International Conference

Semantic Audio

Erlangen, Germany
21–24 June, 2017

CONFERENCE REPORT

Semantic audio is the relatively young field concerned with content-based management of digital audio recordings. The rapid evolution of digital audio technologies—see the ubiquity of data compression and streaming, the emergence of large audio libraries both online and offline, and recent developments in content-based audio retrieval—have significantly changed the way digital audio is created, manipulated, and consumed. The number of available recordings grows greater each day, which is why content-based management of audio data is essential for producers and consumers alike. In addition to audio retrieval and recommendation techniques, the semantics of audio signals are also becoming more important in object-oriented and intelligent production, including recording, editing, and processing. Finally, with an increasing number of devices listening to their environment, "making sense" of audio is more important than ever. Recent product releases demonstrate this, and many more innovations are being unveiled in this area at a very high pace.

The third AES International Conference on the topic of semantic audio took place 22–24 June at Fraunhofer IIS in Erlangen, Germany. Previous editions were the 53rd International Conference in London, UK, January 2014, and the 42nd in Ilmenau, Germany, July 2011, hosted by Fraunhofer IDMT, another institute of the same German research organization. This edition was jointly organized by the hosting institution Fraunhofer IIS, Friedrich-Alexander Universität (FAU) Erlangen-Nürnberg, and their shared Audio Labs, and further supported by gold sponsors Native Instruments and Dolby. The conference was chaired by Christian Uhle (Fraunhofer IIS) and Meinard Müller (FAU, AudioLabs). The scientific program was coordinated by the paper chairs Christian Dittmar (FAU, AudioLabs) and Jakob Abeßer (Fraunhofer IDMT). Between the 76 delegates, 15 different countries from Europe, Australia, Asia, and the Americas were represented, with a 30%–70% split between industry professionals and academics.

The program committee received 38 paper submissions, of which 27 were accepted and presented as lectures (13) or posters (14). Five delegates also responded to the call for Late-Breaking Demos, showcasing recent work alongside the poster presentations. Because of the very nature of semantic audio, the range of topics covered was relatively wide, including work on music, speech, and sound effects.

Organized by:

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

Fraunhofer IIS

AUDIO LABS

Sponsors

DOLBY

NATIVE INSTRUMENTS

TUTORIAL DAY

In keeping with the previous Semantic Audio conference, the main program was preceded by a series of tutorials from experts in the respective topics.

The first session, run by Alexander Lerch and Stefan Weinzierl, was a masterclass on the topic of music performance analysis, offering an accessible overview of the field while also covering specialized techniques.

Given the recent surge in immersive technologies, many delegates were glad to see Stefania Serafin discuss "Sonic Interactions for Virtual Reality Applications." featuring work on using VR in a variety of contexts.

Finally, Christian Dittmar compared several approaches toward achieving "Phase Reconstruction from Magnitude Spectrograms," a significant DSP challenge with applications in coding techniques and source separation.

After these tutorials, delegates were treated to a welcome reception, complete with a mug of typical Franconian beer and a generous buffet. Jürgen Herre, professor at the Audio Labs, proved himself a talented pianist as he accompanied the dinner with renditions of well-known jazz and pop songs.

DAY 1

Conference chair Christian Uhle opened the main program by welcoming the delegates.

Bernhard Grill then gave a presentation on the Fraunhofer Society, demonstrating it has known many more successes than just the invention of the MP3 they are perhaps most famous for among audio engineers.

Bernhard Grill introduces the Fraunhofer Society.

The opening keynote was delivered by Mark Plumbley, professor of signal processing at University of Surrey's Centre for Vision, Speech and Signal Processing. His talk "Audio Event Detection and Scene Recognition" guided the audience through an impressive track record of machine listening research, to finally touch on the increasingly important issue of privacy related to devices recording and analyzing audio. Currently heading the UK Research Council-funded project "Making Sense of Sounds," Mark is particularly well placed to outline the fields of audio event detection and audio scene classification.

Mark Plumbley explains audio event detection and scene classification during his keynote.

After the coffee break, the first lecture session chaired by Jürgen Herre was on the topic of source separation.

Jazz ensemble LINda Capo entertains the delegates on Day 1.

Diego Di Carlo's talk discussed an interference-reduction approach for live sound, based on Wiener filtering and the assumption of independence of signals and inter-microphone phase. In contrast with previous work, the different close-mic recordings were not assumed to have the same energy.

Delia Fano Yela presented advances in vocal-signal extraction, preceded by a remarkably effective, hand-drawn overview of source separation basics. The method of Kernel Additive Modeling was extended by incorporating the temporal context to provide additional information, leading to a significant performance improvement.

The afternoon began with another series of lectures, this time on audio descriptors and features, and chaired by Udo Zölzer.

Athanasios Lykartsis assessed the benefits of using both spectral and rhythm features for speaker identification, employing a corpus of Swiss German speech and based on support vector machines. Overall, the use of rhythm features actually deteriorated the system's performance, despite their suitability for MIR and language identification tasks; some potential reasons for this and suggested directions for further work were ventured.

Andy Pearce then proposed an approach for browsing sound effects libraries using timbral features, part of the Audio Commons project which tries to understand and address the limited use of audio licensed under Creative Commons. Specifically, he determined which English timbral terms were most popular on freesound.org using a part-manual, part-automated approach, and identified a selection of overarching attributes. The corresponding paper was made available to all under an Open Access license.

Udo Zölzer, the 2nd keynote speaker

Finally, Patricio López-Serrano introduced mid-level features for harmonic-residual-percussive source separation. Cascading several HRPSS stages, with different separation factors, he

achieved a finer-grained decomposition that can be used as a preprocessing step for MIR tasks or as identification of drum solos.

The hot weather was offset by ice cream for all delegates, resulting in happy faces at the first poster and late-breaking demo session.

This slot allowed authors to present papers on topics ranging from music production (Nicholas Jillings, Adán L. Benito) over software tools (Zdenek Pruša, Spencer Russell) to machine listening (Gerhard Hagerer, Tom Bäckström, Joren Six). In addition, Alexander Adami and Christof Weiß showcased work on decomposition of applause signals and computational analysis of harmonic structure, respectively.

Still using Fraunhofer IIS's excellent facilities, the delegates were treated to a concert by jazz ensemble LINda Capo, before and during dinner. In between, an award ceremony honored the authors of the best paper submissions, as selected by the program committee.

The title of Best Student Paper was awarded to Rachel Bittner and colleagues from New York University (NYU) and Universitat Pompeu Fabra (UPF), for the paper "Pitch Contours as a Mid-Level Representation for Music Informatics."

Rodrigo Schramm and Emmanouil Benetos from Queen Mary University of London and Universidade Federal do Rio Grande Sul received the Best Paper Award for "Automatic Transcription of a Cappella recordings from Multiple Singers."

Emmanouil Benetos is presented with the Best Paper Award.

Rachel Bittner receives the Best Student Paper Award.

DAY 2

Incidentally, these deserving papers were the first two to be presented in Friday morning's lecture session.

Rachel Bittner (Best Student Paper) kicked off the session on pitch tracking, chaired by Masataka Goto. Her work makes a compelling case for the adoption of pitch contours as a unit of pitch organization, though warns that their estimation is difficult for polyphonic music.

Happy members of the conference committee, having organized a successful event.

Rodrigo Schramm (Best Paper) proposes a post-processing step for the automatic transcription of a capella, multisinger music, accounting for false positive pitch detections such as those caused by overtones.

Skillfully following these award-winning papers, Ashis Pati, standing in for Amruta Vidwans, compared audio features for the automatic assessment of student music performances. Attempting to model a set of human judgments of saxophone recordings from audition tapes, the authors found that score-based features performed better than basic or score-independent features, and a combination of these performed better still, albeit not yet suitable for reliable quality ratings.

The second keynote speaker was Udo Zölzer, professor and head of the Department of Signal Processing at Helmut Schmidt University, and perhaps best known for editing the popular book "DAFX: Digital Audio Effects." In keeping with the morning's theme, the topic of his presentation was "Pitch-based Audio Algorithms," preceded by a bit of a tourist guide to Hamburg, the speaker's home turf. Over the course of the talk, Udo showed off some very exciting pitch-controlled audio effects, but also gave a sneak preview of unpublished work on machine learning-aided pitch tracking, and demonstrated a new modulation-based synthesizer.

After lunch, Alexander Lerch chaired the lecture session on automatic music transcription.

Dasaem Jeong used a pre-aligned score, temporal and harmonic constraints, and the power spectrogram to estimate note intensity through nonnegative matrix factorization (NMF), thus achieving higher accuracy than comparable work.

Following the same theme, Emmanouil Benetos presented recent work on polyphonic music transcription using probabilistic latent component analysis (PLCA). Placing linear dynamic systems before the PLCA, he improved both multipitch detection and instrument assignment.

The technical program of Day 2 concluded with the second and final poster session, featuring new work on various music information retrieval tasks, including instrument (Ashis Pati, Florian Scholz), key (Ángel Faraldo), chord (Filip Korzeniowski), onset (Jose Valero-Mas), frequency (Meinard Müller), and musical structure detection (Frank Zalkow). Two more demos were presented: one by Anna Kruspe on audio-based lyrics alignment and retrieval, and one by Brecht De Man on a web interface to the Mix Evaluation Dataset.

The main social event of the conference took place in a quintessentially German beer garden with dinner and drinks, and a particularly

Enthusiastic discussion around one of the poster presentations.


Informal discussion during a coffee break.

entertaining guided tour through some of Erlangen's historic beer cellars. This led the delegates – some of whom didn't think to pack sweaters and jackets given the sweltering weather – deep into a maze of dark, pre-refrigerator halls that were once stocked with beer. The fantastic reverberation prompted many hand claps and inspired a few hypothetical recording projects.

DAY 3

The concept of deep learning is taking the world of technology by storm, and audio is no exception. Evidence of this was the Saturday morning lecture session, chaired by Christian Dittmar and very well attended despite the merriment of the night before, dedicated exclusively to this topic.

Rainer Kelz shed light on the entanglement problem in neural network methods for music transcription, where systems tend to learn combinations of notes, and do not recognize unseen combinations of notes. The authors suggested a modification of the network's loss function may alleviate this issue.


Masataka Goto gives his invited talk on developing web services for web-native music.

Jakob Abeßer also presented work on music transcription using deep neural networks (DNN), in this case focusing on walking bass transcription in jazz. In the absence of large volumes of annotated training data, unlabeled data was additionally used where pitch estimation exceeded a certain confidence threshold.

The session was completed by Alice Cohen-Hadria, who applied convolutional neural networks to the problem of structural segmentation of music. She proposed the use of square-submatrices, shown to outperform previous attempts using a self-similarity lag matrix.

An invited talk by MIR expert Masataka Goto, prime senior researcher at the National Institute of Advanced Industrial Science and Technology, concluded the technical program. Under the title "Developing Web Services for Web-Native Music," it covered a wide range of web-based music discovery and visualization tools, from linking to derivative consumer-generated versions of songs, to "active listening" augmented by animated computer graphics, chord and structure annotation, and even physical dancing robots.

Conference chair Christian Uhle then closed the conference by naming and thanking everyone who contributed to the impeccable organization, and allowed delegates to announce some related events (such as the 3rd AES Workshop on Intelligent Music Production in Manchester, UK, 15 September) and PhD opportunities. Delegates said goodbye over a light lunch, and thus ended the 2017 International Conference on Semantic Audio.

AES Members can access all papers for free via the AES E-library at http://www.aes.org/e-lib/.


A room full of attentive delegates concentrates hard during one of the numerous paper presentations.