VR For Your Ears

DYNAMIC 3D AUDIO IS KEY TO THE IMMERSIVE EXPERIENCE

BY MATHIAS JOHANSSON • ILLUSTRATION BY EDDIE GUY

PUT ON YOUR VIRTUAL-REALITY HEADSET AND BE TRANSPORTED to a distant planet, ducking the crossfire in a battle between alien species. Laser rifle shots whiz by your head; military shuttles hover before you; the frantic calls of comrades hail from all directions. ¶ *Change the channel.* ¶ Now you are courtside at a basketball game. You hear players trash-talking on the court in front of you and coaches yelling from the bench to your left. Turn your head and the sound spins with you; the announcers in the broadcast booth are in front of you, the court sounds behind. ¶ *Change the channel.* ¶ Now you're at the Gothenburg Concert Hall, whose acoustics are ranked among the best in the world. From your front-row seat, the 109-person orchestra before you simmers quietly at first, then roars to life, its sound enveloping you. Turn your head to the left to hear the violins more strongly; turn it to the right and hear the cellos and brass section a little above the rest.

Today the technology to create the visual component of these virtual-reality (VR) experiences is well on its way to becoming widely accessible and affordable. But to work powerfully, virtual reality needs to be about more than visuals. Unless what you are hearing convincingly matches the visuals, the virtual experience breaks apart.

Take that basketball game. If the players, the coaches, the announcers, and the crowd all sound like they're sitting midcourt, you may as well watch the game on television—you'll get just as much of a sense that you are "there."

Unfortunately, today's audio equipment and our widely used recording and reproduction formats are simply inadequate to the task of re-creating convincingly the sound of a battlefield on a distant planet, a basketball game at courtside, or a symphony as heard from the first row of a great concert hall.

Sure, a stereo recording played through headphones might place the sound of a sports announcer in your right ear and the coaches' chatter in your left. But there they would remain, no matter how much you move around in the virtual environment. For a lifelike experience, engineers need to duplicate the precise directionality and position of every sound—from above and below, far and near, behind and ahead—and update it dynamically as the user moves within the virtual world.

It's a big challenge, but not at all insurmountable. Some virtual-reality producers have already begun using limited, first-generation 3D audio technology to improve on two-dimensional stereo and surround sound. And developments currently in research laboratories, including mine here at Dirac Research, in Uppsala, Sweden, hold out the promise of truly lifelike virtual-reality audio in just a few years. Here's how we think we'll be able to up the ante in virtual reality.

TODAY'S MOST WIDELY USED AUDIO FORMAT is two-channel, or stereo, sound. A stereo system records two signals, left and right; the listener plays them back through a pair of loudspeakers or headphones—again, one for the left, and one for the right. Surround-sound systems go beyond stereo by adding a center front speaker, two or four rear speakers, and a subwoofer for dedicated bass output. Newer approaches, like Auro-3D from Auro Technologies, add speakers at different heights. These start to give a 3D illusion but can't create a virtual world because the speaker positions are fixed. In the real world, you can move, and as you do so your aural experience changes noticeably.

And though these sophisticated surround-sound setups are definitely better than the stereo that preceded them, for most listeners today, evolution has gone backward. Modern portable music devices have moved people away from speakers and into headphones. That's a step away from realistic sound because today's headphones can't even do stereo properly, much less surround sound or anything approaching 3D.

Here's why. With speakers, a sound played only through a left speaker will be heard by your left ear—and also by your right ear, at an almost imperceptibly later instant and with a slight attenuation. Your brain processes this slight delay and attenuation and gives you an instantaneous impression of the direction of and distance to the sound. But when you use headphones, the left channel reaches only your left ear. The experience is artificial, and it causes some odd perceptions. For example, when a voice speaks equally loudly in the left and the right channels of headphones, it will seem to be emanating from inside your head, not from some position in front of you. That's why musical experiences can sometimes seem somewhat odd when heard through headphones. How do we get from there to a fully three-dimensional, interactive, virtual-audio experience?

With just two isolated channels to work with, such a goal may seem unattainable. But in principle it's not: The human auditory system uses just two ears to distinguish between front and back, up and down, and everything in between, so engineers should be able to create a 3D audio experience by carefully controlling the timing, volume, resonance, and echo characteristics of each sound as it reaches the ear. It would take a lot of high-speed calculations to adjust, on the fly, the sound coming through each channel, but fortunately, high-speed calculations are something we can do.

Indeed, researchers throughout the audio industry are working on this approach. One fortunate aspect: Engineers can better control the audio emerging from headphones than from open-air speakers, because they do not need to compensate for the room's shape

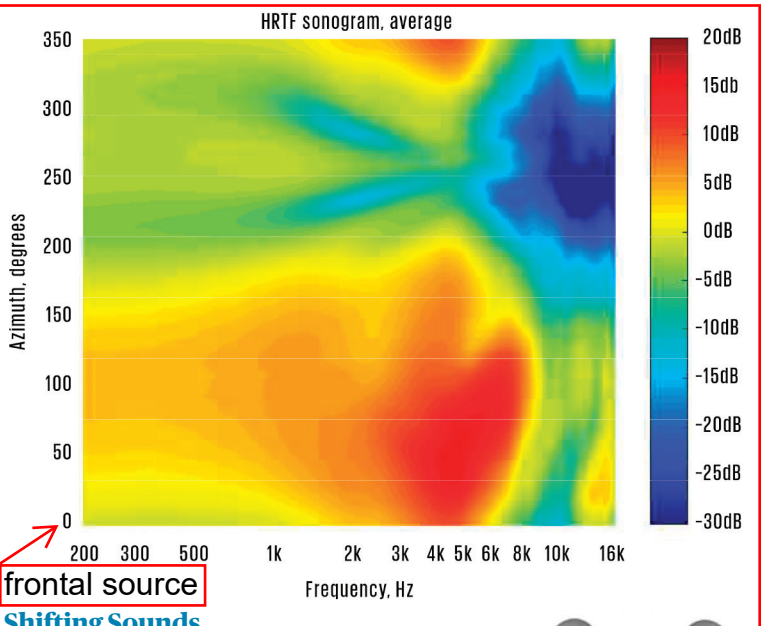
or objects in it, or deal with background or other stray noise or sound reflections that could distract from the virtual environment.

THE SIMPLEST WAY TO CREATE a 3D audio recording is to place microphones in someone's ears and record the scene in stereo from a specific position. In practice, sound engineers usually use a dummy head to do this. This is known as binaural recording, and the process has been around for a long time. Some radio stations have broadcast binaurally recorded concerts, but the technique never caught on widely.

Binaural recordings have several limitations. But for virtual reality, the main problem is the fixed position of both the listener and the sound sources.

Nevertheless, we can learn a lot from binaural recordings and apply what we learn to creating interactive audio for virtual reality. A binaural recording captures the varying sound intensities and times of arrival at two ears from each sound source, and it also captures changes to the sound that occur due to reflections and shadows caused by the shape of the head, ears, and torso. Scientists call these latter effects head-related transfer functions (HRTFs).

With a powerful computer and digital-signal-processing software, along with headphones equipped with a position-tracking device, we can create a standardized HRTF, using a dummy head or live models, and then adjust it in real time during playback according to the listener's head orientation and the original direction of the sound sources.



Shifting Sounds

Head-related transfer functions (HRTFs) are measurements of how ears respond to sounds; they are affected by the shape of the head and ears, the position of the ears in relation to other body parts, and the position of the sound source. For this sonogram, which demonstrates how HRTFs for the left ear change with the orientation of the head, researchers measured six test subjects and averaged the responses.

Here's where we run into a speed bump in the development process, however. Both stereo and surround sound are what audio engineers call channel-based formats: They encode the audio for a certain speaker configuration for playback. Ordinary stereo recordings just have basic left/right information. They don't contain detailed directional information about the recorded sound. Surround sound does a bit better; the most well-known surround-sound format, 5.1 sound, allows the mixing engineer to position sound relative to five different reference locations and adds a low-frequency channel played back through a subwoofer. The format assumes that speakers are positioned to the front left, front center, front right, rear left, and rear right of the listener. This scheme gives us more information than plain stereo does, but it still isn't good enough for convincing VR.

A newer approach is the object-based format. Instead of assuming a certain playback system, an object-based recording encodes the sound field by tagging sound sources. For example, a cello, a piano, and a vocalist are identified with information about their positions, intensities, and other data. This method then relies on smart playback devices to interpret the tags according to their capabilities and emit the sounds in a manner consistent with the tags. Dolby Atmos, introduced in 2012, and DTS:X, introduced in 2015, both use this approach.

The object-based formats were initially created to improve the home theater experience. Dolby Atmos–enabled home theater receivers, for example, incorporate ceiling loudspeakers. But the formats can potentially be adapted for virtual audio.

A third approach is the scene-based format. Ambisonics, developed in the 1970s by Peter Fellgett, Michael Gerzon, and other researchers sponsored by the National Research Development Corp., in the United Kingdom, takes this approach. Scene-based encoding creates a spatial representation of the recorded sound field as seen from a specific position. In its basic configuration, an Ambisonics recording uses four microphone capsules arranged in a tetrahedral pattern (higher-resolution recordings use more). Think of scene-based encoding as sorting the sound surrounding the listener into a number of preset directional bins, compared with object-based formats, which don't predetermine the bins but instead let each sound object be positioned at any arbitrary spot.

Current Ambisonics technology has significant weaknesses, particularly for use in real-world recording. Spatial resolution is low, and the recording microphones tend to blur the directionality somewhat. But it is a convenient way to record. And because

DIRAC RESEARCH

it is an open-source format that is readily available, it is being used by players big and small. These include Facebook, which incorporates it in the company's 360-degree videos, and Google, which uses it in its VR audio technology.

To date, two major commercial audio companies have released encoding formats that support 3D audio incorporating some of these techniques. The MPEG-H 3D Audio System, developed by the Fraunhofer Institute for Integrated Circuits, in Erlangen, Germany, supports object-based, channel-based, and scene-based audio, as well as combinations thereof. Dolby AC-4 supports object-based and channel-based audio. While all of these schemes have found some success, none of them has risen above the pack, and it is not clear if one of them will eventually dominate. And these approaches, to date, have been focused on encoding audio for reproduction through speakers; moving to headphones presents bigger challenges.

All of this activity is good news for those of us trying to create true virtual audio. However, though this work has established a good foundation, it's unlikely that any one of the existing approaches is going to develop into a robust 3D audio technology. A fresh approach is needed.

RESEARCHERS ARE GETTING CLOSER. If we combine object- and scene-based encoding with HRTF processing, we theoretically should be able to render 3D audio over headphones for head-mounted VR, and adjust it interactively as the listener moves through virtual worlds.

But, to date, applying this technology has been a struggle. A key shortcoming of HRTF playback is front/back confusion. Here's the problem. The placement of a human's ears means that a sound produced dead center in front of or behind (or above or below) the listener has the same time of arrival and intensity at each ear. So in order to determine the position of the sound, the brain exploits tiny changes caused by the shape of the ear, head, and torso. The fine details in this anatomy, the shadows and reflections they create, preferentially amplify certain frequencies in relation to others, depending on the direction from which the sound comes. These details vary from person to person, mostly because of the distance between our ears but also because of other anatomical differences.

Many researchers believe that the only way to solve the front/back problem is to use individualized HRTFs, that is, personalized acoustic body maps. More on that in a moment.

Another challenge is that the sound processed by HRTFs often sounds unnatural. Certain pieces of the audio spectrum are inevitably amplified too much or too little. These inconsistencies can be easily perceived by a trained listener. A casual listener might not identify these colorations directly, but would hear something amiss, like the difference between a cheap audio speaker and an expensive one.

The reason for this second problem is unclear. Some argue that the ways we measure HRTF are flawed, causing our existing HRTF databases to be inadequate. Others believe the coloration is inevitable unless HRTFs are individualized. The reality is likely a mix of both: Even when measuring an individual HRTF, the coloration does not completely vanish, though the location accuracy is much improved over a generic HRTF; it's likely that the techniques used to measure an HRTF are at least partly to blame.

Researchers are currently testing different measuring ideas for creating personalized HRTFs. Some are using microphones placed in a listener's ears, recording test signals played from various directions; this is a time-consuming and error-prone method. Others are trying to model individual ears using scanning and computer-graphics analysis tools, like ray tracing, to code how they reflect sound from various angles.

While personalized HRTFs might go far to fix some of the current difficulties with VR sound, they have other problems. For one thing, they assume a static head and torso that move in tandem, while VR users frequently turn their heads without moving their bodies. But the biggest difficulty is simply the time and cost of creating them, which means they aren't going to be practical for the average VR user.

HERE'S WHERE MY COMPANY'S RESEARCH comes in. For the last 10 years, Dirac, of which I am CEO and cofounder, has investigated various approaches to improve HRTF processing. Our researchers found that head movements, in particular, have a substantial impact on the HRTFs.

To understand why, imagine tilting your head toward one of your shoulders. As your ear nears your shoulder, the reflection of sound from the shoulder arrives at the ear more quickly, while the corresponding reflection at the other ear gets additional attenuation and delay.

Acting on this observation, we have built a set of what we are calling dynamic HRTFs. We based these on measurements of 30 people. We oriented the listeners' heads in a variety of yaw, pitch, and roll positions in relation to their torsos, with one degree of resolution in three dimensions, and tested sounds played from in front of the listeners, from both sides, from above and below, and from behind. We ended up making several hundred measurements for each subject. (The actual number of HRTFs measured depended on each subject's range of motion.)

To avoid having to individualize the model, and the expense that would incur, we focused on the common aspects of the HRTFs. If a certain ridge or valley in the frequency response of an HRTF was common (within a tolerance limit) to all people tested, we made it part of our generic model; if a characteristic was uncommon, our algorithms made sure that this HRTF left no sonic trace on the processed sound.

This approach won't be perfect for every person. But we have studied it and believe that there are a few strong shadowing effects and a few strong reflections having to

Moving Through a Virtual Audioscape

In the real world, sounds come from all directions. To simulate that effect for virtual reality, the sounds heard by each ear must change as the person moves within the virtual world, while the sound sources maintain their virtual positions. In this illustration, blue dots indicate sounds heard most strongly by the left ear, red dots most strongly by the right ear, and purple dots equally by both ears.

do with the head orientation in relation to the body that, if modeled well enough, capture the essential information needed by a human auditory system to determine the direction of a sound. And, as long as the HRTFs capture a motion of the head or audio object in a consistent way, small discrepancies between a generic model and an individual's HRTF will be ignored by the listener's auditory system. As a result, the audio experience will be realistic enough for most listeners. For those whose individual HRTF for a certain direction differs dramatically from the model, the sound will still seem reasonably natural, though not 100 percent directionally accurate.

Our first commercial implementation of our dynamic HRTF technology, Dirac VR, will start appearing in products by gaming headset manufacturers later this year.

Fixing the dynamic HRTF problem doesn't get us to a truly realistic virtual sound experience, however. The HRTFs give us a way to mimic sounds from any direction, but sounds are also affected by more than just the physical characteristics of the listener. A person talking to you in an open field sounds very different from a person speaking in a room. Even within a room, the location of walls and other objects strongly affect sound.

So, for virtual environments, we have to take into account how the shape of the virtual room and the objects within it—or the road, cliff, or battlefield—affect sounds. That involves modeling reflections and standing waves, the diffusive properties of walls, and the effects of interior objects when we create the sounds.

Then, on playback, we have to consider both the virtual room and the actual listener, passing not just each sound but each reflection of a sound wave—off the floor, the ceiling, and other objects—and through the appropriate HRTF. This procedure quickly becomes hugely complex and computationally intensive.

Right now, for interactive applications like games, developers simplify the acoustic information. They separate the sound into a set of directional sound sources along with a combined ambient sound field, rather than simulating the acoustic characteristics of an entire scene. The directional sounds can then be processed by HRTFs, while the ambient sound is assumed to come with equal intensity from all directions. For most people, this technique produces reasonably convincing 3D sound in some virtual environments. Eventually, more-realistic acoustic simulations of virtual rooms will evolve, improving the authenticity of the audio experience in a wider range of challenging environments.

I expect that within a few years researchers will create convincing 3D audio experiences for VR streams of, say, a basketball game or a concert. Then the big challenge will be fine-tuning the HRTF algorithms to get the computational and memory requirements down to the point where they can run on portable, battery-operated devices. Once this final barrier is overcome, immersive 3D audio for virtual reality will be ready for mass adoption.

In less than a decade, 3D audio over headsets with head-tracking capabilities will allow us to have remote meetings in which you can move about an actual room, having sidebar discussions with one colleague or another as you huddle close or step away. We will be able to have the experience of sitting courtside at the NBA finals. And we'll be able to enjoy the music of Johann Strauss Jr. at the best seat in the Vienna Musikverein. For me, that last possibility alone is worth the investments required in solving the engineering challenges that remain to fulfill the promise of 3D audio. ■

↗ POST YOUR COMMENTS at https://spectrum.ieee.org/3daudio0219

ISTOCKPHOTO