvements, especially in the reverberant environment compared to traditional DSP approaches [35]. Recently, Conv-TasNet was proposed as an end-to-end low latency time-domain DNN, which is now used as a baseline model in many single-channel source separations works [30]. Unlike previous DNN approaches, Conv-TasNet can run in real-time with a frame size of 2ms and a much smaller model size, making it highly suitable for hearables, especially for speech enhancement. Conv-TasNet employed trainable encoder-decoder (pre-processing and post-processing blocks in Fig. 6) and mask network (core-processing in Fig. 6). The mask network comprises stacked dilated convolutional layers with exponentially increasing dilation factors.

A real-time binaural speech separation technique aimed towards hearables has been proposed recently using Multiple-

Input-Multiple-Output (MIMO) TasNet [36]. TasNet concatenated with binaural features derived from Interaural Phase Difference (phase difference between two ears), and ILD led to improved signal separation performance and reduced ITD/ILD error. However, the separation performance of TasNet and other DNN based sound separation algorithms varies with the type of dataset used to create the sound mixtures, reverberation, and background noise. Thus, real-world source separation algorithms for hearables are a topic of ongoing research [30], [36].

2) Sound Source Localization: In the context of hearables with AR/MR audio, as both sound sources and user move dynamically, sound source localization aims to estimate current spatial positions (referred to as spatial status) with direction (θ_k, ϕ_k) and distance d_k corresponding to individual primary sources $s_k(n)$. Sound source localization methods can be categorized based on a) microphone array geometry- planar/non-planar, and b) input features. It is imperative that non-planar microphone array configurations with at least four microphones would be required to localize the sound sources in 3D. Given the small form-factor of today's hearables, it is a critical practical challenge to incorporate non-planar microphone configurations in hearables for 3D source localization. Although most early research focused on 1D/2D source localization using a planar array (including circular, linear, or binaural), few past studies have tried to solve the 3D source localization using spherical microphone array or other non-planar configurations.

Another important factor in sound source localization is the input features, which must include the spatial characteristics of the sound sources captured by the microphones array. Spatial characteristics can be derived from the relative phase and magnitude information as they are strongly associated with the 3D spatial position of sound sources. Standard input features for sound source localization include TDOAs, cross-correlation audio features, phase transform, log-Mel spectrogram, intensity vectors, eigenvectors, and raw temporal features. TDOA based sound source localization is the most widely used DSP method, which estimates TDOA $\hat{T}_{m1,m2}$ between any pair of microphone array ($m1$,$m2$) and is usually computed using Generalized Cross-Correlation with Phase Transform (GCC-PHAT) $C(\tau)$:

$$\hat{T}_{m1,m2} = \arg \max_{\tau}, \quad \text{where}\, C(\tau) = F^{-1}\left[\frac{X_{m1}(f)X^*_{m2}(f)}{|X_{m1}(f)X^*_{m2}(f)|}\right]. \tag{6}$$

Once the pairwise TDOA estimate is computed, the source position in 3D space is estimated by minimizing the error between true and estimated TDOAs for all discrete spatial positions in space. True TDOA for any given pair of microphones is computed as the difference of arrival time of sound waves reaching two microphones from a given source position. There are many methods to minimize this error, and some of the standard methods include least-squares and maximum-likelihood. Most of the traditional methods, including TDOA, suffer from performance degradation due to strong reverberation and ambient sounds corrupting the peak of cross-correlation-related features. Recently, reverberation robustness of source localization is shown to be improved by using a pre-processing step to select direct-path TF bins using methods like direct-path dominance test, space-domain based selection [37], [38]. Tourbabin et al. [39] further improved the robustness by using a soft-binary mask with weighted schemes to apply to all the TF bins instead of binary selection and thereby, reducing the localization error by 4 to 6°. In the last few years, source localization using DL has demonstrated high localization accuracy in 3D space for multiple sound sources overlapping in time [31]. Unlike DSP methods, where hand-crafted features like GCC-PHAT are susceptible to environment characteristics

in the presence of other interference signals, most of the DL methods mainly use generic features like spectrogram. One of the baseline DL methods in [31] employed both magnitude and phase spectrogram of audio data captured from a four-channel microphone array arranged in the tetrahedral shaped configuration. It employed a Convolutional Recurrent Neural Network (CRNN), which consists of stacked CNN layers followed by RNN layers, to extract and refine the features to achieve better localization performance. CRNN based DL model outperformed the DSP method, especially in reverberant conditions with localization error reduced by more than 25°. Motivated by this work, various sophisticated models have emerged, significantly reducing localization error within 5° of true spatial positions evaluated on a reverberant dataset with multiple overlapping sources [40]. However, a key practical challenge remains to optimize the model to fit into the small form-factors of hearables.

3) Acoustic scene and event classification: Users could move from one environment to another in real-life scenarios, with various sound events happening around them. Therefore, dynamic characterization of user scene consists of two main tasks, 1) acoustic scene (or environment) classification, i.e., identifying what environment the user is in, and 2) sound events detection & classification, i.e., recognizing the type of sound events in the scene along with temporal start (onset) and end (offset) times. Both are of particular importance in AR/MR audio for hearables to effectively control the real sound and render the virtual sound. Thus, the sensing block must detect and classify the scene from the mixture audio $x_m(n)$ along with primary sound sources $s_k(n)$. The critical challenges in acoustic event scene and event classification in real-world scenarios are the vast numbers of sound events, as in urban sounds and indoor environment sounds, diverse inter/intra variation among sound events, model complexity, and real-time processing constraints, especially for hearables.

Büchler et al. [41] studied several traditional ML methods, like Markov models, Bayes classifier, 2-layer perceptron network for environmental sound classification in hearing aids. Lately, different state-of-the-art DL models have been introduced for acoustic scene and events classification, largely thanks to the Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge and associated workshop, which has resulted in many state-of-the-art works and open audio datasets in this field as highlighted in [42]. Hershey et al. [43] evaluated various CNN-based architectures for large-scale acoustic event classification evaluated on AudioSet dataset with ResNet model outperforming all the models. One of the baseline models [44] in acoustic scene classification uses log Mel spectrogram as input feature with a simple model combining two CNN layers and one FC layer. In another work, motivated by the fact that sound events are key constituents of an acoustic scene, multi-task learning has been utilized to jointly solve the acoustic scene and events classification problem with shared learned representations [42]. Recently, the DCASE challenge on low complexity acoustic scene classification has led to various optimized DNN networks with a dual focus on low model computational complexity and efficacy, making them suitable for deployment in embedded devices, like hearables.

4) Integrated and multi-modal sensing: Integrated sensing refers to integrating multiple sensing functions into one to exploit correlations between sensing outputs. In contrast, multi-modal sensing refers to the fusion of different sensing modalities like audio-visual and jointly processing multi-modal inputs providing stronger discriminative power. Most of the sensing problems are traditionally treated as standalone and single modality problems, resulting in high computation

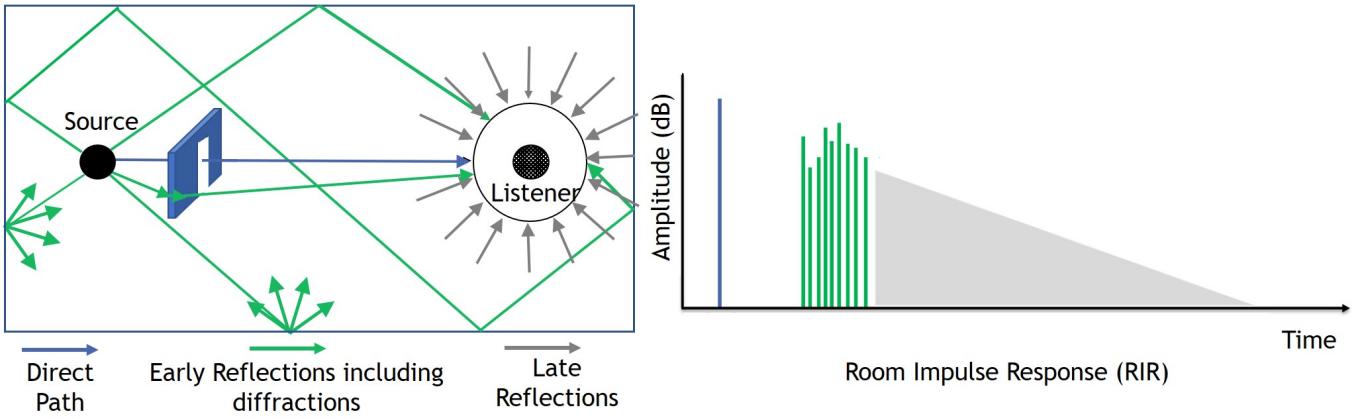

Fig. 7. Illustration of construction of typical room impulse response in an environment from source to listener via direct path (blue), early reflections (green), and late reflections (gray)

load and sub-optimal performance.

As some of these models share similar input features and model attributes, they can be integrated to reduce computational complexity significantly. The integration can either jointly train a single DL model on the same audio recordings with combined ground truth labels to simultaneously predict multiple sensing outputs or use cascade models. For example, sound event detection and localization (SELD) can be jointly solved to tell when, where, and type of the active sound events [31]. Sound source separation can utilize sound localization cues to enhance separation quality. Separated sound sources can also be directly associated with their classification and localization using the joint model of SELD. In recent years, there has been focused attention to jointly process the audio-visual information together for sensing features like scene understanding, events classification, source separation, and scene captioning. Other type of sensing modality can also be combined depending on the user scenarios; for example, GPS, accelerometer data, and acoustic scene analysis can give more accurate predictions about user environment and activity.

C. Environment reverb estimation

Reverb significantly impacts human hearing, especially indoors, since it helps us identify different enclosed environments and the sound sources' distance. Estimating reverb is particularly useful for AR/MR audio to render virtual sound sources in an environment (either real or virtual) and alter real acoustic environment reverb as per the user's desired experience. RIR that characterizes the environment reverb consists of three components, namely, 1) direct path, 2) early reflections, and 3) late reflections or reverberations. Direct paths account for time delays corresponding to the direct line of sight between source and receiver. Early reflections account for the first few reflections from the surfaces, diffractions, and scattering. Diffraction of sound is a phenomenon when a direct sound path is obstructed by wall and bends around the wall before reaching listener. Scattering of sound is a phenomenon when multiple reflections originate due to an incident sound wave hitting a rough surface. Late reflections are a set of multiple dense reflections reaching listener much later than the direct path. Early and late reflections are especially crucial for indoor or enclosed environments as they strongly influence the acoustic environment's overall perception. However, for outdoor environments, which lack enclosed walls, diffraction is of greater significance due to sounds propagating

through various physical obstructions. Fig. 7 illustrates an RIR constructed from various reflections in an environment. In addition to RIR, reverb attributes and geometric 3D mesh can also characterize the environment reverb. In general, the space, location, and reflection coefficients of physical objects affect the reverb. The following subsections describe reverb estimation in terms of RIR, reverb attributes, and geometric 3D mesh, directly and indirectly, using a combined approach.

1) RIR estimation: Acoustic recordings from one microphone or microphone arrays can be used to estimate the RIR of an environment directly. For offline and accurate estimation of RIR, recordings can be done in advance in various environments using different types of microphone arrays. In contrast, recordings from individual microphones in hearables need to be used for in-situ estimation of RIR [45]. RIR can be estimated from spectrograms of the audio recordings from individual microphones using standard DSP methods like dereverberation. Typically, for accurate identification of RIR of any environment, excitation signals with all the frequencies in the audible range from 20 Hz to 20 kHz, like maximum-length sequence (MLS) or sine sweep, are used. However, such excitation signals may not be practically feasible for the in-situ acquisition of RIRs. The advantage of using microphone array-based RIR estimation is to capture information of the sound field's spatio-temporal structure encoded in an RIR database. A spherical array is one such array geometry that is used to capture directional RIRs. A recent work [46] based on the phase-aligned transform of spatial correlation matrix showed remarkable improvements as compared to traditional algorithms, like MUSIC in DOAs and delays estimation of the early reflections from acoustic signals simulated with a spherical microphone array. This method can be further extended to estimate reflections magnitudes, and thus, a full RIR can be derived. However, spherical microphone array-based method is limited to the microphone array's construction constraints. The number and configuration of microphones used on the array determine the overall spatial resolution for sampling the sound field. Due to its large size, a spherical microphone array cannot be easily embedded into the hearables. However, it can be used for offline measurement and estimation of RIRs in a real acoustic environment.

2) Reverb attributes extraction: Key reverb attributes, such as reverberation time (e.g., RT60), room volume, and DRR can characterize an environment's reverb. Reverberation attributes can be estimated from microphone (one or more) recordings using DSP and DL methods. RT60, for example, can be obtained by first locating a sound segment using short-time energies and inter-aural coherences and then applying line fitting on energy envelope followed by statistical analysis. RT60 can also be estimated as a function of frequency by modeling an energy decay relief (EDR) [47]. Room volume can be predicted by extracting and feeding acoustic features and into a GMM model. DRR is typically computed from microphone array recordings by estimating the power of direct and reverberant sound using a spatial correlation/coherence model. For example, in a recent work [48], DRR was blindly estimated using magnitude-square coherence between binaural microphone signals and fitting a beta distribution. Alternatively, reverberation parameters like room volume, RT60 can also be estimated using CNN-based networks by training speech recordings with known reverberation attributes [49], [50]. Speech is the most commonly used signal for reverberation attributes estimation, although most non-transient signals can be used.

3) Geometric 3D mesh acquisition: Unlike the approaches mentioned above, this approach extracts the visual environment using depth or stereo cameras to construct a geometric 3D mesh of the environment. The detailed 3D mesh can be used to compute accurate RIRs for all source-listener positions in such an environment. However, it is challenging to capture a perfect 3D mesh of a real environment, mainly using HMDs cameras. Thus, errors such as holes in 3D mesh must be repaired using hole-filling algorithms [51]. Furthermore, perceptually insignificant details in the 3D mesh should also be simplified to reduce the complexity of computation as shown in [52].

4) Indirect reverb estimation using combination of methods: Reverberation attributes can be indirectly estimated from an environment using already estimated RIR. For the basic estimation of RT60, first, an energy decay curve (EDC) is derived from RIR as the remaining signal energy at any time, and then, RT60 is estimated as the time when EDC crosses -60 dB. EDC can be further generalized to the time-frequency domain using linear curve fitting, which is then used to derive frequency-dependent RT60. Estimating room volume from RIR is not straightforward, and usually, detailed feature analysis is required. In one illustration, room dimensions are obtained by training a GMM model with feature sets and using maximum-likelihood criterion [53]. Alternatively, room dimensions can be accurately derived from the 3D mesh of the environment if available. On the other hand, RIRs can also be indirectly estimated from reverb attributes or 3D mesh. As shown by Jot et al. [54], frequency-dependent RT60 and room volume (referred as *reverberation fingerprint*) were used to adjust the EDR and initial power offset of a pre-measured reference RIR resulting in approximated real environment RIR. Acquired 3D mesh can be used to either generate dynamic RIRs between source and listener position in real-time or can be used to generate RIRs with a pre-defined grid of 3D spatial positions in the scanned environment. Computation acoustics methods, like wave-based methods, ray-tracing, image-source methods, make use of the 3D mesh along with surface materials knowledge to simulate RIR between any source and listener positions in an environment. Similar to spherical array-based directional RIRs, computational acoustics can also be used to determine individual directions for each sound wave reaching the listener's position along with RIR. Furthermore, user scene information predicted from real acoustic scene analysis coupled with estimated reverberation attributes can help find the closest RIR dataset matched with the user environment. DNN models can identify the user environment from acoustic recordings by classifying it into one of the pre-determined environment types, as discussed in scene classification. A large cluster of databases can be maintained to store mappings of acoustic environments categorically in the form of pre-determined RIRs collated from various databases with known environment types, conditions (e.g., empty/partially filled/filled, presence of other types of objects), and reverberation attributes.

D. User activity and state sensing

User activity and state sensing are important for hearables to customize audio experiences according to the user's movement and biological states. For instance, head tracking provides dynamic binaural cues for virtual sound rendering. Once the listener's head position is known, it is possible to make sure that the rendered sound location does not shift with the head's movement. Typically, head tracking is done using inertial measurement unit (IMU) sensors, along with accelerometer and gyro meters [55]. Besides, translational movements must also be tracked to provide dynamic

distance cues for rendering AR audio. Thus, tracking the translations and rotational head movements is referred to as 6 Degree of Freedom (6DoF) user tracking. Today, some headsets use infrared LED constellations, external cameras, and Simultaneous Location And Mapping (SLAM) for translational tracking. Besides, the user's physiological parameters such as heart rate, oxygen saturation, and body temperature can also be tracked by hearables equipped with biosensing hardware such as LEDs, photodetectors, and thermocouples [56]. The physiological data can be used to estimate user's emotional states and subjective preferences, which can, in turn, be harvested to personalize the most preferred AR/MR audio experiences [2]. The user's preferences about the desired experience are also captured here through direct input or hand gestures. These preferences inform the inputs to real sound control and virtual sound rendering in terms of targets for real and virtual sources.

IV. Real sound control: modification of real acoustic environment

The goal of real sound control (or control for short) in AR/MR audio is to transform the existing real sound into a specific auditory experience to match the user's explicit/implicit preference derived from sensing block. For real sound, humans are used to a natural experience of listening with open ears. In AR/MR processing of real sound, we consider the real sound's desired auditory experience to modify the real sound for the open ear case. Though many non-linear modifications (such as dynamic range compression, time-scale modification) can be applied and play an important role in AR/MR audio experience, this section focuses on more basic and common linear effects. Such linear modifications range from attenuation, amplification, and more generally, frequency-dependent equalization (EQ) of the real sound level to suit different listening modes. These linear modifications can shift the sound sources to other locations for a preferred spatial experience and modify the ambient sound environment's existing reverberation to make it sound like another environment. These linear control processes can generally be achieved using a filtering process known as the control filter.

A. Real sound control framework

As explained in Section III, the real sound field could consist of multiple sound sources and ambient sounds for both left and right ears. Techniques such as beamforming and/or source separation make it possible to control each sound source independently and for either ear independently. This framework with independent control of multiple sources is illustrated in Fig. 8. Based on this framework, the following discussion will be based on a single source and can be extended to any number of sources. Desired real sound experience, such as frequency-dependent modification of sound, can be described in the form of several control targets, which are used to derive the respective control filters for both ears. As shown in Fig. 8, the process of deriving the control filter consists of four steps: 1) spatial control filter derivation; 2) reverb control filter derivation; 3) desired filter derivation with inputs from the previous two steps and the EQ target; and 4) the control filters calculation for both ears. These steps are discussed in more detail in the following subsections, which aims to provide a simplified understanding of the real sound control problem and could differ in the actual implementation of the control filters.

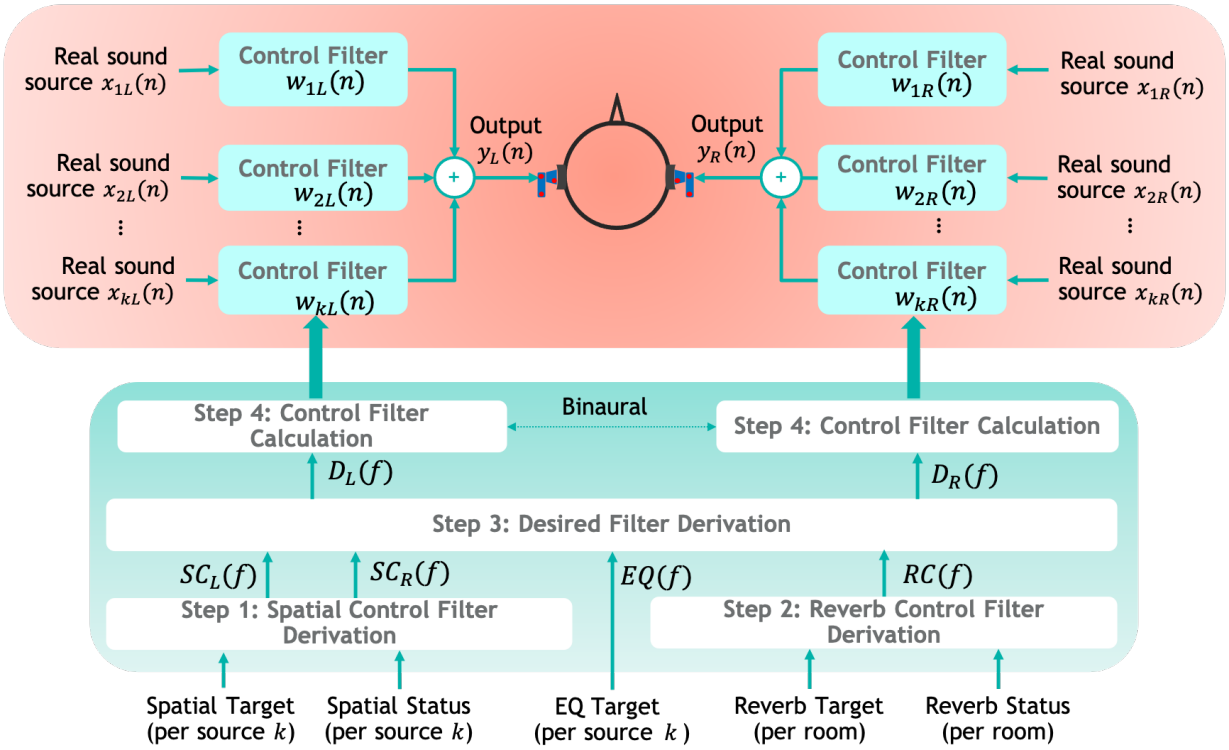

Fig. 8. Real sound control of multiple sources with derived control filters (shown inside orange box). A control filter is derived for each channel (left/right), and each real sound source. The four steps for computation of control filters are shown (inside green box).

The spatial status (current spatial position of the source with respect to the listener) and spatial target (desired spatial position of the sound source), which are obtained from sensing block, are used to derive the left and right spatial control (SC_i) filter for a sound source to ensure the desired response in the left and right ear is spatially coherent. The spatial attributes used for spatial target include the cues associated with direction and distance [16]. These spatial attributes of the current status can be estimated in Section III. To render the sound source at a particular spatial location, we need to consider the near or far-field HRTF of the target location, HeMTF of the current location at the external and/or internal microphones, and RTFs of both the current and desired locations in the same room. Conceptually, the spatial control filter (in the frequency domain) can be derived as:

$$SC_i(f) = \frac{HRTF_{i,target\ location}(f) RTF_{target\ location}(f)}{HeMTF_{i,current\ location}(f) RTF_{current\ location}(f)}. \tag{7}$$

All real sound environments have specific reverberation characteristics, typically represented by RTF, obtained through techniques described in Section III. Note that this reverberation is already recorded in the microphone signals, which needs to be taken care of. Reverberation control aims to alter the perception of the acoustic sound environment characteristics, for example, make a less reverberant room sound more reverberant, and vice versa. Altering reverberation could affect the distance perception discussed above and other perceptual attributes of the sound sources, such as source width. Reverberation control can be categorized into two basic types. One is to add artificial reverberation, which introduces early reflections or late reverberation or both [57]. The other type requires

reducing or removing reverberation, also known as dereverberation, which is a challenging problem but has attracted significant interest in audio and speech processing [58]. However, the goal of dereverberation for speech processing is usually to optimize speech intelligibility, where reverberation control of real sound concerns the overall sound quality. By reducing the current reverberation and adding target reverberation, a generic new reverberation can be created. In the frequency domain, the reverberation control filter can be derived as

$$RC(f) = \frac{RTF_{target\,room}(f)}{RTF_{current\,room}(f)}. \tag{8}$$

In addition to the spatial and reverberation control, the sound source's loudness level, and timbre can also be controlled using an EQ target. Three common listening modes are being identified in today's hearables, as shown in Fig. 9. When the hearables are worn on the head and no active processing enabled, known as passive mode (center of Fig. 9), the physical structure of the hearables introduces a low pass filtering effect to the real sound entering the ear (as shown in HePTF response in Fig. 5). The ANC mode (left of Fig. 9) is used to attenuate low-frequency real sound from getting into the hearable. Acoustic transparency (right of Fig. 9) aims to produce an ear response similar to the open ear case. Here, we introduce the EQ target to represent such frequency-dependent level adjustment for real sound with the open ear case as a reference. Generally speaking, three ranges of EQ target can be defined to cover and extend the above listening experiences. By setting the EQ target $EQ(f)$ less than 1, attenuation of the real sound is often realized using passive and ANC techniques. Acoustic transparency requires the desired ear signal to be the same or very close to open ear reference, with the EQ target set to 1. Furthermore, to create an amplified real sound experience, the EQ target is set larger than unity. In step three of the control filter derivation, a common EQ target $EQ(f)$ is applied to the spatial and reverberation control filters in the two ears to derive the combined desired filters as

$$D_i(f) = EQ(f)SC_i(f)RC(f) \tag{9}$$

B. Signal processing methods for control filter computation

The desired filter (from step 3 in Fig. 8) serves as the input for each ear's control filter calculation (step 4). The left and right ear control filter calculation can be separated or linked to ensure a coherent experience for both ears, which can be termed as "bilateral control" or "binaural control", respectively [34]. The signal flow of the control filter calculation of one ear is illustrated in Fig. 10. Given the input sound $x_k(n)$ that is typically measured from the internal/external microphones, the desired filter needs to be applied to yield the desired signal at the ear ear_d_k. However, when the hearable is placed on the human's ear, the device (for most *occluding-type* hearables) modifies the sound path from the sound source $x_k(n)$ to the eardrum, through a Hearable Passive Impulse Response (time-domain representation of HePTF, denoted by $HePIR_k(n)$), resulting in a leaked real signal that must be taken into account to yield the desired secondary path signal $ear\_d_{sk}(n)$. Real sound control is implemented by applying a control filter

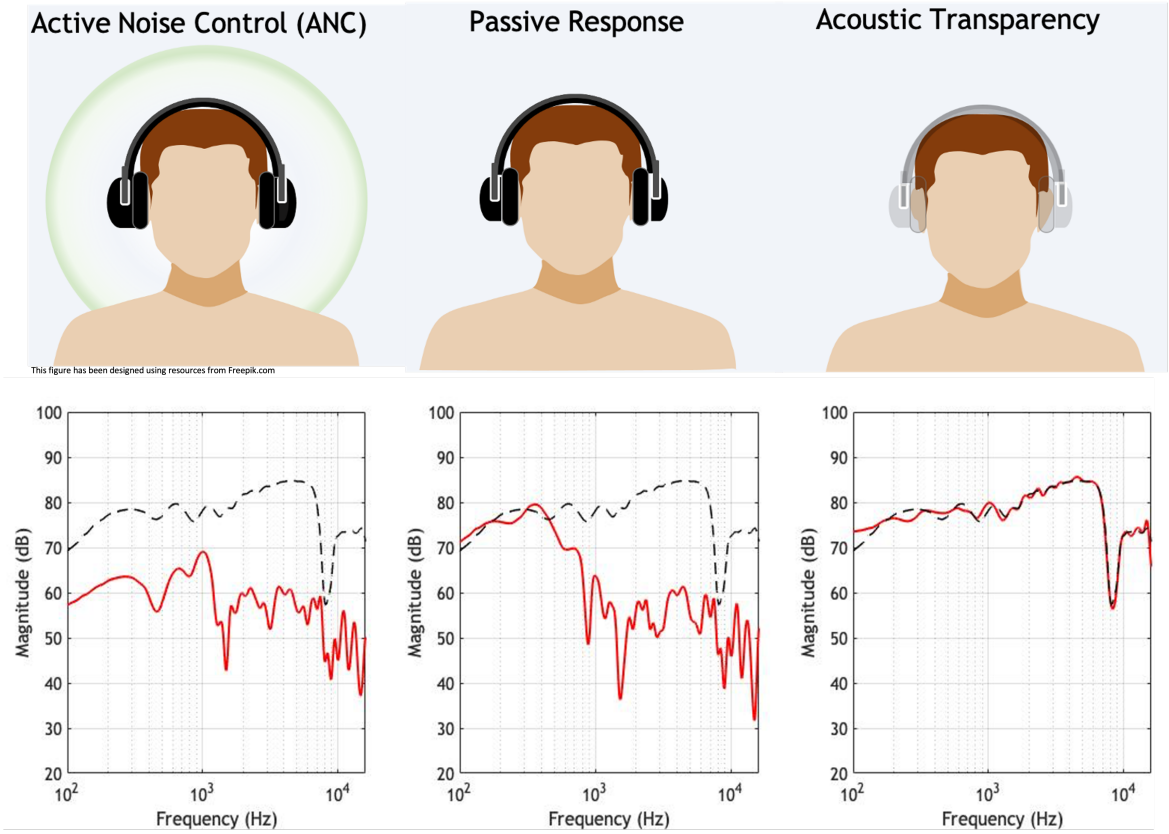

Fig. 9. Real sound control listening modes with examples of the responses at ear under these modes (solid red curve) in comparison to the open-ear response (dashed black curve).

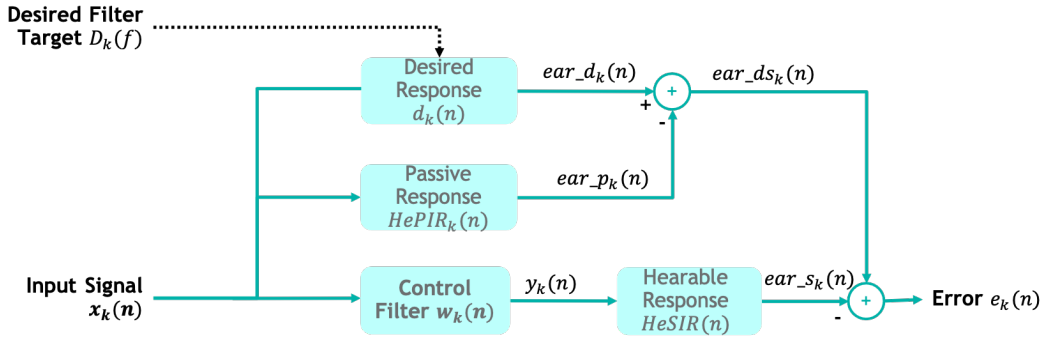

Fig. 10. Control filter calculation for a single source. The summation shown in figure is an acoustic summation.

$w_k(n)$ to the real sound, resulting in the hearable speaker output $y_k(n)$ that undertakes a secondary path of the Hearable Speaker Impulse Response (time-domain representation of HeSTF, denoted by $HeSIR(n)$) before reaching the eardrum as $ear\_s_k(n)$.

In certain cases, the speaker output could also be captured by the microphones, resulting in feedback signals. Such signals must be compensated using mechanical design and/or feedback neutralization algorithms [59] to realize desired control. For simplicity of derivation, we assume that this feedback effect has been compensated using these algorithms if needed. Based on the above consideration, the main goal of the real sound control can be stated to make the secondary ear signal $ear\_s_k(n)$ as close to the desired secondary ear signal $ear\_ds_k(n)$ as possible, which can be expressed by

minimizing the following error signal:

$$e_k(n) = ear\_ds_k(n) - ear\_s_k(n). \tag{10}$$

The above can also be expressed in the frequency domain with the error spectrum derived as

$$E_k(f) = (D_k(f) - HePTF_k(f) - W_k(f)HeSTF(f))X_k(f), \tag{11}$$

where $D_k(f)$, $HePTF_k(f)$, $W_k(f)$, $HeSTF(f)$, and $X_k(f)$ denotes the frequency response for desired, HePTF, control filter, HeSTF, and the spectrum of the real sound source, respectively. While this works well for bilateral control, additional considerations are needed for binaural control to ensure a more coherent experience. In this case, the derivation of the left and right control filters must be linked, where the performance of overall error and spatial error needs to be specified.

Next, we discuss a few computational methods for control filter calculation. The aim is to compute one control filter for each channel, which minimizes the norm of the error signal, as shown in (10) and (11). The computational methods have been broadly classified into three main categories; namely, manual methods, automatic methods, and hybrid methods [60].

Manual calculation of the filters occurs during the hearable system design phase rather than when users are using the hearables.

In general, the manual methods can be realized in three major steps. The first step is problem formulation for the control filter calculation, typically in the frequency domain as shown in (11) along with a criterion for determining the optimal solution (e.g., based on the upper bound of error spectrum of magnitude and/or phase). Next, the filter parameters are manually adjusted to yield the optimal solution. The filter parameters can be adjusted depending on the filter structure. Infinite Impulse Response (IIR) filters, especially biquad filters in the cascade or parallel form, have been commonly used in manual methods. In the last step, the control filter's objective performance is derived using the optimal solution criterion. If the criterion is satisfied, the filter coefficients are computed from the filter parameters and stored for real sound control. Otherwise, these steps need to be iterated until the criterion is met.

The second class of control filter calculation methods, where the optimal filter parameters are calculated automatically, using algorithms to minimize the error in (10) [7]. This method, referred to as the automatic method, contrasts with the manual methods, where human inputs are constantly engaged for calculating control filters. Possible approaches to solve this problem include, but are not limited to, using analytic approaches which yield closed-form solutions or iterative numerical methods. Ordinary Least Squares algorithm is one of the most popular analytic methods when the residuals (i.e., the sum of squared errors) have a linear relationship with each of the parameters in (10). Another class of automatic methods is based on a simplified gradient descent-based iterative approach, such as the least mean squares (LMS) algorithm. In LMS, the adaptive filter minimizes the instantaneous squared error in (10) using stochastic gradient descent. Many variants of the classical LMS algorithm, such as Filtered-X Least Mean Squares, Filtered reference leaky

LMS, and Warped LMS, have been successfully used in ANC, HT, and virtual sound applications [6], [7], [59]. These variants, along with the LMS algorithm, are popular due to their simplicity and good steady-state performance.

In practice, it is possible that the automatic methods may not yield optimal performance, and the manual methods may require a considerable amount of human effort. Previous studies have proposed hybrid methods that combine manual and automatic methods to circumvent these issues. In this approach, a candidate control filter is calculated using automatic methods, based on measurements in certain conditions (usually ideal and simplified, e.g., using dummy heads in anechoic conditions) [5]. These candidate control filters can subsequently be manually selected or adjusted (e.g., in certain frequency bands) to arrive at a more accurate control filter.

C. Real sound control in real-life scenarios

Real-life scenarios of AR/MR audio in hearables span a wide variety of uncertainties, including but not limited to the fitting of hearables (depends on anthropometry and hearable form factors), real acoustical environment, changing user preferences, wear and tear. Usually, the hearable manufacturers characterize the microphone's and speaker's acoustic response for a few common use-cases. However, the acoustic environment, such as the spectrum and direction of the sound source(s), changes rapidly in real-time, and the default fixed settings may not be ideal for all use-cases. As a result, such changes in real-life will alter the transfer functions used in calculating the control filter, making it necessary to tune or update the control filter in real-time.

1) Choosing the single optimal control filter: Fixed filter implementation for real sound control is most attractive in terms of computational complexity and battery savings for hearable devices. In this case, the most practical method is to design a single control filter that yields the best overall performance considering the possible cases and the relative importance of each case. This consideration of the possible cases refers to how the hearable device will be used, including anthropometric features representative of diverse human populations and associated acoustic coupling, acoustic environments, and desired user experiences.

2) Updating the control filter in real-time: While a fixed optimal filter yields an elegant solution, a better AR/MR performance requires the hearable to update the control filter in real-time. We consider control filter updates from the following four aspects: source directions, individualized responses, minimal error in real-time, and desired control mode.

As most hearable responses are directional dependent [61], the control filters should be directional dependent. Instead of using one single control filter, we can update the filter based on the source direction provided by AR/MR sensing block. Here, a direction resolution might be defined to balance the number of filters that need to be stored, and the rate of the update depends on the use case.

As explained in Section III, hearable responses vary significantly based on the user's anthropometry, resulting in both fixed and dynamic variations. For the fixed variations, an offline personalization process can be adopted to obtain each user's control filter. For instance, in-ear headphones have a blocked ear canal resonance effect (characterized by a shift in the peak typically at 2-4 kHz for the open ear to 5-10 kHz for blocked ear canal) that varies from person to

person due to the differences in the ear canal length. On the other hand, dynamic variations can occur in individual user responses depending on how the device is worn on the user's ear. These changes mainly include HeSTF response, which changes with the device's coupling and user's ear shape (anthropometric features), and shall be accomodated with corresponding filter updates [7]. The other type of personalization is the individual hearing profile [62], i.e., perceptual sensitivity to sound across frequency. A hearing test can be used to obtain the hearing profile, and the EQ target can be compensated correspondingly.

Instead of using above additional cues to guide us to adjust the control filter, we can apply an online adaptive filter to tune the control filter based on the error from the actual response and desired response as expressed in (10). Depending on the use case, this error might be obtained directly from the error microphone in the device (such as in ANC [59]) or by estimation or prediction of such error (such as in ANC and acoustic transparency [63]). For example, several virtual sensing techniques have been used in ANC [64] to predict and optimize the eardrum signal based on the error mic signal.

The above discussion was focused on updating the control filter to create the desired listening experience. In practice, the desired listening experience or control mode can be frequently updated based on the changing acoustic environment and users' needs. The desired control mode can be manually provided by the user using physical buttons on the device, through the connected app, direct voice command, or in the future, a neural input through brain-computer-interface. Alternatively, automatic mode control can also be enabled to update the control filter based on an intelligent sensing block that takes both the environment and users' needs into consideration. For example, when the hearables sense the varying acoustic environment from quiet to noisy and/or the user's biophysical states indicating the annoyance due to noise, we can automatically update the control mode from transparency to ANC.

D. Case Study: Active noise control

To attenuate the ambient sound/noise, ANC is the most commonly studied and applied feature of real sound control in today's hearables. Typically, an ANC systems sets the EQ target as close to zero as possible, as illustrated in Fig. 9, which implies that the secondary sound aims to mimic passive leakage sound in similar amplitude but with opposite phase so that their superposition yields reduced sound pressure inside the ear canal [59]. ANC is an almost century-old technique that has been widely used not only in hearables but also in larger spaces like car/aircraft cabins, homes, and factories. It has recently gained momentum in these applications due to the advent of low-cost, low-power, and high-performance processors and electro-acoustic components. ANC works better in lower frequency and smaller space due to the requirement of sound pressure matching. The general ANC performance, denoted as noise reduction β, can be expressed in terms of amplitude error ΔA and phase error $\Delta\theta$ between original noise and anti-noise:

$$\beta(dB) = 10\log_{10}[1 + (\Delta A)^2 - 2\Delta A\cos\Delta\theta] \qquad . \tag{12}$$

Based on (12), we can compute the ANC performance contour in terms of amplitude and phase error, as shown in Fig. 11. As illustrated, small phase differences could yield to large degradation in ANC performance. As a result, smaller

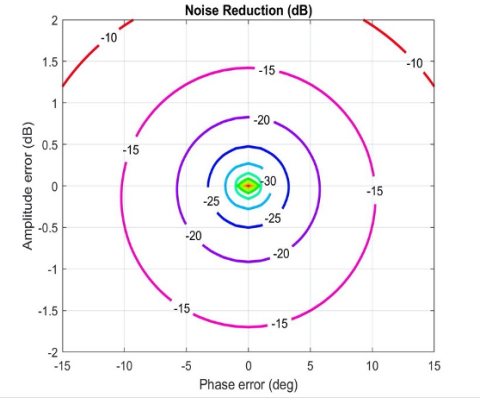

Fig. 11. ANC Performance as a function of amplitude error (in dB) and phase error (in deg)

phase differences require a super low latency system. For example, to achieve at least 20 dB theoretical noise reduction for frequency up to 1 kHz, an additional delay of about 20 microseconds or lower, equivalent to a phase difference of less than 7 degrees, is needed. For most hearables, the noise propagation time from the hearable microphone to the hearable speaker can be as low as about 30 microseconds, which indicates that the latency of the hearable processing system to generate anti-noise needs to be under about 50 microseconds. That is why in earlier days, ANC headphones were realized using an analog system. However, recent years have seen digital systems achieve similar low latency and enable more advanced digital algorithms.

Typical ANC hearables use two microphones at the hearable body's external and internal positions (i.e., feedforward and feedback microphone) to realize a hybrid ANC system for superior performance. As seen in Fig. 9, the noise reduction performance due to passive attenuation and ANC, which are effective at high frequency and low frequency, respectively, are shown. Most of the hearables with ANC employ fixed optimal filters and only require microphones on the same side of the earcup for efficient processing. In recent years, to further improve the ANC performance, we have seen extended studies and commercial adoption that involve adaptive algorithms [59] and using multiple microphones from both sides of the earcup [65].

E. Case Study: Hear-through for acoustic transparency

Acoustic transparency is a popular AR/MR audio experience under real sound control [1]. Most *occluding-type* hearables modify the sound reaching the ear due to the hearable earcup's physical presence. Acoustic transparency requires the passive response to be compensated so that the real sound can be perceived as indistinguishable from the open ear reference, as illustrated in Fig. 9. The techniques to achieve acoustic transparency are referred to as Hear-Through (HT) techniques, which often involve capture, processing, and playback of real sounds in the environment [1]. Typical HT techniques require at least one external microphone on each ear for the sensing block to capture the real sound signals and their metadata information (such as sound event class, direction, SPL), which are relayed to the real sound control block to derive the control filter for hear-through, with EQ target $EQ(f) \approx 1$ for all frequencies.

There are two main challenges to derive the HT filters. The first major challenge is minimizing the delays between

the HT output signal played back through hearables, the passive leaked real signal, and an open ear reference real signal (the desired target) [3], [5]. Excessive delay difference can cause comb-filtering effects and audio-visual mismatch, respectively. In the case of acoustic transparency, comb filtering occurs when the AR/MR audio signal played from the hearable gets added to the passive leaked real signal [5]. The difference in group delays between the two signals causes audible distortions. One possible solution for minimizing the comb filter effect is to reduce the processing delay for the playback of the HT output signal. Section VII-A on real-time implementation discusses possible strategies for minimizing processing latency for HT and other applications. Rämö and Välimäki [5] in their previous study concluded that the comb filtering effect did not depend on delay if the leaked real sound was attenuated by more than 20 dB as compared to the playback signal. A few past studies have proposed the integration of ANC in HT to achieve over 20 dB real sound attenuation, which can reduce comb effects [66], [67]. If there is an excessive delay between the open ear reference signal and HT output signal, the presence of visual stimuli in the user's field of view associated with the open ear reference signal can lead to an audio-visual mismatch. While the perceptual effects of temporal audio-visual mismatch has been investigated in some past studies [68], further investigation is required to determine the impact of delay on the audio-visual mismatch in the context of HT.

The second challenge is to derive the HT responses accurately. Ideally, we need the user-dependent responses such as HeMTFs, HRTFs, and HeSTFs measured in-situ for each individual, each source position, and each hearable fitting to derive the most accurate control filter for HT. However, as mentioned in Section III-A on user response estimation, this is infeasible in practice. Thus, other methods which can compensate for variations in the user-dependent responses have been proposed to design more accurate HT filters. One example is the methods that consider DOA as input (obtained from sensing block) [8] to derive directional HT. Estimation of real sound leakage in-situ proposed in past studies [7] can help to estimate variations in HePTF. The HePTF estimates can be used to adjust filter parameters to design more accurate HT filters. [5], [69].

V. Virtual sound rendering: synthesis of plausible spatial auditory illusions

Binaural rendering needs to mimic the acoustical cues humans use for spatial hearing to create an authentic perception of virtual sound, as explained in Section II-A. In AR/MR audio, the virtual sound information can come from three main formats: object-based, channel-based, and scene-based. Object-based audio consists of dry audio objects and metadata, such as the spatial positions that can be static or dynamic. Channel-based audio contains the number and positions of loudspeaker playback channels and the corresponding audio signal for each channel, such as 5.1, 7.1. The scene-based or ambisonic-based representations usually represent the sound field in terms of its spherical harmonic components, which can usually be decoded into channel-based representation. We consider each virtual sound source an audio object or audio channel with a desired spatial position. Fig. 12 illustrates the framework for virtual sound scene rendering using virtual sound sources as objects. Two main steps are generally required to generate a virtual sound scene with a single or multiple virtual sound sources. First, room auralization using a rendering filter for each source and each channel (left/right) is performed. Room auralization can typically be realized using the convolution of BRIR with the

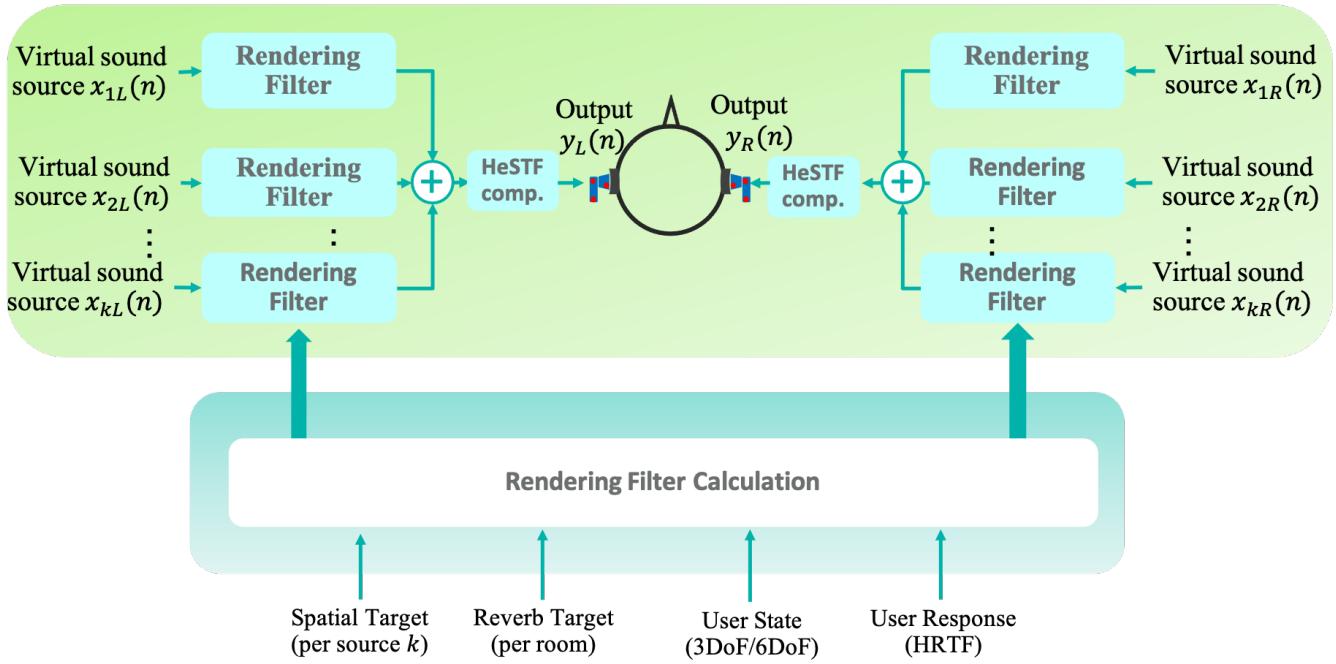

Fig. 12. Basic framework for virtual sound rendering in AR/MR audio. Rendering filter block performs room auralization using BRIR or delay networks, based on the inputs from Sensing.

sound source or using delay networks to model the room reflections. Second, as will be explained in Section V-B, to avoid undesired coloration due to HeSTF, hearable compensation must be performed using a compensation filter before playback of virtual sound signals through hearable speakers. The following subsections discuss room auralization and hearable compensation, followed by the fusion of virtual sounds with the real acoustic environment.

A. Room auralization using rendering filter

Room auralization aims to recreate the virtual sound sources' perception in a particular room with specific reverb characteristics [16]. The most common approach for room auralization is using convolution techniques. The rendering filter BRIRs need to be calculated based on the virtual sound information, room reverberation information, and user states. Static and dynamic binaural rendering are two notable scenarios for virtual sound rendering, where static rendering requires a fixed set of BRIRs per sound source (one for each left/right ear), and dynamic rendering must update the BRIR on the fly based on the source positions, user position, and orientation. BRIRs used in the auralization can be selected from a pre-computed BRIR dataset at the cost of memory or computed on the fly based on a small dataset and other necessary information.

BRIR represents the acoustic response between a sound source in the room and the eardrum(s) of the user. Therefore, BRIR depends on the source position, room, user position, and user anthropometry, represented as the spatial target, reverb target, user state, and user response, respectively, in Fig. 12. The most accurate way to obtain BRIR is by directly measuring impulse responses for the desired environment at the user's ear. However, this approach is infeasible for virtual sound rendering in hearables since the measurements must be conducted in the same acoustic environment with the desired reverberation properties for each user and many combinations of the positions, making it time-consuming

and tedious. Thus, a more practical approach can be used to calculate BRIRs using HRTF and directional RIRs, which can be obtained using sensing techniques. According to past studies on perceptual mixing time [70], listeners cannot discriminate between position-dependent differences above a certain threshold of reflections, for example, 50-150 ms for 3DoF user movements. Thus, RIR can be decomposed into directional (direct sound and early reflections), and diffuse portions [71]. For direct sound and each reflection, HRTFs with the corresponding directions can be convolved with the corresponding samples of the RIR.

These output responses and the diffuse portion of RIR (fixed for all positions) are summed to yield the final BRIR. Furthermore, a simple approximation can be employed by convolving HRTF corresponding to the direction of source relative to the listener with a fixed RIR for all positions in the same environment. Neidhardt et al. indicated that the adaptation of only the direct sound with corresponding HRTFs could be used to create convincing virtual sound illusions [72].

Another method for room auralization is based on artificial reverberation techniques, such as delay-based networks [57]. Delay-based networks recirculate the input virtual sound source signals through delay lines to simulate the effect of increasing arrival density of reflections at the listener's position over time. The gains and filters used in the recirculation path of the filters can be used to control the decay properties of the desired reverb. Two popular methods for delay-based networks are Feedback Delay Networks (FDN) and Scattering Delay Network (SDN). The FDN consists of several parallel delay lines interconnected through a feedback mixing matrix which defines the gains for each connection to generate the desired artificial reverberation [57]. SDN, on the other hand, consists of a network of delay lines connected via scattering junctions or nodes (typically one for each wall) to simulate the first-order reflections accurately and approximates higher-order reflections [73]. SDN uses parameters derived from room geometry models from sensing block to create the desired reverberation.

1) Static binaural rendering: One of the earliest approaches used for room auralization was aimed towards creating a static virtual sound scene [16]. For the static virtual scene, both the virtual sound sources and the user are assumed to be fixed. For the computation of rendering filters, in this case, there are two possible approaches. The first approach is to calculate the BRIR for each source and use this BRIR to simulate the direct path from sources, as well as the reflections and scattering of sound waves. The second approach is to generate the source at the required spatial location using HRTFs and use the reverb estimate (RIR or reverb attributes or room model) to simulate the reverberant characteristics. Another use-case for a static virtual sound scene includes headphone-based binaural reproduction of surround sound setups. In such cases, HRTF/BRIR based filtering can be used to create the impression of a virtual scene playback through loudspeakers in a room, and the room reverb can be added as desired.

2) Dynamic binaural rendering: In binaural listening, the source and/or listener movements alter the sound pressures at the user's eardrums, which must be taken into account to create an accurate and convincing dynamic binaural rendering for AR/MR audio. These user movements can be limited to rotational head movements in three degrees-of-freedom (i.e., 3DoF) or include both rotational and translational movements (i.e., 6DoF). These movements result in changes in the listener's relative direction, orientation, and distance to the sound source, the different walls, and

objects. This affects the overall level as well as the frequency-dependent interaural level and time differences in the direct sound. Likewise, the spatio-temporal structure of the early reflections will vary. To a certain extent, these changes can be perceived by the listener. For example, these perceptual changes due to head movement are utilized by our auditory system to help us locate the position of the sound source and avoid front-back confusions. With increasing reflection density, the differences between different listening positions are less and less audible. In a dynamic virtual sound scene, the rendered virtual sound adapts to the user's movements by updating the BRIRs for convolution. This update can be performed by selecting the BRIR from a pre-measured BRIR database. However, due to the memory limitation in practice, the BRIR database cannot store all possible combinations of orientations and positions. An alternative is to calculate the BRIR filters (or the RIR filters used to synthesize BRIR, as discussed in Sensing) in real-time, which usually incurs high computational cost. Thus, a common practice is to start with a moderate size of the BRIR database and apply interpolation and extrapolation. BRIR interpolation and extrapolation for different orientations and positions are still challenging due to the spatial and temporal complexity of the reverberant part of the BRIRs. A common strategy to facilitate efficient BRIR interpolation and extrapolation is to focus more on preserving the psychoacoustic relevant cues rather than aiming for precise physical accuracy. The method by Bruschi et al. [74] proposed to use a novel peak detection algorithm to match individual reflections, which preserves the structure of the early reflections in a subsequent linear interpolation stage. Another important consideration in dynamic rendering is the position resolution for the BRIR update. Efficient dynamic rendering can be achieved by updating BRIRs based on the required minimum grid resolution of the user's positional changes [72]. The minimum grid resolution can be derived based on perceptual thresholds such as minimum audible angle for head rotations and the just noticeable differences of distance changes induced by the user's head/body translation. Different positional resolutions can be employed for direct sound, early reflections, and the later part of the BRIR.

B. Hearable speaker response compensation

Accurate sound reproduction requires a particular desired hearable speaker response, HeSTF, making it necessary to compensate for the existing HeSTF. A common compensation approach is to design an equalization filter using estimated HeSTF, as shown in Section III-A. A simple inversion is usually infeasible as it could incur a strong amplification in the very low and very high frequencies. Schaerer et al. [75] evaluated seven methods for creating headphone compensation filters. The evaluated techniques include frequency domain pre-processing-based techniques such as octave band smoothing, and regularization-based methods, both in time and frequency domain. The evaluation metrics include high-frequency artifacts, timbre, and localization. The least-squares approach with regularization was one of the methods which showed superior performance and have been widely used in other studies. The selection of the regularization parameter is often critical to minimizing error in this scenario. A study by Bolaños et al. [76] proposed a method to estimate the regularization parameter by comparing the measured response before and after half-octave smoothing. This method showed better perceptual performance as compared to equalization using a fixed regularization parameter. Furthermore, most techniques used for calculating control filters (described in Section IV)

TABLE I
EXAMPLES FOR AR/MR AUDIO USE-CASES AND THE CORRESPONDING INTEGRATION OF SIGNAL PROCESSING TECHNIQUES. THE CHECK SIGN INDICATES THAT THE TECHNIQUE CAN POTENTIALLY BE USED TO REALIZE THE USE-CASE

Use-case examples	Sensing				Real sound control			Virtual sound rendering	
	User and device response estimation	Acoustic scene analysis	Environment reverb estimation	User activity and state sensing	EQ control (ANC/HT)	Spatial control	Reverb control	Static binaural rendering	Dynamic 6DoF rendering
Music listening in noisy environment	✓				✓			✓	
AR/MR gaming	✓		✓	✓	✓				✓
Teleconferencing with real and virtual talkers	✓	✓	✓	✓	✓	✓	✓		✓

can be adopted for hearable compensation.

C. Fusion of virtual sounds into the real acoustic environment: Perceptual aspects

In the case of AR/MR audio in the real environment, the virtual sound rendering output must be seamlessly fused with the existing real sounds. Perceptual interactions between the rendered virtual sounds and the real acoustic environment are important for providing plausible experiences to the user. Plausibility refers to an agreement of the AR/MR experience with the user's expectations from the everyday listening experience [45]. Plausibility for the fusion of virtual sounds with the existing real environment has two important aspects. These are, namely, the context provided by the existing real acoustic environment and content of the raw audio for virtual sounds. Audio and visual cues of the existing real environment such as cues for real sound source position, current environment geometries, and real physical objects like walls provide context to the user and form the listener's expectations [45]. The plausibility of rendered virtual sounds in the existing real environment may also be influenced by past real-world listening experiences of the user. Therefore, user expectations can be highly individualized, and ongoing research efforts are aimed towards exploring the important contextual parameters for achieving plausibility. Similarly, the role of content in the perception of virtual sounds in the real environment is also a subject of ongoing research. However, certain audio content carries explicit or implicit information about the origin of a sound source and thus, sets user expectations. For example, certain sound sources such as planes and birds are expected to be located above the head, while vehicles such as cars are expected to be close to the sagittal plane. Hence, while virtual sounds can be placed at any arbitrary position, in certain cases, the user expectations (e.g., based on audio content) can play a major role in the desired location of the sound source.

VI. INTEGRATION

Delivering a plausible and seamless AR/MR audio experience requires the integration of signal processing techniques from sensing, real sound control, and virtual sound rendering. Despite being non-exhaustive, a few use-case examples are highlighted in this section to explain the integration of the techniques in these three blocks, as summarized in Table I.

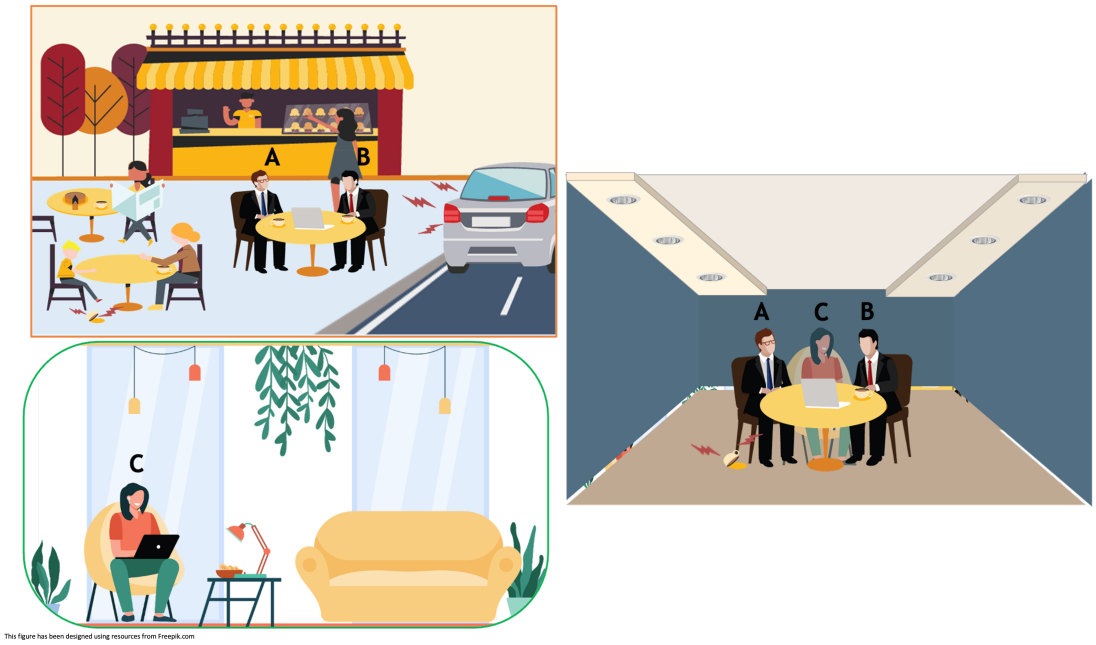

Fig. 13. Teleconferencing using AR/MR audio. The example shows a scenario where three collaborators (A and B in a busy cafeteria and C in a living room) are having a meeting, and how AR/MR audio techniques integrated in the hearables improve the presence and listening experience of the teleconferencing.

A. Listening to music in a noisy environment

In a noisy environment, users typically have to increase the volume or adjust the spectra of music dynamically [77] to mask out the noise, which poses a high risk to their hearing. The hearable can capture and reduce traffic noise with EQ control set to ANC mode to protect our hearing. Furthermore, static binaural rendering can be applied to create a more immersive music listening experience (to provide a feeling of "being there" at a concert, for instance, using the reverb of the concert hall).

B. AR/MR gaming

Creating an immersive AR/MR gaming experience requires a natural interaction between the gamer and the gaming scene. The dynamic processing of virtual sounds renders an interactive gaming scene by tracking the gamer's rotational and translational 6DoF movements. For immersive gaming in VR, where gamers only want to hear the virtual scenes, ANC can attenuate real sounds. On the other hand, for AR gaming, the acoustic environment's reverb must be estimated for the 6DoF dynamic rendering. At the same time, HT can be used for gamers to feel an unaltered presence in the original real environment.

C. AR/MR teleconferencing

Compared to face-to-face interaction, conventional teleconferencing systems usually lack the feeling of presence. Fig. 13 illustrates an example of AR/MR teleconferencing experience with a significantly enhanced feeling of presence. In this example, we explain the AR/MR teleconferencing experience for one of the participants (i.e., the host person A) who wears the hearables. Considering a collaborative meeting scenario, both the host A and his colleague B meet in

a nearby outdoor cafeteria close to a noisy street in Singapore. Simultaneously, a remote collaborator, C, joins them from her living room in Germany. For the meeting to be productive, the host A needs to hear colleague B's voice clearly and perceive collaborator C's voice as if collaborator C was in the same environment. The hearables can make this happen by integrating the techniques described in this paper.

Ambient noise in the cafeteria environment makes conversation much more challenging. The hearables can be configured to a speech enhancement mode to facilitate the conversation with the intended groups. This mode allows the user to listen to the desired voice/speech by setting the EQ control to HT, suppress ambient noise such as traffic noise and voice from other diners using ANC, and reduce the reverberation for better speech intelligibility. The separate voice output is provided by acoustic scene analysis, identifying the desired voice's spatial location and activating the voice event. Besides, the environment's reverb characteristics such as RIR need to be estimated to perform dereverberation of speech signals. Being outdoors also means the host A might need to attend to certain alert sounds such as a coffee cup dropping on the floor or a fire alarm. The hearables can activate the detection, localization, and extraction of the alert sound events and apply HT for host A to be aware of them. Reverberation control techniques can modify the cafeteria's reverb characteristics and even teleport host A and colleague B's voice to a meeting room. On the other hand, collaborator C's voice, transmitted through the internet, can be rendered using the BRIRs of the host A's real or modified environment so that it fuses with colleague B's voice seamlessly. When the host A moves around in the scene, 6DoF rendering ensures collaborator C's voice adapts to the positional change and does not sound unnatural. When more people join the meeting, the hearables make it possible to arrange the voices from both the local and remote people spatially for a more engaging teleconferencing experience. In the future, sensing the user's biological state would allow the hearables to understand the user's intention and needs better for more intelligent control of real sounds in such conversation scenarios. Other participants can also enjoy similar AR/MR teleconferencing experiences with an authentic feeling of presence using hearables.

VII. Challenges

As the AR/MR audio experience and most enabling techniques are relatively recent, there are still considerable challenges in implementing these techniques and understanding the perceptual performance.

A. Real-time implementation considerations

To achieve a natural AR/MR listening experience, we need to ensure both the real and virtual sound output arrives at the listener's eardrum to avoid any perceptual mismatch. However, any real-time signal processing system within the hearables, such as ADC (Analog-to-Digital Converter), DAC (Digital-to-Analog Converter), filtering, and wireless transmission, incurs a latency that needs to be managed carefully for a seamlessly synchronized AR/MR audio playback. We need to manage the processing latency through the hearable for all three processing blocks, namely sensing, control and rendering. The overall end-to-end delay of the system is referred to as Total System Latency (TSL). Perceptual tests are used to derive the permissible TSL for AR/MR audio.

The AR/MR sensing block should minimize latency for the online methods to provide timely input to the other blocks for real-time operation. Latency for sensing block comprises delay due to the sensors such as the IMU and cameras for user-tracking, and haptic sensors for tracking user's gestures. The user tracking must be fast and accurate enough to avoid any perceptual errors, such as the perceived location of the virtual sound source. For example, the human localization is very accurate for the frontal positions, with an angular speed of head movements up to 900° per second [78]. If the user's head-tracking is too slow, static sources may change their relative position when the user's head moves, causing localization errors in a dynamic virtual sound scene. The online acoustic environment sensing algorithms, such as source separation and sound event detection usually requires frame-based processing, which can incur processing latency dependent on the size of the input data frame. While DL techniques have become popular for acoustic scene analysis, limited computational capability on hearables requires the use of lightweight models. The derivation of robust generalized DL models also requires training with a large dataset with diverse acoustic environments, comprising mixtures of multiple stationary and moving sound sources, different reverberation characteristics, and types of sound events. These datasets are often derived based on real recordings and often require human effort and data augmentation techniques. Thus, the synthesis of training and evaluation datasets for DL models also presents a significant challenge.

The latency for the real sound control block is usually the strictest among other processing blocks in hearables since a superposition with the leaked real sound is needed to reduce sound level in the ANC case or to ensure acoustic transparency using HT. Typically, ANC in hearables requires TSL below 0.05 milliseconds, whereas HT use cases can allow up to few milliseconds [5]. For virtual sound rendering, studies investigating acceptable TSL values report a range between 30 and 110 ms [78], varying with the stimuli and test subject's responses. Note that the TSL value was derived for a dynamic virtual sound scene, based on the perceptual threshold of Minimum Audible Movement Angle (MAMA), which denotes the smallest perceptually noticeable localization error of a sound source. Therefore, optimal low-latency hardware and software must be chosen for the hearable. Moreover, if hearables are used along with other modalities, such as a visual display or a haptic feedback system, the latency for AR/MR audio must be equal to the latency of such systems to deliver a plausible experience to the user. Due to the minimum processing latency requirement, as discussed above, most of the signal processing must be handled using the hardware embedded in the hearables. However, the small form factor of the hearables restricts the size of hardware, such as processing chips and batteries that can be used. Often, the manufacturers' design Application-Specific Integrated Circuit (ASIC) chips to incorporate some of the methods described in Sections IV and V for popular AR audio use cases, such as ANC, hear-through, and spatial audio. On the other hand, cloud computing and 5G have gained popularity in recent years, especially for deep learning applications. These can be used to perform tasks with high computational complexity but can tolerate a slightly longer latency, such as estimation of real sound environment, computation, or tuning of existing models/filters based on changing user preferences. Low-power ASICs with optimized signal processing algorithms can also help to increase battery efficiency in case of stringent form factor requirements.

The hearables for AR/MR audio must realize the filtering tasks for real sound control and virtual sound rendering with minimum latency. To achieve efficient digital filtering, computationally intensive operations such as convolutions

must be optimized. Typically, FFT-based fast convolution methods can be used for real-time filtering. However, for cases where filter length is similar to the audio data frame size (such as those associated with RIR or BRIR), partitioned convolution approaches can be used to reduce the computational complexity. The basic principle of this method is to partition the audio signal and filters before convolution, followed by an overlap-add or overlap-save step to combine these partitioned convolution outputs [57].

B. Perceptual performance

Perceptual accuracy is an important goal for delivering AR/MR audio experiences over hearables. The plausibility of AR/MR audio experience is one of the crucial ways in which perceptual performance can be assessed [3]. There are several important aspects of real/virtual sounds that must be considered to achieve plausibility:

- The perceived location of the sound sources should match the user's expectations. Errors in the estimation of individualized transfer functions such as HRTF, HeSTF, and HeMTF may result in poor spatial perception leading to in-head localization and front-back confusions.
- The timbre of the sounds played back through the hearable must be free of undesired artifacts. The artifacts could be produced by any component in the signal processing system of the hearables, including software (e.g., errors in filter derivation) or hardware (e.g., self-noise of microphones).
- Visual cues of both real and virtual scenes are essential for providing context. The ventriloquist effect refers to the perception of sound coming from a plausible visual object rather than the actual sound source location. Therefore, the user may ignore minor inaccuracies in audio-visual modalities due to the ventriloquist effect, but perceptually noticeable errors may result in non-plausible experiences.

Evaluation of plausibility typically involves subjective psychoacoustic experiments, such as Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) test [79] and A/B tests [3], which can be tedious and time-consuming. Some past studies have investigated the derivation of objective metrics such as ITD and ILD errors to predict the perceptual performance [45]. These predictions can be helpful to design better AR/MR audio experiences without requiring detailed subjective tests.

VIII. Conclusions and future trends

With the increasing popularity of AR/MR applications and hearable devices, there is an opportunity to deliver realistic, immersive audio content to the user. This feature article provides an overview of the major developments in this area, such as acoustic transparency, which allows the user to hear transmitted virtual sounds and surrounding sounds simultaneously. This paper includes the most common possible use-cases for augmented listening scenarios, including the modification of real sound input, along with a seamless fusion of real and virtual sound. The techniques are described in three major signal processing blocks: sound sensing, real sound control, and virtual sound rendering. Using ideas from psychoacoustics, control, signal processing, and Artificial Intelligence (AI), we envision that these techniques can be used for delivering a wide variety of AR/MR audio experiences. Finally, we summarize the challenges of realizing

this system in practice for real-time implementation and perceptual performance. Additional considerations, such as security and privacy of user's data, although not described here, are critical aspects of the AR/MR audio systems. We hope that the increasing user's needs of the augmented listening experiences we described and their associated challenges will inspire the next wave of research and development of AR/MR audio technologies and products from academia and industry. We believe that such a system can be seamlessly integrated with the AR/MR system for other sensory modalities, such as haptics and vision, to deliver a truly immersive augmented experience to the user.

Acknowledgment

The authors would like to thank Santi Peksi for her assistance in drafting some of the figures used in this paper. R. Gupta and W. S. Gan would like to acknowledge the support by the Singapore Ministry of Education Academic Research Fund Tier-2, under research grant MOE2017-T2-2-060. The work of V. Välimäki is part of the activities of the Nordic SMC Network funded by NordForsk (Aalto University project 86892). The work of A. Neidhardt, C. Schneiderwind, F. Klein, and K. Brandenburg was funded by DFG (Projects BR1333/14-1 and BR 1333/18-1) and the Free State of Thuringia, Germany (FKZ: 5575/10-16).

References

[1] V. Välimäki, A. Franck, J. Rämö, H. Gamper, and L. Savioja, "Assisted listening using a headset: Enhancing audio perception in real, augmented, and virtual environments," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 92–99, 2015.

[2] P. Crum, "Hearables: Here come the: Technology tucked inside your ears will augment your daily life," *IEEE Spectr.*, vol. 56, no. 5, pp. 38–43, 2019.

[3] A. Härmä, J. Jakka, M. Tikander, M. Karjalainen, T. Lokki, J. Hiipakka, and G. Lorho, "Augmented reality audio for mobile and wearable appliances," *J. Audio Eng. Soc.*, vol. 52, no. 6, pp. 618–639, 2004.

[4] M. Tikander, "Usability issues in listening to natural sounds with an augmented reality audio headset," *J. Audio Eng. Soc.*, vol. 57, no. 6, pp. 430–441, 2009.

[5] J. Rämö and V. Välimäki, "Digital augmented reality audio headset," *J. Electr. Computer Eng.*, vol. 2012, p. 457374, 2012.

[6] R. Ranjan and W. S. Gan, "Natural listening over headphones in augmented reality using adaptive filtering techniques," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 11, pp. 1988–2002, 2015.

[7] J. Liski, R. Väänänen, S. Vesa, and V. Välimäki, "Adaptive equalization of acoustic transparency in an augmented-reality headset," in *Proc. Audio Eng. Soc. Int. Conf. Headphone Technology*, 2015.

[8] R. Gupta, R. Ranjan, J. He, and W. S. Gan, "Parametric hear through equalization for augmented reality audio," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 1587–1591.

[9] K. Brandenburg, E. Cano, F. Klein, T. Köllmer, H. Lukashevich, A. Neidhardt, U. Sloma, and S. Werner, "Plausible augmentation of auditory scenes using dynamic binaural synthesis for personalized auditory realities," in *Proc. AES Int. Conf. Audio for Virtual and Augmented Reality*, 2018.

[10] S. Tamura, "Bose Quiet Comfort 1," *Wikipedia Commons, CC BY-SA 1.0* https://creativecommons.org/licenses/by-sa/1.0. Accessed Aug. 30, 2021. [Online]. Available: https://upload.wikimedia.org/wikipedia/commons/c/c5/BOSE_QuietComfort.JPG

[11] Ramadhanakbr, "Microsoft Hololens," *Wikipedia Commons, CC BY-SA 4.0* https://creativecommons.org/licenses/by-sa/4.0. Accessed Aug. 30, 2021. [Online]. Available: https://upload.wikimedia.org/wikipedia/commons/0/02/Ramahololens.jpg

[12] M. Pesce, "Photo of Apple AirPods wireless headphones and their case," *Wikipedia Commons, CC BY 2.0* https://creativecommons.org/licenses/by/2.0. Accessed Aug. 30, 2021. [Online]. Available: https://flickr.com/photos/30364433@N05/28954822254

[13] A. Müseler, "Apple Airpods Pro," *Wikipedia Commons, CC BY-SA 3.0 DE* https://creativecommons.org/licenses/by-sa/3.0/de/deed.de. Accessed Aug. 30, 2021. [Online]. Available: http://www.arne-mueseler.com

[14] B. Roesmann, "Sony WH-1000XM3 wireless bluetooth headset with active noise cancelling," *Wikipedia Commons, CC BY 4.0* https://creativecommons.org/licenses/by/4.0. Accessed Aug. 30, 2021. [Online]. Available: https://www.digitalpush.net/mediathek/

[15] A. Müseler, "Apple Airpods Max," *Wikipedia Commons, CC BY-SA 3.0 DE* https://creativecommons.org/licenses/by-sa/3.0/de/deed.de. Accessed Aug. 30, 2021. [Online]. Available: http://www.arne-mueseler.com

[16] J. Blauert, *The Technology of Binaural Listening.* Springer, 2013.

[17] H. Fastl and E. Zwicker, *Psychoacoustics: facts and models.* Springer Science & Business Media, 2006, vol. 22.

[18] B. Xie, *Head-related transfer function and virtual auditory display.* J. Ross Publishing, 2013.

[19] J. He, R. Ranjan, W. S. Gan, N. K. Chaudhary, N. D. Hai, and R. Gupta, "Fast continuous measurement of HRTFs with unconstrained head movements for 3D audio," *J. Audio Eng. Soc.*, vol. 66, no. 11, pp. 884–900, 2018.

[20] W. Kreuzer, P. Majdak, and Z. Chen, "Fast multipole boundary element method to calculate head-related transfer functions for a wide frequency range," *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. 1280–1290, 2009.

[21] B. F. Katz, "Computational model of an individual head-related transfer function using the BEM," *J. Acoust. Soc. Am.*, vol. 105, no. 2, pp. 1193–1193, 1999.

[22] B. F. G. Katz and G. Parseihian, "Perceptually based head-related transfer function database optimization," *J. Acoust. Soc. Am.*, vol. 131, no. 2, pp. EL99–EL105, 2012.

[23] H. Møller, D. Hammershøi, C. B. Jensen, and M. F. Sørensen, "Transfer characteristics of headphones measured on human ears," *J. Audio Eng. Soc.*, vol. 43, no. 4, pp. 203–217, 1995.

[24] D. Pralong and S. Carlile, "The role of individualized headphone calibration for the generation of high fidelity virtual auditory space," *J. Acoust. Soc. Am.*, vol. 100, no. 6, pp. 3785–3793, 1996.

[25] B. Masiero and J. Fels, "Perceptually robust headphone equalization for binaural reproduction," in *Proc. Audio Eng. Soc. 130th Conv*, 2011.

[26] M. Tikander, "Modeling the attenuation of a loosely-fit insert headphone for augmented reality audio," in *Proc. Audio Eng. Soc. Int. Conf. Intelligent Audio Technology*, 2007.

[27] K. Drossos, S. Adavanne, and T. Virtanen, "Automated audio captioning with recurrent neural networks," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 374–378.

[28] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, 2012.

[29] J. Hirschberg and C. D. Manning, "Advances in natural language processing," *Science*, vol. 349, no. 6245, pp. 261–266, 2015.

[30] Y. Luo and N. Mesgarani, "Conv-Tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[31] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 1, pp. 34–48, 2018.

[32] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press Cambridge, 2016, vol. 1, no. 2.

[33] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S. Y. Chang, and T. Sainath, "Deep learning for audio signal processing," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 2, pp. 206–219, 2019.

[34] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 18–30, 2015.

[35] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 9, pp. 1652–1664, 2016.

[36] C. Han, Y. Luo, and N. Mesgarani, "Real-time binaural speech separation with preserved spatial cues," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6404–6408.

[37] B. Rafaely and K. Alhaiany, "Speaker localization using direct path dominance test based on sound field directivity," *Signal Processing*, vol. 143, pp. 42–47, 2018.

[38] V. Tourbabin, D. L. Alon, and R. Mehra, "Space domain-based selection of direct-sound bins in the context of improved robustness to reverberation in direction of arrival estimation," in *Proc. 11th European Congress and Exposition on Noise Control Engineering (EURONOISE18)*, 2018, pp. 2589–2596.

[39] V. Tourbabin, J. Donley, B. Rafaely, and R. Mehra, "Direction of arrival estimation in highly reverberant environments using soft time-frequency mask," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 383–387.

[40] A. Politis, A. Mesaros, S. Adavanne, T. Heittola, and T. Virtanen, "Overview and evaluation of sound event localization and detection in DCASE 2019," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 684–698, 2021.

[41] M. Büchler, S. Allegro, S. Launer, and N. Dillier, "Sound classification in hearing aids inspired by auditory scene analysis," *EURASIP Journal on Advances in Signal Processing*, vol. 2005, no. 18, pp. 1–12, 2005.

[42] J. Abeßer, "A review of deep learning based methods for acoustic scene classification," *Applied Sciences*, vol. 10, no. 6, 2020.

[43] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 131–135.

[44] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," *arXiv preprint arXiv:1807.09840*, 2018.

[45] J. Blauert and J. Braasch, *The Technology of Binaural Understanding.* Springer, 2020.

[46] T. Shlomo and B. Rafaely, "Blind localization of early room reflections using phase aligned spatial correlation," *IEEE Trans. Signal Process.*, vol. 69, pp. 1213–1225, 2021.

[47] J. M. Jot, "An analysis/synthesis approach to real-time artificial reverberation," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 1992, pp. 221–224 vol.2.

[48] P. Calamia, N. Balsam, and P. Robinson, "Blind estimation of the direct-to-reverberant ratio using a beta distribution fit to binaural coherence," *J. Acoust. Soc. Am.*, vol. 148, no. 4, pp. EL359–EL364, 2020.

[49] A. F. Genovese, H. Gamper, V. Pulkki, N. Raghuvanshi, and I. J. Tashev, "Blind room volume estimation from single-channel noisy speech," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 231–235.

[50] S. Deng, W. Mack, and E. A. Habets, "Online blind reverberation time estimation using CRNNs." in *Proc. INTERSPEECH*, 2020, pp. 5061–5065.

[51] W. Zhao, S. Gao, and H. Lin, "A robust hole-filling algorithm for triangular mesh," *Visual Comput.*, vol. 23, no. 12, pp. 987–997, 2007.

[52] S. Siltanen, T. Lokki, L. Savioja, and C. Lynge Christensen, "Geometry reduction in room acoustics modeling," *Acta Acustica united with Acustica*, vol. 94, no. 3, pp. 410–418, 2008.

[53] N. R. Shabtai, Y. Zigel, and B. Rafaely, "Room volume classification from room impulse response using statistical pattern recognition and feature selection," *J. Acoust. Soc. Am.*, vol. 128, no. 3, pp. 1155–1162, 2010.

[54] J. M. Jot and K. S. Lee, "Augmented Reality headphone environment rendering," in *Proc. AES Int. Conf. Audio for Virtual and Augmented Reality*, 2016.

[55] W. Hess, "Head-tracking techniques for virtual acoustics applications," in *Proc. Audio Eng. Soc. 133rd Conv.*, 2012.

[56] M. Masè, A. Micarelli, and G. Strapazzon, "Hearables: new perspectives and pitfalls of in-ear devices for physiological monitoring. a scoping review," *Frontiers in Physiology*, vol. 11, p. 1227, 2020.

[57] V. Välimäki, J. D. Parker, L. Savioja, J. O. Smith, and J. S. Abel, "Fifty years of artificial reverberation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 5, pp. 1421–1448, 2012.

[58] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation.* Springer Science & Business Media, 2010.

[59] S. M. Kuo and D. R. Morgan, "Active noise control: A tutorial review," *Proc. IEEE*, vol. 87, no. 6, pp. 943–973,

1999.

[60] V. Välimäki and J. D. Reiss, "All about audio equalization: Solutions and frontiers," *Appl. Sci.*, vol. 6, no. 5, 2016.

[61] S. Liebich, J. G. Richter, J. Fabry, C. Durand, J. Fels, and P. Jax, "Direction-of-arrival dependency of active noise cancellation headphones," in *Proc. ASME Noise Control and Acoustics Division Session presented at INTERNOISE*, 2018.

[62] A. C. S. Kam, J. K. K. Sung, T. Lee, T. K. C. Wong, and A. van Hasselt, "Clinical evaluation of a computerized self-administered hearing test," *Int. J. of Audiology*, vol. 51, no. 8, pp. 606–610, 2012.

[63] F. Denk, H. Schepker, S. Doclo, and B. Kollmeier, "Acoustic Transparency in hearables—technical evaluation," *J. Audio Eng. Soc.*, vol. 68, no. 7/8, pp. 508–521, 2020.

[64] N. Miyazaki and Y. Kajikawa, "Head-mounted active noise control system with virtual sensing technique," *J. of Sound Vib.*, vol. 339, pp. 65–83, 2015.

[65] J. Cheer, V. Patel, and S. Fontana, "The application of a multi-reference control strategy to noise cancelling headphones," *J. Acoust. Soc. Am.*, vol. 145, no. 5, pp. 3095–3103, 2019.

[66] R. Gupta, R. Ranjan, J. He, and W. S. Gan, "On the use of closed-back headphones for active hear-through equalization in augmented reality applications," *Proc. AES Int. Conf. Audio for Virtual and Augmented Reality*, 2018.

[67] V. Patel, J. Cheer, and S. Fontana, "Design and implementation of an active noise control headphone with directional hear-through capability," *IEEE Transactions on Consumer Electronics*, vol. 66, no. 1, pp. 32–40, 2020.

[68] M. Zampini, S. Guest, D. I. Shore, and C. Spence, "Audio-visual simultaneity judgments," *Perception & psychophysics*, vol. 67, no. 3, pp. 531–544, 2005.

[69] V. Riikonen, M. Tikander, and M. Karjalainen, "An augmented reality audio mixer and equalizer," in *Proc. Audio Eng. Soc. 124th Conv.*, 2008.

[70] A. Lindau, J. Estrella, and S. Weinzierl, "Individualization of dynamic binaural synthesis by real time manipulation of ITD," in *Proc. Audio Eng. Soc. 128th Conv.*, 2010.

[71] J. Ahrens, "Auralization of omnidirectional room impulse responses based on the spatial decomposition method and synthetic spatial data," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 146–150.

[72] S. Werner, F. Klein, A. Neidhardt, U. Sloma, C. Schneiderwind, and K. Brandenburg, "Creation of auditory augmented reality using a position-dynamic binaural synthesis system—Technical components, psychoacoustic needs, and perceptual evaluation," *Appl. Sci.*, vol. 11, no. 3, 2021.

[73] E. De Sena, H. Hacıhabiboğlu, Z. Cvetković, and J. O. Smith, "Efficient synthesis of room acoustics via scattering delay networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1478–1492, 2015.

[74] V. Bruschi, S. Nobili, S. Cecchi, and F. Piazza, "An innovative method for binaural room impulse responses interpolation," in *Proc. Audio Eng. Soc. 148th Conv.*, 2020.

[75] Z. Schärer and A. Lindau, "Evaluation of equalization methods for binaural signals," in *Proc. Audio Eng. Soc. 126th Conv.*, 2009.

[76] J. G. Bolaños, A. Mäkivirta, and V. Pulkki, "Automatic regularization parameter for headphone transfer function inversion," *J. Audio Eng. Soc.*, vol. 64, no. 10, pp. 752–761, 2016.

[77] J. Rämö, V. Välimäki, and M. Tikander, "Perceptual headphone equalization for mitigation of ambient noise," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 724–728.

[78] A. Lindau, "The perception of system latency in dynamic binaural synthesis," *Proc. 35th DAGA*, pp. 1063–1066, 2009.

[79] H. Schepker, F. Denk, B. Kollmeier, and S. Doclo, "Acoustic transparency in hearables—perceptual sound quality evaluations," *J. Audio Eng. Soc.*, vol. 68, no. 7/8, pp. 495–507, 2020.