

AES 43rd INTERNATIONAL CONFERENCE



Audio for Wirelessly Networked Personal Devices

Pohang, Republic of Korea

2011 September 29 – October 1

Technical Sessions

This preliminary program is accurate as of press time. See updates at www.aes.org/events/43

Thursday, September 29

09:30

OPENING REMARKS AND KEYNOTE SPEECH

[Keynote] MPEG Unified Speech and Audio Coding—
Schuyler Quackenbush, Audio Research Labs, Scotch
Plains, NJ, USA

Unified Speech and Audio Coding (USAC) is the newest MPEG audio standard, published in late 2011. It achieves consistently state-of-the-art compression performance for any mix of speech and music content. USAC incorporates several perceptually-based compression techniques developed in previous MPEG standards: perceptually shaped quantization noise, parametric coding of the upper spectrum region, and parametric coding of the stereo sound stage. However, for the first time in MPEG it combines the well-known perceptual techniques with a source coding technique: a model of sound production, specifically that of human speech. The paper gives an overview of the architecture of the Unified Speech and Audio Coding algorithm and a brief description of the tools giving the greatest compression performance. Finally, it presents results of subjective listening tests showing the performance of the standard relative to state-of-the-art benchmark coders.

Thursday, September 29

11:00

PAPER SESSION 1: INTERACTIV AUDIO—PART 1

1-1 **[Invited] Living with Net Lag—***Chris Chafe*, Stanford
University, Stanford, CA, USA

Internet latency can be ignored, tolerated or exploited in making music together at a distance. The premise that these are distinct ways of relating to lag is examined in case histories of recent projects. Scores of examples of split ensembles collaborating remotely in real time can be cited from the last few years. Five scenarios from this musical world have been selected to look closely at music being made across networks and the differing relationships to lag. Each involves large or multi-site en-

sembles. The first four represent academic/ contemporary idioms (involving jacktrip software, and advanced university networks) and the last is a working band that uses commodity tools to rehearse pop music covers (using jamLinks and standard residential networks).

1-2 **Symphony Orchestra Recording for Interactive Opera Performances—***Lars Hallberg, Jan Berg*, Luleå University of Technology, Luleå, Sweden

Network technology and its applications open the possibilities for conventional art forms to find alternative ways of expressions leading to new user experiences. A significant feature is interactivity, which enables the user to become a co-creator of the experience. One of the more traditional art forms is the opera, which thus can be enhanced and developed. In order to provide content that allows for interaction, specific strategies and techniques have to be utilized from the start of the production. When user influence over a symphony orchestra sound is desired, the different parts of the orchestra have to be recorded separately. In this paper considerations regarding the recording process involving a real full-scale symphony orchestra are reported. It is concluded that such a production requires either compromises in the recording process or new ways of doing the recordings in order to fully utilize the potential of interactivity.

Thursday, September 29

13:15

PAPER SESSION 2: NEXT-GENERATION AUDIO CODING—PART 1

2-1 **Enhanced Stereo Algorithms in the Unified Speech Audio Coding—***Eunmi Oh, Miyoung Kim*, Samsung Electronics, Suwon, Korea

The unified speech and audio coding (USAC) is under the ISO/MPEG standardization and is supposed to complete the standardization process in 2011. In the perspective of technical merits that the USAC brings to

audio and speech compression, this paper highlights stereo algorithms that perform best depending on bit rates and audio content types in various use case scenarios. At very low rates (e.g., 16 to 48 kbits/sec), MPEG Surround tool with phase parameters can be used to deliver high stereo quality in a bit-efficient way. At the mid range of bit rates such as 48 and 64 kbits/sec, a prediction-based phase coding can be utilized with residual signals with little burden of complexity. Finally, at higher bit rates, a complex stereo prediction in MDCT domain can be used through the real-to-imaginary transform in addition to the conventional Mid/Side stereo. Over broad range of bit rates, the newly developed methods in the USAC result in excellent stereo quality encompassing a variety of audio contents in a bit-efficient way with little cost of complexity.

2-2 LPD Single Mode MPEG-D USAC Technology—Taejin Lee,¹ Seungkwon Beack,¹ Kyeongok Kang,¹ Whan-Woo Kim²

¹ETRI, Electronics and Telecommunications Research Institute, Daejeon, Korea
²Chungnam National University, Daejeon, Korea

As mobile devices become multi-functional, and multiple devices converge into a single device, there is a strong market need for an audio codec that is able to provide consistent quality for mixed speech and music content. To satisfy this, a process was initiated by ISO/MPEG, aiming to standardize a new codec with consistent high quality for speech, music, and mixed content over a broad range of bit rates. In this paper we propose LPD single mode MPEG-D USAC structure that could provide consistent quality compared with current FD/LPD dual mode MPEG-D USAC. For the evaluation of the new codec architecture, we followed MPEG audio listening test procedures. Three categorized (music, speech, mixed) items are used and audio experts are joined for the evaluation. The listening test results show that the objective and subjective performance of proposed single mode USAC system is statistically comparable quality with that state of the art FD/LPD dual mode MPEG-D USAC.

2-3 Discrimination Module for Voice/Audio Signals Based on Wavelet Ridges Analysis—Daniel Saucedo-Peña, Alfonso Prieto-Guerrero, Universidad Autónoma Metropolitana-Iztapalapa, México D.F., México

Low bit-rate at high quality perception is the aim of coding schemes. Traditionally audio and voice coders have evolved as different paradigms: the state of the art voice-coders cores resides in Algebraic Code Excited Linear Prediction (ACELP) technologies while audio coding has its core in Transform Coding (TC). The Unified Speech-Audio Coding (USAC) scheme has become a new paradigm where the principal goal is to choose between the ACELP or TC to reduce the bit rate and increase the high quality perception. This modern coder is based in a module that decides which core coder to use on a specific signal frame. This paper proposes a decision module based on ridges detection in the wavelet transform of the input signal. Wavelet ridges permit tracking of the instantaneous frequencies contained in the analyzed signal. These instantaneous frequencies, linked to the signal pitch and its harmonics, permit the establishment of a module for determining whether it is a voice signal or audio.

Thursday, September 29

14:45

INDUSTRIAL SOLUTION INTRODUCTION

Organizer: Hyun-o Oh

Participants: Samsung Electronics, LG Electronics, and audio solution providers

Mobile device manufacturers including Samsung and LG Electronics, and audio solution providers will introduce their latest developments of audio technologies. Exhibition and demo booths will be located in the conference site.

Thursday, September 29

15:45

PAPER SESSION 3: IMMERSIVE AUDIO—PART 1

3-1 [Invited] The 22.2 Multichannel Sounds and its Reproduction at Home and Personal Environment—Kimio Hamasaki, Kentaro Matsui, Ikuko Sawaya, Hiroyuki Okubo, NHK, Tokyo, Japan

A 22.2 multichannel sound system has been developed for a sound system of Super Hi-Vision (SHV) by the Japan Broadcasting Corporation (NHK). SHV comprises 7680 x 4320 pixels (16 times the total number of pixels of an HDTV) and 60 Hz frame rate with progressive scanning. SHV also employs the 22.2 multichannel sound system consisting of vertical three-layers-layout channels including 9 channels at top layer, 10 channels at middle layer, and 3 channels at the bottom layer with 2 Low Frequency Effects (LFE) channels. The 22.2 multichannel sound system provides three-dimensional sound that enables viewers or listeners to feel as if they were in a realistic world of three-dimensional sound. SHV has been currently developed for a next-generation TV for home and personal viewing environment. This paper reviews the 22.2 multichannel sound system and proposes various sound reproduction systems for home and personal hearing environment.

3-2 Two-Band Approximation of Virtual Sound Imaging Systems—Jae-woong Jeong,¹ Tacksung Choi,² Se-Woon Jeon,¹ Young-cheol Park,³ Dae-hee Youn,¹ Seok-Pil Lee⁴

¹Yonsei University, Seoul, Korea

²LG Electronics Inc., Seoul, Korea

³Yonsei University, Wonju-city, Kangwon-do, Korea

⁴Korea Electronics Technology Institute (KETI),

Bundang-gu Seongnam, Gyeonggi Province, Korea

This paper presents an approximation model for virtual source imaging systems based on crosstalk cancellation technique. The proposed model provides simple but very effective approximations in low- and high-frequency bands. At low frequencies, the approximation results in the well-known amplitude panning. On the other hand, the high-frequency approximation is expressed using representative interaural level and time differences. The effectiveness of the model for various applications was confirmed via computer simulations and listening tests.

3-3 Adaptive Crosstalk Cancellation Using Common Acoustical Pole and Zero (CAPZ) Model—Hanwook Chung, Sang Bae Chon, Nara Hahn, Koeng-Mo Sung, Seoul National University, Seoul, Korea

This paper introduces an adaptive crosstalk cancella- ➤

tion that uses a Common Acoustical Pole and Zero (CAPZ) model for a Head Related Transfer Function (HRTF). As the CAPZ model for HRTF provides an interpretation of the HRTF wherein zeros describe the spatial difference caused by the acoustical propagation path and common poles describe the characteristics of human auditory system, we designed the proposed model to follow the zero components of the CAPZ model. Through simulations, it was verified that the proposed model provides enhanced performance compared with a conventional Finite Impulse Response model for HRTF.

Friday, September 30

09:00

PAPER SESSION 4: IMMERSIVE AUDIO—PART 2

- 4-1 [Invited] Design and Implementation of “Spatial Equalizer”**—*Yang-Hann Kim, Min-Ho Song*, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea

Three dimensional sound field (3-D sound field) is what human beings have dreamed of having. It is interesting, however, 3-D sound is not scientifically well defined. Therefore, there is no absolute measure that can determine the quality of 3-D sound. People may want to define his/her own 3-D sound! There have been many attempts that claim some success or practical achievements of implementing a 3-D sound field by using multiple loudspeakers or speaker array. Wave field synthesis, Ambisonics, and what are based on HRTF (Head Related Transfer Function) are some of them. The methods have tried to regenerate the sound field in time and space as they wish. In other words, they wish to make sound field exactly same as what they designed. These approaches, however, ignore what we want to have in a “3-D sound field” at their home or the listening space: ignoring their acoustic sensation. It is noticeable that very primitive stereo systems came with a knob that was simply designed to control the volume of two loudspeakers. The knob, which is generally called a “balance knob,” essentially provided the means that controlled the sound from two loudspeakers so that it meet what we wanted to hear. It is therefore our ultimate goal to have a similar human-sound interface for 3-D sound realization. As for the case of a conventional stereo system, how well the interface achieves this will be evaluated by who really controls the interface. We developed “the knobs” that can really implement the desired 3-D sound in space and time. This objective can be achieved simply by introducing a means that can make sound field that sounds like what user or operator wants. This can be done by attempting to make a point focusing or several focused points in the space of interest. A very simple case can be done by making two focused sound sources in front of us, and one focused sound source along a line passing our head perpendicular to our ear-to-ear line. Using these three points, we are ready to hear the sound field that can be made. This arrangement allows us to have a “spatial equalizer,” analogous to a frequency equalizer or simply an equalizer that has been used for most of the radio or audio system. We select the magnitude of each frequency band and listen until we are satisfied. Very similar things can be done by using the three point focused sound.

- 4-2 Acoustic Measure of Causality Artifacts for Generating Focused Source**—*Min-Ho Song, Yang-Hann Kim*, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea

A focused source is used to present a physical realizable monopole that can be made by a set of loudspeakers. If we attempt to make such a sound source by using a set of loudspeakers, then the sound waves by each loudspeaker converge toward a desired focus point and diverge after having passed the point. The diverging part of the focused source is what is expected to be generated. It could mimic the sound field of a monopole sound source. This makes it possible to have a virtual sound source. However, it is inevitable to have a focused source without undesirable artifacts because of the direct waves from secondary control sources to any listening point. They always arrive before the waves from a focused source arrive. Artifacts can trigger the “precedence effect” and lead listeners to localize virtual sound source in a wrong and undesirable direction. The objective of this paper is to predict the effect of focused source with its artifacts at selected listening region. Assuming that the driving function of secondary sources and array distribution are given for selected location of the focused point, the magnitude of focused source and artifacts are predicted at different listening locations. A novel acoustic measure is defined for such focused source with respect to listening location in order to compare the effect of focused source and artifacts. Finally, a proposed measure used to test simple driving function that can generate focused source and analysis is given.

Friday, September 30

10:45

PAPER SESSION 5: IMPLEMENTATIONS

- 5-1 Approximation of a Virtual-Reverberation Filter for Handheld Devices**—*Kwang Myung Jeon, Hong Kook Kim*, Gwangju Institute of Science and Technology (GIST), Gwangju, Korea

A virtual reverberation process requires excessively high computational complexity to generate a reverberant sound due to the long impulse response of the reverberation filter. Simple approximation by eliminating some impulse responses of a filter is possible to reduce the complexity. However, approximation without considering overall shape of impulse responses could distort the original reverberant characteristics of the given filter. In this paper we propose an approximation method of the reverberation filter without quality degradation by incorporating the overall shape of the filter’s impulse response. In order to demonstrate the effectiveness of the proposed method, we compare the computational complexity of the reverberation filter approximated by the proposed method and that without any approximation. In addition, we compare the quality of sound signals processed by the approximated reverberation filter with that by the original reverberation filter.

- 5-2 System Approach to Avoid Audio Amplifier Oversizing in Mobile Phone Application**—*Eric Sturtzer,¹ Gaël Pillonnet,¹ Nacer Abouchi,¹ Frédéric Goutti²*
¹Lyon Institute of Nanotechnology (INL) UMR CNRS 5270, Lyon, France
²STMicroelectronics, Grenoble, France

This paper underlines the oversizing of audio amplifiers compared to microspeaker performances and proposes some guidelines to solve this problem. The power efficiency of the audio reproduction chain has less than 0.01% mainly due to the low efficient electromechanical conversion of the loudspeaker. Furthermore the ampli-

er power efficiency is not optimized at nominal output power level. Thus, some significant system level improvements are possible. Regarding the audio quality, the loudspeaker performance is almost ten times lower than the amplifier's one. By being less drastic on some electrical specifications (noise, linearity, and frequency range), the amplifier design trade-off could be redefined to target a better nominal efficiency (without disturbing the acoustical performances).

5-3 Audio and Control: Simulation to Embedded in Seconds—*Nathan Bentall*, Oxford Digital Limited, Stonesfield, Oxfordshire, UK

In real-time audio or control applications, algorithms are typically developed using simulation tools such as Matlab/Simulink. Traditionally, this development is followed by a conversion to fixed point, and a labor-intensive manual optimization and conversion to the assembly code of a signal processor or ASIC/FPGA core. The prudent will also run a series of test vector comparisons to check the conversion process. By using custom simulation models, and an automated process of net-list extraction and code generation, these time consuming steps may be eliminated, going from simulation to running embedded DSP code in seconds.

Saturday, October 1 09:00

PAPER SESSION 6: INTERACTIVE AUDIO—PART 2

6-1 [Invited] Portable and Networked Devices for Musical Creativity—*Juha Backman*, Nokia Corporation, Espoo, Finland

The various networks, starting from dedicated local interfaces to the Internet, offer together with personal portable devices vast creative potential for all kinds of musical performance. This paper approaches the technological potential for shared music creation from different points of view: performances using otherwise conventional technology, but over remote connections; the use of wireless technology to ease performances on stage or in studio; and using the potential of personal portable devices to bring connected performances to every user, including non-musicians.

6-2 Gaussian Mixture Model for Singing Voice Separation from Stereophonic Music—*Minje Kim, Seungkwon Beack, Keunwoo Choi, Kyeongok Kang*, Electronics and Telecommunications Research Institute (ETRI), Daejeon, Korea

This paper presents an adaptive prediction method about source-specific ranges of binaural cues, such as inter-channel level difference (ILD) and inter-channel phase difference (IPD), for centrally positioned singing voice separation. To this end, we employ Gaussian mixture model (GMM) to cluster underlying distributions in the feature domain of mixture signal. By regarding responsibilities to those distinct Gaussians as unmixing coefficients of each mixture spectrogram sample, the proposed method can reduce artificial deformations that previous center channel extraction methods usually suffer, caused by their imprecise or rough decision about ranges of central subspaces. Experiments on commercial music show superiority of the proposed method.

Saturday, October 1 10:45

PAPER SESSION 7: NEXT-GENERATION AUDIO CODING—PART 2

7-1 Enhanced Interchannel Correlation (ICC) Synthesis for Spatial Audio Coding—*Dong-il Hyun,¹ Young-Cheol Park,² Seok-Pil Lee,³ Dae Hee Youn¹*

¹Yonsei University, Seoul, Korea

²Yonsei University, Wonju, Korea

³Korea Electronics Technology Institute (KETI), Seongnam, Korea

In spatial audio coding, Interchannel Correlation (ICC) synthesis is implemented in two different ways. One uses both ICC and phase parameters, and the other uses only ICC. In the latter, ICC is estimated as a real part of the normalized cross-correlation 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 anti-phase, while the primary signals are assumed to be in-phase. When a negative-valued ICC is encountered, this assumption can cause excessive ambient mixing. To solve this problem, we propose a new ICC synthesis method based on an assumption that the primary signals are in anti-phase when negative ICCs are indicated. We first investigate problematic cases of negative ICC synthesis in the conventional methods. Later, we propose a new 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.

7-2 A Unified Coding Approach for Wireless Audio Streaming between Networked Personal Devices—*David Trainor, Neil Smyth*, Cambridge Silicon Radio (CSR), Belfast, UK

An area of increasing relevance for developers, manufacturers, and users of wirelessly-networked personal devices capable of real-time transfer of coded audio is the concept of coding scalability. In the most general sense, scalable coding implies the existence of a unified audio coding scheme that is flexible enough to allow different codec performance measures to be traded off to some degree and offers good performance over a wide variety of audio material. This paper discusses the use of dynamic composition and adaptation of a 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.

Saturday, October 1 11:45

WORKSHOP

Audio in the Future: Networked Personal Devices

Organizers: Youngcheol Park, Eunmi Oh

Participants: Invited speakers

Invited speakers will present short talks and have a discussion about the audio in the future networked personal devices.

Saturday, October 1 13:00

CLOSING REMARKS