Francisco Kiss, University of Stuttgart
Sven Mayer, Carnegie Mellon University
Valentin Schwind, Frankfurt University of Applied Sciences

Audio VR—Did Video Kill the Radio Star?

Insights

- Human-computer interfaces have been historically based on visual media.
- Current trends suggest the deployment of ubiquitous AR in the near future.
- We argue that audio-based AR has great potential, particularly to overcome limitations inherent to visual interaction.

Today we see a global trend emphasizing the visual over other media. Visual media pervades our lives in the form of text, images, and video. Already in the mid-18th century, humans were fascinated by visual and often animated projections, such as magic lantern shows. With the advent of cinematography, silent films pulled crowds into the first movie theaters. However, the dominant media for information exchange was the ubiquitous newspaper, created in the 17th century. Starting in the 1950s, television gained dominance in media for both entertainment and information exchange, its dominion challenged only recently by digital streaming platforms. Thus, visual information is a longstanding, major source of knowledge transfer for a large portion of humankind, and both its ubiquity and the public it reaches grow as technology advances.

Human-computer interaction (HCI) is not impermeable to this cultural phenomenon—visuals remain the most frequent medium in computer interfaces. Technological advances have empowered interaction techniques, moving away from punched cards to screens displaying text in real time, and further to graphical depictions and even video and animations. This

Audio interaction got left behind—but it's time for a comeback.

PHOTO BY ANTONIS SPIRIDAKIS ON UNSPLASH

trend helped to make interaction with machines more human friendly: In making computers use human language, humans were no longer forced to learn computer languages. The technical constraints that once determined interaction to be visual are mostly gone. Nonetheless, as of today, most interfaces are chiefly visual, with sound or haptics playing a support role in the interaction. This legacy bias toward visual interfaces still differentiates human-computer interaction from natural human-human interaction and, further, privileges its application in computer-mediated human interactions.

Virtual and augmented reality (VR/AR) are designed using the same paradigm of reliance on visual interfaces. The results are amazing: VR technologies enable us to experience landscapes and places beyond our imagination, while AR technologies allow us to seamlessly overlay the digital world on top of the real world. Current levels of achievable realism and immersion already excel in conveying anything visual.

The benefits of video-based VR/AR have been clear for a while, but the drawbacks are starting to become evident too. When we leave the realm of entertainment and super-immersive applications, we find that daily activities are not compatible with current approaches. Immersive systems captivate users' attention, rendering them unable to perform normally on activities beyond the interaction. Day-to-day interactions with the real world very often rely on vision, and both reality and virtuality compete for attention. In application scenarios where VR and AR try to aid a visual task, designers need to invest extra care in not turning a single task into two, forcing the user to switch attention back and forth from the task to the VR/AR aid. For some activities, like driving a car, the risk of distracting the user is

The current AR/VR interaction paradigm strongly relies on visual interfaces.

too high, and therefore visual VR/AR is too risky.

THE VERSATILITY OF AUDIO

Sound has been the primary means of communication between humans since time immemorial. It plays a crucial role in how we perceive the world, and it provides rich information about our environment and the events occurring around us. In the early stages of cognitive development, we learn to relate sounds to individual events and phenomena. This evolutive feature helped our ancestors detect environmental threats; teaching children animal sounds is ongoing anecdotal proof of its importance. Even in present times and with urbanized lifestyles, recognizing the sound of an oncoming car will likely enhance your probability of survival as a pedestrian.

Humans are not only capable of identifying and classifying sounds, but also of recognizing the relative position where they originate and other spatial features of the surrounding environment. Binaural hearing, a hearing skill analogous to stereoscopic vision, enables the human brain to estimate even the speed of incoming sound sources with an accuracy high enough that we can find a ringing phone in the pocket of a coat—or hunt down that annoying mosquito that keeps us from sleeping. Hearing sound is also a fundamental component of spoken communication, the most efficient means of communication that humans ever developed. Typically learned at a very young age, for most people this skill constitutes the foundational basis for socialization. Reading, writing, and even thinking owe their existence to language, and language owes its existence to sound.

It appears you are driving on snow. Would you like me to show you how to put chains on the wheels?

Even if helpful in some cases, visual interfaces are not adequate in many scenarios.

Sound can also convey feelings and evoke emotional states even beyond words, whether through intonation of speech or other complex mechanisms like music. Rhythm, harmonies, and melodies affect the irrational part of the human mind. Minor chords sound universally sad, while some types of music make people feel like dancing. Specific types of sound can relax or annoy with more ease than images, and the use of sound to convey situational cues has been widely used in cinema, even during the silent-movie era.

Given the importance and versatility of audio in human activities, it is striking how, in modern times, it remains relegated to a support role in HCI. Besides consumer media such as music, audio is used mostly to reinforce visual stimuli, for example, to provide situational cues.

A NEW APPROACH: AUDIO-CENTERED INTERACTION

As new trends arise, audio has been taking a more central role in applications such as the smart home and in-car control, using, for instance, Siri, Alexa, Bixby, or the Google Assistant. Smart assistants are in fact the ambassadors of a new interaction paradigm. HCI has been getting closer to human-like interaction, not only by allowing speech control but also by anthropomorphically personifying system interfaces. However, at their current stage, these interfaces perform actions suboptimally, as they often misunderstand the users' pronunciation, wording, or context—sometimes all three. While the purely technical limitations can be overcome by improving voice- and context-recognition algorithms, it will then be the burden of the HCI community to allow users to intuitively interact with audio-centered assistants.

Current digital assistants are already heading in the right direction, moving away from forcing users to communicate with computers using a mouse and keyboard. Already today, assistants are allowing users to communicate with computers in a more human-like manner using voice and simple gestures. In addition, current research seeks to make

TOP PHOTO BY LUCREZIA CARNELOS ON UNSPLASH; BOTTOM PHOTO BY JAROMÍR KAVAN ON UNSPLASH, EDITED BY THE AUTHORS

conversations with digital conversation partners indistinguishable from human conversations. This will help them to correctly interpret subtleties, intonations, and maybe even gestures and facial expressions, as well as read the mood and intent of users. Digital assistants would thus become capable of understanding users and really helping them in their daily activities. For example, a proactive assistant could suggest that tired users take a break, warn users when they are about to hit their head on a hanging lamp, play soothing music when they are stressed, and remind them of things only if they have really forgotten them. Thus, it is potentially only a matter of time before digital assistants can become the ubiquitous assistants, secretaries, personal trainers, and potential confidants suggested by Joseph Weizenbaum's ELIZA. Of course, these good audio-centered assistants would render most conventional smartphone interfaces redundant, limiting the need for a screen exclusively to consuming visual media or making a video call.

The possibilities of audio VR/AR are not limited to voice assistants. For instance, current technologies enable us to cancel many sounds with ease. Active noise-canceling (ANC) technologies have quickly become a popular feature in headphones, even without supporting much user customization. ANC is the gateway to empowering users to shape their own soundscapes. Similarly, we want to empower users to adapt what they hear according to their needs and taste in any given situation. For example, if someone wanted to relax in a coffee house, they could mute the traffic noise coming from the street. At home, parents could turn up the volume of the sound coming from the next room in case the baby wakes up. Or at a dinner party, hosts could set the sound level of the dinner-party playlist to always be half as loud as the guests' voices. Finally, being able to mute the person next to you who is snoring loudly on a plane or train would be a highly desirable feature.

Implementing such a system is not as challenging as it may seem; most of the needed hardware is readily available and the use of headphones is already widespread. We envision that these extensions are bound to happen in the coming years, given the potential of modern tools such as machine learning and high-performance, energy-efficient processors. Thus, the urgency of understanding the effects of deploying such technologies in human societies becomes evident, particularly in light of the negative consequences of social media usage that we are currently witnessing. Our vision for customizable soundscapes aims to reduce sources of stress or distraction, hence improving users' well-being and quality of life.

Another hearing-augmentation possibility is three-dimensional audio. Rendering binaural spatial audio unlocks a new dimension for interaction design since it enables us to augment every possible object with sound. The potential to give each object its own voice offers a totally different approach to digital assistants, replacing the metaphor of the invisible butler with that of the wizard—a powerful framework for interaction with ubiquitous systems, in which it is possible to address objects individually and command them to perform actions. This can result in a completely different design space for both general user interfaces and more specific ones such as those for games, accessibility, augmented reality, home automation, and industrial installation maintenance. Video games are obvious beneficiaries of realistic 3D sound, particularly for applications that aim to submerge players into immersive scenarios, such as first-person games. Being able to position sound sources in space credibly can greatly improve the user experience for both single and multiplayer games. AR can also benefit from three-dimensional sound,

Talking, singing, or laughing—sound plays an important role in social interaction.

Headphones: the platform of future ubiquitous computer interfaces?

TOP PHOTO BY FOTOGRAFIA.GES ON UNSPLASH; BOTTOM PHOTO BY SVEN MAYER

Headphones and headsets are already ubiquitous and socially accepted in many scenarios.

since positioning sounds or voices in overlap with objects can be used to convey identity and agency. This can also reduce the number of visual stimuli required to bring the attention of users to a particular area or object, as well as help users become aware of spatially relevant information outside their field of view. Along these lines, home automation can apply the same concept to smart homes, assigning a voice to home appliances and adding agency to the home as if it were an invisible butler. A similar approach can also benefit industrial installations, helping technicians locate components faster or monitor complex processes by using 3D sonification. All these application areas rely on the precise tracking of users and a detailed mapping of their surroundings. But once these technical aspects are solved, it is HCI's turn to postulate a theoretical and practical framework to enable the development of applications for every possible scenario: from a cup of coffee warning you that it is hot, to a car key calling from its hiding place, to an invisible bee guiding a person through a building—imagination is the only limit.

But even before we let our imagination run wild, binaural spatial audio also enables a rich set of interaction elements from a more classical HCI point of view. Placing distinct sounds in distinct positions provides a larger design space than the visual space on a screen, while enabling a less invasive, more natural interaction. For instance, representing notifications with the sound of raindrops would be a sound that is easy to ignore but that would remain constant. Different levels of urgency or volume could be assigned according to frequency. Dismissing such notifications with a midair gesture in the direction of the originating sound would result in natural and intuitive interactions. Sound metaphors to describe events and phenomena are already part of our language and culture—putting sounds in space makes them almost tangible.

Due to the use of separate modalities, audio interfaces can be used while performing another visually dominant task. A good example is driving a car. While talking to someone in the car or on the built-in phone is allowed and may pose arguably little risk, texting while driving is extremely dangerous; thus, it is also prohibited in most countries. The same is true for watching videos, even though modern cars' embedded displays can easily stream visual media. Every task that requires visual attention is harder to do when it competes with other visual signals and cues, but not necessarily with auditory information. Most drivers manage to enjoy music or have conversations with passengers while driving without causing a traffic accident. An audio-centered user interface can potentially replace graphical interfaces when visual attention is already occupied, unlocking the possibility of performing tasks in parallel.

Sound metaphors to describe events and phenomena are already part of our language and culture—putting sounds in space makes them almost tangible.

CHALLENGES AND OPPORTUNITIES FOR VIRTUAL AUDIO

The technical feasibility of our vision of audio virtual reality is beyond question. From a hardware perspective, everything necessary already exists. Hear-through headphones, active noise-canceling, frequency separation, and even the bone conduction of sound are well-established technologies. Additionally, consumer-grade applications are available at stores at increasingly reasonable prices. Rendering spatial sound to a good-enough quality requires some processing power, but processors on mobile phones are becoming more and more powerful. The hardware required to produce realistic audio VR/AR is smaller and less expensive than a comparable setup for video VR/AR. If current trends continue, we can expect an audio VR/AR system to be embedded in normal-looking glasses within only a few years. From a software point of view, algorithms and filters need to be created. A development toolkit for interactive audio AR/VR would be very helpful. Moreover, these would all benefit from a predefined but flexible modular structure and standard. We envision these research opportunities being the focus of efforts in the very short term.

The real challenge comes from an HCI perspective: How will this technology be implemented? How can the interaction space be designed in a way that enhances our lives and aids in our activities without interfering with our goals or interactions with other people? The bottleneck for pervasive

PHOTO BY HENRY BE ON UNSPLASH

audio VR/AR is interaction design, since current design strategies are the product of years of research and work on a very different medium. A shift in paradigms is necessary: We must think out of the box and develop novel interaction concepts around sound and hearing. Manipulating the way humans hear also requires a profound understanding of human perception and cognition. Subtracting and adding sound sources can have effects on the human psyche, including affecting orientation in space, increasing the baseline cognitive load, and even affecting the user's emotional state. Further, it is clear that multiple clusters of problems will arise from areas like privacy, social acceptability, regulations, and ethics. Augmenting human hearing can also result in detrimental effects for users from a social perspective. It can enable malicious behavior, such as augmented eavesdropping, or social conflicts, such as mobbing or segregation by muting individuals or social groups. It would be very helpful to foresee the problems to come and think in advance of how to shape interaction design to guarantee a positive technology deployment.

There is another factor that will influence how this technology advances, coming from a different sphere of life: Sound-based interfaces may become an alternative, tangential development direction, and thus move away from the well-established visual culture. The impact on the technology industry and corporate interests is easy to imagine, because at the end of the day, research needs funding. Consumption, the ultimate force behind the modern economy, depends fundamentally on visual media. Branding, advertising, click-baiting, and political propaganda rely quite heavily on the innate susceptibility of the public to appealing imagery. Huge businesses like data mining, marketing, and consumer models have a sole, single goal: the creation of visual content to persuade people into doing or believing something. These concrete applications are one of the main forces pushing technological development, the other being the production of consumer goods. Both of them likely have a lot to lose if the all-pervading screens are simply gone. This would likely promote resistance to a change in paradigm, or lobbying to gain control of how it is regulated. As we have witnessed many times, the regulation of technology comes only after misuse and abuse of the possibilities it creates. In the case of pervasive auditory augmentations, the results could be appalling: Imagine hearing voices in your head all the time, constantly whispering slogans.

Thus, we conclude it is important for the scientific community to take a timely stance on the subject and proactively propose a framework for the positive, ethical development of audio-centered interaction design, taking into account the possible future development scenarios.

WHAT THE FUTURE WILL BRING

Our vision lies ahead and will unfold in a gradual, incremental manner in the upcoming decade. Noise-canceling headphones became widely available in just a couple of years, but they did not trigger a revolution like the Sony Walkman did in the 1980s. A new era of audio is due, and the paradigm shift in audio interaction is happening at this very moment. In the shorter term, we can expect to see some pioneering development in the field of augmented audio and audio VR. Microsoft's HoloLens already provides some basic rendering of spatialized audio, and Bose's AR glasses aim to provide an audio AR experience, whenever they are released. Yet these products remain as developer tools, mainly because they offer only a hardware framework with some raw toolkits—there is still no proven workflow or design guidelines.

The continuous development of machine learning and distributed computing, combined with 5G, can enable the deployment of audio VR/AR applications on existing hardware—for example, mobile phones—without forcing users to invest in extra infrastructure. The Internet of Things (IoT) might see a revival, since audio VR/AR may solve the problem of how to address individual components with ease. We expect digital assistants to get smarter in the coming years and, combined with current trends in machine learning and computer vision technologies, to be able to understand people better and thus provide smoother interactions.

In the medium term, we will be able to buy soundscape configurators—devices similar to current noise-canceling headphones—with the ability to control individual sound levels of individual or general sound sources. This kind of technology can help people cope with living and working in loud and busy places without being completely isolated from their surroundings and other people. Such devices have more specific hardware requirements, but we expect to see them in the coming years.

In the longer term, audio-centered interaction is the way Weiser's dream will come true. This goes beyond opening doors with voice commands or shaping soundscapes to experience the peace of a Zen garden while traveling on overcrowded public transportation. Audio-based VR and AR will revolutionize interaction design, changing the way we interact with machines, moving away from the interaction shaping us, toward an interaction design that supports our natural ways of communicating and experiencing reality, even in augmented ways. Video may have killed the radio star, but audio VR has the potential to once again dominate both entertainment and information exchange.

Francisco Kiss is an HCI researcher currently finishing his Ph.D. at the University of Stuttgart. He has a background in electronics and his main focus of research is sensory augmentation—the use of technology to enhance how humans perceive and experience reality.
→ franciscokiss@acm.org

Sven Mayer is a postdoctoral HCI researcher at Carnegie Mellon University. His research focuses on future interface design. While his main interest is in enhanced interaction in a mobile context, he also focuses on AR/VR applications. Here, he investigates multimodal interactions such as touch and voice.
→ sven-mayer@acm.org

Valentin Schwind is a professor for HCI at the University of Applied Sciences in Frankfurt, Germany. He has a background in computer graphics focusing on augmented and virtual reality. Before his academic career, he worked as a technical director for CGI/VFX animations and video games.
→ valentin.schwind@acm.org.

DOI: 10.1145/3386385