Audio for Wirelessly Networked Personal Devices

AES 43rd International Conference

September 29 to October 1, 2011

The AES 43rd International Conference, *Audio for Wirelessly Networked Personal Devices*, was held at the POSCO International Center at Pohang University of Science and Technology (POSTECH), Pohang, Republic of Korea, September 29 to October 1, 2011. Participants from around the world listened to 18 paper presentations, including the keynote lecture, four invited lectures, and a closing panel discussion.

The conference focused on state-of-the-art audio technologies for wirelessly networked portable devices. It brought together experts in both audio and systems engineering to discuss recent developments in this area. Very broadly, the conference paper topics covered audio compression, immersive audio, audio over networks, and systems-level issues in networked portable devices.

The conference was held at the POSCO International Center, which is on the campus of Pohang University of Science and Technology (POSTECH). Pohang became a major industrial center after steelmaker POSCO choose to locate its first mills in Pohang in 1972. Subsequently, growth of heavy industry in the city brought to the local economy to a blend of iron, steel, shipbuilding, and fisheries by the end of the 20th century. POSTECH has strong programs in electrical engineering and communications, including the Educational Institute of Future Information Technology, and has the unique Graduate Institute of Ferrous Technology, the only graduate program for the study of steel science and technology.

DAY 1

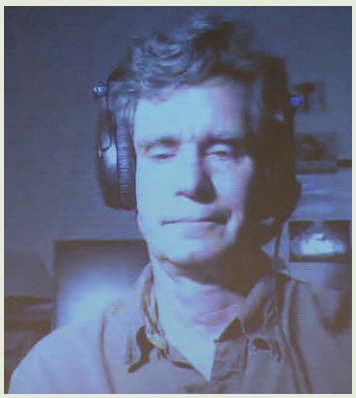
The conference cochairs, John Oh of Pulsus and Kyungwhoon Cheun of POSTECH, opened the conference and welcomed the attendees. John Oh noted that this is the third in a series of conferences on this topic. Over this time span there has been significant change in the landscape of portable devices, which are now more computationally capable and have relatively high-bandwidth wireless channels.

The technical program began with the keynote lecture, "MPEG Unified Speech and Audio Coding," delivered by Schuyler Quackenbush of Audio Research Labs. Unified speech and audio coding (USAC) is the newest MPEG audio standard, published in late 2011. It is able to achieve consistently state-of-the-art compression performance for any mix of speech and music content. Quackenbush noted that USAC exploits both models of sound perception and models of speech production in order to achieve a very high level of compression. He gave an overview of the architecture of the USAC algorithm and how the various compression tools operate in response to the instantaneous statistics of arbitrary mixed-content signals. The architecture combines transform-coding tools from MPEG Advanced Audio Coding (AAC), transform-coded-excitation (TCX) tools, analysis-by-synthesis speech coding tools, MPEG spectral band replication (SBR), and MPEG Surround spatial audio coding tools. He briefly described the tools that give the greatest compression performance. He presented the results of a comprehensive subjective listening test showing that the new standard performed better

Conference cochairs John Oh (left) and Kyungwhoon Cheun

Schuyler Quackenbush, keynote speaker

than either of the state-of-the-art benchmark coders—HE-AAC v2 and AMR-WB+—over the entire range of tested bit rates of 8 kb/s for mono to 96 kbps for stereo. In the USAC "sweet spot" of 16 to 32 kbps stereo, this performance advantage was large. This superior performance was maintained regardless of the type of content coded, e.g., speech, music, or a mix of speech and music.



Invited speaker Chris Chafe addressed the attendees via a teleconference link.

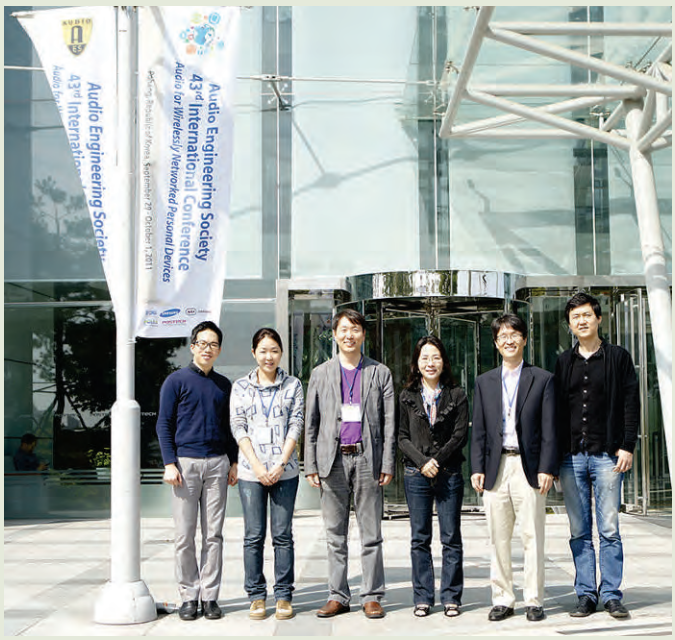
Following the keynote, the first paper session was on interactive audio.

The first paper in the session was the invited paper "Living with Net Lag," presented by Chris Chafe, cofounder of MusicianLink and director of Stanford's CCRMA. This talk was delivered via video conference, with the presenter at Stanford University. His talk opened with a video of an example application in which a "garage band" used laptops to collaborate from geographically disparate locations via networked audio and video links. He then showed a "jamLink" hardware module that can compress up to four channels of audio and connect directly to the user's network router, thereby achieving the lowest possible latency in the audio link. Chafe emphasized that a high-quality experience imposed several requirements on the networked audio: that the total one-way audio latency be less than 25 ms, that the network throughput be greater than 1 Mbps, that the audio compression scheme (if any) have low latency, and that the delivery of packets over the network have low jitter. The system supports both audio and video, but does not attempt to synchronize audio and video streams. Rather it delivers audio with the lowest possible latency and video with much higher latency. Several examples of collaborative music sessions were given. Finally, Chafe noted that while tightly-synchronized "jamming" does require less than 25 ms latency, other types of collaborative music may be amenable to much higher one-way latency, permitting a much wider geographic area for collaboration.

The next paper was, "Symphony Orchestra Recording for Interactive Opera Performances," presented by Lars Hallberg of Lulea University of Technology. The presenter gave a brief overview of music studies at Lulea University. One recent project at the university was the creation of an opera; the music of the opera, performed as an interactive opera or iOpera, was discussed in the paper. The concept of iOpera is that the sound from each section of symphonic instruments and the vocalists can have its level adjusted interactively according to personal preference. In order to create the audio signals to support such interactivity, the symphony was recorded using relatively few microphones with an emphasis on capturing both the instrument sound and the associated hall reverberation. This was done by recording each section of instruments in the symphony (e.g. violin, viola, cello, woodwinds) separately and thus having the component signals available to be manipulated interactivity. The result was a rich set of symphonic signals that could be the basis for many Internet-based interactive presentations.

The second session in the conference "Next-Generation Audio Coding ," was the first of two sessions on this topic.

The first paper was, "Enhanced Stereo Algorithms in Unified



Conference committee and staff in front of POSCO

Speech and Audio Coding," presented by Eunmi Oh, Samsung. This built on the Keynote presentation by describing the stereo coding components present in Unified Speech and Audio Coding (USAC). In USAC, stereo coding tools are found in the frequency-domain (FD) coding tools and again in the MPS212 sound-stage coding tool. The presenter noted that, by her estimation, nearly one-third of the core-experiment activity in USAC related to improvements in stereo coding. The result is that USAC has significant improvements in coding performance for the stereo signal. In the lowest range of 16 to 24 kbps MPS212, with its parametrically coded stereo phase information can deliver high stereo quality in a very bit-efficient manner. In the mid range of 48 to 64 kbps, a new unified stereo scheme predicts the residual signal from the downmix (i.e. sum) signal. This reduces the energy in the residual and hence the bits needed to code it. Finally, at higher bitrates, where the core coder is operating at the full sampling rate (i.e. no SBR) with a full-band residual, complex stereo prediction in the MDCT domain can be used with a novel real-to-imaginary transform. This is in addition to the conventional Mid/Side stereo coding. Over a broad range of bit rates, the newly developed methods in the USAC give bit-efficient stereo coding with little additional complexity resulting in excellent quality for any audio content.

The second paper was, "LPD Single Mode MPEG-D USAC Technology," presented by Keunwoo Choi, ETRI. The presenter reviewed the goals of the MPEG Unified Speech and Audio Coding (USAC) work and noted that the USAC Frequency-Domain (FD) coding mode is best for music-like signals and that the LPD mode is best for speech-like signals. Since the USAC TCX windows with "flat tops" have higher spectral side-lobes as compared to a sine window, the paper proposes a modified architecture using only sine windows in the TCX tool. This modification does not alter the overall one-way latency and does not increase the computational complexity of the proposed system as compared to that of the USAC standard. Subjective quality tests indicate that the quality of the proposed scheme is not different from that of the USAC standard.

The third paper was, "Discrimination Module for Voice/Audio Signals Based on Wavelet Ridges Analysis," presented by Daniel Sauceda, Metropolitan Autonomous University, Mexico City. The

Daniel Sauceda (left) and John Richards engage the audience on Day 2.

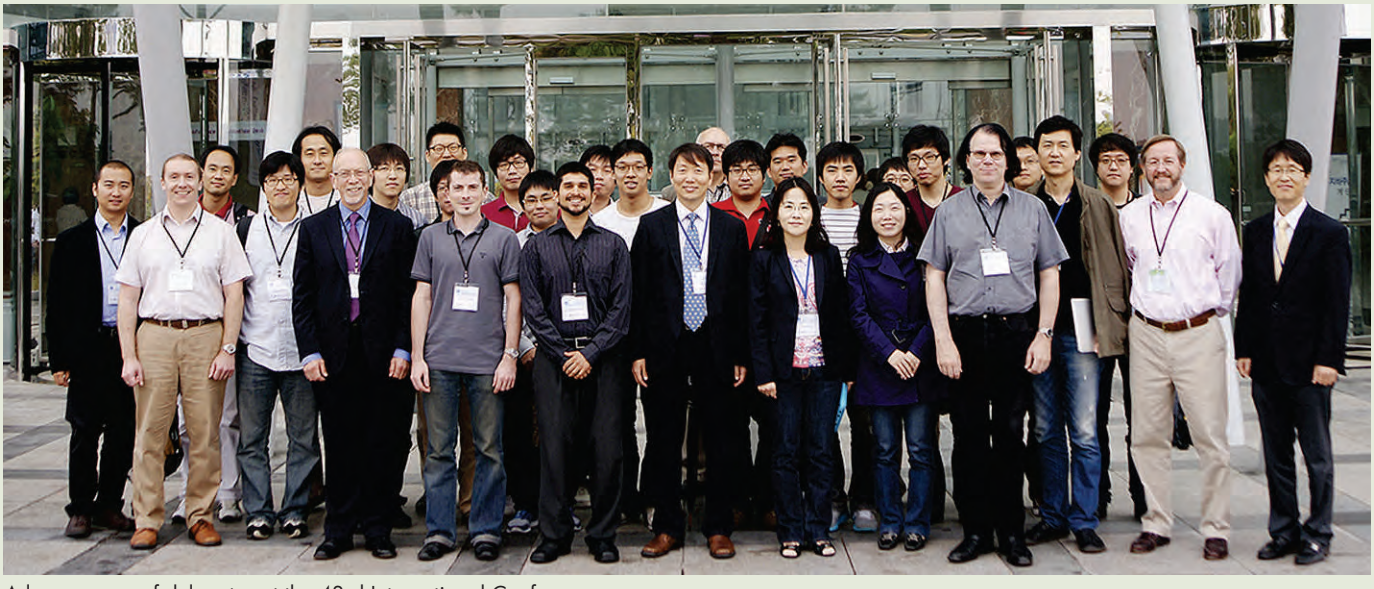
paper presented a method for categorizing the type of input signal for the purpose of configuring the set of encoder tools in Unified Speech and Audio Coding (USAC), i.e. for selecting the TCX or ACELP coding modes. The method is based on the continuous wavelet transform using a Morlet complex wavelet. Only the low band of the input signal (up to 1200 Hz) was considered. The discrimination criterion involved identifying "wavelet ridges" in the transform domain, and a signal is classified as "speech" if the ridges have more than a minimum time duration, less than a maximum frequency and less than a maximum standard deviation over frequency. When analyzing a large set of signals, the algorithm was able to correctly classify a signal as speech in 60% of the cases.

In the next session the companies that have brought "Industrial Solution" demonstrations gave a brief introduction to their technology.

John Richards, Oxford Digital, gave an overview of two Oxford products and applications. The first, Sonic Tuning Solutions, is used to remove loudspeaker and cabinet resonances, extending bass frequency response, dynamic range control, and increasing signal loudness under the constraints of low power consumption. The second, Audio DSP Core and Toolset, address product production workflow and permits changes in algorithms on programmable processors even as products enter the production phase. The DSP can be programmed using various methods, ranging from source code (e.g. assembly or C-code) to using optimized graphical block diagram "linkers."

The day was concluded with the session "Immersive Audio," which was the first of two sessions on this topic. The first paper was "Acoustic Measure of Causality Artifacts for Generating Focused Source," presented by Min-Ho Song, Korea Advanced Institute of Science and Technology (KAIST). He presented the concept of a focused source, which is a virtual sound source located inside a listening area bounded by loudspeakers. The focused source can only be perceived by a listener located in the "diverging region" which is beyond the location of the virtual source with respect to the generating loudspeakers. Causality artifacts are undesired wavefronts that arrive at the listener's ear before those of the focused source. The paper proposes an acoustic measure that can be used to quantify the amount of causality artifacts for a given virtual source and listener position. Audio rendering could be via higher order ambisonics (HOA), wavefield synthesis (WFS) or Acoustic Focusing. The acoustic measure uses a "threshold of arrival time difference" which is the difference in wavefront arrival times between the virtual source and an artifact wavefront. The acoustic measure of causality artifacts, Q, is defined as the ratio of energy of the virtual source wave to the artifact wave. Simulations were presented that gave the value of Q for virtual sources generated by linear arrays of 11 and 21 loudspeakers and circular arrays of 70 loudspeakers.

The second paper was "Two-Band Approximation of Virtual Sound Imaging Systems," presented by Young-cheol Park, Yonsei University. Virtual sound imaging is simply how to localize an arbitrary virtual sound in a virtual space. One method to localize a virtual sound is to use pair-wise panning with stereo loudspeakers, where the amplitude difference between loudspeakers is equivalent to inter-aural time differences (ITD). A second method is to use head-related transfer function (HRTF) based panning. The paper proposes a trade-off between pair-wise panning and HRTF panning by treating the panning problem in frequency bands: in the lowest band, (below 700 Hz) use pair-wise amplitude panning and the highest band (1000 to 5000 Hz) use new method based on an inter-aural level differences (ILD) and ITD model. The intermediate band uses a "cross-faded" linear combination of the two methods. It was found that, for localization, the proposed model was better than amplitude panning and comparable to the HRTF-based method,



A large group of delegates at the 43rd International Conference

however, in terms of sound quality, the proposed model was much better than HRTF-based panning. The proposed method uses only a few parameters, which leads to a low-complexity implementation.

The final paper in the session was, "Adaptive Crosstalk Cancellation Using Common Acoustical Pole and Zero (CAPZ) Model," presented by Hanwook Chung, Seoul National University. The author presented the crosstalk cancellation problem in the context of stereo loudspeakers and the conventional LMS solution to the problem. Next, the Common Acoustical Pole and Zero (CAPZ) model was presented. In the CAPZ model for HRTF, the zeros describe the acoustical propagation path and common poles describe the characteristics of the ear's auditory system. Based on simulations, the proposed model provides better crosstalk cancellation and shows lower contralateral response. The proposed model also provides a linear phase response.

Lars Hallberg asks a question during one of the conference workshops.

Young-Cheol Park delivers his paper on a virtual sound imaging system.

DAY 2

The first session of the day was devoted to immersive audio. The opening paper was an invited paper, "Spatial Equalizer—Design and Implementation," presented by Yang-Hann Kim of Korea Advanced Institute of Science and Technology (KAIST). The presenter began by proposing that a spatial equalizer is a device to manipulate sound in space and time and that "realistic sound" is reproduced sound that can not be distinguished from the original live sound. He gave several examples of manipulating sound in space, for example creating a very focused "private" audio listening zone. Several methods of controlling a set of loudspeakers for the purpose of sound manipulation were reviewed. Wavefield synthesis (WFS) is one means, and the general WFS case is amenable to simplification from a surface of loudspeakers to a line of loudspeakers. However, WFS is subject to several drawbacks due to the simplifications, such as frequency aliasing and effects due to array length truncation. Another means is higher-order ambisonics (HOA). This method reproduces the pressure at a point in space as a sum of spherical harmonic basis functions. The greater the number of HOA modes, the larger the "sweet spot" in the reproduction system. A third method is "acoustic brightness and contrast control." This is inherently an ill-posed problem, but can have a unique solution if additional constraints are imposed. Each method and the desired manipulation can lead to preferred loudspeaker arrangements, for example, linear array, circular array, or arbitrary locations. Given that there are a number of methods, the next question is: how does one evaluate which method is "best," where evaluation could be some objective measurements or a subjective preference. If one goes the route of subjective preference, then it may be appropriate to impose on the sound generation method a "spatial equalizer" that is tuned to personal preferences. Investigating such a spatial equalizer is Kim's research problem. He concluded with an overview of his laboratory for investigating these issues.

Invited speaker Yang-Hann Kim discusses spatial equalizer design.

The second paper in the session was another invited paper, "The 22.2 Multichannel Sounds and its Reproduction at Home and Personal Environment," presented by Kimio Hamasaki of the NHK Science and Technology Laboratory. This presentation was via video conference with the presenter in London. NHK is keenly interested in a more immersive presence in audiovisual presentations. For video, immersion is directly related to the viewing angle subtended by the visual display. A good impression of visual immersion results from at least a 120-degree display field of view. For audio, a three-dimensional soundfield is essential, with stable and precise sound localization in all directions. Hamasaki noted that a stereo pair of loudspeakers can create a virtual center sound source, but as the separation between speakers becomes greater, the virtual source becomes less stable. Adding a physical center speaker can stabilize the sound image. This argues for a large number of loudspeakers in a 3-dimensional arrangement. NHK has conducted subjective tests that suggest that in the horizontal plane at least 8 loudspeakers are required for good image stability across all frequencies. Tests of vertical sound localization suggest that the lateral direction is most discriminating. Front and back discrimination is much less. Such experiments led NHK to propose a 22.2 channel loudspeaker configuration as optimal for its Super-High Vision (SHV) system. Hamasaki showed an NHK "speaker array frame" video display in which there are a large number of loudspeakers incorporated into the peripheral of the visual display, which permits a large number of loudspeakers to easily be accommodated into the home environment. Finally, Hamasaki discussed the problem of downmixing from M to N channels (e.g. from 22.2 to 5.1 channels). He proposes to use binaural reproduction via headphones for personal devices, including the possibility of a listener's exact head-related transfer functions (HRTF).

The next session was on implementations. The first paper was "Approximation of a Virtual-Reverberation Filter for Handheld Devices," presented by Kwang Myung Jeon of the Gwangju Institute of Science and Technology (GIST). The motivation for the paper is to reproduce the reverberation present in real acoustic spaces using a low-complexity method appropriate to hand-held devices. Complexity is reduced by means of an approximation of the reverberation filter and via indexed-based convolution. A combination of a constant value and a decaying exponential is used as a threshold, and impulse response values below the threshold are reassigned to zero. A table of indices associated with nonzero values is used to implement the convolution. In general, the fixed threshold results in retention of early reflections while the subsequent decaying exponential selects components of the reverberant sound. Jeon gave an audio demonstration of the reverberation with various levels of simplification. The paper gives results of a MUSHRA

subjective quality evaluation and complexity evaluation for the various levels of simplification. The subjectively preferred simplified method reduced the complexity by several orders of magnitude relative to the full impulse response convolution.

The second paper in the session was "System Approach to Avoid Audio Amplifier Oversizing in Mobile Phone Applications," presented by Eric Sturtzer of the Lyon Institute of Nanotechnology (INL). He noted that conversion from digital signal to acoustic energy in a hand-held device requires two groups of experts: analog electronics experts and acoustic experts. Typical mobile phone designs achieve no more than 0.01% power use efficiency, and amplifier power efficiency is not optimized at nominal output power levels. The paper proposes a process to find a globally optimal solution to the audio output problem that maximizes quality and minimizes power consumption. The acoustic conversion efficiency is typically no more than 0.0006% for headphones and 0.01% for hands-free loudspeakers. The electrical conversion efficiency is approximately 70% for class A/B linear power amplifiers (which are most appropriate for headphones) and more than 90% for class D power amplifiers (which are most appropriate for loudspeakers). By using the proposed global approach to system design and by being less strict on some electrical specifications, such as THD+N level or the frequency range of the audio amplifier, it is possible to improve the nominal system efficiency without significant reduction in audio quality.

The final paper in the session was "Audio and Control: Simulation to Embedded in Seconds," presented by John Richards of Oxford Digital. He gave an overview of "graphical design" tools in which the graphical building blocks are all "prevalidated." The implication of this is that validation is at the system level (i.e. of the system-level connections between blocks) and can be done in the simulation environment. No subsequent validation is needed at the realization level (e.g. at the source code or hardware level). This permits rapid workflow from product specification, design, redesign, to realization.

The afternoon of Day 2 attendees took a break from the technical sessions for an excursion followed by the conference social event. For the excursion, the bus first took the group to the Bulguksa Temple in the foothills of Toham-san (Toham Mountain). The temple is considered a masterpiece of the golden age of Buddhist art in the Silla kingdom and dates from eighth century, with construction completed by Kim Daeseong. Most notable are its two stone pagodas (Dabotap and Seokgatap), stone bridge staircases (Cheongun-gyo), and two gilt-bronze statues of Buddha. Next was a visit to Seokguram Grotto, a man-made grotto containing a stone Buddha statue. Legend has it that Bulguksa Temple was dedicated to Kim's parents in his present life while the Seokguram Grotto was dedicated to Kim's parents from a previous life. It is very popular to view sunrise over the sea from the grotto entrance. Bulguksa and Seokguram Grotto are UNESCO World Heritage sites. The next stop was a visit to the Cheonmachong Tomb, one of several very large earthen tomb mounds from the fifth or sixth century. This tomb was excavated, permitting visitors to see the wooden tomb chamber and the various precious artifacts that were buried with the king. As sunset approached, the group visited Anapji, part of the palace complex of ancient Silla. It was constructed in the seventh century. The site consists of several ornate pavilions on the shore of a beautifully landscaped lake. As the sun set, the group was able to snap photos of the pavilions perfectly reflected in the lake.

The banquet, held at the Gyeongju Hilton, featured a delicious traditional Korean meal and a performance of traditional music and dance.

DAY 3

The first session on Day 3 was on interactive Audio. The session started with the invited paper "Portable and Networked Devices for Musical Creativity," presented by Juha Backman of Nokia. He noted that musical performance as both an individual or a group activity is highly valued in any human society. Technological advancements that facilitate cooperative music creation in which artists are not physically colocated can open new avenues of collaboration between different musical cultures. The key challenge of distributed performances over any network, long-distance or local, is latency. Studies of performance in acoustic spaces indicate that sound from fellow musicians must reach a performer within 30 ms to provide support for ensemble playing. Longer delays cause confusion, even for experienced performers in acoustical spaces. Even when latency issues are addressed, music collaboration platforms should be accessible to persons having a range of musical experience. New consumer platforms such as tablets and smartphones offer great promise in that they have built-in network connectivity and can offer intuitive interfaces.

The second paper in the session was "Gaussian Mixture Model for Singing Voice Separation from Stereophonic Music," presented by Keun-woo Choi of ETRI. The paper proposed an adaptive prediction method for binaural cues such as interchannel level difference (ILD) and interchannel phase difference (IPD) for signals having centrally positioned vocals. Based on the assumption that the target source has a specific position in the stereophonic soundfield, such as centrally positioned vocals, binaural cues of input mixture signals were clustered using GMM. Experimental results on commercial music showed improvement in separation performance as compared to ordinary hard-decision methods.

The next session was on next-generation audio coding. The first paper was "Enhanced Interchannel Correlation (ICC) Synthesis for Spatial Audio Coding," presented by Dong-il Hyun of Yonsei University. He noted that in spatial audio coding, interchannel correlation (ICC) may be estimated as the real part of the normalized crosscorrelation coefficient between two channels and thus can result in a negative value. Conventional methods assume that ambient components mixed to two output channels are in antiphase, while the primary signals are assumed to be in phase. When a negative-valued ICC is encountered, this assumption can cause excessive ambient mixing. The paper proposes to solve this problem with a new ICC synthesis method based on an assumption that the primary signals are in anti-phase when negative ICCs are indicated. The solution uses an upmix matrix that satisfies the assumption for the primary components in a negative ICC environment. The effectiveness of the proposed method was verified by computer simulations and subjective listening tests.

The second paper of the session was "A Unified Coding Approach for Wireless Audio Streaming Between Networked Personal Devices," presented by David Trainor of Cambridge Silicon Radio. This paper discussed the use of dynamic composition and adaptation of a

Invited speaker Juha Backman introduces the topic of musical creativity on portable networked devices.

Delegates visit the Bulguksa Temple in the foothills of Toham-san.

base library of signal-processing functions as a means to provide a scalable and unified audio codec for real-time wireless audio streaming that can be practically implemented on networked personal devices. For example, a scalable codec could entail dynamic and joint optimization of the following audio codec performance characteristics based on the performance priorities and constraints supplied by the wider system: coded audio quality, coded audio bit rate, audio coding delay, codec computational complexity, and coded audio robustness to network errors or other data loss.

The final event of the conference was the workshop "Audio in Future Networked Personal Devices." All the invited speakers and the attendees participated in the discussion. John Oh addressed end-user benefits such as new experience, high quality, and easy accessibility that should be provided by future technologies. He also stated the importance of experiential innovation and interoperability of future mobile devices. Juha Backman addressed the possible use of networked mobile terminals for collecting and representing sound scenes. Young-cheol Park remarked that the technologies we have now might be enough to find innovative applications, and engineering creativity would be more important than technological development. Lars Hallberg mentioned ideal environments for creation of interactive opera, and other attendance shared their thoughts on future of immersive audio environments at home and office.

The success of this third AES conference on the topic of wireless personal devices and the explosive growth in the use of these devices worldwide make it a good bet that there will be a fourth conference in the not-to-distant future.

Editor's note:The CD-ROM of conference papers can be purchased at www.aes.org/publications/conferences. Individual papers can be purchased at www.aes.org/e-lib.