InSitu Sound™

A Study in Spatially Relevant Audio

Steve Curd
Scaeva Technologies, Inc.
August 1, 2020
All Rights Reserved ©

TABLE OF CONTENTS

ABSTRACT

This paper examines the behavior of sounds within physical spaces, the mechanics of how we perceive those sounds, and explores new digital techniques to shape sound such that the brain perceives an accurate three-dimensional sound stage from stereo headphones. We have accomplished this through capturing the detailed acoustic characteristics of any speaker or physical space with extraordinarily high precision using a set of mathematical transforms we call Quadvolution™. This led to our development of Scaeva InSitu™: a new way to reproduce music and other content to authentically simulate the way humans perceive sound with spatial relevance.

BACKGROUND

We hear in three dimensions, yet headphones only deliver a single acoustic dimension per ear. This causes an array of practical problems due to *perceptual misalignment,* such as the illusion that sounds originate from inside of our head, other spatial anomalies which affect stereo separation and frequency-dependent aberrations. But before diving into the technical details of why this is the case, it is helpful to first consider how we perceive our environment.

To ensure survival, we have redundant sets of sensing organs for both eyesight[1] and hearing[2]. This allows our brain to construct a real-time *cognitive map*[3] of our environment and is an essential element to survival. For example, as it relates to vision, a single eye cannot detect the distance to an object. With two separate eyes located some distance apart, our brain is able to process dual sets of visual images presented by the two eyes, then correlate these parallel images in real-time. Through stereoscopic vision, our two eyes (coupled with our brain's processing) adds a third dimension to our sense of sight: the dimension of distance[4].

1

Read, J. Stereo vision and strabismus. *Eye* 29**,** 214–224 (2015). https://doi.org/10.1038/eye.2014.279

[2] Avan P, Giraudet F, Büki B: Importance of Binaural Hearing. Audiol Neurotol 2015;20(suppl 1):3-6. doi: 10.1159/000380741

[3] Lennox, Peter. (2013). Presentation: Cognitive Maps and Spatial Sound. Audio Engineering Society 52nd International Conference, Surrey UK

[4] Nityananda, V., & Read, J. (2017). Stereopsis in animals: evolution, function and mechanisms. *The Journal of experimental biology*, *220*(Pt 14), 2502–2512. https://doi.org/10.1242/jeb.143883

Similarly, since we possess two ears, we also *hear* in three dimensions[5]. Just as our brain correlates images from our eyes, it also correlates audible signals from our two separate ears, enabling us to perceive a spatially relevant acoustic environment. The sound waves approaching each of our two ears are modified by the physical environment, thus they differ in terms of timing, loudness, and frequency spectrum. Our brain then interprets these minute differences and maps sounds within a virtual three-dimensional space through what is known as *psychoacoustics*[6].

Environmental Influence on Perception

The perception of hearing is one of our most complex senses. Hearing is also our fastest sense, helping us to construct a mental image of our environment before we even see it.[7]

This is supported by the fact that sound always occurs within a context: the characteristics of the sound source, the environment within which the sound is being generated and heard, and the physical characteristics of our head and ears all contribute to the way our brain perceives sound. Although some of the perception-shaping effects are generally linear, such as phase delays[8], others are non-linear, such as driver distortion[9].

We will begin by looking at specific ways in which the environment influences our perception of sound, with a focus on digitally simulating physical environments with high precision.

The Source

"Sound" is how we describe the perception of vibrating air. The distance between two crests or two troughs is determined by the wavelength, or *frequency* of the vibrations. In Figure 1, the object to the left is generating sound by vibrating left and right, creating alternatingly high pressure (dense) clusters of air molecules, and low pressure (less dense) clusters.

---

[5] Culling, John F, & Akeroyd, Michael A. (2010). Spatial Hearing. *Oxford Handbook of Auditory Science: Hearing.*

[6] Moore, Brian. (2007). Psychoacoustics. Springer New York. pp. 459-501.

[7] Horowitz, Seth. (2013). *The Universal Sense: How Hearing Shapes the Mind.* Bloomsbury USA.

[8] Smith, Julius O. III. (2017). *Phase Delay and Group Delay*. Department of Electrical Engineering, Stanford University.

[9] W. Klippel, J. Schlechter, *Distributed Mechanical Parameters of Loudspeakers,* J. Audio Eng. Society 57, No. 9 pp. 696-708 (2009 Sept.).

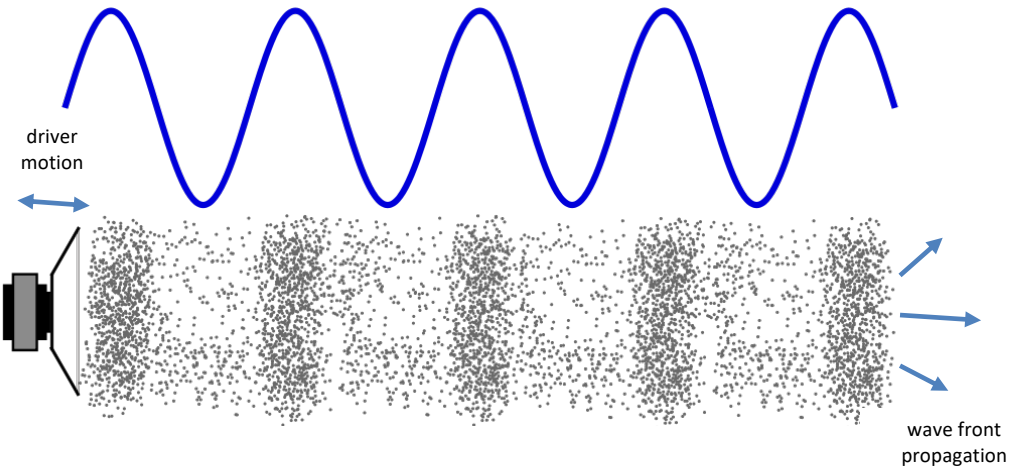

Figure 1: sound waves in air

As the disturbance, or wave, travels through air it strikes (and reflects off of) all surfaces within the physical environment. These waves ultimately reach our ears. For the vibrations that alternate between 20 times and 20 thousand times per second, our brain correlates the two sets of bioelectric signals from our two ears, resulting in the perception of a spatially relevant *sound stage*[10].

If these waves happen to represent music, the source of the vibrations is some form of a loudspeaker, or simply *speaker*[11]. A speaker translates electrical signals into physical air motion[12] as depicted in Figure 1. However, differences in the physical characteristics of the speaker itself result in the creation of waves that do not precisely match the electrical signal. This is one reason speakers sound different from one another: some produce lower frequency sounds more efficiently (often described as "warmer") while others are more efficient at reproducing higher frequencies (or characterized as "brighter").

[10] Enomoto, Seigo & Ikeda, Yusuke & Ise, Shiro & Nakamura, Satoshi. (2008). Three-dimensional sound field reproduction and recording system based on the boundary surface control principle. *Proceedings of the 14th International Conference on Auditory Display*, Paris, France June 24 - 27, 2008

[11] Headphones include at least two small speakers: one located near each ear.

[12] Ballou, Glen. (2008). *Handbook for Sound Engineers, 4th Ed.* Taylor and Francis. p. 597. ISBN 978-1136122538.

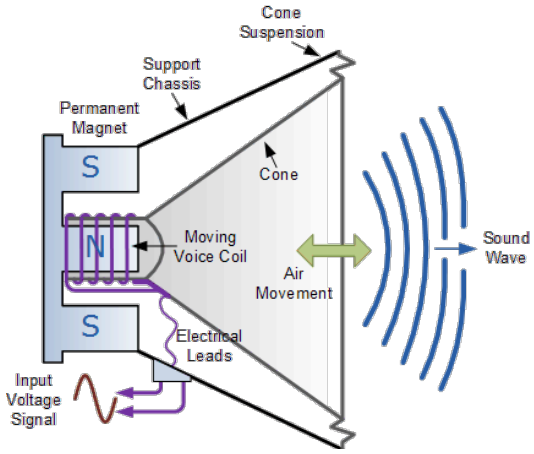

Figure 2: Loudspeaker design

These imperfections are the consequence of the physical construction and performance capabilities of the speaker device itself (as shown in Figure 2), and the electronics which drive it. Physical flaws in terms of frequency response curves, transient response and nonlinear distortion are a few examples of how speakers fail to precisely replicate the motion intended by the electrical signal. Therefore, the way we perceive sound must always begin with an understanding of the speaker's physical performance. This is a critical component in emulating how sound behaves within any physical environment, by contributing to our perception of the intended sound as it is being generated.

Reflections and Phase

As a speaker vibrates, it releases energy in the form of moving air. This energy is transferred throughout the acoustic environment, until which time it strikes surfaces or dissipates. A portion of the wave's energy is partially absorbed by any surface which it strikes, but the rest is scattered, or reflected back into the environment[13] as determined by the *acoustic reflectivity* of the material. As the wave is reflected, it is also modified by the surface material as it changes direction. And of course, this process of modify-reflect is recursive: reflected waves strike other surfaces in the environment and are further modified, just to be reflected and modified again.

The consequence of this theoretically infinite number of reflections is a series of waves that differ from one another in ways that are critical to our audible perception of our physical environment, beginning with variations in timing.

[13] Parker, Barry. *Good Vibrations: the physics of music*. Johns Hopkins University Press. p. 248. ISBN 9780801897078. Retrieved 4 January 2019.

A vibration in air (which moves as a *wave front*) travels at approximately 1,000 feet per second[14]. This means that for every foot that a sound wave travels, it is delayed by about a millisecond (one thousandth of one second). Therefore, a sound wave traveling from a speaker that is 10 feet in front of you reaches your ears in about 10 milliseconds, or one one-hundredth of a second. When that same wave passes you and strikes a wall 10 feet behind you, that reflection reaches your ears another 20 milliseconds after the first wave front as shown in Figure 3.

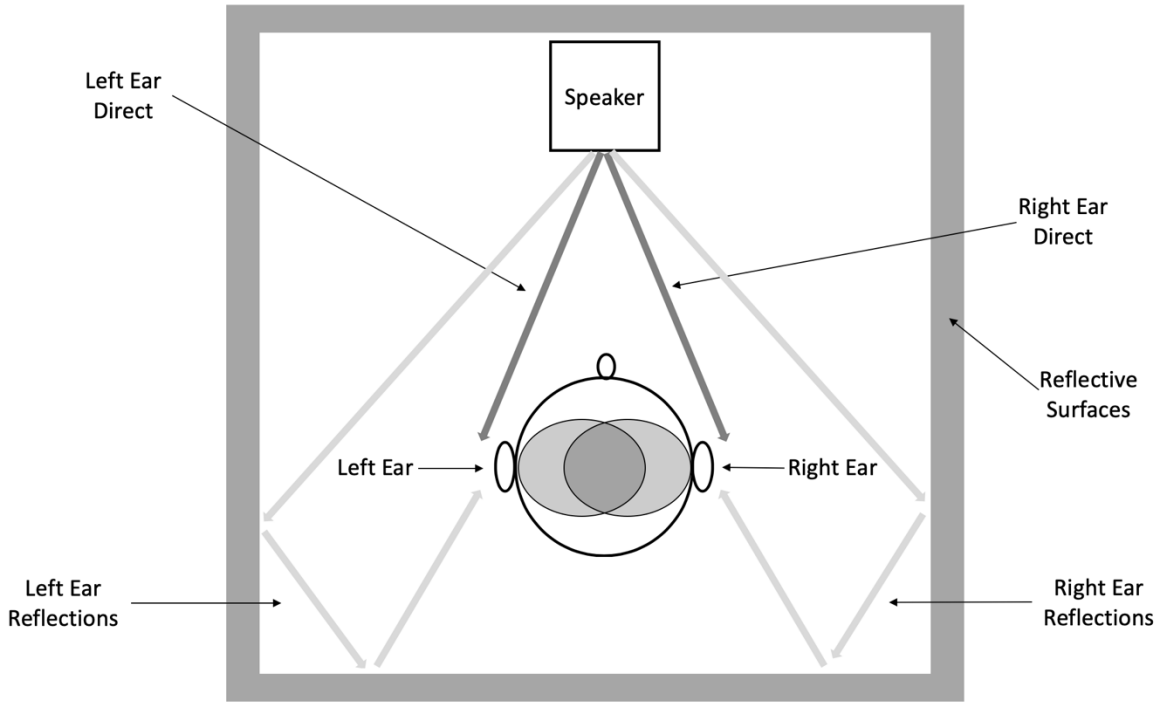

Figure 3: environmental reflections

Since your ears are separated by the size and shape of your head, the primary wave fronts and all the subsequent reflections reach each of your two ears with subtle differences in timing[15]. Also referred to as *phase changes*, these minute timing differences provide an immense amount of information to support your brain's mapping of its physical environment. It is also important to note that each reflection, and even your head itself, modifies the frequency spectrums of each wave front. This provides another set of vital environmental clues, which we will examine next.

[14] At sea level (1 atmosphere of pressure) and 70 degrees Fahrenheit, a sound wave travels through air at approximately 344 meters per second (1,128 feet per second).

[15] Moore, B. C. (2004) *An Introduction to the Psychology of Hearing*. 5th Edition London: Elsevier Academic Press.

FREQUENCY, LOUDNESS, AND THE SPACE-RELATED TRANSFER FUNCTION (SRTF)

When we think of our sense of hearing, we think first of our ears. After all, our ears actually detect audible waves. It is important to note that our anatomy, particularly the physical shape of our ears and head, modifies the characteristics of the sound waves as they reach each of our two ears independently. For example, sounds are modified as they reflect off of the creases and valleys and physical shape of each of our ears, and the hair and skin of our head absorb and scatter higher frequencies differently than lower frequencies. People with relatively larger heads will experience sounds differently than others with smaller heads. The mathematical description of how our physical biology affects the way we hear is referred to as the head-related transfer function, or HRTF[16]. HRTF is instrumental in our understanding of the consequences of our physical biology, however it is important to note that any given individual's HRTF is largely a constant[17].

While the ears do convert air motion into electrical signals, it is actually our *brain* that processes these signals and constructs the perception of sound. In fact, processing for hearing requires considerably more brainpower than the sense of smell, or even the part of your brain responsible for all of memory[18]. The reason for this is that hearing requires precise and rapid signal processing — after all, it could be fatal if the sound of a hungry carnivore is delayed!

Due to this exceptional real-time signal processing capability, our brain has developed an exquisite set of spatial-sensing capabilities. We discussed one analytical dimension of spatial sound earlier: phase and timing. But while our brain is actively computing the minor timing differences between our two ears, it is also simultaneously comparing average differences in the loudness (volume), as well as the relative loudness across the audible frequency spectrum – in other words, how we perceive differences in frequencies between our two ears.

Since reflected sound waves are always modified from the original signal[19], namely changes in phase, loudness and frequency spectrum, these variations are critically important in our brain's ability to construct an accurate acoustic map of our physical environment. Just as the head-related transfer function (HRTF) describes how our anatomy affects how we hear; we extended

---

[16] Xie, Bosun. (2013). *Head-Related Transfer Function and Virtual Auditory Display.* J. Ross Publishing. ISBN 13: 978-1-60427-070-9.

[17] Berger, Christopher C., Gonzalez-Franco Mar, Tajadura-Jiménez Ana, Florencio Dinei, Zhang Zhengyo. (2018). *Generic HRTFs May be Good Enough in Virtual Reality. Improving Source Localization through Cross-Modal Plasticity*. Frontiers in Neuroscience. Vol. 12. https://www.frontiersin.org/article/10.3389/fnins.2018.00021

[18] Chaplin TA, Rosa MGP, Lui LL. Auditory and Visual Motion Processing and Integration in the Primate Cerebral Cortex. *Front Neural Circuits*. 2018; 12:93. Published 2018 Oct 26. doi:10.3389/fncir.2018.00093

[19] Garai, Massimo. (1993). Measurement of the sound-absorption coefficient in situ: The reflection method using periodic pseudo-random sequences of maximum length. *Applied Acoustics,* Vol. 39, Issues 1-2, pp. 19-139.

this thinking by developing the *spatially related transfer function* (SRTF) to mathematically describe how our physical environment affects the sounds we hear.

SRTF AND DIGITAL PERCEPTION MODELING

Much work has been done to explore HRTF[20], and, how the shapes of the outer ear flaps (known as *pinnae*[21]) modify our perception of sound. Some companies have even introduced solutions that require the user to send a digital photo of their ears[22]. These photos are then used to modify, or *filter*, the sound to accommodate the effect of unique ear shapes. Paradoxically, our pinnae are permanently attached to our heads, so every sound we hear in nature (or even through over-the-ear headphones) is inevitably pre-filtered by our biological pinnae. After considerable research, we have learned that adjusting for HRTF may actually *degrade* the perception of virtual spatial environments, by an artificial HRTF filter on top of our intact ears and pinnae.

Consequently, much of our research shifted to advancing our understanding of the effects of physical environments on how we perceive sound, as interpreted by cognition. In other words, we focus on representing how moving air is modified by the physical environment, and ultimately how it interacts with an individual's static head-related transfer function. These new space-related transfer functions encapsulate the way humans perceive speaker performance, source location, and acoustic reflection and scattering caused by spaces and objects within the physical environment. However, unlike traditional immersive audio, SRTF required significant advancements in digital signal chain management, leading to new ways to digitally model human perception.

SPATIAL CAPTURE: TRANSFER FUNCTIONS AND IMPULSE RESPONSE

A transfer function is defined as the ratio between the output to the input of a system in the Laplace domain, with the initial conditions and equilibrium point set to zero. Although the theory is based on complex and somewhat opaque math (thanks to the French mathematician and astronomer Pierre-Simon Laplace), the goal of a transfer function is actually quite simple: to enable the mathematical representation of a complex linear system for which the timing relationship between the system's input and its output is fixed (known as *time-invariant*).

---

[20] Algazi V.R., R. O. Duda, D. M. Thompson and C. Avendano. (2001). "The CIPIC HRTF Database," Proc. 2001 IEEE Workshop on Applications of Signal Processing to Audio and Electroacoustics, pp. 99-102, Mohonk Mountain House, New Paltz, NY, Oct. 21-24, 2001.

[21] Purves D, Augustine GJ, Fitzpatrick D, et al., editors. Neuroscience. 2nd edition. Sunderland (MA): Sinauer Associates; 2001. *The External Ear*. https://www.ncbi.nlm.nih.gov/books/NBK10908/

[22] Examples: Embody Immerse, Sony 360 Reality Audio

Thankfully, the physical properties of sound meet this requirement, so the use of transfer functions allow us to apply digital-friendly algebraic equations to shape perception of sound, as opposed to relying on more complicated (and energy-consuming) differential equations.

As the first step in emulating any specific space, we must first gather all of the acoustic properties of the transducer (speaker) and of all of the surfaces within the physical space itself. This allows us to create arrays of mathematical terms, or coefficients, which we can subsequently utilize in proprietary algebraic equations. The result allows us to shape any recorded sound as if it had been generated and heard in the space being emulated, with an unprecedented level of precision.

To understand how air behaves within any physical environment, it is constructive to start with the notion of an *impulse response, or IR.* In general, an IR defines the reaction of any dynamic system when presented with an external change. IR has been historically used in music performance and production to emulate the tonal response of guitar cabinets, or to replicate reverb effects. Figure 4 demonstrates the acoustic impact of an impulse event within a large space. On the left of the graph (at zero milliseconds), we introduce a rapid surge of energy, or an impulse, into the space as shown by the vertical spike on the graph at 0.

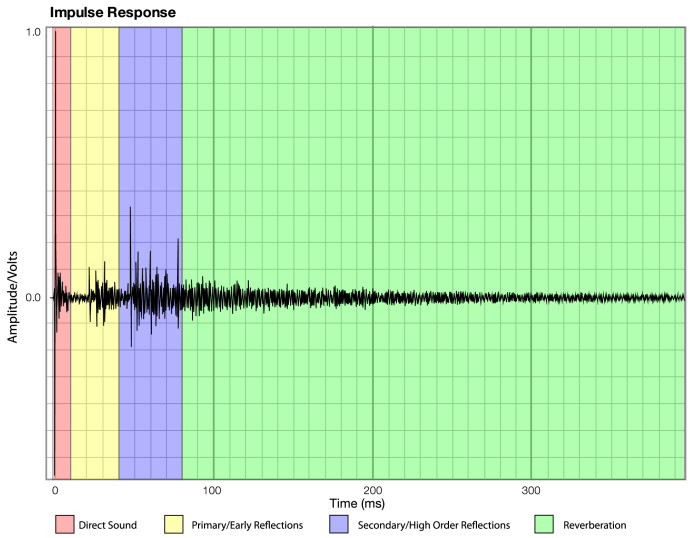

Figure 4: environmental response to an impulse

Theoretically this pulse must impart energy into the environment across all audible frequencies, with an infinitesimally small duration. This is impossible in practice, which presents us with our first challenge. An impulse is often approximated by popping a balloon or clapping hands, but unfortunately, we were unable to derive results with sufficient precision using these traditional techniques. After months of research, we had to develop new ways to introduce sound energy into the sampled space to achieve a higher precision model. It is also necessary to capture the environmental response to the impulse with very high accuracy, utilizing a proprietary high-resolution microphone array.

To demonstrate this, Figure 4 also illustrates how the environment responds to the energy imparted by an impulse. After the initial *Direct Sound* of the impulse from the speakers is captured (pink), the first set of *Primary/Early Reflections* are captured (yellow). This maps all surfaces that are perpendicular to the direction of travel of the initial wave front – the shape of the reflection captures the audio absorptive and reflective properties of each surface, and the distance from the source is derived from the timing.

This is then followed by a myriad of *Secondary/High Order Reflections* (violet), which captures the effects of the side walls, ceiling and floor, and any objects in the room which scatter or deflect the wave front. Finally, a relatively long *Reverberation* tail (green) captures the total size and "liveliness" of the space.

To summarize, we were required to innovate across two related dimensions to achieve a high precision model of a physical environment:

- impulse response methodology: a new way to inject energy into the air of any physical space being modeled
- impulse capture methodology: a new way to accurately capture the environmental response to the energy impulse

After digitally capturing the comprehensive spatial profile of a physical space, the final step in executing accurate spatial emulation is reducing the digital spatial capture into multiple sets of quadratic coefficients. These are the "magic coefficients" that are then applied to the signal stream in real-time, utilizing ultra-high precision convolution processing which we call *Quadvolution*.

QUADVOLUTION™ AND HIGH-PRECISION PROCESSING

The most challenging step in emulating the effects of physical space is processing the stereo signal in real-time. The demanding goal is to re-introduce the spatial effects of the emulated space with high precision, as if the listener were hearing the content within that space. This required innovation on three fronts.

First, the coefficients derived from the spatial capture process must be handled carefully and with high precision. Digital processing engines are notorious for "losing bits" through rounding-off, scaling, and transcendental approximations. We learned early in our research that the math libraries used to process a high-precision audio signal stream are frequently lacking in accuracy, requiring a particular focus on the underlying algorithms used to re-create a believable perception of the physical space. While developing these algorithms, we also discovered that we needed to perform the complex mathematical transforms in real-time, with high precision and exceptional power efficiency.

Second, we learned that it is critically important to manage the integrity of the entire signal chain – from the capture of physical spatial characteristics, to the computation of quadratic coefficients, to the real-time transformation of the digital audio stream. However, our research told us that the amount of computational horsepower necessary to execute these functions in real-time was impractical for battery-powered devices such as headphones. Since our ultimate goal was a self-contained headphone, much of our research was driven by intelligent algorithm optimization and power management innovation. Quadvolution packages all of these intelligent signal processing capabilities into a comprehensive digital signal chain.

Finally, extensive testing led to the conclusion that high-precision InSitu Sound requires optimally tuned headphone hardware and electronics, to pair with the finely tuned digital signal chain. Since InSitu Sound emulates intricate and complex acoustic spaces, the Quadvolution algorithms are more effective if they understand the unique physical performance characteristics of the headphone drivers or speakers used to reproduce the sound. Stated another way, InSitu Sound creates an impressive multi-dimensional effect with any sound reproduction device; however, listener experience is improved by tightly integrating the high-resolution, high-precision processing of Quadvolution with known, specifically tuned headphone hardware. To prove this, we developed a high-performance headphone platform and trained our Quadvolution algorithms to accommodate for its unique performance – by controlling for transient and frequency response performance we demonstrate excellent InSitu Sound results.

"It Just Sounds Better"

By integrating Quadvolution as a system-on-a-chip (SoC) platform, we achieved advanced spatial emulation in low-power portable headphones, capable of delivering a strikingly precise feeling of three-dimensional space. A listener can be virtually transported to an unlimited number of different physical locations, including control rooms of iconic recording studios, legendary dance venues, high-end car interiors, and even replicating $50,000 premium speakers, all in a lightweight set of headphones.

This is made possible by a small, easy to integrate microchip which incorporates Quadvolution – delivering our landmark InSitu Sound™. By reintroducing the spatially relevant sound stage that is aligned with our perceptual expectations, the result is organic and palpable. In fact, every listening test of InSitu Sound has evoked a similar response: "It just sounds better!"

Applications of InSitu Sound

Headphones have presented acoustic challenges since their invention in the late 1950's, exacerbated by widespread acceptance after Sony introduced the Walkman 20 years later. By isolating each ear with its own loudspeaker, headphones eliminate all environmental effects, resulting in an unnatural perceptual misalignment that our brain must attempt to adjust for.

Although this real-time adjustment is subconscious, it interferes with a listener's connection to the content due to an artificial, detached perception. These shortcomings also frustrate professionals who attempt to mix commercial content in headphones. This leads to the requirement to either lease recording studio facilities, or to acoustically treat their home studio, typically with mediocre results.

With an accelerating trend of knowledge workers working remotely, music and film professionals have been tethered for far too long to studios and stages, plus they must jump through hoops to test their mixes in home theatres, cars, and Bluetooth speakers. Further, professional DJ's often don't know how their mixes will sound until they are in a venue, with a live audience.

An early high-value application of InSitu Sound™ is to replicate the precise acoustic performance of key environments to facilitate the creation of high-value commercial content while eliminating the need to be physically present in those environments. For example, we have confirmed that a skilled recording engineer can mix a music track *as if* he or she were sitting in the control room of a $1 million studio environment, using only a laptop and headphones with InSitu Sound. That same mix can also be "heard" in a virtual SUV, high-end sedan, smart phone, audiophile listening room, and even a mastering studio, all in the same session using the same headphones. With InSitu Sound, many of the clear physical benefits of an expensive, extensively treated control room evaporate, and in fact we are aware of cases where the InSitu Sound mix is superior to a studio mix. Similarly, an EDM DJ can test mashups and mixes in their home studio during the production process, as if they were on the dance floor in a popular Ibiza club. This can eliminate surprises during the actual live performance and assist with perfecting the mix during production.

More broadly, InSitu Sound can also affect the perceptual impact of *any* recorded or streamed content such that it sounds more natural. We are not fans of "artificial" after-the-fact modification of commercial content; however, InSitu Sound instead can place any listener inside of a virtual world-class recording studio, enabling them to enjoy extensive acoustic treatment and professional-grade speakers, from anywhere. The fan essentially hears what the producer heard and intended, without having to visit the studio where the content was mixed. The result is truer to the originally produced content than non-InSitu Sound reproduction in standard headphones.

InSitu Sound is also applicable to AR/VR solutions. By "re-inflating" a virtual physical environment through Quadvolution, perceptual queues are inserted into the audio stream which enhance the listener's natural ability to sense direction, distance, and the physical configuration of the environment itself. For gaming, this could include real-time updates of the perceptual maps such that they adapt to the action or visual representation of the game. This enables a much more immersive environment, and reduces the perceptual misalignment between the visual and audio action.

Finally, although our focus to-date has been on replicating acoustic spatial environments in headphones, InSitu Sound technology can also be applied to sound bars and traditional loudspeakers. Of course, speakers already exist within a physical space such as a theatre room

or living room. However, InSitu Sound with Quadvolution is powerful enough to change the acoustic performance of the actual physical space to more closely replicate an ideal listening environment. Although it is not yet feasible to fix all acoustic problems, this does enable lower-cost speaker systems in substandard environments to perform as well as more expense systems in carefully treated spaces.

THE FUTURE OF 3D AUDIO

Given advancements in our understanding of how humans perceive sound, and improvements in low-power, real-time processing technologies, we are rapidly approaching a time when we will no longer be satisfied to hear music or film content through the native single-dimensional electro-acoustic transducers we refer to as speakers.

We are three-dimensional beings, living in a three-dimensional world. After extensive research, development, and validation, InSitu Sound with Quadvolution technology is proven to deliver high-precision emulation of valuable physical spaces such as recording studios and mastering rooms. With the immediate and recognized need for industry professionals to produce content remotely, the first wave of InSitu Sound will change the lives of those that produce music or film sound. As we look to the future, there is no question that every music fan will demand the organic, natural connection with their favorite tracks when they experience the remarkable difference in headphones with InSitu Sound enabled.