Technical Requirements for a Personalized Auditory Reality

Karlheinz Brandenburg[1],[2], Estefanía Cano Cerón[2], Florian Klein[1],
Thomas Köllmer[2], Hanna Lukashevich[2], Annika Neidhardt[1],
Johannes Nowak[1],[2], Ulrike Sloma[1], Stephan Werner[1]

[1] *Technische Universität Ilmenau, 98693 Ilmenau, Deutschland, Email: karlheinz.brandenburg@tu-ilmenau.de*

[2] *Fraunhofer Institute for Digital Media Technology IDMT, 98693 Ilmenau, Deutschland*

Abstract

At DAGA 2018, we introduced the concept of a Personalized Auditory Reality (PARty) [3], a new research field that investigates methods for manipulation of acoustic surroundings. Within such an auditory reality, users would be able to freely move around, modify their acoustic scene by enhancing relevant sounds, suppressing irrelevant ones, or adding new ones. The perceived acoustic environment will follow the paradigm of augmented and mixed realities where sounds from actual surroundings are edited and combined with added sound sources. In order to accomplish the main tasks of PARty via a wearable device, a number of requirements need to be met. This contribution outlines necessary characteristics of both hardware and software components. We see, among others, the following questions: Which existing components and systems can be used? How do we measure the quality of prototypes? What computing power is available?

A closer look is taken for some of the key components of a PARty system: Decomposing a real-world acoustic scene with the help of a small microphone array, real-time object classification and integration of virtual sound sources in the actual environment.

What is a PARty

As introduced at DAGA 2018, a Personalized Auditory Reality (PARty) is a mixed reality system enhancing our way to listen to sounds of all kinds. Many of us are used to wearing glasses. We take their capability to enhance our visual sense for granted. Many of us have experienced a situation where we search for the glasses only to find them just on the nose.

PARty systems should do the same for hearing: We expect total transparency in "normal" circumstances. If we find noises around us too loud, we want to listen to music (actual or virtual), we want to concentrate our listening to understand certain persons, we want to just lower the voice of a certain group of loud bystanders, we could do that with a PARty system.

This sounds like a science fiction scenario, but at the same time it is the name for a number of research activities.

Main ingredients of a Personalized Auditory Reality

The following paragraphs shortly describe technical subsystems necessary to implement a PARty. To implement such an audio mixed reality system, real world input, user interaction and improved binaural rendering, all done in real-time is needed so that it feels perfectly plausible.

Scene decomposition

Humans can "selectively" hear by nature and consciously focus on individual sound sources in their environment. An automatic system for selective hearing using artificial intelligence (AI) must first learn the underlying concepts. The automatic decomposition of acoustic scenes first requires recognition and classification of all active sound sources followed by a separation in order to process, amplify or attenuate them as separate audio objects.

The research field auditory scene analysis tries to detect and classify both time-localized sound events such as footsteps, clapping or screaming as well as more global acoustic scenes such as concerts, restaurants or supermarkets on the basis of a recorded audio signal. Current methods exclusively use AI and deep learning techniques. This involves data-driven training of deep neural networks, which learn to recognize characteristic patterns in the audio signal on the basis of large datasets of training [19]. Inspired by progress in the research areas of image processing (computer vision) and speech processing (natural language processing), mixtures of convolutional neural networks for two-dimensional pattern recognition in spectrogram representations and recurrent neural networks for the temporal modelling of sounds are usually used.

For audio analysis, there are a number of specific challenges that need to be overcome. Deep learning models are very data-hungry due to their complexity. Compared to the research areas image processing and speech processing, only relatively small data sets are currently available for audio processing. The largest dataset is the AudioSet dataset from Google [11] with about 2 million sound samples and 632 different sound event classes, most of which with less data. This small amount of training data can be addressed, for example, by transfer learning, in which a model pre-trained on a large data set is then fine-tuned to a smaller data set intended for the application with new classes (fine-tuning) [1]. Furthermore, methods from semi-supervised learning are used to

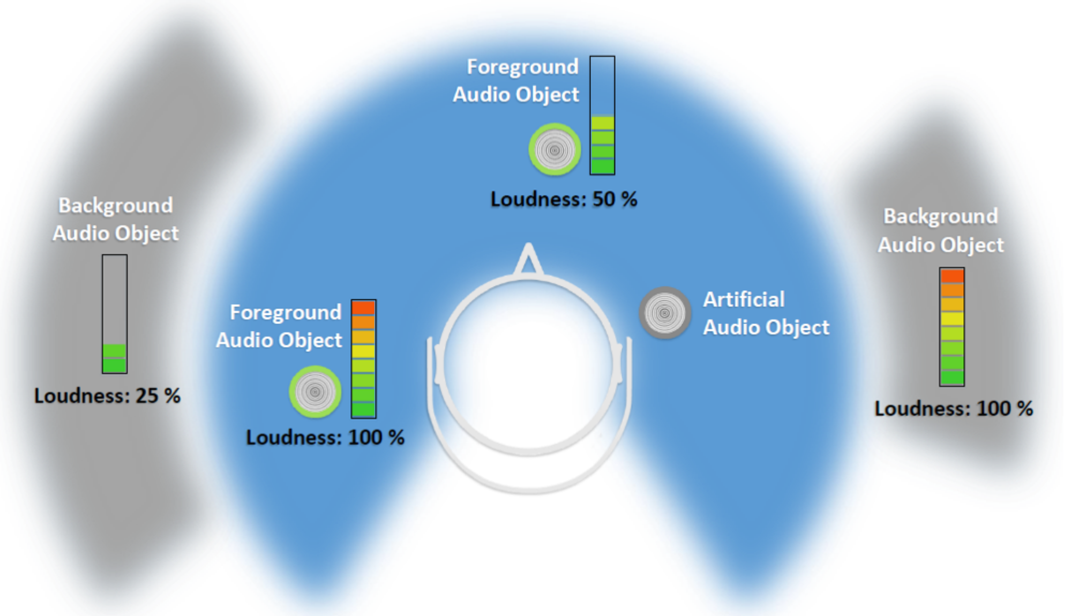

Figure 1: Exemplary use case of PARty: The user can modify (supress or accentuate) real sound sources or add artificial ones.

include the generally large amount of unannotated audio data in the training.

Another essential difference to image processing is that simultaneously audible acoustic events do not cover sound objects (as with images) but result in a complex phase-dependent superposition. Current algorithms in deep learning use so-called "attention" mechanisms, which enable the models, for example, to focus on certain time segments or frequency ranges during classification [21]. The recognition of sound events is further complicated by the high variance in their duration. Algorithms should be able to detect both very short events such as a pistol shot and long events such as a passing train.

Due to the strong dependence of the models on the acoustic conditions during the recording of the training data, they often show unexpected behavior in new acoustic environments, which differ e.g. in room reverberation or microphone setup. Various solutions have been developed to alleviate this problem. Data augmentation, simulation of different acoustic conditions and artificial overlapping of different sound sources are used to achieve a higher robustness of the models [16]. Furthermore, the parameters in complex neural networks can be regularized, so that overtraining & specialization on the training data is prevented and at the same time a better generalization on unseen data is achieved. In the last years different algorithms for "domain adaptation" [12] have been proposed to adapt already trained models to new application conditions.

In the application scenario planned in PARty, a real-time capability of the sound source recognition algorithms is of elementary importance. In this case, it is inevitable to weigh up the complexity of the neural network against the maximum possible number of computational operations on the underlying computing platform. Even if a sound event has a longer duration, it must still be detected as quickly as possible in order to start a corresponding source separation.

Several sound sources must be assumed and their number and type is initially unknown and can change constantly. For the separation of sound sources, several sources with similar characteristics, such as several speakers, are particularly challenging [13]. In order to achieve a high spatial resolution, several microphones in the form of an array must be used [4, 2]. In contrast to usual audio recordings in mono (1 channel) or stereo (2 channels) such a recording scenario allows an exact localization of the sound sources around the listener.

Source separation algorithms usually cause artifacts such as distortion and crosstalk between sources [9], which are generally perceived as annoying by the listener. By remixing the tracks, however, such artifacts can be partially masked and thus reduced [15]. To improve blind source separation, additional information such as the number and type of sources detected or their estimated spatial position is often used (Informed Source Separation [14]). For meetings, in which multiple speakers are active, current analysis systems can simultaneously estimate the number of speakers, determine their temporal activity, and then isolate them by source separation [18].

In recent years, Fraunhofer IDMT has conducted multiple studies on the perception-based evaluation of sound source separation algorithms [6]. In the field of music signal processing a real-time capable algorithm for the separation of the solo instrument and the accompanying instruments was developed, which uses a basic fre-

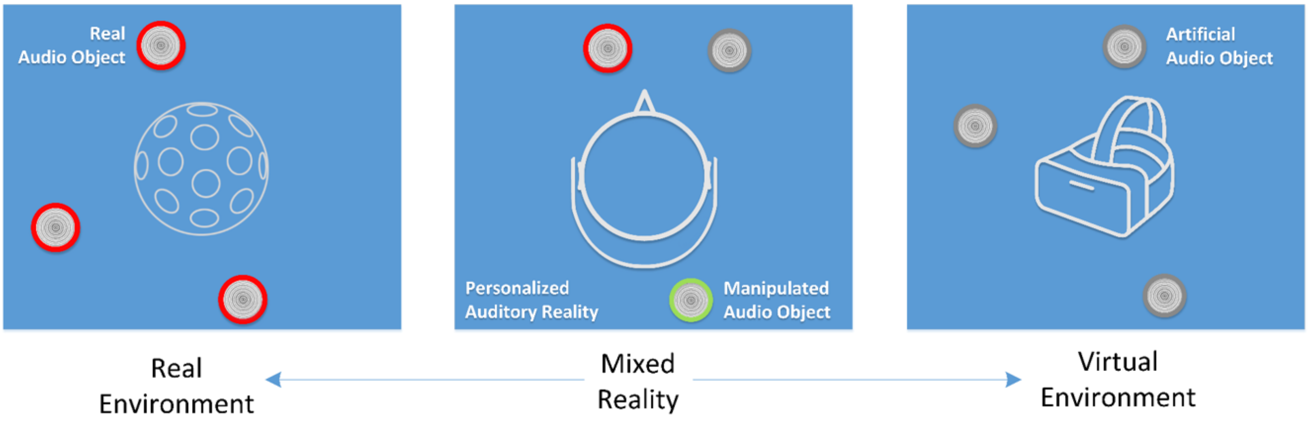

Figure 2: Personalized Auditory Reality is an extension of mixed reality by the possibility to add and suppress sound sources in real time.

quency estimation of the solo instrument as additional information [8]. An alternative approach to vocal separation from complex pieces of music based on deep learning methods was presented in [17]. Specialized source separation algorithms were also developed for use in industrial audio analysis [7].

Modification of the sound scene

In the next step, the sound scene is modified according to the input of the listener. While global noise cancellation (e.g. for all outside noises) is well understood, a PARty system is required to do active noise control (ANC) for some sources and feed through or enhancement of sounds for others. Another (not so difficult) option is to add new sound sources to the scene, like the voice of somebody on the phone with the listener.

The scene representation in a PARty system needs to find the balance between high audio quality and efficient representation for both foreground and background audio objects. Techniques to be employed are known in audio coding. This includes current standardization work on MPEG-I (Immersive) done by the MPEG group.

Dynamic binaural rendering adapted to the actual room

The scene recomposition and rendering part of the algorithms enable an efficient and effective reconstruction of the personalized audio experience. The algorithms need to be adapted to the actual room the listener is in. This is necessary to enable the transparent sound experience and to counter room divergence effects [20]. The binaural rendering has to be adapted to the position and the head pose of the listener to create a convicing experience and to avoid in-head localization [5].

Technical Requirements

The final system must provide high authenticity and plausibility. When somebody puts PARty devices in-ear, the sound from the environment first (before sound modifications are requested) should not change at all compared to the “no device in-ear” situation [10].

The hardware for development purposes will have the computing power we can expect from portable devices (smartphones etc.) around 5 years from now. Special DSP architectures optimized for AI (as already found on some current devices) will help to enable the machine learning parts of the algorithms.

For the software part, a special focus will be on delay-optimized algorithms, both for general signal processing (beam forming, binaural rendering) as well as for the psychoacoustic models employed.

More details on current work

Work is currently under way for several of the major road blocks towards a PARty system. This includes improved binaural rendering in an 6DOF scenario, research on sound object detection and work on source separation algorithms. An complete plan for implementing a PARty system is currently prepared.

Current work includes

Algorithm development

- Synthesis of new artificial source positions based on measurements
- In-Situ estimation of room acoustic characteristics and generation of BRIRs
- Steering of movable audio objects

User behavior analysis

- Movement analysis for behavior estimation
- Attention estimation and guidance

Real-time rendering optimization

- Real-time calculation of BRIR filters for moving listeners
- Predictive BRIR calculation based on user movements
- Combined real-time and offline rendering

Conclusions

The vision of PARty are devices which help to make our lives more enjoyable. There will be less distractions, less noise, more of the sounds we would like to hear.

As the acronym (PARty) says, the original target application for such devices is for social interaction at too loud places. There are many more potential applications: Traffic and mobility, medical applications, education, art and culture. We think of this work as a first step to build a platform for better sound for everybody.

References

[1] J. Abeßer, S. Balke, and M. Müller. "Improving bass saliency estimation using label propagation and transfer learning". In: *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*. Vol. 448, 306–312.

[2] A. Avni et al. "Spatial perception of sound fields recorded by spherical microphone arrays with varying spatial resolution". In: *Journal of the Acoustical Society of America* 133.5 (2013), pp. 2711–2721. DOI: 10.1121/1.4795780.

[3] K. Brandenburg et al. "Personalized Auditory Reality". In: *44th Annual Meeting on Acoustics (DAGA), Garching by Munich, Germany.* Deutsche Gesellschaft für Akustik (DEGA). 2018.

[4] M. Brandstein and D. Ward. *Microphone Arrays: Signal Processing Techniques and Applications.* 5th edition. Frankfurt/Main: Springer, 2010. ISBN: 9783642075476.

[5] W. Brimijoin, A. Boyd, and M. Akeroyd. "The Contribution of Head Movement to the Externalization and Internalization of Sounds". In: *Plos One, Vol. 8* (2013).

[6] E. Cano, D. FitzGerald, and K. Brandenburg. "Evaluation of Quality of Sound Source Separation Algorithms: Human Perception vs Quantitative Metrics". In: *Proceedings of the 24th European Signal Processing Conference (EUSIPCO).* Budapest, Hungary, 2016, pp. 1758–1762.

[7] E. Cano, J. Nowak, and S. Grollmisch. "Exploring Sound Source Separation for Acoustic Condition Monitoring in Industrial Scenarios". In: *Proceedings of the 24th European Signal Processing Conference (EUSIPCO).* Kos Island, Greece, 2017.

[8] E. Cano, G. Schuller, and C. Dittmar. "Pitch-informed Solo and Accompaniment Separation: Towards its Use in Music Education Applications". In: *EURASIP Journal on Advances in Signal Processing* 23 (2014), pp. 1–19. ISSN: 1687-6180.

[9] E. Cano et al. "Musical Source Separation: An Introduction". In: *IEEE Signal Processing Magazine* 36 (Jan. 2018).

[10] F. Denk et al. "An individualised acoustically transparent earpiece for hearing devices". In: *International journal of audiology* 57.sup3 (2018), S62–S70.

[11] J. F. Gemmeke et al. "Audio set: An ontology and human-labeled dataset for audio events". In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE. 2017, pp. 776–780.

[12] S. Gharib et al. "Unsupervised adversarial domain adaptation for acoustic scene classification". In: *arXiv preprint arXiv:1808.05777* (2018).

[13] J. R. Hershey et al. "Deep clustering: Discriminative embeddings for segmentation and separation". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* 2016, pp. 31–35.

[14] S. Marchand. "Audio scene transformation using informed source separation". In: *The Journal of the Acoustical Society of America* 140.4 (2016), p. 3091.

[15] D. Matz, E. Cano, and J. Abeßer. "New Sonorities for Early Jazz Recordings using Sound Source Separation and Automatic Mixing Tools". In: *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR).* Malaga, Spain, 2015, pp. 749–755.

[16] A. Mesaros, T. Heittola, and T. Virtanen. "A multi-device dataset for urban acoustic scene classification". In: *arXiv preprint arXiv:1807.09840* (2018).

[17] S. I. Mimilakis et al. "Monaural singing voice separation with skip-filtering connections and recurrent inference of time-frequency mask". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE. 2018, pp. 721–725.

[18] T. von Neumann et al. "All-neural online source separation, counting, and diarization for meeting analysis". In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).* Brighton, UK, 2019.

[19] T. Virtanen, M. D. Plumbley, and Ellis, D. P. W, eds. *Computational Analysis of Sound Scenes and Events.* Springer International Publishing, 2018.

[20] S. Werner et al. "A Summary on Acoustic Room Divergence and its Effect on Externalization of Auditory Events". In: *Proceedings of 8th International Conference on Quality of Multimedia Experience (QoMEX).* 2016. DOI: 10.1109/QoMEX.2016.7498973.

[21] Y. Xu et al. "Large-scale weakly supervised audio classification using gated convolutional neural network". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE. 2018, pp. 121–125.