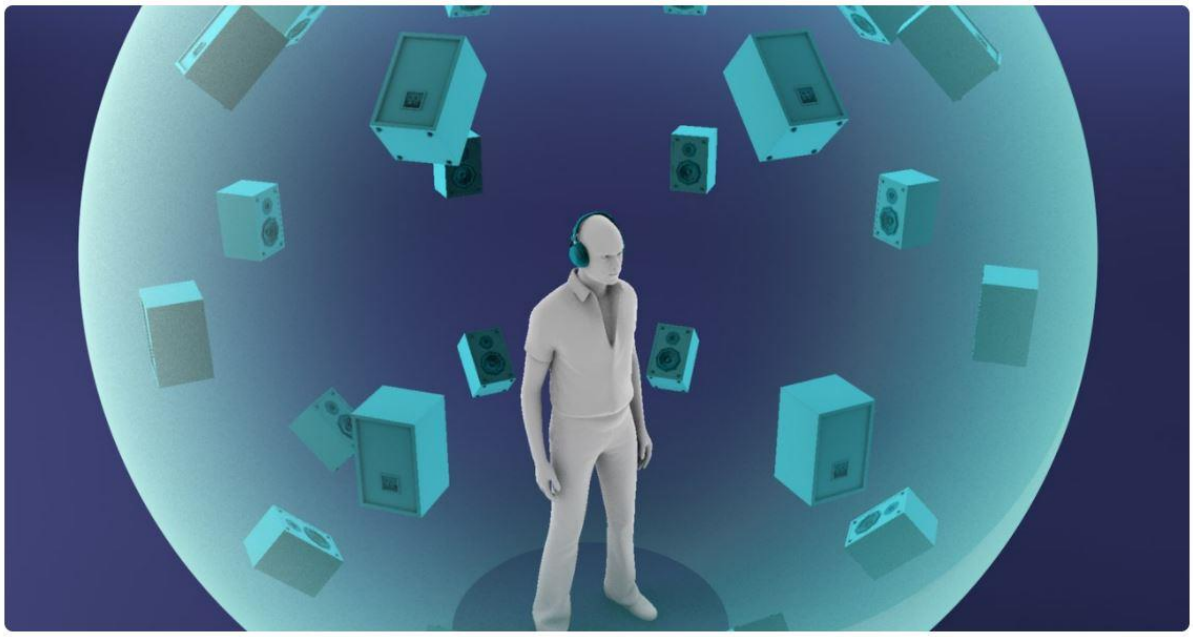
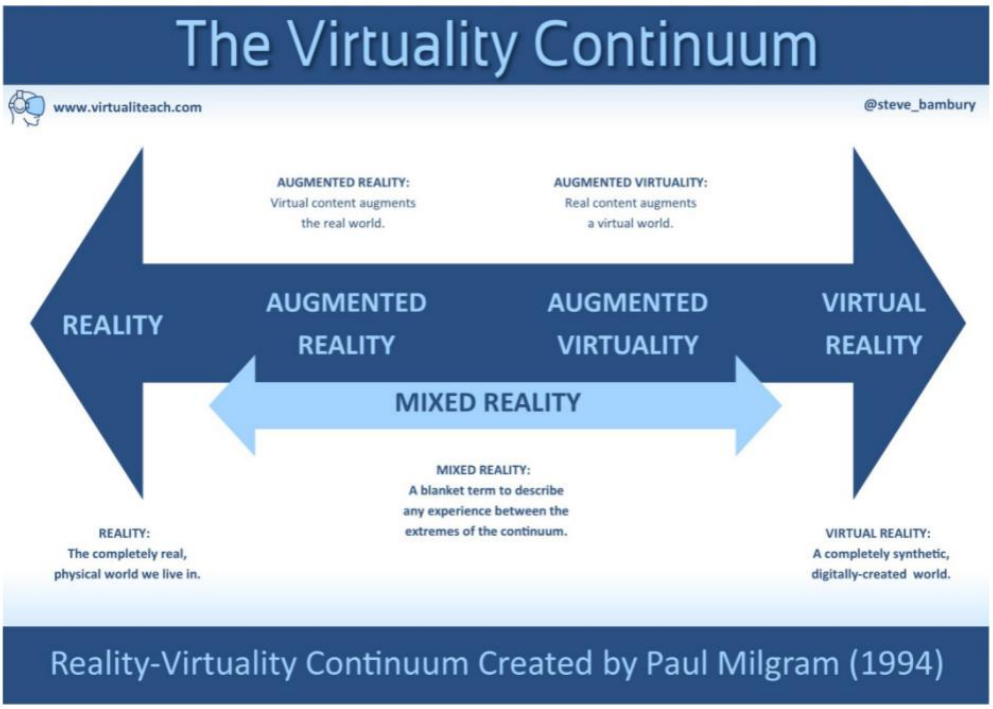
Designing sound for mixed reality

Iain McGregor
Online Sound Design MSc Programme Leader, Edinburgh Napier University

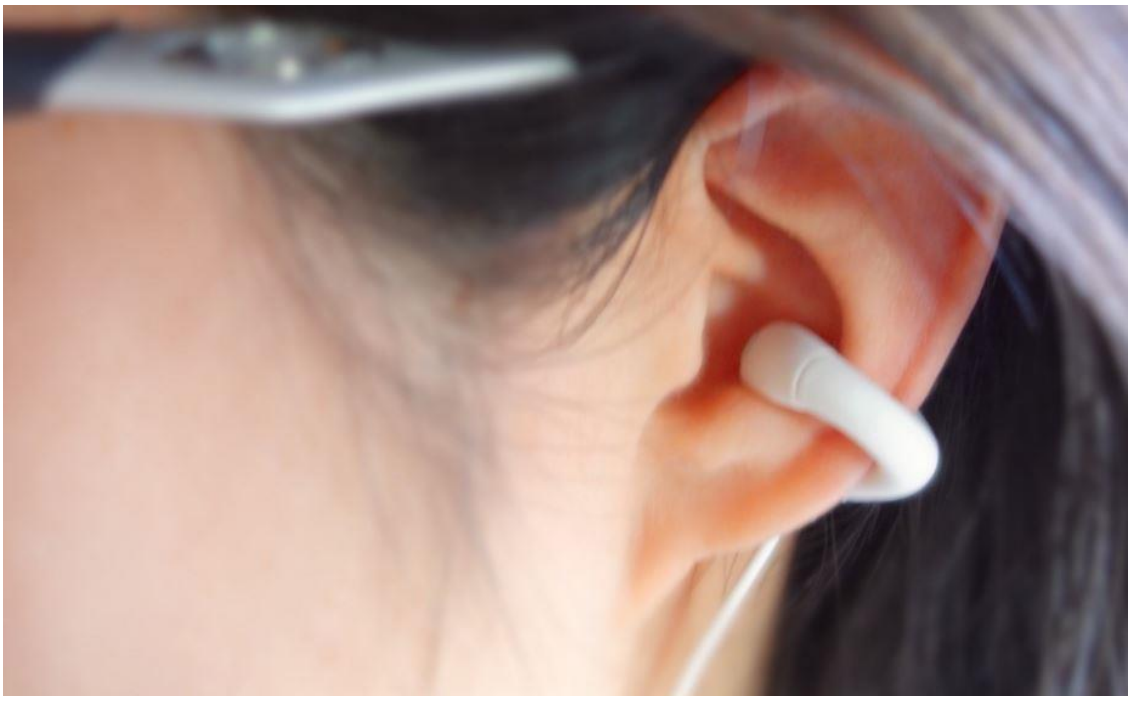


https://resonance-audio.github.io/resonance-audio/

Mixed reality refers to technologies which can be used to overlay virtual artefacts onto the physical 'real' world. The physical world can be augmented with content to provide either an enhanced, or fantastic experience, such as guiding a lifesaving operation, or being chased by a dinosaur. Currently there are two main platforms: Head Mounted Displays or glasses, and handheld devices, such as smart phones or tablets. Both formats utilise binaural stereo sound with separate headphones/earbuds, or speakers/bone conduction built into the hardware. Mixed reality offers a wide range of natural movement, which means that users expect much more variation when it comes to sound than is associated with other more established forms of interactive media.



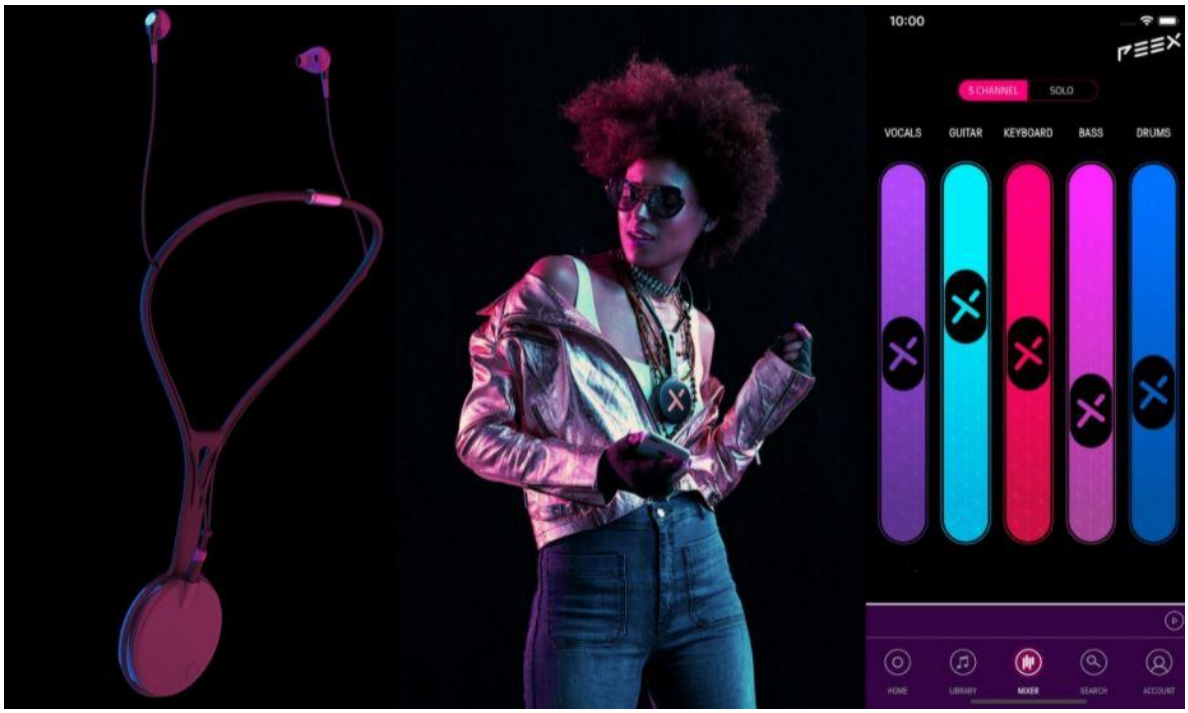
The mixed reality continuum dates back to the 1990s and incorporates Human Robot Interaction (HRI), Augmented Reality (AR), Augmented Virtuality (AV) and Virtual Reality (VR). There have been multiple iterations, with varying levels of success, but when sonic augmentation of the physical world is considered the tradition is probably as old as humanity itself. Those who had parents read them a bedtime story have experienced augmented reality. The narrative, sometimes with vocal sound effects, created a whole new world for us to inhabit, overlaid onto the pre-existing physical world. Most of us have continued to use some form of auditory augmentation well into adulthood, by listening to some form of audio (Audiobooks, Music, Podcasts, Radio, Telephone, turn by turn navigation) while performing another task. Typically, a concurrent rich understanding about both the physical and virtual auditory worlds is possible, even when the transmission medium is poor quality, such as in a mobile phone.



The first consideration for any mixed reality application is the hardware. Irrespective of the device, whether it is a mono loudspeaker, or a fully personalised binaural rendering over headphones, listeners can still become immersed, attribute meanings and interpret spatial cues. Headphones are a popular option for mixed reality, and they come in many forms, from fully enclosed, designed to isolate a listener from their acoustic environment, through to fully open making use of bone or cartilage conduction. This is similar to VR and AR. One of the best analogies for VR is that of enclosed headphones, with AR being open-backed, and one of the proposed solutions to improve VR adoption is the inclusion of a built in camera, that could allow users to quickly view their environments when desired without having to remove their headset. Enclosed headphones can also make use of this same principle by subverting microphones designed for active noise cancellation so that they pass through the sonic content unaltered, on request.



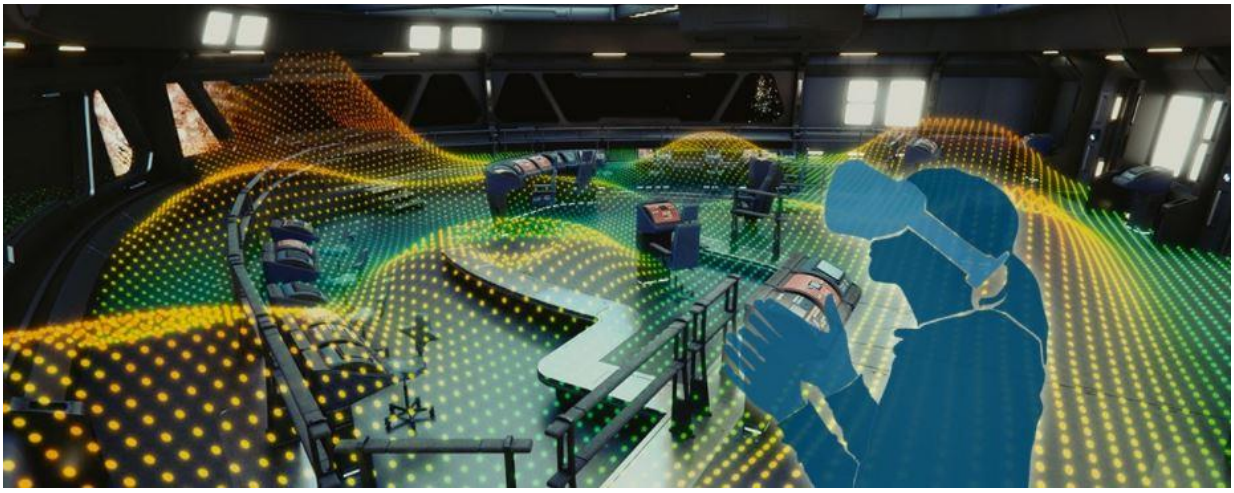
Humans expect to hear much more than they can see. One of the theories about the evolutionary loss of control of the extrinsic auricular muscles is that it allowed humans to more immediately react to a threat, by forcing them to look at a sound source when a head naturally rotated to maximise the audible input to both, forward facing, ears. The aural technologies of stethoscopes and hearing aids were both invented in the 19th century and can provide superhuman hearing. These devices can enhance or interpolate sounds that might be misheard, or even provide access sounds that would normally be inaudible, and both can be easily incorporated into Mixed Reality.



Bluetooth stethoscopes are commonly available, with active noise cancellation to improve clarity, allowing spectral filtering and amplitude control of typically inaudible objects. Combining hearing aids with Bluetooth has enabled them to be used for widespread AR, especially with the inner ear discrete models. Any internet of things (IoT) device with a microphone can now potentially become a sound source. Combined with technologies that correct for time alignment, such as those used in concert halls to select discrete mixes, clarity of real-world sounds can be attained even in highly problematic environments. This allows for not only the introduction of new sonic elements but also decisions to made about what pre-existing sounds remain audible, become enhanced or even inaudible. A virtual slider can be created between the real and the virtual in a manner similar to the mixed reality continuum itself.



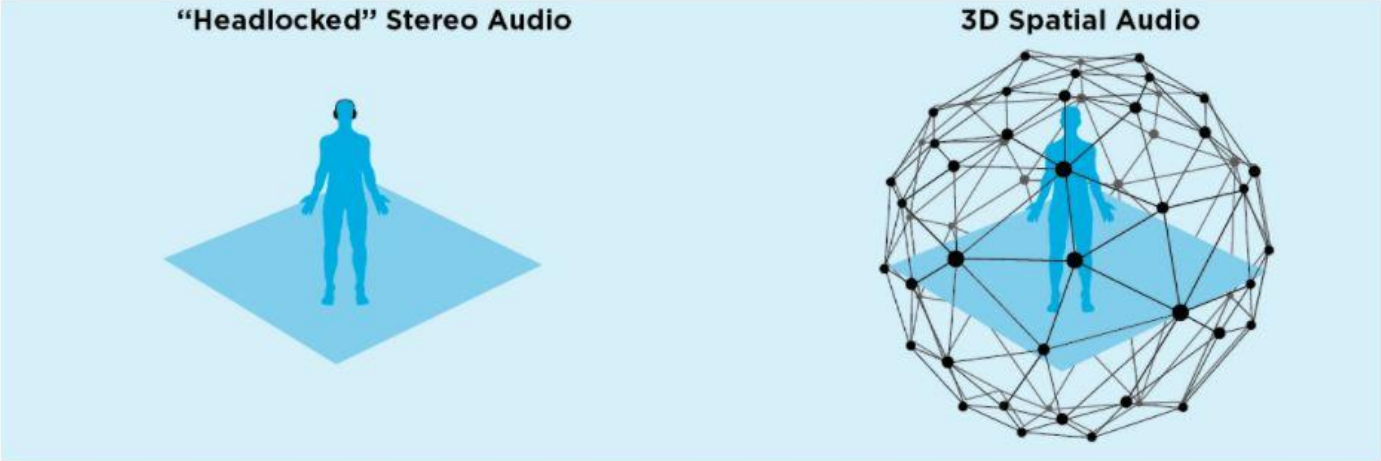
When using headphones, binaural reproduction has become the norm for reducing the effect of Inside the Head Locatedness (IHL), which makes audio content sound like it is being generated between the ears. Head Related Transfer Functions (HRTFs) can be measured and generic models supplied in software to give the sonic illusion that audio content was generated outside the head. The success of each system depends upon how closely the end user's head matches the provided model. A large head will have gaps where the sound jumps from left to right, whereas a small head will have an overlap between left or right. In mixed reality, auditory cues can be deliberately introduced to identify which HRTF set gives the greatest accuracy for a specific end user. This is preferable to relying on generic, overly averaged responses, which do not sufficiently take account of anatomic variations as well as reproduction variables, levels of listening or even distraction due to context. A similar concept is the chaos mode in education tablets, where the abilities of the user are established through simple tasks in order to identify which content to include. Tailored HRTFs can be trialled and updated in real-time when a new model is needed.



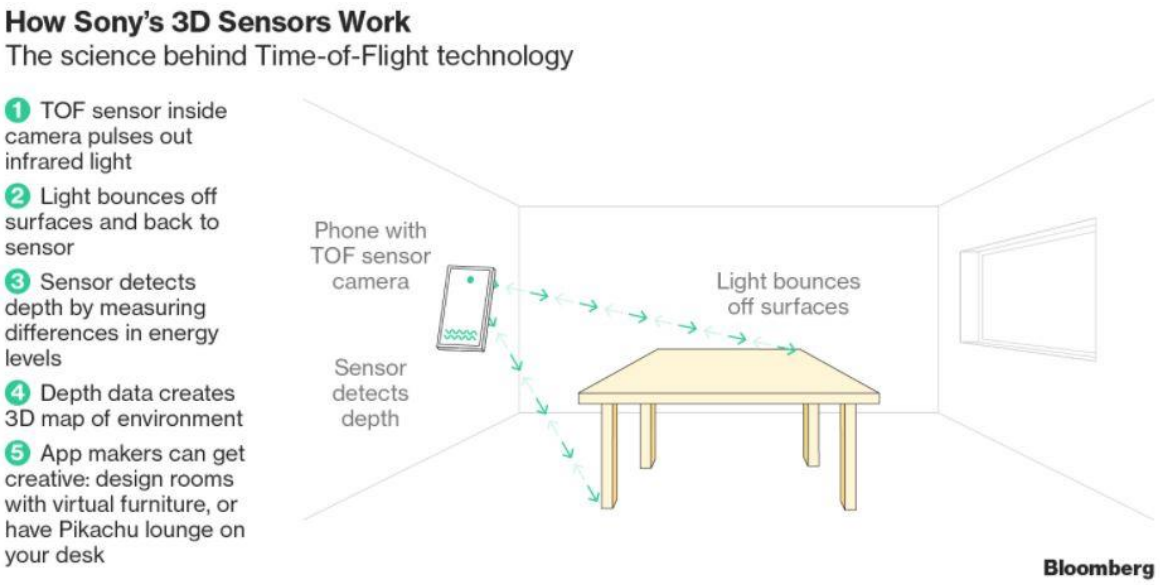
One technique to create a sense of immersive content without relying entirely on HRTFs is to use simple mono point source sounds with panning and frequency-based height cues. Bass is associated with objects being lower than the horizon, and high frequencies are often experienced as being above the horizon. There are good practical reasons for this natural association. In nature high frequency sources can often be found in trees (song birds), or at the very least above the ground (vocalisations), whereas impacts, which cause a significant amount of low frequency content, typically occur more closely to the ground, where high frequencies are usually absorbed or refracted due to obstacles. The direct sound can be treated in this way, with indirect reflections sent to the HRTF model, in which both are easily overlaid.



A video game approach can be adopted for the design of mixed reality audio, ensuring that elements are either figure or ground. A figure represents something that can be interacted with, whereas ground denotes an acoustic backdrop to provide context. Foreground (figure) sounds are actively attended to, whereas the background (ground) sounds are typically ignored. Midground sound events are often omitted as they can easily be confused as figure, with gamers become frustrated when they cannot interact with them fully. There is an expectation that everything that is clearly audible can be engaged with. A similar assumption is made by those experiencing mixed reality. A well-designed system allows sonic space for the pre-existing auditory physical world to be the midground, so that users can allow elements to be interpreted as foreground or background according to need.



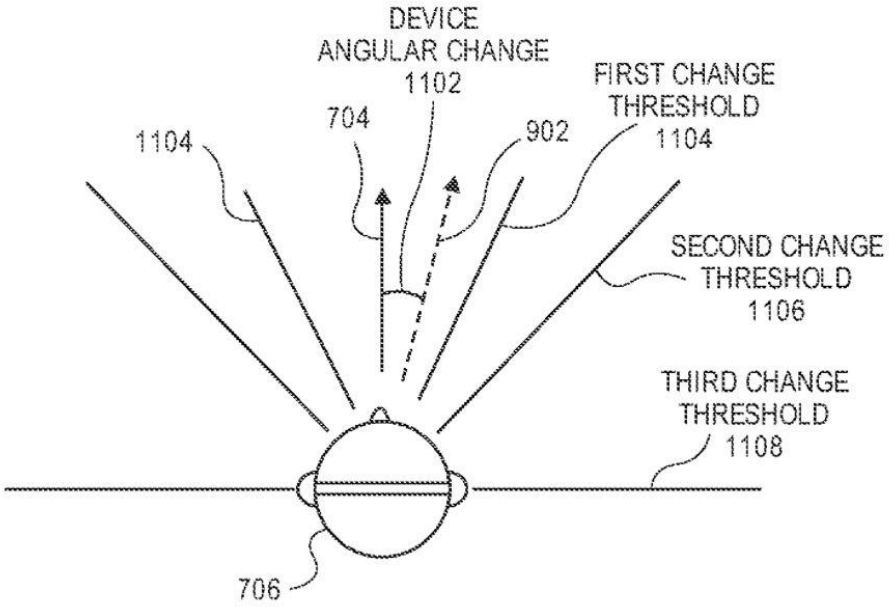
Two of the most common forms of sound design for auditory interfaces are auditory icons and earcons. The former is a kind of caricature of the action being conveyed sonically, which aids interpretation, whereas the latter is closer to a type of musical symbolism that has to be learned. This approach is similar to the sounds found in cartoons, where realistic and musical sounds happily coexist to represent a character's actions. We can use these tropes to make it easy for users to understand when a sound represents something real, or conceptual, by varying the extent to which a sound falls into either of these categories. The decision as to whether to head lock a sound source can also be used to convey a sense of reality. Sound sources that stay in a fixed relative spatial orientation as a head moves will be perceived as existing in the real world, whereas those that are not affected by head movements will be experienced as existing only in the augmented world.



When entirely believable sounds are wanted, the inclusion of time-of-flight (ToF) cameras in mobile devices can allow a level of prediction with regards to the acoustic dimensions of a small space, as well as objects that might refract or reflect sound. These data can be used to auralise any audio content, including the user's speech, to provide a sense of the acoustic colouration that would be present in an environment. When coupled with a service that recognises objects and matches them to an absorbency database, even more accurate results can be obtained. Using these technologies, augmented sound sources can be made to seem like they are present in the real environment, as their auditory characteristics can be altered in real-time, as an end user moves around the virtual source.



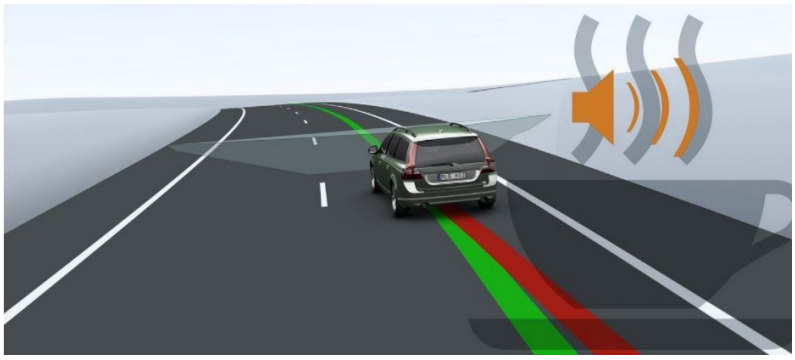
Accurate audio spatialisation is not always important, the ventriloquist effect is used routinely in cinema to allow even coverage of dialogue, in sometimes suboptimal seating positions. The sound that is coming through the centre speaker can be perceived as originating from the spatial location of its attributed visual source. This is an essential skill for humans who spend most of their time indoors, as otherwise room reflections would make it potentially difficult to identify the origin of a sound. Echo suppression is a further factor that aids interpretation, so that the direct element of a sound source, which arrives before the reflections is given priority.



Head or device movements could be tracked in order to affect the audio content. The motions that represent when a user is unsure of what they are trying to find, such as sudden movements in any direction could change the content so that only essential isolated foreground sounds are perceived. Whereas more continuous controlled sweeps provide a contextual background. A similar approach can be used to simulate selective auditory attention and habituation. Sonic cues which instigate a response, such as a user's head turn could become clearer instantly, whereas those that do not can be attenuated. Clarity could be emphasising presence frequencies (1 – 4 kHz), reducing reverberation, increasing volume, or reducing other sounds in a similar spatial location and frequency range. Increasing the temporal gaps between sound events can also assist users make sense of what they are experiencing. Short repetitive sounds are easier to locate than continuous ones, as are those with higher frequencies.



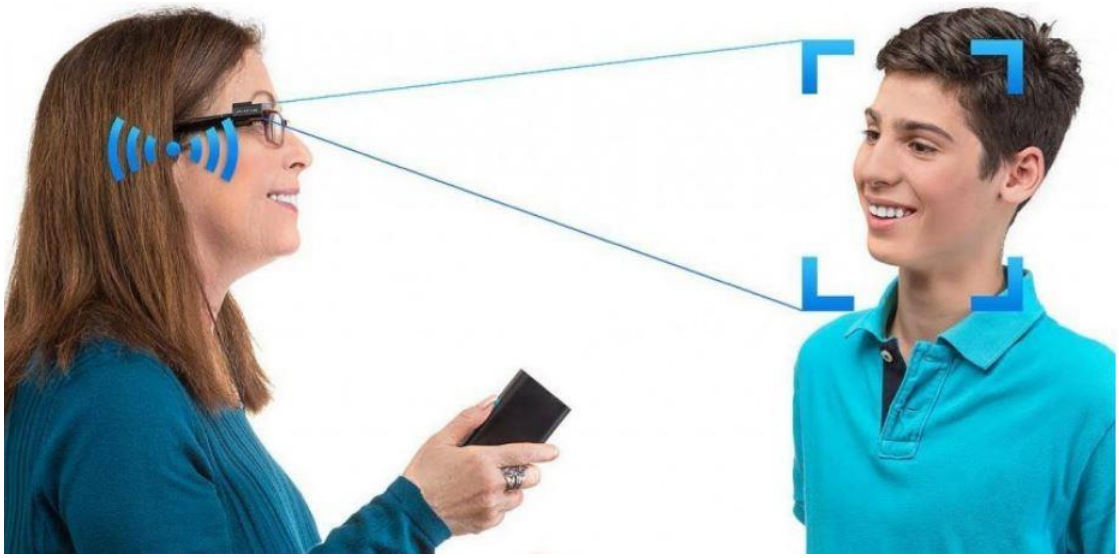
The speed of travel might also be used to alter the sonic output, driving mode is a standard function in smart phones, but it can be extended so that according to the relative speed more or less information is conveyed. A pedometer can monitor whether the form of transport is walking, running, bus, car, cycling or even horse riding. Combining these data with previous preferences, location and time of day means that the audio content could automatically transition from in-depth to more superficial accordingly. In a car if speed limiting technologies are being ignored, then audible content could be slowed down to encourage entrainment, or even made quieter so that the audible aspects of travelling at speed are more pronounced to the driver.



Safety is an ideal aspect of auditory augmentation in mixed reality. If a user is not paying sufficient attention to a task, then sonic prompts can encourage them to perform the desired action. If the audio itself is the cause of the distraction, then it can be removed, reduced or masked. A phone call can be interrupted, with the person at the other end of the line then a simple announcement of the converse is not currently able to talk. Safety alerts, such as emergency vehicle sirens, or fire alarms can be reinforced in the correct orientation to make it easier to identify their source, and act upon them, if necessary. Accurate sound source identification services are now available and becoming increasingly accurate due to the popularity of smart speakers.



Comparing the auditory cue with the pre-existing auditory environment in real time can ensure that a pertinent sound is almost always heard. If it is not attended to, simple things like introducing variation or novelty can make a sound more noticeable. For spatial guidance a sound might also be introduced immediately in front of a listener and then quickly panned or tilted to the correct location, in a sort of 'follow the arrow' manner. Similarly, a 'hotter/colder' volume and spectral clarity approach may be adopted, so that when the source is positioned correctly it reverts to mono without spatialisation and is head locked so that minor head movements no longer have any effect upon the sound until the desired task has been completed. This mimics the effect of when a sound source is being fully attended to, and becomes cognitively isolated from the acoustic environment, only the content is being conceived. Such as when talking to someone on the telephone, and no longer hearing the auditory artefacts. Listeners do not need spatialisation cues for long on familiar auditory cues, which become habituated, and subsequently ignored. With any sound if the content is perceived as having significance, then listeners quickly move from causal (source identification) to semantic (meaning), and only rarely to reduced (sonic traits) listening mode.



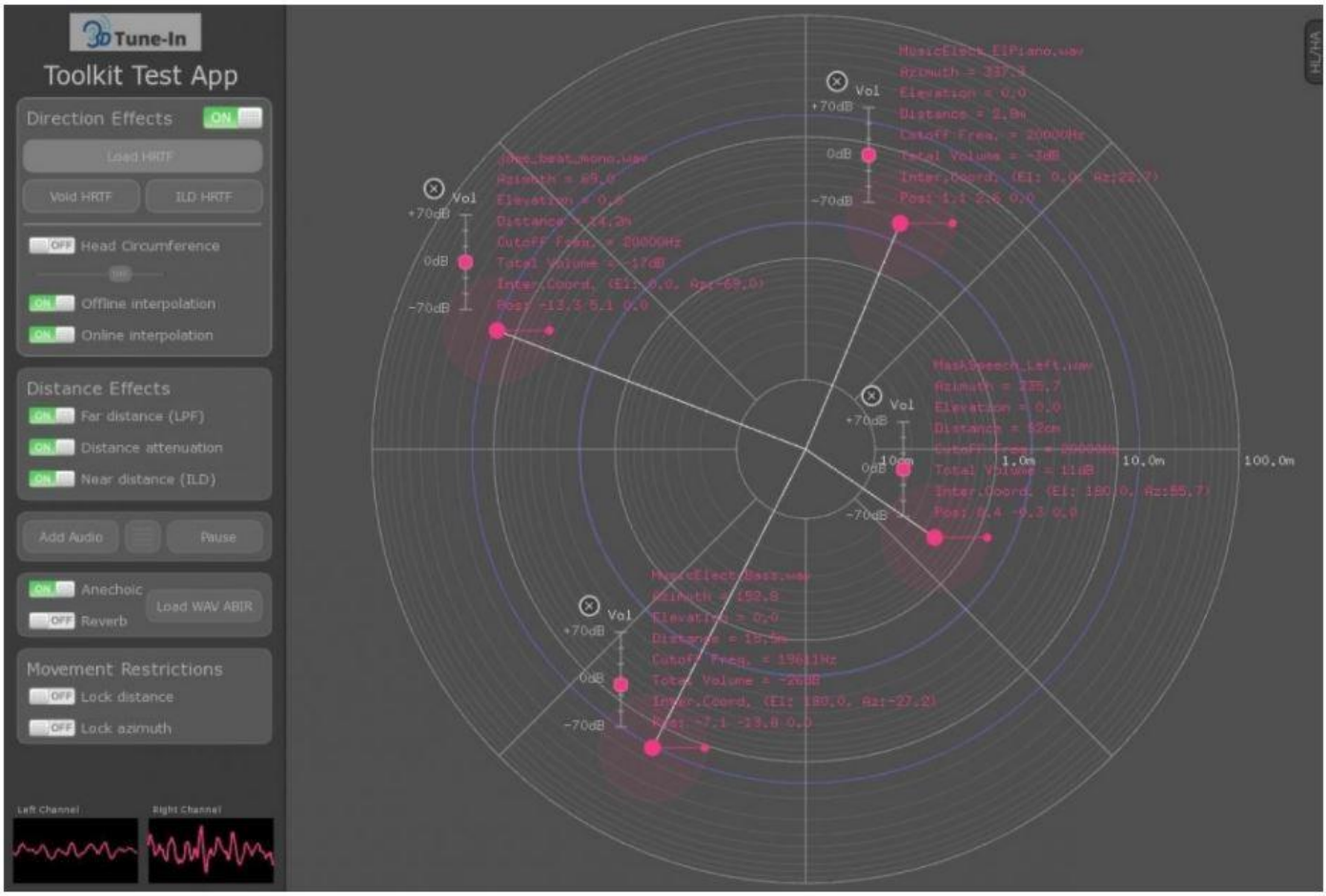
Audio content for mixed reality can use all three of the stems found in other media: dialogue, music and sound effects. Technologies already exist for the blind to inform them who is located in front of them, and this approach can be applied to not only remind a user of a person's name, but any other pertinent information, even if it is only an aural reminder of a question that they wished to ask. Background noise cancellation, such as developed for hearing aids, could then be applied to enhance speech if desired. Or even one of the ever-developing real time translation applications. Technologies such as Amazon's Whispersync could seamlessly switch between the electronic book and the audio version according to the relative proximities of the user to devices with different functionalities. Music can be altered according to the acoustic environment, or the tasks in hand. It can either become spatialised to match the physical room or be adjusted so that it is always louder than the noise floor. It might even be paused when a listener starts to speak, falls asleep or is within range of an audible emergency signal.



Audio guides are a longstanding example of Mixed Reality and have been in continuous use since the 1950s and can now be linked to both Wi-Fi and GPS to provide accuracy. When coupled with a compass, orientation becomes possible as well, which allows any large, fixed object or location to have an easily accessible mixed reality presence. Businesses and other institutions can decide how they wish to be portrayed sonically, a concert hall might relay a rehearsal, whilst a restaurant might blend together its kitchen and dining area to convey how busy it is. Other forms of regularly updated commonly available data can also be made use of. In turn by turn navigation systems, the sound of traffic movement might help drivers, or even those on foot decide which route to take based on the spatial information that conveys the level of congestion, pollution or even the speed limit.

Many forms of media now have audio descriptions, and these can be automatically selected when available. Similarly, a more pronounced dialogue mix could be transmitted, should a factor be identified which makes it difficult to hear, such as hearing loss or a high noise floor. Translations can be included should the content be in a language not spoken by the user.

Legislation has been around for at least a decade in many countries that treats hearing loss as a disability, which means that institutions often have to provide a hearing loop system. The telecoil, or T setting on hearing aids allows access to this audio feed, and many entertainment, commercial and educational environments already have active microphones as part of their standard operating practice in order to amplify voices. These can be as simple as microphone on stage feeding the sound back to a green room, where actors are waiting to be called on stage, through to every sound source being captured and mixed through a desk. Even when experiencing pre-recorded content such as films or television, the centre channel of a surround sound mix is usually dedicated to speech, with limited sound effects. These pre-existing technologies can easily be used to augment the hearing experience of any person within range.

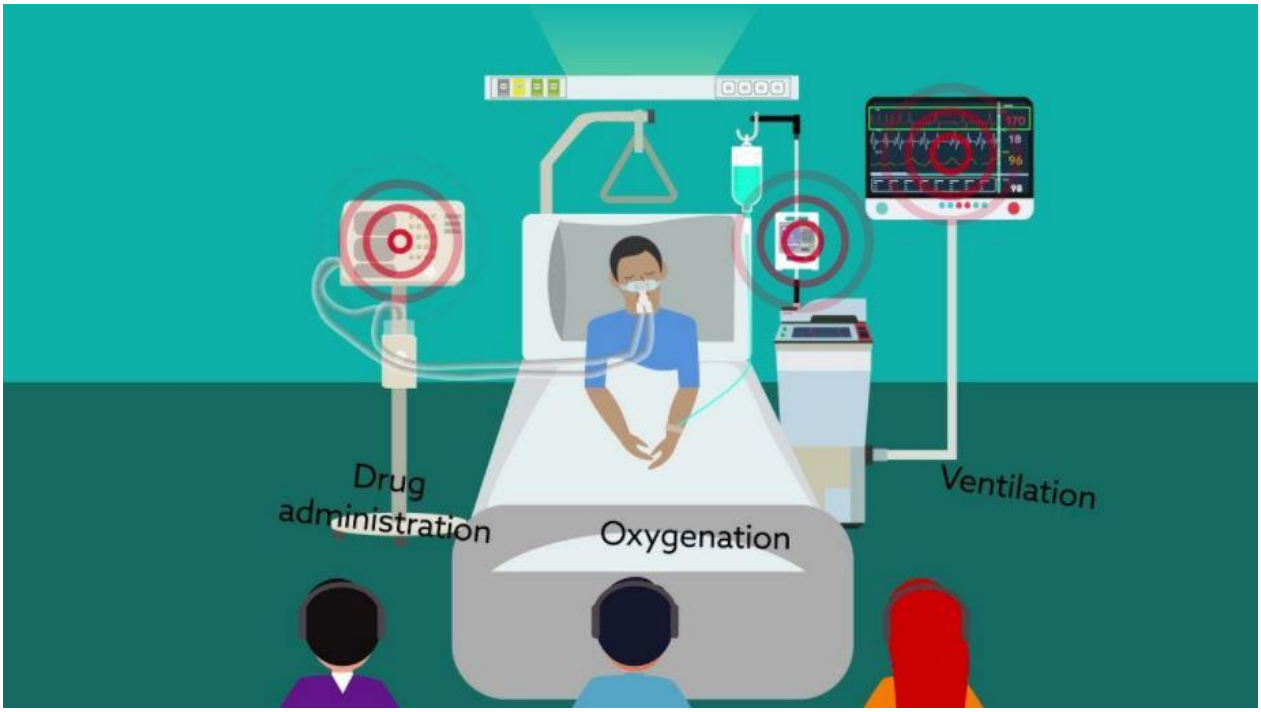


Using stereo based audio technologies means that simple techniques such as spatialising alert sounds can provide information about where the originator of a message is in relation to the user. The perceived level of importance, based upon the response times for previous messages, can also be encoded. Louder is typically more important, as is full bandwidth, with richer harmonic content. Natural associations with the content of the message or the sender can be incorporated so that the equivalent of the header is immediately comprehended. These could be taken from cartoons to convey humorous content, vehicle

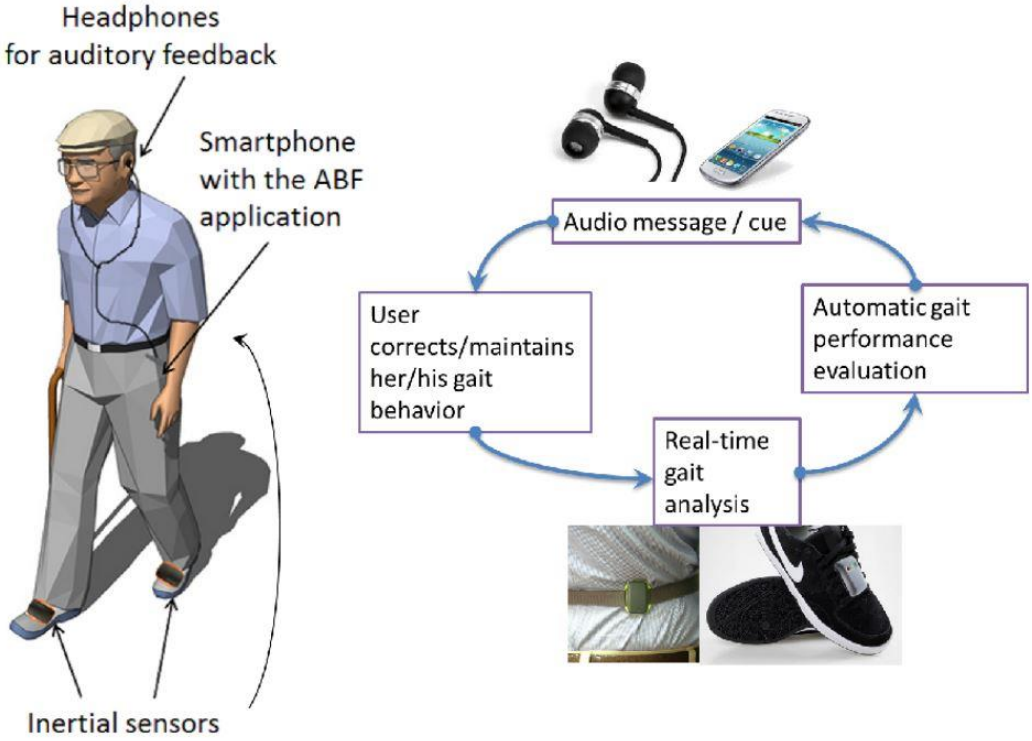
engines for deliveries, or for more serious topics, simply mask or mute existing sounds to facilitate a more considered reflection upon the content. The principle can also be used to automatically inform a parent when a child has wandered beyond a desired range, so that they hear the child's voice calling out the guardian's name in the correct orientation. This can be customised according to the time of day, location, previous settings, and even extended so that any time the user's name is spoken within a specified range their attention is directed to the identified source, if desired.



Augmented sounds can be used to mask unwanted ones to make an environment more palatable, such as many listeners regularly do with music. YouTube is already a popular resource for birdsong, café environments, and even log fires, but any sound with appropriate associations could be used, such as a cat purring, children playing, or a sporting event. New auditory elements might also provide context to make repetitive mundane sounds more interesting through increased variation. The continuous white noise of distant traffic may be modulated to transform it into waves on a beach. An air conditioner could have varying levels of ice cracking sounds to indicate the effect it was having on the air passing through it. Sounds can easily be modified to develop a different narrative, such as voices hidden in wind, or cars transforming into spaceships, in a similar manner to the multitude of AR camera effects currently available for smart phones.



The real power of sound in mixed reality comes to the fore in environments such as hospitals. Out of necessity most devices in a hospital make a sound, whether intentionally or not. The requirement for easily cleaned surfaces, and open plan layouts can make for a highly stressful acoustic environment, that affects staff, patients and visitors alike. Auditory alerts could be transferred to a virtual medium, where they only become audible in the physical world if they are not attended to quickly enough. Staff can be equipped with open ear headsets to monitor all of the required technologies, spatially represented in the correct orientation. Reaction times are improved as complex reverberations will not impede accurate interpretation of the spatial location of the source. Expanded auditory content can be represented to guide actions when closer to the sound source, which will help prevent errors in use, as well as ensuring privacy, and reduced noise pollution. Sonic representations of a patient's condition should be just as private as their medical notes. In some Intensive Care Units (ICUs) the average sound pressure level is always at least 5 dB above the World Health Organisation's recommendations, irrespective of the time of day, leading to patient sleep deprivation, which is known to impact recovery.



Other medical applications include physiotherapy and chronic diseases such as Parkinson's. In both situations sound can be used to aid movement. Gait issues associated with Parkinson's can be partially alleviated through real time synthesized cues that represent footsteps from healthy adults, which helps overcome issues with a patient's internal clock. For physiotherapy, techniques from sport can be adopted. Optimal movements made by therapists wearing sensors can be sonified, so that patients also wearing sensors can try and replicate the movements and hear how closely the sounds they produce match the optimum actions. This replicates music lessons where a teacher will play in synchrony to encourage immediate improvement. Any condition that employs sensors for monitoring can have data sonified to encourage patients to alter their behaviour, from staying stationary too long through to warning about over exertion. Reminders can be provided for actions, such as medication, or basic bodily functions, and whilst these could be done with traditional loudspeakers, knowing that these are discrete sounds, adoption is more likely. Even pain and stress levels can be reduced by the introduction of appropriate sounds when needed, which can be monitoring by measuring skin salinity, with options cycled through automatically until the most effective sound source is identified.



All sports benefit from participants listening to the actions taking place, and data capture has become standard practice in almost every discipline. Traditionally only the coaching team have access to the information, but it can be easily shared with the athletes in real-time. One-way radio communications between coaches and specific players have been standard in sports like American Football for decades. Sonification can be utilised to represent any form of measurable action. The field of wearable computing is gradually expanding, with whole body suits now available with multiple sensors to monitor discrete movement in the desired parts of the anatomy. Pressure sensors are available for ski boots to provide real-time auditory coaching as a form of gamification to improve performance. Further enhancements can be provided such as tracking relative position to an avalanche and advising skiers accordingly. This can be as simple as altering the panning of any audio content, such as music or a telephone call to encourage movement in a sideways direction. Verbal recommendations, such as swimming or creating an air pocket can also be automatically made, based upon the motions being measured.



Mixed reality allows listeners to choose whether they wish to augment or replace their auditory environment in order to create a truly unique soundscape. They can use this to hear more than has ever been possible by making use of the microphones built into the ever-expanding number of Internet of Things devices. Alternatively, end users can choose to transform their acoustic experience by masking so that they hear much less. The difference between merely isolating themselves with passive headphones, is that of context, listeners' own actions, as well as the events taking place around them can alter the audio. Any form of real time data captured by sensors can be used to affect auditory content. These can vary from simple entertainment, efficiency improvements through to life saving actions, whilst still allowing the pre-existing auditory environment to be perceived in conjunction with all of the other available senses. The sonic aspects of mixed reality can not only facilitate superhuman abilities, but also respite and safety, and most importantly it is already a principle that we are all familiar with.



© Iain McGregor All Rights Reserved Find more at www.soundformovingpictures.com