45th International Conference

Applications of Time–Frequency Processing in Audio

Helsinki, March 1–4, 2012

CONFERENCE REPORT

Ville Pulkki, conference chair, opens the event.

An attentive and enthusiastic audience during one of the technical presentations.

Andrew Goldberg (center), in charge of sponsors, with sponsor John Richards of Oxford Digital (left) and Juha Backman (right), in charge of public relations.

Tapio Lokki, technical program chair

Ville Sivonen (left), papers chair, explains a technical point to a delegate.

Marko Takanen, in charge of the social program, prepares to lead orange-hatted delegates on a trek across the snow.

The AES 45th International Conference, *Applications of Time–Frequency Processing in Audio*, continued the series—16th, 22nd, and 30th—of AES conferences organized by the AES Finnish Section. About 80 audio researchers were gathered at the Dipoli Congress Center in Espoo, which is located in the Helsinki capitol area. The venue has a unique architecture designed by Reima and Raili Pietilä, incorporating excellent facilities for a truly international conference, just next door from Aalto University (formerly known as Helsinki University of Technology).

In his opening remarks, Ville Pulkki, conference chair, began the discussions by observing that all audio processing is performed in time and frequency. Indeed, this was a good starting point for a wide variety of 30 oral and 10 poster presentations, selected by papers chair Ville Sivonen. The conference program also included four invited presentations and a good selection of academic spatial sound demonstrations. The technical program, organized by Tapio Lokki, was offered together with a sporty outdoor event and cozy social program in the evenings.

The conference opened on Thursday afternoon, 1 March, with academic demonstrations at the facilities of the Department of Signal Processing and Acoustics at Aalto University. Delegates had an opportunity to visit three anechoic rooms, one fascinating audio-visual room, and a standard listening room to get an inside view on the work done at the university. Tapani Pihlajamäki showed how B-format recordings can be projected onto virtual 3-D surfaces. The system enables interactive walk-throughs "inside" spatial sound recordings. Jussi Rämö presented a real-time perceptual frequency response simulator for music in varying background noises. In an audio-visual system, consisting of three large back-projected screens and 29 loudspeakers, Olli Rummukainen played real audio-visual recordings. The visuals were captured with a commercial Ladybug-3 video camera and spatial sound was captured with an A-format Soundfield SPS200 microphone. The spatial sound rendering was performed with directional audio coding (DirAC) to the

Surround sound demonstrations were set up in one of the anechoic chambers.

Karlheinz Brandenburg (left), a keynote speaker, with Bernd Edler

Anssi Klapuri gives his keynote address on the first day.

Torsten Dau, the third keynote speaker, discusses human auditory signal processing.

A demo of 3-D back-projected video with spatial sound reproduction.

loudspeakers. In the corridor, Marko Hiipakka displayed the pressure–velocity sensors that he has used to measure HRTF functions. With such a device the pressure at the eardrum can be reliably estimated by measuring energy density at the ear canal entrance. In two small anechoic rooms the conference attendees listened to DirAC-processed binaural head-tracked reproduction of B-format signals and a demo of upmixing stereo music to 5.0 by Mikko-Ville Laitinen and Juha Vilkamo, respectively. Finally, the participants were amazed in a large anechoic room by 8.0 surround sound renderings of real concert halls. Sakari Tervo had simulated a symphony orchestra with 33 loudspeakers emitting anechoic recordings on the stage of six concert halls. Spatial sound in five locations in each hall was captured with a microphone array. Thus, in the demo, listeners could jump from hall to hall and from seat to seat in a blink of an eye to compare details of the acoustics of real concert halls.

Thursday evening was dedicated to the mathematical background for time–frequency processing of audio. A tutorial given by Bernd Edler gave an extensive overview for time and frequency domains and to different ways to move from a domain to another. He did not hesitate to challenge the audience with equations. All in all, the tutorial formed a good basis for the whole conference.

KEYNOTE ADDRESSES

The conference included three distinguished keynotes. The first, presented by Anssi Klapuri, opened the conference on Friday morning. Klapuri highlighted the CQT (Constant Q transform) and how it is very well suited to music signal analysis as its time–frequency resolution is close to the resolution of human hearing. However, sampling of frequencies in CQT is not constant, thus it needs cumbersome data structures. Even though CQT is not widely used, it enables natural operations such as pitch shifting. Klapuri also explained another time–frequency processing technique to control the ratio of harmonic and transient components in the recorded signals, and he explained how a matrix factorization algorithm can extract magnitude spectrograms from complex music.

In the second keynote, on Saturday morning, Karlheinz Brandenburg presented a very exciting history of time–frequency domain-based audio coding. Since the 1970s a lot has happened, and it was fascinating to see the creativity of engineers over those years. He concluded the talk by saying that unlike video coding, we seem to have some convergence in the use of filter banks, as no real progress has been done in recent years.

The third keynote was given by Torsten Dau on human auditory signal processing in complex acoustic environments. He picked up recent topics, such as how deficits in cochlear processing have major perceptual consequences, particularly in complex acoustic environments. In addition, he proposed that spectro-temporal modulations appear to be crucial for speech intelligibility.

TIME–FREQUENCY PROCESSING OF AUDIO

The first day of the conference was dedicated to time–frequency processing and time–frequency representation of audio. Eric Battenberg started by presenting his work on drum separation, which is a front end to a live drum understanding system. He used a gamma mixture model to train the decomposing system, which then used nonnegative vector decomposition to find the individual drum hits. The results were promising and the sound examples convincing. Automatic recognition of guitar scores was discussed by Fawad Mazhar. His system can recognize almost 70% of the chords, also in quite bad SNR conditions. The system has applications in human–computer interaction, in computer games, and in many other musical applications.

Antti Jylhä talked about sonic handprints, in other words, how we can identify a person based on his hand clapping sound. Such sound-based identification would involve a robust and cheap single-microphone system and does not require the contact with the identification hardware. Sixteen subjects trained the implemented system to extract the individual spectral templates, based on which the system could finally recognize the claps with an accuracy of 64%, which outperformed human accuracy of 46%. The context of smart environments was also the key in Aki Härmä's presentation on footsteps in walking. His algorithm used time–frequency representation of footstep sounds and it clustered the templates. The system precision was 90% and recall 82%. The algorithm was also tested to recognize snare drum hits with precision of 93% and recall 92%. The last talk in the first session was by Ravy Shenoy, from Bangalore, India. His experimental results indicate that the spectral zero-crossing rate of the head-related transfer functions (HRTF) alone contains sufficient information to estimate ITD as reliably as state-of-the-art techniques.

TIME–FREQUENCY REPRESENTATION OF AUDIO

The afternoon session on time–frequency representation of audio was packed full with mathematics, and the speakers were not afraid of filling their slides with equations.

Kensuke Fujinoki discussed evaluation of button sounds by doing the classification in the time–frequency plane obtained with wavelet transform. Distinctive features can be extracted with triangular biorthogonal wavelets. Volker Gnann explained how multiresolution STFT audio processing usually has problems in detecting and separating transients from steady-state signals. His solution was to initialize the phase estimation of the long-window STFT with the result of the short-window STFT and vice versa. The reason behind this approach is that the better temporal resolution of the short-window STFT moves information about the temporal behavior of the signal from the phase spectrum to the magnitude spectrogram, making it accessible to the phase estimator in the initialization step. Ryo Wakisaka proposed a new noise-reduction method for binaural hearing-aid systems that preserves sound localization. The separation of multiple instrumental sources based on semisupervised nonnegative matrix factorization (SNMF) was addressed by Kosuke Yagi, who also proposed a new constrained SNMF.

Spatial perception in stereo and multichannel playback has been identified to depend especially on the signal covariance matrix in perceptually relevant frequency bands. In his paper, Juha Vilkamo presented a method to control the covariance matrix of a signal by optimal crossmixing of the channels. Informal testing suggested that the method performed robustly, and a wide variety of likely applications were identified in the presentation. Finally, Analk Olivero talked about sound-morphing strategies based on alterations of time–frequency representations by Gabor multipliers.

POSTER SESSION

The posters were presented on Friday evening in an informal joint dinner, drinks, and posters session. Ten posters initiated live discussions on various topics and a few demos could be listened to. Topics included multipoint room response equalization; study on audibility of coloration artifacts in HRTF filter designs; complex FM expansion; a prototype audio spatialization system for teleconferencing; modeling of the cocktail party effect; 3-D microphone array for speech enhancement in hands-free systems in cars; blind audio source separation, and parametric spatial audio coding based on spatial auditory blurring.

A spirited exchange during the informal poster session.

Catarina Mendonça explains human adaptation to HRTFs.

SPATIAL SOUND

The papers in the spatial sound session touched both binaural and multichannel reproduction. Catarina Mendonça talked on the human adaptation to nonindividualized HRTF-based auralization. She has found that trained subjects are able to localize better with generic HRTFs and that the ability lasts a long time, even as much as a month. In addition, the externalization was found to be better after training and even more so when more time elapsed from training. Kimberly Fink has investigated how the horizontal plane HRTFs (PCA representation of HRTF database) can be tuned by modifying the principal component weights. Analytical sequential tuning for PC weights shows that the pinna notch can be tuned and the spectral distortion reduced.

Delegates enjoy informal discussions during lunch on the first day.

Aki Härmä gives a lively presentation.

Maximo Cobos discussed the use of a small tetrahedral microphone array in capturing sound for wavefield synthesis (WFS) auralization. Direction of arrival information from recorded signals was estimated with time–frequency-based spatial filtering. Another WFS paper was given by Andreas Franck, who presented efficient rendering of directional sound sources. Tapani Pihlajamäki proposed two methods that use projection of real and virtual auditory environments onto 3-D surfaces in a virtual world. The first method projects B-format recordings onto an arbitrary surface using a directional audio coding approach. The second method similarly projects room reverberation onto a surface in a doorway between two rooms, thus simulating the audible reverberation to the listener through a doorway.

PSYCHOACOUSTICS AND HEARING

The session on psychoacoustics and hearing gathered papers on a wide range of topics. Jukka Ahonen suggested how parametric spatial audio processing can be applied with bilateral behind the ear hearing aids. Direction of arrival and diffuseness of sound were analyzed in time and frequency channels using two microphones in each behind-the-ear device. It was concluded that an improvement of the speech reception threshold can be obtained with the method, which is comparable to published improvements with typical bilateral signal processing schemes. Kai Siedenburg presented an audio denoising problem from the viewpoint of sparse atomic representation. He proposed a generalized framework for time–frequency thresholding. His approach was competitive with respect to signal-to-noise ratio, and the results showed a reduction of one decade in computational cost.

Jean-Francois Sciabica proposed a method to model a dissimilarity test by comparing the time–frequency representations of car engine sounds. A perceptually motivated cochleagram enabled the emphasis of perceptual attributes in the time–frequency domain. In addition, he showed a robust evaluation framework, based on subjective sensory evaluation.

Nicola Prodi continued the listening test methodology development by presenting a listening efficiency method that was applied to the study of the speech comprehension inside a conference room with a panel of listeners. Furthermore, Jussi Rämö talked about a real-time demonstrator, which has been developed to simulate the perceived music in a noisy listening environment. The system implemented masking threshold and partial masking and is used as a tool to demonstrate how background noise alters the timbre of the music.

Andrea Primavera presented an automatic audio morphing technique applied to percussive musical instruments. The aim of audio morphing algorithms is to combine two or more sounds to create a new sound with intermediate timbre and duration. In the session's final presentation, Rafael de Paiva proposed a method for modeling nonlinear audio systems. The model is based on the swept-sine measurement technique to obtain the time–frequency representation of a nonlinear system. In addition he used principal component analysis to reduce the complexity of the model. Based on this reduction, the computational cost can be reduced by 66% when compared to traditional swept-sine models.

MEMORIAL SESSION FOR PROFESSOR MATTI KARJALAINEN

The last conference day had only one session with six presentations. The papers selected for this session received the highest scores from the papers' reviewers. The session was dedicated to Professor Matti Karjalainen (1946–2010), who was the father of modern acoustics and audio DSP research in Finland. Session chair Vesa Välimäki opened the session with a brief introduction to the highlights of Matti Karjalainen's academic career, including AES Fellow (1999) and AES Silver Medal (2006) awards.

Vesa Välimäki introduces a special session in memory of Matti Karjalainen.

Adrien Sirdey discusses his work on environmental impact analysis.

Tobias Bliem explained his robust sparse multicarrier audio watermarking system. The listening test results prove that the watermarked audio is hardly distinguished from the reference and the system works well in various environmental conditions. Adrien Sirdey talked about his work on the environmental impact of sound analysis and synthesis in Gabor frames. His system enables parametric control for sound synthesis and convincing sound examples of hitting a glass, a bell, and a wood block were heard. Brian Hamilton had made an extensive comparison of parameter estimation methods for an exponential polynomial sound signal model. Frank Wefers explained how fast convolution of long impulse responses and signals can be implemented. He had investigated optimal partitions of filters to perform fast, low-latency multichannel convolution. Yasuki Murakami proposed generation mechanisms of two-tone suppression using a cochlear model that he had developed. Murakami suggested that the somatic motility of the outer hair cell generates the two-tone suppression. Finally, Marko Takanen presented a binaural auditory model for the evaluation of reproduced stereophonic sound. He had successfully applied the model to study the width of stereo image with several stereo-widening algorithms.

OUTDOOR SPECIAL EVENT AND BANQUET

The organizing committee had planned a social program with a strong Finnish flavor. The outdoor event on Friday afternoon was organized on the ice of the Baltic Sea. Delegates were equipped with warm orange woolly hats and they could try ice fishing, kick-sledding, and a simple version of curling. In addition, coffee, hot juice, and cardamom bread were served in a cozy tent that was warmed with a campfire.

The banquet was organized at the old castle in downtown Helsinki. The building was originally built for the student union of the Helsinki University of Technology. The delicious food was accompanied with a few songs led by the organizing committee, as the students in Scandinavia are used to drinking schnapps with a song. Musical performance was offered by a professional and very funny duo playing and singing prankish music with a heavy Nordic folk music influence. You can only imagine the music, which was played by one tuba and one portable pump organ. Their music was amusing and helped all the participants to relax, without thinking about time–frequency problems.

The banquet was held in Helsinki's historic castle.

A duo played traditional Nordic folk music after the banquet.

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.

Snowy fun at the 45th International Conference

Delegates in their orange hats prepare to set off for some snowy Finnish fun on the second day of the conference.

The game of curling is given a try.

Attempts are made to bore holes in the ice without falling through.

Some fishing is attempted through the holes bored in the ice.

Delegates learn to get about on snow scooters.

Some of the local beverage is sniffed suspiciously.