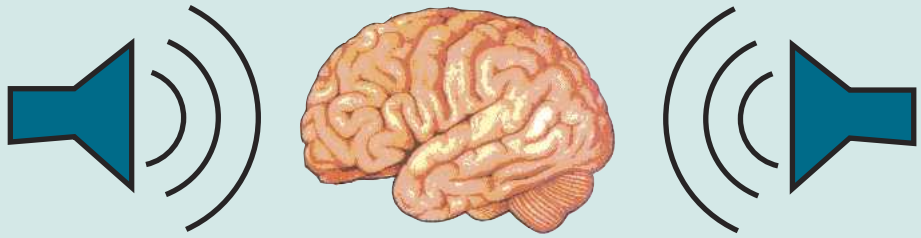
Intelligent Audio Environments



By Staff Technical Writer

The recent AES 30th International Conference, held in Saariselkä in Finland, took *Intelligent Audio Environments* as its theme. Among the research papers presented there were a number that point at important directions for the future, particularly in the fields of personal, mobile, and networked audio, in adaptive sound rendering, and in the decomposition of auditory scenes for surveillance and spatial analysis purposes. In this article we summarize a few of the main themes of these papers in an attempt to show some of the directions inherent in this ambitious field of research.

PERSONAL TRUSTED DEVICES

Huopaniemi, in his invited paper "Future of Personal Audio—Smart Applications and Immersive Communication," pointed to the scope of future interaction research that would be particularly relevant in the field of personal trusted devices. Personal trusted devices are the next step beyond mobile telephones and will be the main way in which people access and consume digital content in the not too distant future, according to Huopaniemi. They will be windows into the future Internet and smart environments, or a gateway between the physical and digital worlds. This is an important area of research, sincs by 2010 there will be an estimated 4 billion mobile phone users globally. In Fig. 1, Huopaniemi shows what he means; we see that the man in the middle interacts with remote data and services that are global in nature. The term "ubicomp" is industry slang for ubiquitous computing, which refers to all-pervasive computing applications in everyday life. Locally the user interacts with his environment involving a number of senses, leading to potentially new human practices. The three areas of technical development involved at the bottom are next-generation user interfaces, immersive communication, and smart applications and services.

In the area of smart applications, metadata is cited as a key enabling feature. Particularly in music-related applications, metadata enables ➥

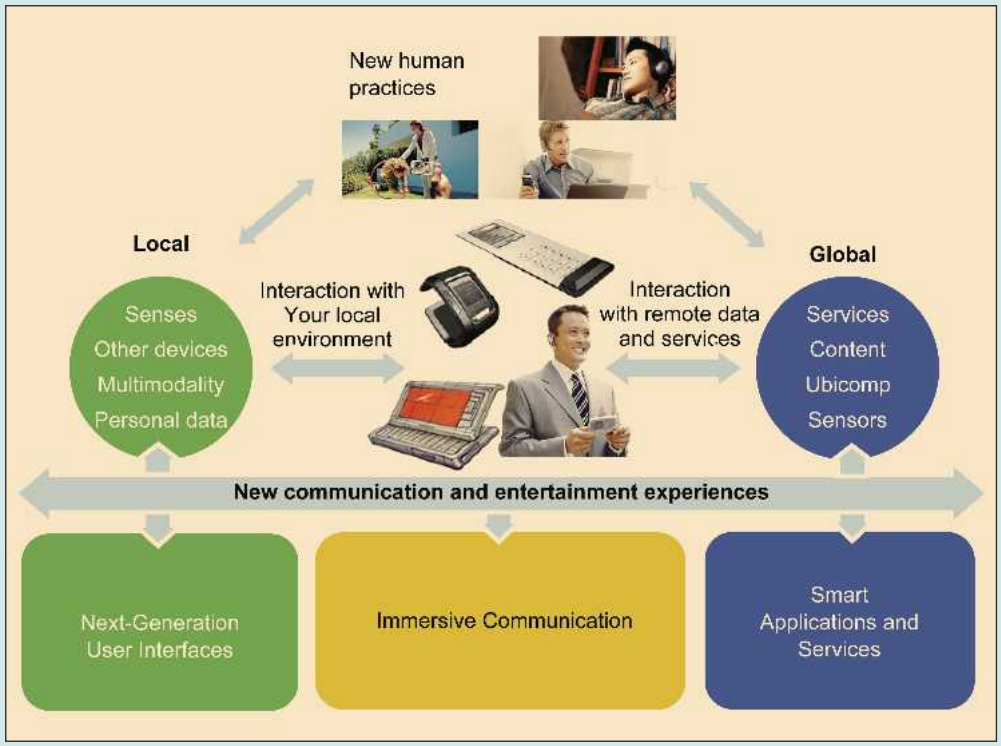


Fig. 1. Scope of future interaction research for mobile computing devices (Figs. 1–3 courtesy Huopaniemi; copyright Nokia)

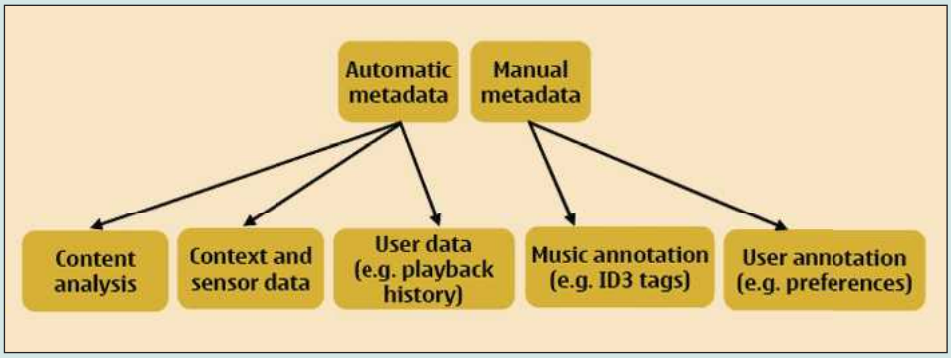


Fig. 2. Typical types of automatic and manual metadata for music

content to be annotated or tagged with additional information that allows it to be searched and managed. Such data can be manually annotated if it requires expert knowledge that cannot be encapsulated in a machine, or might be automatically generated on the basis of usage information or machine content analysis. This is shown in Fig. 2. One idea that might transform the usage of music content is the linking of automatically extracted semantic features with user or device context. For example, Huopaniemi suggests that personalized music recommendations could take into account location, people nearby, or a history of sharing that item. Sensor data from external accessories would be required in order to gather the contextual information needed to inform such processes.

User interfaces will need considerable attention over the coming years. The design of mobile devices presents challenges in terms of size, weight, power consumption, and usability. It has been suggested that the interplay between different modalities and the possibility for hands- or eyes-free operation will be high up the list of priorities. Gesture input, speech and audio interfaces, tactile feedback, and gaze tracking will all contribute to achieving these ends.

One of the main differences between future communication or entertainment environments and current ones lies in the domain of immersion. Here Huopaniemi is referring to a life-like 3-dimensional experience delivered using a mobile device. He sees 3-dimensional audio technology being used to enhance mobile computing in the fields of personal content and entertainment, communication, gaming, and user interfaces. In the first category, spatial cues would probably be mainly static, whereas in gaming they would be more likely to be dynamically changing. 3-D audio could be used for enhanced auditory displays of data and control information. One simple application for 3-D audio in mobile devices is the widening of stereo images so as to give convincing spatial audio from narrowly spaced loudspeakers. Another is to present surround sound over mobile devices using efficient audio coding schemes such as MPEG Surround. Application programming interfaces such as Java JSR-234 and Khronos Open SLES open up the possibility for fully interactive 3-D audio on mobile phones. Binaural audio is also potentially very suitable for immersive teleconferencing environments, which have been hindered to date because of the lack of network support and inherent complexity. However, as Fig. 3 shows, one can envisage such an audio-visual conferencing setup with talkers occupying different auditory spatial locations.

Huopaniemi's vision of future personal audio concludes with the idea that in the future the mobile platform will become a "digital you," capable of storing all your memories and mediating between you and the world of data using nonintrusive displays and headsets, combined with directional and location tracking.

INTELLIGENT AUDIO IN VOIP

Mobile phones notwithstanding, the other major area for communications development at the moment is Voice-over-IP (VoIP). Using the Internet as a medium for voice communications is something that has mushroomed in importance in recent years, exemplified by companies such as Skype. Markus Vaalgamaa, in the invited paper "Intelligent Audio in VoIP: Benefits, Challenges, and Solutions," compares the quality achievable using VoIP solutions to traditional telecommunications, as well as reviewing the benefits and challenges of the technology.

Comparing the basic technical features of traditional telecoms and VoIP, Vaalgamaa points out that packet-switched IP (Internet protocol) networks were not really designed for speech. Sharing bandwidth with other data and variable network conditions make delivery unreliable. In order to get a good quality of service from VoIP, it is therefore necessary to adopt clever design solutions. The typical audio signal path followed in a VoIP application is shown in Fig. 4. Some form of audio enhancement is usually needed to make the audio quality clear at the transmitting and receiving ends. There will also be some data compression and a means of concealing the effects of packet loss. Buffering is needed to account for delays in transmission or reception of packets.

One of the main benefits of VoIP compared with traditional telecoms is the possibility to use wideband speech coding and to grab a high network bandwidth. Many acoustic interfaces are also possible. However, VoIP can

Fig. 3. Vision of immersive conferencing with 3-D audio

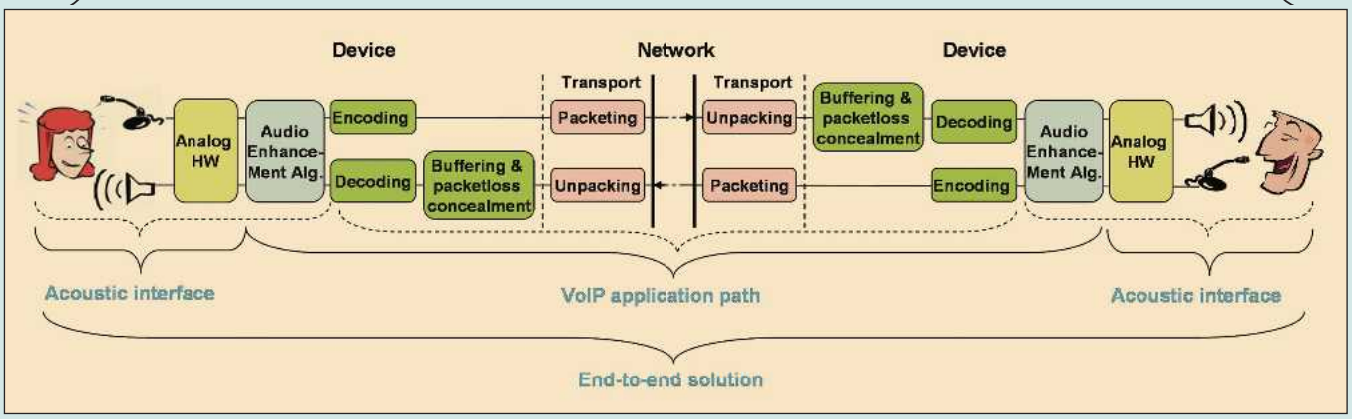


Fig. 4. Audio signal path in VoIP (courtesy Vaalgamaa)

suffer from long delays (often more than 200 ms), packet loss, annoying acoustic echoes, and variable conditions and equipment of all sorts. An intelligent approach to managing delays includes appropriate packet routing via paths that have the lowest delays. While some providers route calls through dedicated servers or nodes, most provide peer-to-peer connections between users where possible. Appropriate drivers can also shorten delays, as can buffering and error-concealment algorithms. One clever way of managing the latter is to use intelligent time compression and stretching of the speech signal without affecting the formant structure, so that the voice appears to retain the same identity throughout. It may also be possible to cut out or add silences between phonemes or words of speech, although this can be noticeable if noise is present. There is presently a lack of appropriately accurate measurement tools for evaluating the performance of these algorithms, which is an important area for further research.

NETWORKED WIRELESS ADAPTIVE AUDIO SYSTEMS

In another invited paper presented at the AES 30th Conference, Mourjopoulus examines the limitations of all-digital, networked wireless, adaptive audio systems, particularly for home-based multichannel applications . He points out that the typical components in a home audio system, such a controllers, preamplifiers, amplifiers, connecting cables, and loudspeakers, tend to restrict their flexibility. Ideally such systems should be capable of adaptation to a wide range of different source formats, involving multiple channels. He sees the key to universal integrated solutions being the wireless coupling of any number of digital acoustic transduction elements to digital audio sources in networked applications. This will enable the evolution of novel, data-independent, intelligent reproduction systems.

The benefits of the approaches he specifies are the result of having an all-digital signal chain with flexible and exact control of operation, as well as smaller size and improved efficiency. The ability to adapt to any audio format or number of channels and the advantages of wireless operation enable adaptation to the listening environment and to different decoder hardware. One could add or remove ➥

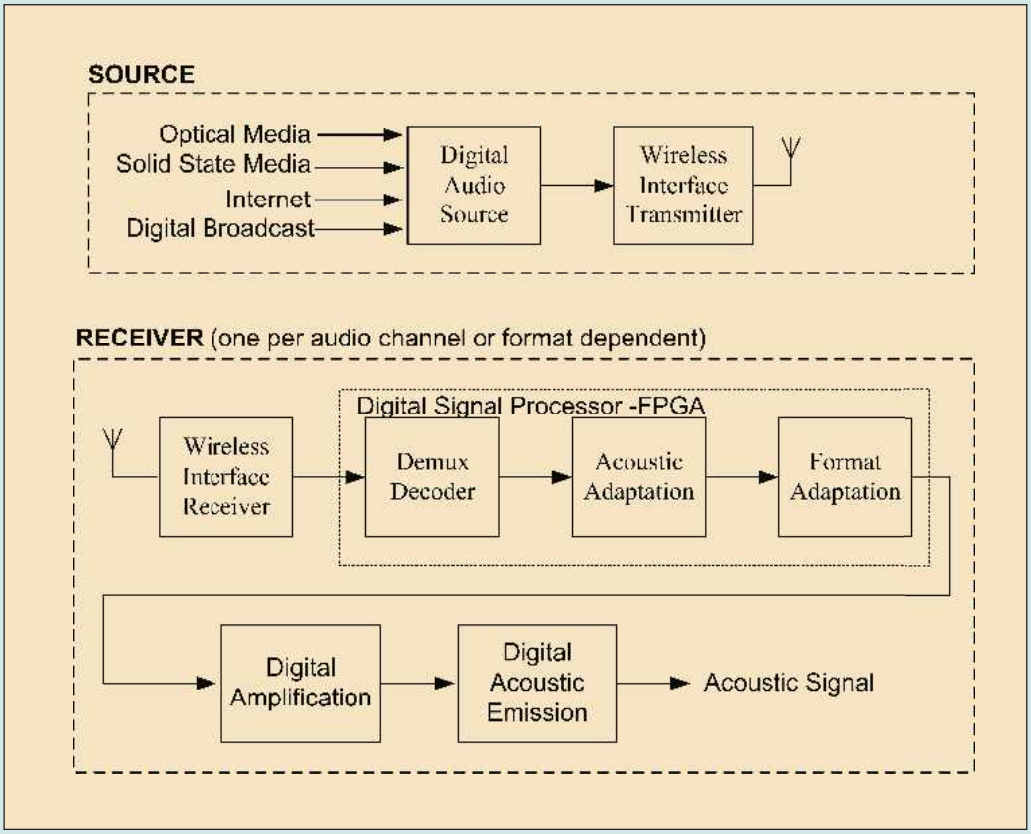


Fig. 5. Possible structure for an all-digital wireless networked, adaptable signal chain (courtesy Mourjopoulos)

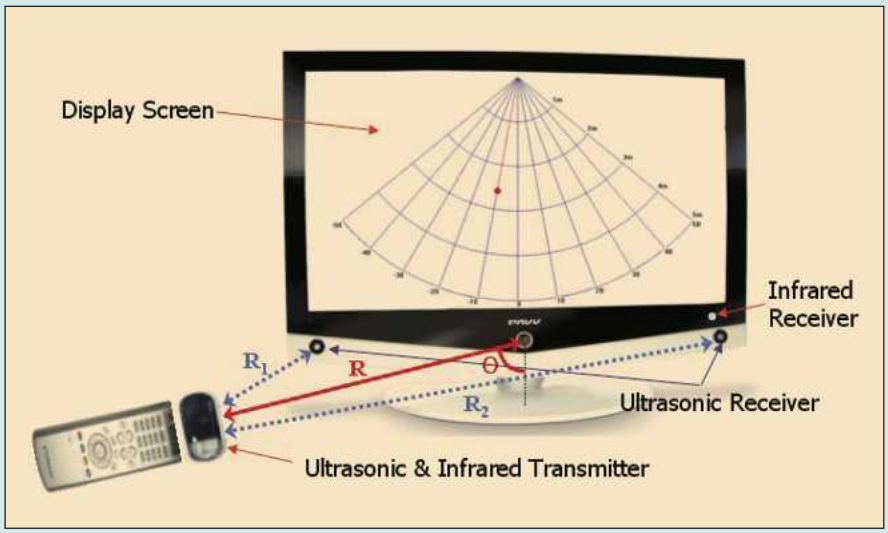


Fig. 6. Digital television with listener-position tracking system (Figs. 6 and 7 courtesy Kim et al.)

devices and serviced zones easily. Fig. 5 shows a possible structure for such an all-digital wireless networked signal chain.

Two different approaches can be taken to the use of a wireless local area network (WLAN) for audio in this case. One is a point-to-point delivery approach whereby an audio server transmits in real time to a number of unsynchronized wireless receivers using an audio access point (AP) device. The number of audio replay devices that can be connected depends strongly on the current conditions on the network, such as interference and signal strength, and has to be controlled using dynamic bandwidth reservation algorithms. While a number of wireless home audio solutions exist, operating in the S-Band (2.4 to 2.8 GHz) and the C-Band (5.725 to 5.875 GHz) and also using Bluetooth, they often use compressed digital audio or even analog modulation schemes and are based on proprietary protocols. This solution may be suitable for the transfer of audio between different rooms or at different times, but not for advanced spatial audio reproduction to multiple devices intended to work together in one room. The alternative approach is wireless point-to-multiple-receivers, which differs principally because it involves synchronization between the receivers so that delivery times and audio sampling rates can be matched. This is important for multichannel audio because of the need to maintain a specified relationship between the channels. Mourjopoulos explains that this needs the widespread IEEE 802.11 standard. One of the features that such systems could take advantage of is a procedure for automatic receiver position adaptation and topology discovery. These could be integrated with enhancements of 802.11 including lower protocol layer metrics for adjusting radio power upon reception and the like. Each receiver (loudspeaker) would need to include a suitable buffer to accommodate network transmission delays, so that channels can be adequately synchronized.

Mourjopoulos also discusses the options for digital transducer arrays and for acoustic adaptation of reproduction systems to room acoustics. In the case of the latter, current objective or system-dependent models for room-dependent equalization can give rise to some unwanted side-effects or distortions dependent on the listening position, program material, or specific loudspeaker–room combinations, whereas perceptually-oriented analysis tools may be able to complement those approaches by predicting the audibility of distortions so that the soundfield at the listening position can be adjusted to conform to some desired characteristic.

ADAPTIVE VIRTUAL SURROUND

When a listener wants to hear 5.1-channel surround sound from only two loudspeakers, it is usually necessary to use some sort of loudspeaker virtualization algorithm employing binaural filters and crosstalk cancelling. Such systems filter the rear loudspeaker signals with head-related transfer functions (HRTFs) and mix them into the front loudspeakers so that they appear to be coming from a position behind the head. Crosstalk cancelling is supposed to deal with the problem that there is crosstalk between each channel's loudspeaker signal and the contralateral ear, which otherwise ruins the binaural effect. The main limitation of these approaches is that they restrict the listener to a small range of locations, because the filters used are only valid for a small range.

Kim et al., in "Adaptive Virtual Surround Sound Rendering Method for An Arbitrary Listening Position," describe an approach to virtual surround that can be used with digital televisions (DTVs) having only two loudspeakers. They offer a solution to the problem that arises when a listener moves from the sweet spot by attempting to track him using ultrasonic and infrared sensors. Essentially this is a static system designed to work at one of a number of fixed listening positions, rather than one that attempts to track the listener dynamically. One also must assume that it is intended to work for a single listener. As shown in Fig. 6, the display screen is fitted with two ultrasonic (US) receivers and one infra-red (IR) receiver (the standard one), while the remote control is fitted with transmitters for both types of signal. The idea is that when the user presses a button on the remote, signals of both types are transmitted and the receiver measures the relative delays between them, using the fast-travelling IR signal as a reference. Since the ultrasonic signal travels more slowly, it is possible to measure the relative delay between the received IR signal and that arriving at the two US receivers. This gives information about the distance and lateral position of the listener. The information so gathered is used to adjust the delay and gain of the audio channels, as well as to select the most appropriate crosstalk cancelling filter, as shown in Fig. 7.

The system was designed using both a "naturalization" filter and a localizing

filter for the virtual surround signals. The former is used to avoid the in-head or center localization of the rear loudspeaker signals, which can happen with a left–right symmetric structure, and was achieved by a combination of channel decorrelation and filtering to imitate a desired soundfield type. In order to make the incorporation of this system economical in a typical DTV receiver, it is necessary to reduce the number and complexity of the filters. This is achieved by a variety of means, including exploiting the similarity between left and right ear HRTFs, filter length minimization, and multirate processing. The latter is based on an evaluation that determines that the low-frequency range is most important for speaker virtualization. The number of different possible localizing filters is reduced from over 16,000 possibilities to just 23, taking into account the fact the delay and gain of channels can be adjusted separately.

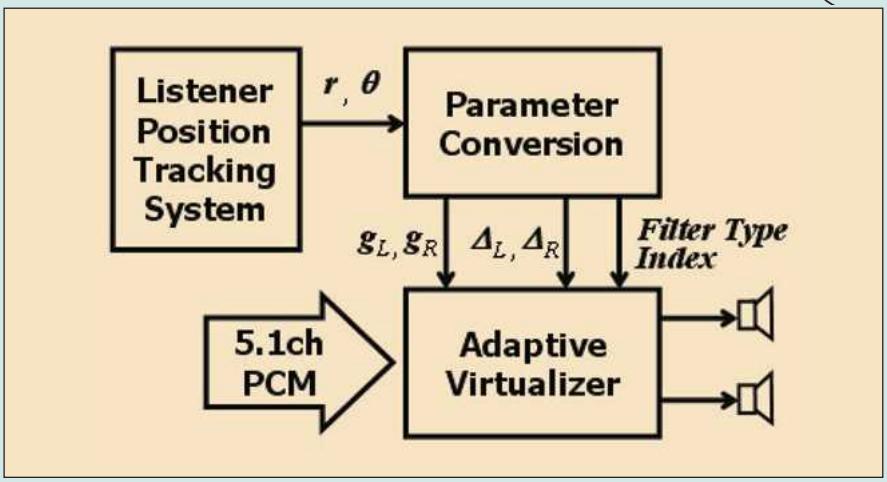


Fig. 7. Block diagram of adaptive virtual surround rendering system. g_L and g_R are channel gains, and Δ_L and Δ_R are channel delays.

STRUCTURED AUDITORY SCENES FROM FIELD RECORDINGS

In interactive spatial audio applications it is often considered important to be able to render a spatial audio scene that adapts to the position of the listener. As Gallo and Tsingos explain in "Extracting and Re-Rendering Structured Auditory Scenes from Field Recordings," scene-authoring models tend to assume that individual sound objects in the scene have to be emitted by monophonic point sources that have to be individually generated or recorded. However, there are limitations to the use of this in practice owing to the fact the real sources are quite complex, microphones have specific pickup patterns, and propagation models are limited. The alternative extreme in terms of spatial scene capture is to record and re-render the scene as a whole, in a manner more akin to conventional stereo recording. However, this allows little or no control over the elements of the scene and interactivity is limited.

The authors propose an intermediate approach that is based on the capture of a scene using a limited number of omnidirectional microphones whose locations are known and quite widely spaced. It is intended mainly for field recordings of real-world signals. The signals from these microphones are used to analyze the locations of sources within the scene, based on measurements of propagation delays. In fact the signals are split up into a number of subbands in the frequency domain, which are separated into background and foreground components. Background components represent the diffuse, stationary elements of the scene such as reverberation and background noise, whereas foreground components represent potentially nonstationary, well-localized sound events. This is graphically depicted in Fig. 8. In order to separate these components the authors employed a denoising model derived from speech processing, which led to a foreground signal with only limited musical noise. They found that this musical noise was usually masked when recombined ➥

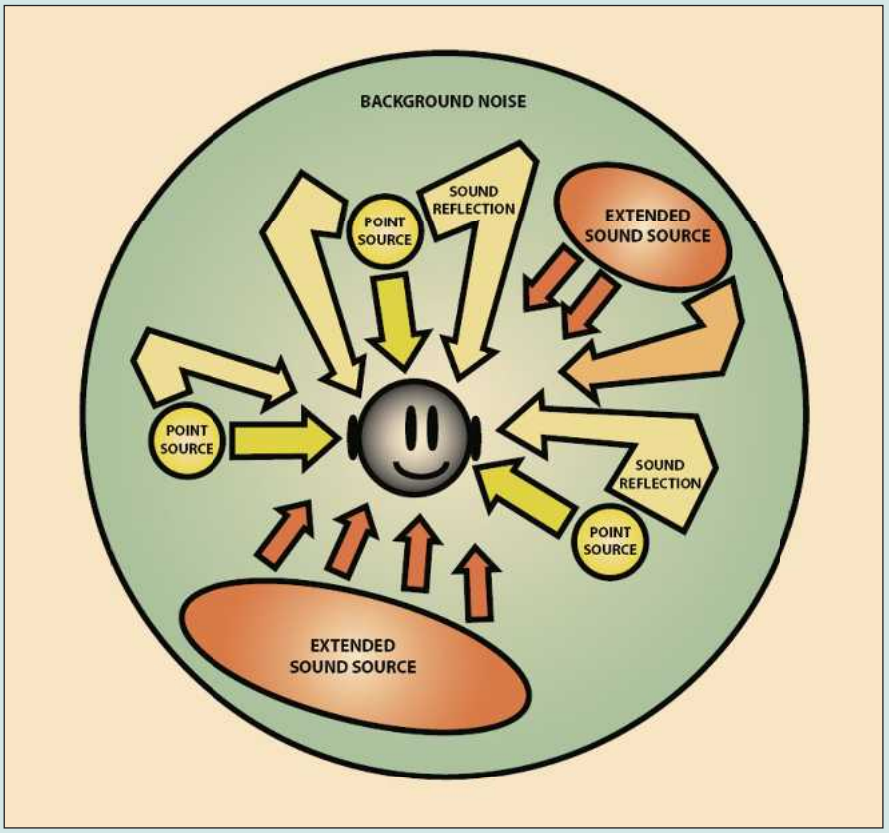


Fig. 8. Typical components of a real-world auditory scene (Figs. 8–10 courtesy Gallo and Tsingos)

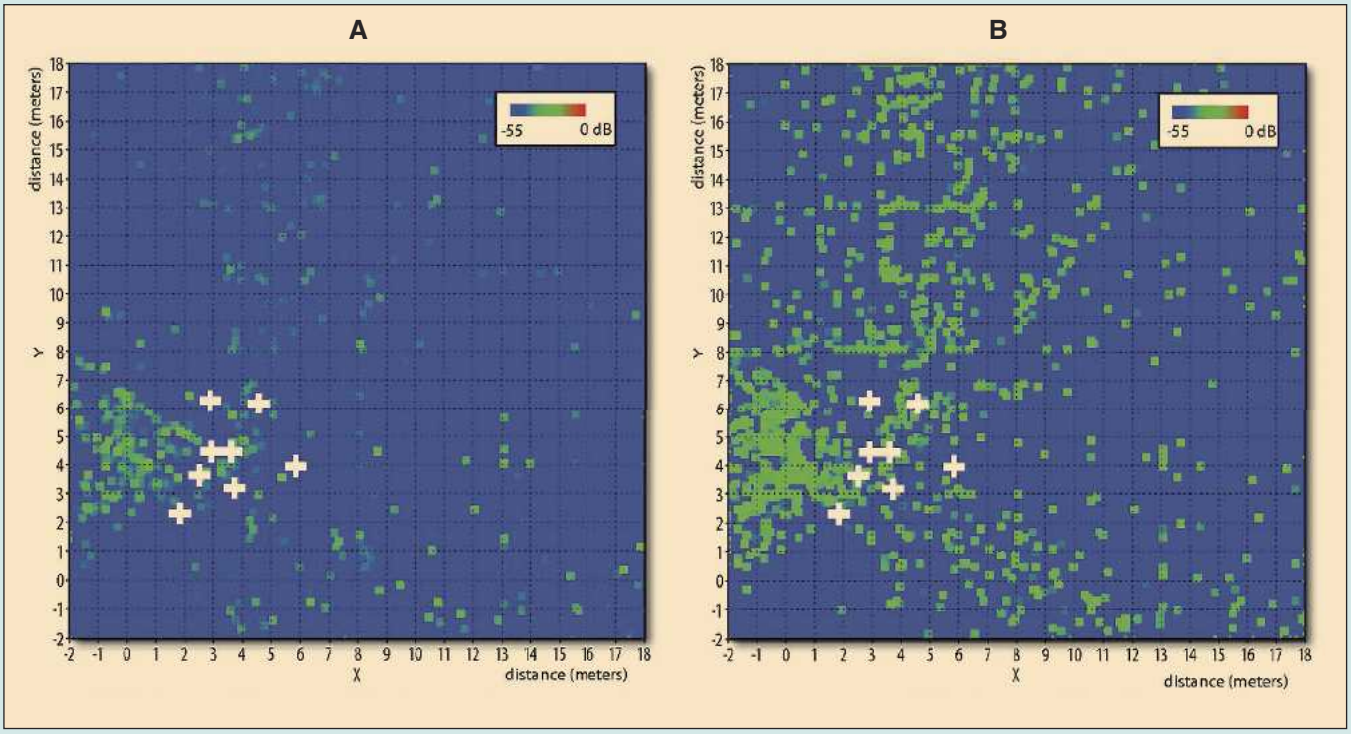


Fig. 9. Energy localization of all subbands integrated over the duration of a seashore recording, showing (A) foreground component only and (B) the unseparated recording. The foreground-only version shows considerably greater sparseness in recognized source locations, owing to the separation from background signals.

with the background component during re-rendering. Sources in the scene are separated partly by assuming that the scene is relatively sparse in the time–frequency plot. In other words, that a single time–frequency element represents information from only one source, and that sources do not overlap substantially. It was found that after background–foreground separation, using the denoising approach just mentioned, there was a greater likelihood of the desired time–frequency sparseness than with the original signal, as shown in Fig. 9.

Rerendering of the scene is undertaken using binaural rendering based on HRTFs of recognized source locations. These are "warped" according to the listener's position in the virtual environment, based on a knowledge of the original microphone positions and their relationship to the listener's position. The closest foreground signal is employed in the rerendering process, as shown in Fig. 10. Foreground events are rendered as point sources whereas background sounds are rendered using low-order spherical harmonic decomposition.

In a subjective evaluation of the system, the authors used a number of different field recordings in three formats. The first format was the new approach described above, using 8 monophonic microphones. This was compared with a reference binaural recording and a B-format version captured using a Soundfield microphone. The spatial analysis of the new system was undertaken in eight subbands, and the final rerendering

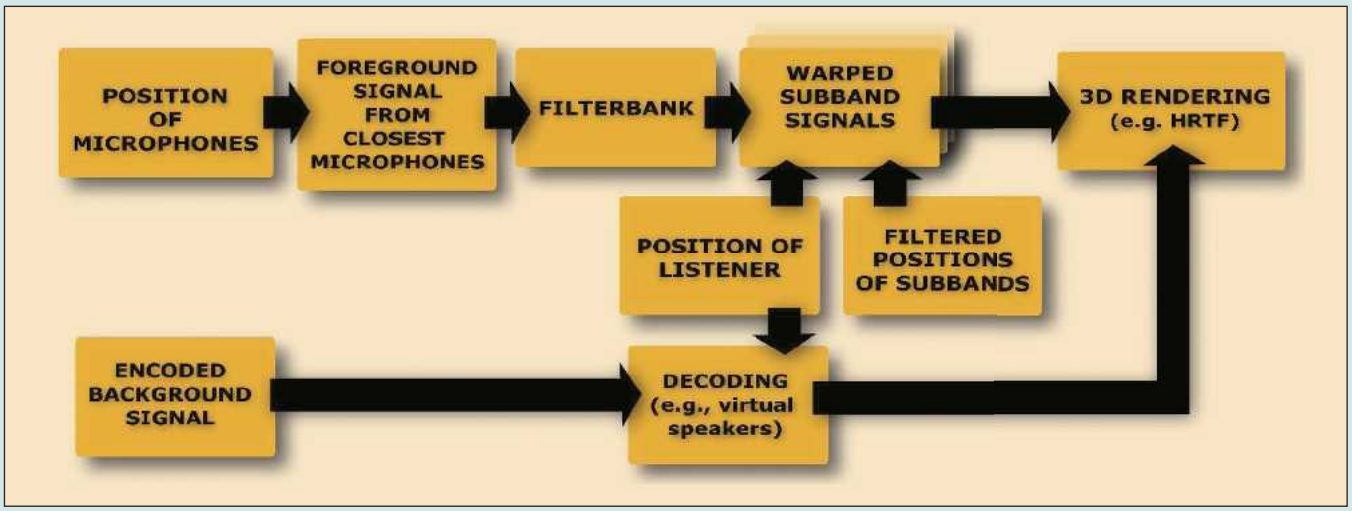


Fig. 10. Block diagram of resynthesis process for structured audio scenes

was undertaken in a binaural form, designed to match the point-of-view of the other two recordings quite closely. The B-format recording was rendered for headphones using a virtual loudspeaker technique, employing HRTFs derived from the LISTEN database developed by IRCAM (http://recherche.ircam.fr/equipes/salles/listen/). This enabled the three versions all to be compared on headphones using forms of binaural rendering. A further version for comparison was created by omitting the background–foreground separation process, considering the entire content to be foreground. Using an approach based on MUSHRA (Multiple Stimuli with Hidden Reference and Anchors), listeners compared the versions using monophonic versions of the recordings as low anchors and the binaural recording as reference. The listeners were instructed to give the lowest possible score to the signal with the worst spatial degradation. General timbre differences between the systems were supposed to be ignored. It was found that the original foreground-only (non-separated) approach generally performed best in terms of spatial quality, although the new approach worked well in terms of overall audio quality. The foreground-only approach suffered from some artifacts with fast-changing sources (that is those whose subband location changes rapidly with time), which were not present so much in the foreground–background approach because of the smoother and slower changing spatial cues in the low-order background signal, which might have masked the foreground. The new approach was rated only slightly better than the B-format recording.

The authors suggest that the use of head tracking and individualized HRTFs might improve the results, as might some improvement in the quality of the segmentation and the use of energy differences between the microphones as opposed to time differences for extracting background localization information.

CONCLUSION

The key themes in all of the developments summarized in this article seem to be "smart" and "adaptive," in other words audio applications are becoming aware of their context and of the user's location or requirements. Flexibility seems to be another watchword, with systems needing to be capable of adapting to suit physical changes in system setup, distributed or wireless connection, and mobile operation. Intelligent audio environments seem poised to become a reality, possibly within the next ten years.

Editor's note: The papers reviewed in this article, and all AES papers, can be purchased online at www.aes.org/publications/preprints/search.cfm and www.aes.org/journal/search.cfm. AES members also have free access to a large number of past technical review articles such as this one and other tutorials from AES conventions and conferences; go to www.aes.org/tutorials/.