Report from Dagstuhl Seminar 13451

Computational Audio Analysis

Edited by

Meinard Müller[1], Shrikanth S. Narayanan[2], and Björn Schuller[3]

1 **International Audio Laboratories Erlangen,**
Universität Erlangen-Nürnberg, DE, `meinard.mueller@audiolabs-erlangen.de`
2 **USC – Los Angeles, US,** `shri@sipi.usc.edu`
3 **TU München, DE,** `schuller@tum.de`

Abstract

Compared to traditional speech, music, or sound processing, the computational analysis of general audio data has a relatively young research history. In particular, the extraction of affective information (i. e., information that does not deal with the 'immediate' nature of the content such as the spoken words or note events) from audio signals has become an important research strand with a huge increase of interest in academia and industry. At an early stage of this novel research direction, many analysis techniques and representations were simply transferred from the speech domain to other audio domains. However, general audio signals (including their affective aspects) typically possess acoustic and structural characteristics that distinguish them from spoken language or isolated 'controlled' music or sound events. In the Dagstuhl Seminar 13451 titled "Computational Audio Analysis" we discussed the development of novel machine learning as well as signal processing techniques that are applicable for a wide range of audio signals and analysis tasks. In particular, we looked at a variety of sounds besides speech such as music recordings, animal sounds, environmental sounds, and mixtures thereof. In this report, we give an overview of the various contributions and results of the seminar. We start with an executive summary, which describes the main topics, goals, and group activities. Then, one finds a list of abstracts giving a more detailed overview of the participants' contributions as well as of the ideas and results discussed in the group meetings of our seminar. To conclude, an attempt is made to define the field as given by the views of the participants.

Seminar 03.–08. November, 2013 – www.dagstuhl.de/13451
1998 ACM Subject Classification I.5.4 Applications—Waveform analysis, H.5.5 Sound and Music Computing, H.1.2 User/Machine Systems, H.5.1 Multimedia Information Systems
Keywords and phrases Audio Analysis, Signal Processing, Machine Learning, Sound, Speech, Music, Affective Computing
Digital Object Identifier 10.4230/DagRep.3.11.1
Edited in cooperation with Jonathan Driedger

1 Executive Summary

Meinard Müller
Shrikanth S. Narayanan
Björn Schuller

With the rapid growth and omnipresence of digitized multimedia data, the processing, analysis, and understanding of such data by means of automated methods has become a

Computational Audio Analysis, *Dagstuhl Reports*, Vol. 3, Issue 11, pp. 1–28
Editors: Meinard Müller, Shrikanth S. Narayanan, and Björn Schuller
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

central issue in computer science and associated areas of research. As for the acoustic domain, audio analysis has traditionally been focused on data related to speech with the goal to recognize and transcribe the spoken words. In this seminar, we considered current and future audio analysis tasks that go beyond the classical speech recognition scenario. For example, we looked at the computational analysis of speech with regard to the speakers' traits (e. g., gender, age, height, cultural and social background), physical conditions (e. g., sleepiness, alcohol intoxication, health state), or emotion-related and affective states (e. g., stress, interest, confidence, frustration). So, rather then recognizing *what* is being said, the goal is to find out *how* and *by whom* it is being said. Besides speech, there is a rich variety of sounds such as music recordings, animal sounds, environmental sounds, and combinations thereof. Just as for the speech domain, we discussed how to decompose and classify the content of complex sound mixtures with the objective to infer semantically meaningful information.

When dealing with specific audio domains such as speech or music, it is crucial to properly understand and apply the appropriate domain-specific properties, be they acoustic, linguistic, or musical. Furthermore, data-driven learning techniques that exploit the availability of carefully annotated audio material have successfully been used for recognition and classification tasks. In this seminar, we discussed issues that arise when dealing with rather vague categories as in emotion recognition or when considering general audio sources such as environmental sounds. In such scenarios, model assumptions are often violated, or it becomes impossible to define explicit representations or models. Furthermore, for non-standard audio material, annotated datasets are hardly available. Also, data-driven methods that are used in speech recognition are (often) not directly applicable in this context; instead semi-supervised or unsupervised learning techniques can be a promising approach to remedy these issues. Another central topic of this seminar was concerned with the problem of source separation. In the real world, acoustic data is very complex typically consisting of a superposition of overlapping speech, music, and general sound sources. Therefore, efficient source separation techniques are required that allow for splitting up, re-synthesizing, analyzing, and classifying the individual sources—a problem that, for general audio signals, is yet not well understood.

In this executive summary, we give a short overview of the main topics addressed in this seminar. We start by briefly describing the background of the participants and the overall organization. We then give an overview of the presentations of the participants and the results obtained from the different working groups. Finally, we reflect on the most important aspects of this seminar and conclude with future implications.

Participants, Interaction, Activities

In our seminar, we had 41 participants, who came from various countries around the world including North America (10 participants), Japan (1 participant), and Europe (Austria, Belgium, Finland, France, Germany, Greece, Italy, Netherlands, Spain, United Kingdom). Most of the participants came to Dagstuhl for the first time and expressed enthusiasm about the open and retreat-like atmosphere. Besides its international character, the seminar was also highly interdisciplinary. While most of the participating researchers are working in the fields of signal processing and machine learning, we have had participants with a background in cognition, human computer interaction, music, linguistics, and other fields. This made the seminar very special in having many cross-disciplinary intersections and provoking discussions as well as numerous social activities including common music making.

Overall Organization and Schedule

Dagstuhl seminars are known for having a high degree of flexibility and interactivity, which allow participants to discuss ideas and to raise questions rather than to present research results. Following this tradition, we fixed the schedule during the seminar asking for spontaneous contributions with future-oriented content, thus avoiding a conference-like atmosphere, where the focus is on past research achievements. The first two days were used to let people introduce themselves, present scientific problems they are particularly interested in and express their expectations and wishes for the seminar. In addition, we have had six initial stimulus talks, where specific participants were asked to address some burning questions on speech, music, and sound processing from a more meta point of view, see also Section 3. Rather than being usual presentations, most of these stimulus talks seamlessly moved towards an open discussion of the plenum. Based on this input, the second day concluded with a brainstorming session, where we identified central topics covering the participants' interests and discussed the schedule and format of the subsequent days. To discuss these topics, we split up into five groups, each group discussing one of the topics in greater depth in parallel sessions on Wednesday morning. The results and conclusions of these group meetings were then presented to the plenum on Thursday morning, which resulted in vivid discussions. Continuing the previous activities, further parallel group meetings were held on Thursday afternoon, the results of which being presented on Friday morning. Finally, asking each participant to give a short (written) statement of what he or she understands by the seminar's overall topic "Computational Audio Analysis," we had a very entertaining and stimulating session by going through and discussing all these statements one by one. The result of this session can be found in Section 6. In summary, having a mixture of different presentation styles and group meetings gave all participants the opportunity for presenting and discussing their ideas, while avoiding a monotonous conference-like atmosphere.

Main Topics

We discussed various topics that addressed the challenges when dealing with mixtures of general and non-standard acoustic data. A particular focus was put on data representations and analysis techniques including audio signal processing, machine learning, and probabilistic models. After a joint brainstorming session, we agreed on discussing five central topics which fitted in the overall theme of the seminar and reflected the participants' interests. We now give a brief summary of these topics, which were addressed in the parallel group meetings and resulting panel discussions. A more detailed summary of the outcome of the group sessions can be found in Section 4.

1. The *"Small Data"* group looked at audio analysis and classification scenarios where only few labeled examples or small amounts of (training) data are available. In such scenarios, machine learning techniques that depend on large amounts of (training) data ("Big Data") are not applicable. Various strategies including model-based as well as semi- and unsupervised approaches were discussed.
2. The *"Source Separation"* group addressed the task of decomposing a given sound mixture into elementary sources, which is not only a fundamental problem in audio processing, but also constitutes an intellectual and interdisciplinary challenge. Besides questioning the way the source separation problem is often posed, the need of concrete application scenarios as well as the objective of suitable evaluation metrics were discussed.

3. The *"Interaction and Affect"* group discussed the question on how to generate and interpret signals that express interactions between different agents. One main conclusion was that one requires more flexible models that better adapts to the temporal and situational context as well as to the agents' roles, behaviors and traits.
4. The *"Knowledge Representation"* group addressed the issue of how knowledge can be used to define and derive sound units that can be used as elementary building blocks for a wide range of applications. Based on deep neural network techniques, the group discussed how database information and other meta-data can be better exploited and integrated using feed-forward as well as recurrent architectures.
5. The *"Unsupervised Learning"* group looked at the problem on how to learn the structure of data without reference to external objectives. Besides issues on learning meaningful elementary units, the need of considering hierarchies of abstractions and multi-layer characterizations was discussed.

Besides an extensive discussion of these five main topics, we have had many further contributions and smaller discussions on issues that concern natural human machine communication, human centered audio processing, computational paralinguistics, sound processing in everyday environments, acoustic monitoring, informed source separation, and audio structure analysis.

Conclusions

In our seminar, we addressed central issues on how to process audio material of various types and degrees of complexity. In view of the richness and multitude of acoustic data, one requires representations and machine learning techniques that allow for capturing and coupling various sources of information. Therefore, unsupervised and semi-supervised learning procedures are needed in scenarios where only very few examples and poor training resources are available. Also, source separation techniques are needed, which yield meaningful audio decomposition results even when having only limited knowledge on the type of audio. Another central issue of this seminar was how to bring in the human into the audio processing pipeline. On the one hand, we discussed how we can learn from the way human process and perceive sounds. On the other hand, we addressed the issue on extracting human-related parameters such as affective and paralinguistic information from sound sources. These discussions showed that understanding and processing complex sound mixtures using computational tools poses many challenging research problems yet to be solved.

The Dagstuhl seminar gave us the opportunity for discussing such issues in an inspiring and retreat-like atmosphere. The generation of novel, technically oriented scientific contributions was not the focus of the seminar. Naturally, many of the contributions and discussions were on a rather abstract level, laying the foundations for future projects and collaborations. Thus, the main impact of the seminar is likely to take place in the medium to long term. Some more immediate results, such as plans to share research data and software, also arose from the discussions. As measurable outputs from the seminar, we expect to see several joint papers and applications for funding. Beside the scientific aspect, the social aspect of our seminar was just as important. We had an interdisciplinary, international, and very interactive group of researchers, consisting of leaders and future leaders in our field. Most of our participants visited Dagstuhl for the first time and enthusiastically praised the open and inspiring atmosphere. The group dynamics were excellent with many personal exchanges and common activities. Some scientists mentioned their appreciation of having the opportunity for prolonged discussions with researchers from neighboring research fields—something which is often impossible during conference-like events.

In conclusion, our expectations of the seminar were not only met but exceeded, in particular with respect to networking and community building. Last but not least, we heartily thank the Dagstuhl board for allowing us to organize this seminar, the Dagstuhl office for their great support in the organization process, and the entire Dagstuhl staff for their excellent services during the seminar.

2 Table of Contents

Executive Summary

Stimulus Talks

Group Sessions

Further Topics

3 Stimulus Talks

3.1 Multimedia Analysis for the Poor

Xavier Anguera (Telefónica Research – Barcelona, ES)

In many areas of multimedia analysis (i. e., when extracting knowledge from multimedia or multimodal data) one usually fist derives models from corpora of annotated training data, which are then applied to some unknown data. Highly performing systems have been built using this methodology in the past. However, it must not be overlooked that producing high-quality annotated data for training takes time and resources, which are not always available. Examples of high-quality labeling scarcity can be seen when trying to analyze highly diverse data like what is found on online media sources such as YouTube or SoundCloud, or with rare languages in speech (i. e. those languages for which the number of speakers is to small to attract commercial interest). For this reason it becomes very relevant to explore new avenues to be able to extract knowledge with no (or very limited) labeled examples. There are already many efforts in this direction within the research community such as:

- Speech: Audio summarization through the analysis of repetitions in the audio stream; query-by-example spoken term detection; training systems (e. g., large vocabulary speech recognition) on little transcribed data, or on low quality transcripts (e.g. close captions) data.
- Music: Structural analysis of songs.
- Image processing: Unsupervised concept extraction (e. g., the system developed by Google and G. Hinton to automatically learn how to recognize cats in Youtube videos).
- Text: Unsupervised document clustering and topic detection.
- Bioinformatics: Unsupervised repetitions/structure detection and finding mutations.

In this inspirational talk, I first motivated the need to do research on unsupervised and semi-supervised algorithms to tackle problems like those mentioned above. Then, after presenting some examples of technologies that are able to perform well with these constraints, I described the task of (query-by-example) spoken term detection. The objective of this task is to find all lexical instances of a spoken term within an audio database of spoken words. Within the Mediaeval 2013 Spoken Web Search evaluation (which I helped organize), we have considered the scenario where nothing is known a-priori about the query or the database (only the fact that the database contains data from nine different languages was known).

Next, I discussed about how low/zero resource techniques can complement high resource systems. For this I hypothesized how babies learn about their surrounding world by registering similar repeating patterns that occur many times. I proposed the discovery of repeating information (e. g., repeating acoustic patterns in speech) to be used as an informed initialization to transcription systems to automatically retrieve sizable training corpora from high-quality seed transcriptions.

Finally, the following set of open questions were posed to the audience to trigger some discussion:

- How to incorporate acoustic modeling into dynamic programming techniques?
- How to describe the acoustic space (or whatever space) in an unsupervised (but robust) manner?
- How do we discriminate between “interesting/relevant” and “filler” events?

- Does it all make any sense? Or will there be a point where we always have enough training data for a given task?

3.2 Interpreting 'Intentional' Behaviour in Audio Scenes

Roger K. Moore (University of Sheffield, GB)

Whilst there is no doubt about the immense practical benefits that could be derived from the automated analysis of audio scenes, it is not clear that the research community has yet developed a sufficiently sophisticated theoretical framework to realise its full potential. Recent years have seen measurable progress in computational approaches to information extraction by applying the latest machine learning techniques to annotated (or even unannotated) data, but most of the focus has been on classifying the surface phenomena associated with acoustic events. Little attention has been given to interpreting the underlying 'intentional' states that are unique to living organisms and which drive the physical actions that are performed (particularly communicative behaviour such as speech). Of course, if there was a simple one-to-one relationship between internal intentional states and the consequent surface behaviour, then interpretation would be relatively straightforward. However, in reality there is significant 'coupling' (i. e., dependencies) between objects, agents and their environment, and this means that interpreting what is happening in an acoustic scene requires a yet-to-be-defined unified computational modelling approach which is capable of integrating the relevant contingencies. This stimulus talk illuminated these issues and raised the following issues for discussion: How important is it that we acknowledge that the world contains intentional agents? Can we envisage a unified computational modelling approach which is capable of integrating the relevant contingencies? What are the implications of modelling self (recursion, context dependency, embodiment)? And can an agent ever understand a natural scene if it is not (or has never been) part of it? The final question is thus: what does an automated agent need to know about the world and the entities it contains in order to make sense of a general audio scene?

3.3 Semantic-Affective Models for Multimodal Signal Processing

Alexandros Potamianos (National TU – Athens, GR)

In this stimulus talk, I reviewed multimodal aspects of audio processing focusing mainly on three areas. First, I considered the area of affective analysis and recognition of audio and multimedia streams. I presented recent results of emotion recognition from audio signals as well as movies, and the interplay between audio events, music and spoken language was outlined. Second, I discussed aspects of saliency and attention for audio and multimedia streams. Recent progress on selectional attention models for speech and audio were reviewed. Furthermore, the role of saliency/attention for audio/speech processing was discussed and future research directions outlined. Third, I addressed the topic of associative and representational models of audio and multimodal semantics. Motivated by cognitive considerations, associative lexical semantic models have been recently proposed. These models have been extended to include also multimodal or crossmodal information such as images. I discussed how such models can be extended for audio, music, and speech content to create multimodal similarity networks as well as how such networks are relevant for inference and classification tasks.

3.4 Exploitation of Human Perception Principles in Audio Processing Systems

Gaël Richard (Telecom ParisTech, FR)

The integration of auditory perception in most audio processing systems remains limited. A number of perceptually-relevant concepts have been exploited in audio processing research. But, for example for audio indexing/classification, it still seems that it is difficult to build a fully perceptually-relevant system that outperforms efficient machine-learning based methods that use only some rudimentary principles of perception. This remains surprising because from a pure acoustical point of view it intuitively appears that it may be unnecessary to capture similarity or dissimilarity information that is not perceived by humans. Should we look for better perceptual features or better perceptual representations such as cortical representations? Should we better model feature dynamics? Should we better model the complex and hierarchical processing information in the brain? In this stimulus talk, I discussed some of these issues. One may draw the conclusion that even though human perception principles are in general seen as important, mimicking the human perception or the functioning of the brain does not seem to be a prerequisite for computational audio analysis systems.

3.5 Stop Listening to Speech, Language, and Vision Research!

Paris Smaragdis (University of Illinois at Urbana Champaign / Adobe, US)

Artificial intelligence, and more recently machine learning, has always been guided by the dream of making machines that can understand the world around them. Unfortunately the lion's share of this activity has been on domains that exhibit few insights on how to make machines that can listen, and this has resulted in an ongoing derailment of how machine learning should be applied on audio problems. A big chunk of machine learning is dealing with problems that stem from vision, language, and speech, which are all domains where we make hard decisions. A pixel either belongs to one object of another, a word is either "cat" or "dog", and a spoken language belongs to one of many different families. As a result, the vast majority of machine learning approaches operate with a winner-takes-all philosophy, where the objective is to find that one solution that is the only correct value. Sound however is different. In real-life we never hear one sound or another. Instead, we hear mixtures of sounds. Such problems cannot be properly treated with tools such as discriminative learning and even many of the common generative models, and we see a need for some fundamental rethinking of how learning algorithms should be applied on sound.

3.6 Sound Event Detection and Recognition in Everyday Environments

Tuomas Virtanen (Tampere University of Technology, FI)

For humans, the most important functionality of auditory scene analysis is to acquire information about our everyday environments: a car approaching from behind, warning beeps, door knocking, door opening and closing and so on. Until now, most of the research on computational audio analysis has been done in the context of speech and music processing. Research on automatic detection and recognition of sound events has mostly been limited to isolated sound events, specific environments (such as meeting rooms), and small number of specific types of events.

Computational audio analysis in everyday environments has applications in areas such as multimedia content analysis, context-aware devices, assistive technologies, and acoustic monitoring. The research in the field has so far approached the task by studying two problems. The first one is context recognition, where a recording is classified into one of predefined contexts. Such contexts may be characterized by locations such as home, office, street, or grocery store or by physical and social activities. Second, sound event or acoustic event detection aims at estimating the start and end times of individual events in a recording as well as estimating a class label for each detected event.

Automatic detection and recognition of events in realistic environments requires addressing many problems (e. g., robustness) that have already been faced in the context of speech and music recognition. Additionally, operating with realistic sounds in everyday environments raises many new questions. Sound events can originate from very different sources and have therefore diverse acoustic characteristics, which may force us to rethink our conventional pattern recognition approaches. The identity of an event can be encoded in many different ways, e. g., by the rough shape of the spectrum, modulations over time, relationships between atomic sound units, or parts of a signal. This affects not only the feature extraction front-end used by an event detection system, but also the architecture of the whole system.

Finally, a single sound can have multiple different interpretations, and it is not at all clear how the event classes to be detected should be defined. Possibilities include definitions about the physical source, semantics, or acoustic similarity. Since it may not be possible to manually define classes for all sounds present in a signal, the use of unsupervised learning techniques needs to be taken into account.

4 Group Sessions

4.1 Small Data (Learning from Few Examples)

Bryan Pardo, Xavier Anguera, Jonathan Driedger, Bernd Edler, Jort Gemmeke, Franz Graf, Gernot Kubin, Frank Kurth, Meinard Müller, and Christian Uhle

We define “small data” in opposition to the current buzzword “big data”. These are cases where there are few labeled examples to learn from. This may be because the labeling requires intense analysis by an expert (e. g., structural analysis of a Beethoven symphony). This may be because there are only a few examples in existence (there are only nine Beethoven symphonies), yet we wish to learn something useful and meaningful for some task. One concrete example of a "small data" situation is that of building an acoustic car crash detector for a particular tunnel. Collecting real data (crashing cars) is expensive. Car crash events are relatively rare in the tunnel (once or twice a year). Yet we want the detector to start working as soon as possible and with as few examples as possible. Another example is voice-controlled home automation for people with unique speech impairments. Here, again, data is hard to collect and each speaker is very different from the rest of the population.

There is an intellectual appeal to learning from small numbers of examples. Humans often learn generalizations from as few as three examples of a class. It would be interesting to learn how to duplicate this kind of performance. The practical appeal of being able to learn in an online manner, before collecting a lot of data (e. g. the car crash example) is also great. Further, when working with systems on small data sets, the researchers themselves can have a much better understanding of the data. Rather than millions of unexamined examples, there are dozens of well-understood ones. Current approaches to "small data" typically use statistical learners that require lots of data to work properly. Therefore approaches tend to look for ways to bridge the gap between learners that need lots of data and tasks that provide small numbers of examples. Approaches fall in the following categories:

- Data Synthesis: Create synthetic data by adding noise (of some expected variety) to the small number of known examples. Alternately synthesize data by using a simulation that can be done more cheaply than collecting real data (e. g., replace cars with garbage cans and crash garbage cans together in the tunnel).
- General to Specific: Start with a generic model, learned from a lot of data. Tweak that model slightly to conform to the particular "small data" case.
- Model Selection: Build several generic models for known generic cases, then collect a small number of data points from the current case to select which generic model best suits the current case. This can then be used in combination with the previous strategy to make it more specific.
- Think Smarter: Offload the learning to the human, who figures out a smarter way of preprocessing data to put it in a format that is very easy to learn from, by using extremely salient features.

The question we had was if there are other approaches we have not encountered. Can we do better than these four approaches?

4.2 Interaction and Affect

Martin Heckmann, Murtaza Bulut, Carlos Busso, Nick Campbell, Laurence Devillers, Anna Esposito, Sungbok Lee, Roger Moore, Mark Sandler, Khiet Truong, and Rita Singh

Interaction is driven by intentional agents. Agents accommodate to the role and capabilities of other agents. The success of the interaction depends on the generation and interpretation of appropriate signals, often across multiple modalities (e. g., bio-signals, image, speech). However, the effective processing of these signals also depends on "rich" information and not just "big" data. This includes the temporal and situational context as well as the role, characteristics, behaviors and traits of the agents. Current systems depend on theory-laden annotations (not capturing the true nature of the interaction), which unnecessarily constrain the learning outcome. We believe that a viable first step is to develop and continuously adapt the model for the agent's world from observed data. One possible means to achieve this is to implement algorithms for detecting changes and deviations from the learned normal/stable points.

4.3 Learning of Units and Knowledge Representation

Florian Metze, Xavier Anguera, Sebastian Ewert, Jort Gemmeke, Dorothea Kolossa, Emily Mower Provost, Björn Schuller, and Joan Serrà

Our group came together to discuss how knowledge could be used to define and infer units of sound that could be used in a portable way for a number of tasks. Participants felt that a top-down approach would be needed, which is complementary to purely data-driven bottom-up clustering approaches, as are currently prevalent in classification experiments. Members wanted to specifically investigate how an attempt to solve multiple problems at the same time ("holistic" approach) could benefit each individual task by exposing and exploiting correlations and complementarity, which would otherwise stay hidden. Members also felt that a sound statistical framework was needed and that a careful modeling of uncertainty and a mechanism to feed back confidences was needed. This would also be beneficial in the presence of multiple, possibly overlapping signals as is typically the case for sounds Finally, members were interested in working on meta-data of speech. First ideas were discussed on how to learn from data units representing emotions that would be both acoustically discriminative and useful in the context of a certain application, or discernible by humans.

Most members had some background in low-level feature extraction and in deep learning. Against this background, members developed an experiment, which they intend to execute in a distributed collaboration over the next couple of weeks. The experiment will be performed on the IEMOCAP database using various existing tools available to the group members. Collaboration tools will be set up at CMU. To establish a baseline, members will investigate the suitability of multi-task learning by training a single deep neural network (DNN) to predict both binary and continuous valued emotion targets on the IEMOCAP benchmark database. The network will be adapted to other databases (most likely AVEC and CreativeIT) to investigate the portability of the learner and to investigate the utility of multi-task learning. These experiments can be performed with feed-forward as well as recurrent architectures. Next, prior knowledge will be incorporated into the classification by adding database information, speaker information, or other meta-data (automatically extracted or manually labeled) as additional inputs to the network training. Finally, the recurrence loop will be optimized by investigating which information should be fed back. This information may comprise the utility of certain features or classes in a certain task, the saliency of some features, or the classification accuracy (posterior probabilities) of some classes on a held-out dataset. Members discussed an uncertainty weighted combination approach that should be able to update the structure and parameters of the classifier so as to improve classification accuracy. The goal will be to optimize the allocation of parameters towards modeling useful target units rather than attempting to accurately model distinctions that will eventually not be used in an application. Results will be published in peer-reviewed literature, and will hopefully lead to follow-up collaborations including organizing future workshops and joint proposals.

4.4 Unsupervised Learning for Audio

Tuomas Virtanen, Jon Barker, Shrikanth Narayanan, Alexandros Potamianos, Bhiksha Raj, Gaël Richard, Rita Singh, Paris Smaragdis, Stefano Squartini, and Shiva Sundaram

After a more fundamental discussion on unsupervised learning for audio, our group decided to focus on the use of unsupervised learning in a concrete application scenario. Even though unsupervised learning could be used in many processing stages of a computational audio analysis system (e. g., to develop the feature extraction front-end), in practical scenarios one often takes advantage of some prior information. In particular, defining a specific application and ways to evaluate the performance of the audio processing system already imposes some prior information.

We considered an application where the goal is to detect car crash sounds from continuous audio recordings. In the unsupervised learning scenario, one has audio recordings that can be used as a training data, but no reference times of crashes are actually annotated. There are several kinds of prior information that can be used in the given application. First, one may assume that the events one is looking for can be characterizing by a specific set of audio features such as MFCCs. Second, one may assume to have a metric that is appropriate for describing distances between acoustic features. Third, we know that events are rare, i. e., only a small number of target events are present in the training data. Finally, we assume that each event is localized in time and the duration of events is approximately known (e. g. one or two seconds). The above prior information can be used for novelty-based audio segmentation using the calculated features and the distance metric. Alternatively, unsupervised learning can be used to learn features and a distance metric. The resulting segments can be assumed to be homogeneous. Segments can then be clustered so that each cluster contains a specific kind of sound. Subsequently, the developer of the system can manually examine each cluster to see whether a cluster contains a sound relevant for the development of the detector. Assuming that the events of interest are rare, the cluster with the largest number of segments need not be examined (containing sounds that do not correspond to a car crash). The system could also work in an incremental fashion, where the clustering may change as new data becomes available. This results in a system that achieves a more knowledgeable perspective on the problem to be solved.

The main benefits of the use of unsupervised learning in this application is the reduction of amount of manual work: the events of interest can be found from the recordings simply by examining a single sample from a cluster. In our discussions, it was also pointed out that the use of unsupervised learning removes a user bias and allows for finding phenomena or concepts that cannot be precisely defined.

4.5 Source Separation

Christian Uhle, Jonathan Driedger, Bernd Edler, Sebastian Ewert, Franz Graf, Gernot Kubin, Meinard Müller, Nobutaka Ono, Bryan Pardo, and Joan Serrà

Our group attracted participants from various research areas to discuss aspects of source separation for audio signals, performed either in a blind way or by using additional knowledge about the underlying sources or the mixing process. In many source separation approaches, one assumes that sources are independent, uncorrelated, and do not overlap with regard to a given representation. Also one often presupposes that the mixing process is linear and time-invariant. However, in practice these assumption are often violated. In addition, sound sources may influence or interact with each other, so that the separated source signals may sound unnatural or different to situation where they occur in an isolated fashion. Examples are the coupling between piano strings and the Lombard effect that describes the adaption of a speaker to noisy environments. Further fundamental problems in source separation are the unmasking of undesired sounds (e. g., FM noise or audio coding artifacts), shortcomings of objective evaluation metrics, or the sound quality (e. g., due to the phase reconstruction problem). Last but not least, even the definition of what to understand by a source is ambiguous: a source can be a physical entity that emits sound, an object or event that is perceived by a human listener (stream), or a musical voice in a polyphonic sound mixture.

There are various applications that motivate ongoing research in source separation including remixing and upmixing, Karaoke applications, speech enhancement for hearing aids and communication, dialogue enhancement, audio editing, and audio content analysis. Besides these applications, source separation is a fascinating, intellectual, and interdisciplinary research area that requires and provides a deep understanding of the underlying audio material with regard to various aspects ranging from physical processes to cognitive aspects.

5 Further Topics

5.1 Engineering Selective Attention into Acoustic Scene Analysis Systems.

Jon Barker (University of Sheffield, GB)

A general goal of acoustic scene analysis is to recover abstract high-level descriptions of the individual sound sources given the raw acoustic mixtures. It is often assumed that a machine scene analysis system should extract some sort of 'complete' description in which all sources are described with equal detail. In certain contrived scenes, for example 'cocktail parties' composed of speakers uttering sentences from fixed grammars, computational systems are able to generate complete descriptions by composing individual source models and performing exact or approximate inference. In such cases machines can outperform human (e. g., [4]). However, it is unclear how such approaches can be usefully applied to handle complex everyday scenes containing unknown numbers of dynamically changing sources with unpredictable onsets and offsets.

In contrast to the above, 'complete description' problem, we can consider machine listening that adopts a more human version of scene analysis where there are favoured 'attended sources' (i. e., a 'foreground') and unattended sources that are allowed to remain unresolved in the background. Such systems would not form complete scene descriptions, but would instead try to mimic the human ability to fluidly switch attention between alternative 'foregrounds', driven by high-level goals or by the saliency of the competing sources (see [3]). A simple version of the approach is exemplified in the fragment decoding technique for robust speech recognition ([1, 2]): simple source-independent models are used to perform a local decomposition into acoustic 'fragments' and then, at a higher level, fragments are integrated over time by composing detailed models of the target speaker mixed with much simpler models of the background. However, within any attention-driven framework, where foreground and background are treated asymmetrically, there exist many unresolved questions. How can the complexity of the foreground and background models be balanced so as to maximise performance at a fixed computational cost? What are the dimensions of auditory salience that drive attention? How to model 'top-down' selective attention? How to model 'bottom-up' reflexive attention? In particular how much processing of the background is required in order to be aware of the salient qualities (particularly with respect to 'top-down salience'?

References

1 J. P. Barker, M. P. Cooke, and D. P. W. Ellis. *Decoding speech in the presence of other sources.* Speech Communication, vol. 45, no. 1, pp. 525, Jan. 2005.

2 N. Ma and J. Barker. *A fragment-decoding plus missing-data imputation system evaluated on the 2nd CHiME challenge.* Proceedings of the 2nd CHiME Workshop on Machine Listening in Multisource Environments, Vancouver, Canada, Jun. 2013, pp. 5358.

3 M. I. Posner and S. E. Petersen. *The attention system of the human brain.* Annual Review of Neuroscience 13: 2542, 1990.

4 S. Rennie, J. R. Hershey and P. A. Olsen. *Single Channel Multi-talker Speech Recognition: Graphical Modeling Approaches.* IEEE Signal Processing Magazine, Special issue on Graphical Models, November 2010.

5.2 Compensate Lexical/Speaker/Environment Variability for Speech Emotion Recognition

Carlos Busso (The University of Texas at Dallas, US)

Affect recognition is a crucial requirement for future human machine interfaces to effectively respond to nonverbal behaviors of the user. Speech emotion recognition systems analyze acoustic features to deduce the speaker's emotional state. However, human voice conveys a mixture of information including speaker, lexical, cultural, physiological and emotional traits. The presence of these communication aspects introduces variabilities that affect the performance of an emotion recognition system. Therefore, building robust emotional models requires careful considerations to compensate for the effect of these variabilities. Important research issues are concerned with normalization schemes that compensate the variability introduced by multiple communication aspects not related to emotions. These approaches include environment, speaker, and lexical normalization.

5.3 Interpretation and Computational Audio Analysis

Laurence Devillers (LIMSI – CNRS, FR)

URL http://www.limsi.fr/Scientifique/emotions/

Most of the research on Computational Audio Analysis has been on classifying the surface phenomena associated with acoustic signals and with speech events. The meaning of these events usually depends on the context in which they occur. The analysis of audio (and video) scenes can help machines to interpret speech of humans or of human-machine interactions. One of the important issues is how to decide which contextual information to acquire and how to incorporate it into machine learning. Machines should be able to deal with interactions with multi-speakers and interpret the relationship between speakers. To give to the machines the capabilities to interpret and generate appropriate signals taking into account the context of the interaction (with multi-sources analysis) is a real challenge.

5.4 Perceptually Appealing Reconstruction of Spectrally Modified Signals

Jonathan Driedger, Meinard Müller (Universität Erlangen-Nürnberg, DE)

In many audio processing tasks such as source separation or time-scale modification, the audio signal is modified in the spectral domain and then resynthesized by applying some inverse transform. Examples are binary or relative masking in source separation procedures or phase propagation techniques as used in the phase vocoder. However, appyling such modifcations typically ignore the complex relationships between phases and magnitudes of superimposed sound components. As a result, besides the intended effects, the reconstructed signals often contain unwanted artifacts. In this seminar, we have raised the question of how to evaluate the quality of reconstructed sigals. Further issues were how artifacts may be reduced using phase adaption strategies or perceptually masked using suitable post-processing techniques. A fundamental observation was that a listener's expectation of how a modified signal should sound often diffes to what is actually contained in the data. This has shown that tasks such as time scale modification or source separation (without any further applications) are highly subjective and ill-posed problems.

5.5 The Situated Multimodal Facets of Human Communication

Anna Esposito (International Institute for Advanced Scientific Studies, IT)

Main reference M. Rojc, N. Campbell, (eds), "Coverbal Synchrony in Human-Machine Interaction," 434 pp., ISBN: 9781466598256, CRC Press, 2013.

Humans interact with each other through a gestalt of emotionally cognitive actions which involve much more than the speech production system. In particular, in human interaction, the verbal and nonverbal communication modes seem to cooperate jointly in assigning semantic and pragmatic contents to the conveyed message by unraveling the participants cognitive and emotional states and allowing the exploitation of this information to tailor the interactional process. These multimodal signals consist of visual and audio information that singularly or combined may characterize relevant actions for collaborative learning, shared understanding, decision making and problem solving. This work will focus on the visual and audio information including contextual instances, hand gestures, body movements, facial expressions, and paralinguistic information such as speech pauses, all grouped under the name of nonverbal data, and on the role they are supposed to play, assisting humans in building meanings from them.

5.6 Bayes and Beyond Bayes: The Integration of Prior Knowledge

Sebastian Ewert (Queen Mary University of London, GB)

To analyse audio recordings using automated methods, one typically makes assumptions about characteristics and properties of the recorded content. Such assumptions can be explicit or implicit, and can exist on various semantic levels. For example, in music processing, methods often exploit that most musical instruments produce harmonic sounds to analyse the musical content or to identify individual sound sources. Similarly, in speech recognition, methods rely on the fact that different utterances of a specific phoneme are in a common manifold of a feature space, which can be described using probabilistic models. Even in methods, which are generally considered as unsupervised, one can find various implicit assumptions. For example, in methods such as NMF, one exploits that many sounds can be approximated by a convex combination of a few fixed spectral templates which would not be true for highly non-stationary sounds or noise. Also the number of templates used in NMF is typically based on some kind of assumption.

All these different assumptions can be considered as a form of prior knowledge and, in this sense, prior knowledge is an essential component in every signal analysis method. Still, it is not always clear how prior knowledge is integrated best. Some types of prior knowledge only loosely correlate with specific signal properties, and it might not be clear whether the integration of such prior knowledge is useful at all. It is also not always clear how the prior knowledge can be integrated. In particular, while prior distributions in Bayesian probabilistic models have been used successfully in recent years in this context, whether they can or should be used to represent a specific type of knowledge. Furthermore, prior knowledge can be available on various semantic levels. For example, a musical score provides high-level information about pitch and timing of note events, which can be used to simplify extremely complex problems such as source separation.

In this seminar, I asked and discussed with other participants the following questions. What kind of implicit and explicit prior knowledge are you facing in your work? How are you using prior knowledge in your methods? What kind of general strategies exist to integrate prior knowledge in front end transformations, in signal and acoustic models, in backend and machine learning components? What is your experience with prior knowledge on various semantic levels? What strategies do you employ to integrate knowledge beyond Bayesian modelling?

5.7 NMF meet Dynamics

Cédric Févotte (JL Lagrange Laboratory – Nice, FR)

Over the last ten years nonnegative matrix factorisation (NMF) has become a popular unsupervised dictionary learning and adaptive data decomposition technique with applications in many fields. In particular, much research about this topic has been driven by applications in audio, where NMF has been applied with success to automatic music transcription and single channel source source separation. In this setting, the nonnegative data is formed by the magnitude or power spectrogram of the sound signal and is decomposed as the product of a dictionary matrix containing elementary spectra representative of the data times an activation matrix which contains the expansion coefficients of the data frames in the dictionary.

In my own research, I have worked on model selection issues in the audio setting, pertaining to the choice of time-frequency representation (essentially, magnitude or power spectrogram), and to the measure of fit used for the computation of the factorisation. Driven by a probabilistic modelling approach, I came up with arguments in support of factorizing of the power spectrogram with the Itakura-Saito (IS) divergence [1]. Indeed, IS-NMF is shown to be connected to maximum likelihood estimation of variance parameters in a well-defined statistical model of superimposed Gaussian components and this model is in turn shown to be well-suited to audio. In my work, I have also addressed variants of IS-NMF, namely IS-NMF with temporal regularisation of the activation coefficients [2], automatic relevance determination for model order selection [3] and multichannel IS-NMF [4].

Recently, I have started to look into dynamical variants of NMF [5], in which structured transitions occur from spectral patterns to others. This is a desirable property for example for speech signals, for which some temporal correlation (or anti-correlation) is expected to occur between subset of speech patterns. Introducing dynamics into NMF is a challenging task at the modelling and estimation levels. To put it simply, one might say that NMF has superseded the traditional GMM. If HMM is the natural dynamical extension to GMM, what is the natural dynamical extension to NMF?

References

1 C. Févotte, N. Bertin, and J.-L. Durrieu. *Nonnegative matrix factorization with the Itakura-Saito divergence with application to music analysis.* Neural Computation, 21(3):793–830, Mar. 2009.

2 C. Févotte. *Majorization-minimization algorithm for smooth Itakura-Saito nonnegative matrix factorization.* In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, May 2011.

3 V. Y. F. Tan and C. Févotte. *Automatic relevance determination in nonnegative matrix factorization with the beta-divergence.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(7):1592–1605, July 2013.

4 A. Ozerov and C. Févotte. *Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation.* IEEE Transactions on Audio, Speech and Language Processing, 18(3):550–563, Mar. 2010.

5 C. Févotte, J. Le Roux, and J. R. Hershey. *Non-negative dynamical system with application to speech and audio.* In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, Canada, May 2013.

5.8 Features beyond Machine Learning

Martin Heckmann (Honda Research Europe, DE)

A large part of my research concentrates on the extraction of features from speech signals: on the one hand for recognizing what was said and on the other hand how it was said. In the development of these features I try to combine three ingredients: first, the usage of domain knowledge; second, taking inspirations from what is known of the processing in the human brain (e.g. high dimensional sparse representations); and, third, machine learning approaches. One example is a set of hierarchical spectro-temporal features, which build on a perceptual representation and form sparse and high-dimensional features learned from unlabeled speech data [1]. Currently, I am particularly interested in the extraction of more subtle prosodic variations which play a very important role in human communication. This includes back-channels which indicate how the listener is following the conversation as well as the prominence different words receive which is related to the importance a speaker attributes to a word [2, 3]. Here, the domain knowledge is one of the key ingredients so far. However, in recent years I experience a trend away from extensive domain knowledge and psychophysical inspirations more towards approaches based on machine learning. The different paralinguistic challenges at INTERSPEECH by Schuller et al. are a prime example as how the same set of features can successfully be applied to many different tasks with the right machine learning backend [4]. Related but a bit different is the tremendous success of Deep Neural Networks in the last two years. Currently they are used as a powerful and versatile tool of machine learning which is particularly suited to exploit the rich information provided by very large datasets. Furthermore, researchers have also started trying to integrate inspirations from the processing of the human brain in this approach such as convolutional networks. In this seminar, I have discussed ideas for methodologies to fruitfully integrate the rapid advances in machine learning with processing principles in the brain and domain knowledge to come up with better features.

References

1 M. Heckmann, X. Domont, F. Joublin, and C. Goerick. *A hierarchical framework for spectro-temporal feature extraction.* Speech Communication, Vol. 53, No. 5, pp. 736–752, 2011.

2 M. Heckmann. *Audio-visual evaluation and detection of word prominence in a human-machine interaction scenario.* In Proc. INTERSPEECH, Portland, OR, 2012.

3 M. Heckmann. *Inter-speaker variability in audio-visual classification of word prominence.* In Proc. INTERSPEECH, Lyon, France, 2013.

4 B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan. *Paralinguistics in speech and language–State-of-the-art and the challenge.* Computer Speech & Language, Vol. 27, No. 1, pp. 4–39, 2013.

5.9 Detection of Repeated Signal Components and Applications to Audio Analysis

Frank Kurth (Fraunhofer FKIE – Wachtberg, DE)

Main reference F. Kurth, "The Shift-ACF: Detecting Multiply Repeated Signal Components," in Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'13), pp. 1–4, IEEE, 2013.
URL http://dx.doi.org/10.1109/WASPAA.2013.6701820

Our work in the last two years was mainly concerned with the detection of structured audio components within source signals. In this, an important type of structure are repetitions such as repeating bird calls or percussive elements in music. A few months ago, we have proposed a novel technique for detecting multiply (i. e., more than once) repeated signal components within a target signal. For such cases, we were able to improve classical autocorrelation techniques. In our experiments, we up to now have successfully considered applications in bioacoustics and in speech processing. It was interesting to discuss the topic within an interdisciplinary community as it was present at the Dagstuhl seminar and to learn about further possible applications—and existing solutions—from other domains, especially when dealing with noisy or distorted signals. For me, related interesting questions are both how to automatically separate, or even extract, all structured signal parts from the residual signal and how to do this efficiently for large scale signal scenarios.

As a first follow-up activity to the Dagstuhl seminar, I am organizing a special session on "Audio Signal Detection and Classification" covering topics such as audio monitoring, signal detection, segmentation and classification, audio fingerprinting, matching techniques, and audio information retrieval. The special session, which will be held at the IEEE Workshop on Cloud Computing for Signal Processing, Coding and Networking (IWCCSP) on March 11, 2014, aims at bringing together experts from the audio signal processing area with the cloud computing community.

5.10 Informed Source Separation for Music Signals

Meinard Müller, Jonathan Driedger (Universität Erlangen-Nürnberg, DE)

Joint work of Sebastian Ewert

Main reference S. Ewert, B. Pardo, M. Müller, M. Plumbley, "Score-Informed Source Separation for Musical Audio Recordings," IEEE Signal Processing Magazine, 2014, to appear.

One central problem in music processing is the decomposition of a given audio recording of polyphonic music into components that correspond to the various musical voices or instrument tracks. The main challenge arises from the fact that musical sources are highly correlated, share the same harmonies, follow the same rhythmic patterns, and so on. Musicians play together, follow the same lines, and interact with each other. As a consequence, different musical voices often do not differ statistically from each other, which makes the separation of musical sources or voices infeasible and and even ill-defined problem. Therefore, when processing music data, music-specific techniques are needed that exploit musical knowledge or music-specific constraints. For example, to support the separation of musical voices, one strategy is to use additional cues such as the musical score or user input [1]. In this seminar, we have discussed questions such as: Is source separation of music signals a meaningful problem? What is it good for? What are possible applications? How can one measure the success and the complexity of the task? How can one integrate additional knowledge? Where does one obtain such knowledge from? How can this knowledge be learned from example data? What can be learned from the field of speech processing?

References

1 S. Ewert, B. Pardo, M. Müller, M. Plumbley. *Score-Informed Source Separation for Musical Audio Recordings*. IEEE Signal Processing Magazine, 2014.

5.11 Approaching Cross-Audio Computer Audition

Björn Schuller (TU München, DE)

Main reference B. Schuller, "Intelligent Audio Analysis," 345 pp., Signals and Communication Technology Series, Springer, 2013.
URL http://www.springer.com/engineering/signals/book/978-3-642-36805-9

Substantial progress has been made over the last years in a number of intelligent audio analysis sub-disciplines that lead closer to the realisation of genuine cross-audio computer audition. This includes in particular advances in blind audio source separation such as by Non-negative Matrix Factorisation variants, but also in the feature extraction and computational intelligence parts, e. g., by feature brute-forcing, or context-sensitive deep neural networks and tandem architectures with graphical model topologies, or recent transfer learning approaches. By these and further means, the community is at a point where we are able to shift more into handling complex compounds of speech, music, and sound simultaneously as this is how they appear in the real world [1]. In this seminar, we have discussed important tools and inspirations on how to proceed on this avenue.

References

1 Björn Schuller. *Intelligent Audio Analysis.* Signals and Communication Technology, Springer, 2013.

5.12 What can we Learn from Massive Music Archives?

Joan Serrà (IIIA – CSIC – Barcelona, ES)

Music is an extremely powerful means of communication that shapes our brain in intricate ways, unique to mankind, and transversal to all societies. As a scientific community we are slowly but steadily progressing towards the availability of massive amounts of music and music-related data for research purposes. The Million Song Dataset, Peachnote, the Yahoo! Music Dataset, the Last.fm API, Musicbrainz, or Wikipedia are just but some examples. Certainly, such big data availability will shift the perspective in which we approach many (if not all) of the traditional music information retrieval tasks. From genre or mood classification to audio or cover song identification, practically all tasks will experience a change of paradigm that frame them under more realistic, large-scale scenarios. In this seminar, we discussed new avenues for research that are awaiting for us. In particular, future work will be concerned about extracting and using knowledge that can be distilled from such massive amounts of data—not only knowledge about music itself (rules, patterns, anti-patterns, and their evolution), but also knowledge about ourselves, as music listeners, users, or creators.

5.13 Acoustic Monitoring in Smart Home Environments: A Holistic Perspective

Stefano Squartini (Polytechnic University of Marche, IT)

Joint work of Emanuele Principi
Main reference E. Principi, S. Squartini, F. Piazza, D. Fuselli, M. Bonifazi, "A distributed system for recognizing home automation commands and distress calls in the Italian language," in Proc. of the 14th Annual Conf. of the Int'l Speech Communication Asssociation (INTERSPEECH'13), pp. 2049–2053, ISCA, 2013.
URL http://www.isca-speech.org/archive/interspeech_2013/i13_2049.html

In recent years, there has been significant interest around the "Smart Home" paradigm, a scenario where several research fields seem to naturally converge. One of the most relevant objectives consists in monitoring the activity of inhabitants for different purposes: emergency state recognition and fall detection (especially for elderly people), intrusion or theft detection, people localization, usage of appliances, or power consumption besides the more common home automation commands, which have been already implemented in many commercial entertainment-oriented devices. In this context, acoustic monitoring techniques play an important role. Even though many scientific studies have been conducted so far, the results do not yet seem to match the market expectations.

Our research group is developing distributed system for recognizing home automation commands and distress calls in Italian language. The system integrates the automatic recognition of emergency states and home automation commands with remote assistance and hands-free communication. The ITAAL database has been developed for this purpose and a preliminary prototype is already available. Nevertheless, many issues still need to be addressed in order to make the system more appealing, reliable and useful for exploitation in real domestic environments. This typically requires dealing with heterogeneous acoustic data, which must be treated by looking at them from a holistic perspective, also taking other types of sensing activity into account. Some of these issues are reported here as open challenges to be addressed in future research:

- How to integrate speech and sound analysis for activity monitoring? Utterances spoken by a user, even if not really related to specific commands devoted to activate certain smart functionalities, can be useful to understand what the user is doing and in which part of the house he is located, specially if adequately integrated with no-speech sounds related to his activity.
- How to integrate information coming from infra- and ultra-sound sensors? Spanning the frequency range beyond the audible range can be very useful (e. g., subsonic sounds for fall detection and ultrasonic sounds for localization), especially in an integrated fashion with the "real" acoustic information. Therefore, unsupervised learning techniques can be implemented to find out and efficiently use cross-domain relationships.
- What is the role of paralinguistic features? In emergency state recognition, for instance, the capability of detecting the presence of paralinguistic features in the vocal activity,

and likely understanding their meaning, can have a substantial impact in the overall performance and asks for consideration in smart home environments.

- How to deal with minimum a-priori knowledge? In several practical smart home scenarios, the adaptation of automatic recognition systems to a speaker's characteristics is not allowed, since the provided technology should be as transparent as possible to the final user.
- How to deal with the "novelty" issue? One of the objectives of acoustic monitoring consists in automatically recognizing a novel event with respect to the "usual" ones, in order to take adequate actions (e. g. in case of thefts).

5.14 Sound Processing in Everyday Environments

Emmanuel Vincent (INRIA – Nancy – Grand Est, FR)

I am interested in an efficient integrated approach to sound processing in everyday environments. The various relevant tasks are often treated one after another: source localization, source separation, speaker/event identification, speech recognition. This "pipeline" approach yields suboptimal results due to the propagation of errors from one step to the next. Our approach is to propagate not only deterministic signals and values but a full posterior distribution (which is approximated as a Gaussian) from one step to the next. Some techniques exist to estimate this distribution but they are not very accurate yet. Burning questions in this context are: How to accurately estimate and propagate uncertainty? How to use it in combination with state-of-the-art ASR and speaker/event identification systems?

References

1 N. Duong, E. Vincent, and R. Gribonval. *Spatial location priors for Gaussian model based reverberant audio source separation.* EURASIP Journal on Advances in Signal Processing, Springer, 2013.

2 A. Ozerov, M. Lagrange, and E. Vincent. *Uncertainty-based learning of acoustic models from noisy data.* Computer Speech and Language, Elsevier, 2013, 27 (3), pp. 874-894.

3 J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green. *The PASCAL CHiME Speech Separation and Recognition Challenge.* Computer Speech and Language, Elsevier, 2013, 27 (3), pp. 621-633.

4 E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni. *The Second 'CHiME' Speech Separation and Recognition Challenge: An overview of challenge systems and outcomes.* IEEE Automatic Speech Recognition and Understanding Workshop, 2013.

6 Towards a Definition of Computational Audio Analysis (CAA)

Towards the end of the seminar, each participant was asked to give some kind of definition for a research field we coined "Computational Audio Analysis" (CAA). The following list gives an overview of the various statements which, as a whole, also give a good impression about the range of topics we have discussed at our Dagstuhl seminar.

- Computational audio analysis provides quantitative methodologies that enable detailed analyses of human behavior and interaction.

- The goal of computational audio analysis (CAA) is to understand the underlying structure of a given audio recording using computational methods in order to extract information and higher level semantics. CAA is a highly multidisciplinary field bringing together researchers in computer science, digital signal processing, machine learning with domain experts in, for example, speech and music processing, human computer interaction, biology, medicine, and acoustics.
- In the field of artificial intelligence (AI), people have tried to realize machines that simulate the role of a human. Examples are the Deep Blue system in chess or the Watson system for answering questions posed in natural language. Computational audio analysis aims at realizing machines that hear and understand sounds like a human. However, it seems to be a long way to realize such system, so that CAA remains an exciting research area.
- Computational audio analysis aims at inferring meaningful structures from audio signals, finding hidden relationships in heterogeneous collections of acoustic data from multiple perspectives, as well as detecting and understanding the meaning of events as occurring in natural environments
- The objective of computational audio analysis is to give a generative explanation of a sound complex, where a soundscape is decomposed with sufficient fidelity to meet the needs of particular applications.
- Computational audio analysis aims at extracting information from audio signals using techniques from signal processing, machine learning, information retrieval, and related fields. One central objective is to segment, structure, and decompose audio signals into elementary units that have some semantic (e. g. linguistic, musical) meaning. These units not only serve as basis for higher level analysis and classification tasks but also deepen the understanding of the underlying acoustic material.
- Computational audio analysis refers to the modeling and analysis of audio, in particular the voice, with the goal to extract 'meaningful information' from audio. What 'meaningful information' means, depends on the respective application. Inferring interactive events or states from audio, classifying environmental sound events, or separating sources to improve ASR are such examples tasks. Furthermore, CAA provides us with techniques to automatically attribute labels or perceptual characteristics to sounds.
- Computational audio analysis—in an utmost compact description—essentially focuses on extracting information from audio using computational methods.
- Computational audio analysis is the processing of audio signals in order to characterize or decode them in a way humans can understand. It incorporates signal processing techniques as well as models of perception and cognition. The main difficulty is that it needs to model a highly complex system with large inter-subject variability: human listeners.
- Computational audio analysis means resolving audio into machine understandable constructs.
- In computational audio analysis, annotation-flexible models that adapt to new conditions are developed in order to achieve a more representative (machine) learning outcome. Furthermore, the interplay between speech signals, other human-produced signals such as physiological signals (heart rate, skin conductance, activity, gestures), and non-speech audio signals (e.g. cough, snoring, sneezing) are explored. The understanding of how and why audio influences the human mind using low-level features can open possibilities for new application.

- Computational audio analysis means extracting knowledge from audio and making sense of it.
- Computational audio analysis provides computational methods for finding relevant structure (pertinent features and class labels as well as appropriate decompositions) in acoustic data, where relevancy, pertinence and appropriateness are usually defined in a task-dependent way. Methods are not limited to acoustic data, but can use multi-modal input, as long as audio data is among the considered modalities.
- Computational audio analysis aims at the detection, separation and description of acoustic objects via computational means. Descriptions can be of a qualitative (e. g., "warm") and quantitative (e. g. "70 dB SPL") nature. The source of these acoustic objects can be a human speaking, real or virtual musical instruments being played, or other vibrating physical objects such as loudspeakers. In addition to the analysis of separable acoustic objects, CAA also targets at holistic descriptions of acoustic scenes or parts of a scene (e. g., being at a train station).
- Computational audio analysis is the automated analysis of acoustic signals (whether natural or man-made) in order to perform some task that has utility to humans. There are no restrictions on the task: the setting may be online/offline, unimodal/multimodal, passive/interactive and may involve any form of acoustic signal including speech, music or environmental audio. The analysis may use perceptually motivated features (e. g., MFCCs) or perceptually motivated processing. However, in contrast to computational audition, the processing does not need to follow human audio processing, i. e., it does not explicitly model human hearing and the field is not concerned with learning about human hearing from human/machine comparisons.
- Computational audio analysis is a way to describe the effect that audio (both naturally occurring and artificially synthesized) has on humans, independent of language, linguistics, or phonetics. Due to the difficulty of describing its "targets" with words, or measure its physiological effects exactly, labels are very hard to get by. This makes CAA a challenging combination of fields such as computer science, musicology, psychology, or physiology.
- Computational audio analysis is the analysis and interpretation of an acoustic scene by a machine. This analysis can be either obtained in a supervised way, which is guided by human perception or sound production mechanism when known, or it can be unsupervised with the aim, for example, to discover new concepts (such as sound objects or sound primitives) not necessarily formally defined in advance by humans.
- Computational audio analysis is about machine-assisted extraction of information from sound. It can be either fully automatic (unsupervised) or user-guided (semi-supervised).
- Speech conveys information beyond verbal message including intentions, emotions, and personality traits that influence the way we communicate with others (people, robots, computers, devices). Computational audio analysis offers the opportunity to develop tools for learning and inferring these traits. The challenge in building such systems is to capture the temporal dynamic and situational context of behaviors.
- Computational audio analysis is the processing and modeling of the inherently heterogeneous general audio signal to uncover latent structures, to derive mappings to and between representations of interest, and to empower target applications such as summarization, retrieval, synthesis, and categorization. As a special case, the computational representations and formalism of CAA can benefit from human audition principles.
- Computational audio analysis is about processing audio data with respect to a specific application scenario and domain knowledge in order to extract task-specific information.
- Computational audio analysis is concerned with the extraction of a parametric description

for an audio signal from its waveform (and possibly other additional representations). The type of the description varies depending on the requirements of the desired application.

- Computational audio analysis aims at understanding audio by means of computational means. This could mean being able to build a model of the source (source modeling and separation) extracting relevant messages (speech recognition), or understanding the environment the sources are in (e. g., room ID through reverb). CAA is open to any and all computational methods to do so (including semantic web, crowd sourcing).
- Computational audio analysis involves the processing of audio signals by the help of computers with the objective to obtain information from it. Such information can refer different levels of abstraction ranging from basic signal measurements and low-level features to semantically meaningful information such as words, emotions, or melodies.
- Computational audio analysis is the intersection of audio analysis by digital means (i. e., digital signal processing) with computer science. It therefore might include any relevant aspect of computer science, including but not limited to logic, inference, representation (ontologies), HCI, information retrieval, machine learning, cryptography and encryption, autonomous agents, communication (not telecommunication) theory, and so on. It should develop computational means and mechanisms for transitioning from audio data, to audio information, to (audio) knowledge and understanding for all forms of audio, i. e., speech, music, environmental, making that information and knowledge useable in a wide variety of application domains, including creative activity. It does not exist in isolation and has close ties to other sensory and affective data/modalities. It embraces the representational power of Semantic Web technologies which empowers many of the areas of computer science above in the linked data world of the future.
- By audio, we deal with mechanical waves, i. e., a complex series of changes in or oscillation of pressure as compound of frequencies within the acoustic range available to humans and at sufficiently intense level to be perceived, i. e., audible by them. The analysis of audio aims at the extraction of information and, on a higher level, attachment of semantic meaning to audio signals. Computational audio analysis typically includes the involvement of computational intelligence algorithms as provided by the means and methods of machine learning going beyond signal processing.
- Computational audio analysis deals with rich (audio) data and a complex (audio) signals. It encompasses a variety of aspects such as the analysis of spoken language, the mood of a song, and the human interaction including feelings and emotion.
- Computational audio analysis is the engineering approach to reproduce the human capability of processing sounds to understand acoustic scenes and respond appropriately to the environment.
- Computational audio analysis deals with the analysis of audio in combination with other sensor information such as video, body sensors, GPS, and so on. The analysis of such data is generally statistical, deep, atheoretical, and hard for people to understand. CAA should be time- and context-dependent. It may involve continuous adaptation and may incorporate protension.
- Computational audio analysis is the use of computers (from microprocessors over smartphones to supercomputers) for the analysis of audio signals (acquisition and storage, feature extraction, model building and interpretation) with applications in telecommunications, multimedia, automotive, industry, biomedicine, performing arts, forensics, human curiosity, and science.

Participants

- Xavier Anguera
Telefónica Research – Barcelona, ES
- Jon Barker
University of Sheffield, GB
- Stephan Baumann
DFKI – Kaiserslautern, DE
- Murtaza Bulut
Philips Research Lab. – Eindhoven, NL
- Carlos Busso
The University of Texas at Dallas, US
- Nick Campbell
Trinity College Dublin, IE
- Laurence Devillers
LIMSI – CNRS, FR
- Jonathan Driedger
Univ. Erlangen-Nürnberg, DE
- Bernd Edler
Univ. Erlangen-Nürnberg, DE
- Anna Esposito
Intern. Institute for Advanced Scientific Studies, IT
- Sebastian Ewert
Queen Mary University of London, GB
- Cédric Févotte
JL Lagrange Laboratory – Nice, FR
- Jort Gemmeke
KU Leuven, BE
- Franz Graf
Joanneum Research – Graz, AT
- Martin Heckmann
Honda Research Europe, DE
- Dorothea Kolossa
Ruhr-Universität Bochum, DE
- Gernot Kubin
TU Graz, AT
- Frank Kurth
Fraunhofer FKIE – Wachtberg, DE
- Sungbok Lee
Univ. of Southern California, US
- Florian Metze
Carnegie Mellon University, US
- Roger K. Moore
University of Sheffield, GB
- Emily Mower Provost
University of Michigan – Ann Arbor, US
- Meinard Müller
Univ. Erlangen-Nürnberg, DE
- Shrikanth S. Narayanan
Univ. of Southern California, US
- Nobutaka Ono
National Institute of Informatics – Tokyo, JP
- Bryan Pardo
Northwestern University – Evanston, US
- Alexandros Potamianos
National TU – Athens, GR
- Bhiksha Raj
Carnegie Mellon University, US
- Gaël Richard
Telecom ParisTech, FR
- Mark Sandler
Queen Mary University of London, GB
- Björn Schuller
TU München, DE
- Joan Serrà
IIIA – CSIC – Barcelona, ES
- Rita Singh
Carnegie Mellon University, US
- Paris Smaragdis
University of Illinois at Urbana Champaign / Adobe, US
- Stefano Squartini
Polytechnic Univ. of Marche, IT
- Shiva Sundaram
Audyssey Laboratories, US
- Khiet Truong
University of Twente, NL
- Christian Uhle
Fraunhofer IIS – Erlangen, DE
- Emmanuel Vincent
INRIA – Nancy – Grand Est, FR
- Tuomas Virtanen
Tampere University of Technology, FI