ISSN 2464-4617 (print)
ISSN 2464-4625 (CD-ROM)

Personalized Sound Zoning for Communication means - user studies and evaluation

Ferdinand Fuhrmann, Clemens Amon, Christina Leitner, Anna Maly and Franz Graf
JOANNEUM RESEARCH Forschungsgesellschaft
Institute for Information and Communication Technologies
Steyrergasse 17
Austria, 8010, Graz
{forename.surname}@joanneum.at

ABSTRACT

We present a new audio user interface for communication means based on automatic sound zoning (ASZ). Highly directive audio reproduction and capturing devices are combined with the output of a newest-generation depth sensor. We particularly use parametric loudspeaker and microphone arrays for the spatial filtering of sound playback and recording. This enables device-less, mobile and ergonomic communication. We evaluated the system using subjective experiments assessing, on the one hand, the immersion properties of the system and, on the other hand, general usability aspects. We could show that users are highly impressed by the capability of such a system and would greatly benefit from the inherent properties.

Keywords

audio interface, sound zoning, device-less communication

1 INTRODUCTION

Unlike visual and haptic interfaces, the audio interface has not gone through radical changes across the last decades. Starting from the telephone, the communication principles and sensor basics have mostly stayed the same; a loudspeaker – as close as possible to the ear of the user – is used for the playback channel while a microphone – as close as possible to the user's mouth – captures the acoustic speech signal. This configuration leads to several drawbacks, outlined in the following.

In office scenarios, different audio emitting and receiving devices often act together, producing distracting noise. Headsets as a solution suffer from ergonomic deficits, especially during prevalent use and long working hours. Moreover, tethered headsets restrict the user's mobility. Further, audio interfaces are usually distributed; a single interface is used for each application (e.g. radio handset, telephone, and internet communication).

In this work we propose a new audio user interface for communication means, applicable in both home and office environments. The aim is to eliminate ergonomic problems while reducing the noise-level, and guarantee greatest mobility in a multi-purpose interface. The presented system combines motion tracking technology with directional audio reproduction and capturing devices. This enables automatic sound zoning (ASZ) where the information about the position of the user is used to automatically steer the sound zones for audio playback and capturing. Figure 1 shows an illustration of the presented interface. The advantages of this approach are twofold; (i) the user can move freely in the room while communicating device-less, (ii) the cross-talk to other persons in the room is reduced.

Figure 1: Illustration of the presented acoustical interface, generating sound zones.

We specifically use the newest generation of Microsoft's Kinect sensor[1] in combination with a

[1] https://dev.windows.com/en-us/kinect/hardware

Figure 2: Hardware prototype of the presented acoustic interface, consisting of the Kinect 2 depth sensor, a 13 channel microphone array, and two parametric arrays mounted on automatic alignment units.

directional sound reproducing system and a microphone array. Figure 2 shows the resulting hardware prototype. By using such techniques in communication systems or gaming applications, the immersion of the user is supported. In a professional environment such a system can be applied to provide cognitive relief, reduce stress, and increase mobility while reducing the overall noise level, or simply improve ergonomic aspects by freeing the user of a worn headset.

The presented audio interface is part of an interaction concept originally designed for modern control centers [Kai14]. The principles and technology can however be translated in several other domains, where audio interaction is indispensable (e.g. home, office, public facilities, production site, etc.). In this paper we describe the main technical components, both hard- and software, and present a thorough assessment of the usability of the audio interface. We thereby conducted subjective experiments to involve the user in the design and evaluation process of the system. In an accompanying work [Fuh16] we also provide objective evaluation results related to the applied technical components.

2 RELATED WORK

Microphone arrays are widely used for teleconferencing. They are often combined with video systems to improve the accuracy of the beam steering for better sound quality [Hua11] or are used in automatic speech recognition systems to improve recognition rates [Asa04]. Other application fields include control centers such as air traffic control [Gul09]. A different approach for audio capturing are microphone domes that have been investigated for the application in multimodal interfaces for crisis management [Sha03].

Directional sound reproduction systems are realized by either loudspeaker arrays or parametric loudspeakers. They have been applied to create an immersive audio environment for single users in desktop applications [Str03] or for gaming [Gan11]. Besides fixed installations, there exist automatically aligned systems that not only steer towards, but are able to follow a user [Kas14]. For multiple users, the concept of sound zones has been developed; each user is provided with a personal sound zone directed at her location [Bet15]. This can be applied in different environments such as shared offices or exhibitions. For communication applications, directional sound reproduction systems are combined with microphone arrays [Hua11, Gul09] to create a hands-free interface without headset.

3 CONCEPT

The presented audio interface consists of an audio reproduction and capturing system. Both components are characterized by a narrow spatial focus, which is automatically steered towards the user's head. Here, the alignment information is provided by a depth sensor (Kinect 2), enabling head-tracking in a triangular area in front of the sensor with an opening angle of 70° and depth to 3.5 m. More information about the sensor's capabilities can be found in [Amo14].

The reproduction system consists of two parametric loudspeaker arrays, emmitting spatially focused sound. By applying the principle of ultrasonic sound reproduction and the coupling of many parallel transducers, the emission area of these loudspeakers can be limited to a very narrow sound beam [Pom09]. More precisely, an inaudible ultrasonic carrier signal is first used to generate a spatially narrow radiation pattern. Modulation of the audio signal onto this ultrasonic carrier causes the generation of differential frequency components between the two signals. If the carrier's level is high enough – typical parametric arrays use sound pressure levels of 120 dB and more – this differential components become audible while the spatial emission pattern is preserved [Gan12]. The audio signal is hence demodulated along the radiation pattern of the ultrasonic carrier.

In our implementation we use two Acouspade ultrasonic[2] speakers. Special ultrasound preamps are used to generate a carrier frequency of 40 kHz and modulate their amplitude with the desired playback signal. The speakers are placed on the desk of the work station in a typical stereo arrangement and are automatically aligned to the respective ear of the user. For this purpose, the speakers are mounted on engine-driven automatic alignment units. We built prototypes of these alignment systems using two stepper motors – one for the azimuth and one for the elevation angle. The motors are run by two motor drivers and a micro controller, processing the control data for steering.

The audio capturing system consists of a microphone array followed by a beamforming algorithm. The beam-

[2] http://www.ultrasonic-audio.com/products/acouspade.html

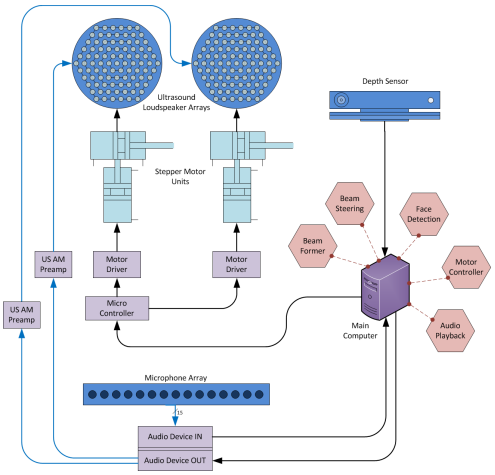

Figure 3: System architecture overview. Hexagonal objects represent software components while rectangular shapes correspond to hardware modules.

former's main directivity is enforced towards the position of the user, enhancing the target speech while suppressing noise and interfering sounds. We constructed a linear array consisting of 13 omnidirectional microphones with an inter-distance of 2 cm. This ensures an effective signal bandwidth up to approximately 8.5 kHz, which is sufficient for the transmission of speech and other acoustical messages (e.g. sound notifications).

Figure 3 illustrates an overview of the system architecture. Here, software components (hexagonal objects) such as head tracking, beamforming, and the motor controller of the loudspeaker alignment system are processed by the main computing unit (PC), which is directly connected via external hardware modules (rectangular objects) with the sensors and actuators.

4 EVALUATION AND RESULTS

4.1 Auditive interference

Here we evaluated the acoustic cross-talk of the parametric arrays in a subjective listening test. The focus of the evaluation was to estimate the cognitive relief with respect to the acoustic information emitted by the proposed system. Overall 24 subjects – 15 male and 9 female – took part in this study, the average age was 32.9.

We considered a two-workplace-scenario where two persons work side-by-side while performing telephone calls or talking via a radio handset. The test subject was seated at one desk, a second person was simulated by an artificial head with torso in front of the second desk. We asked the test subject to read a text (about 500 words) aloud while at the same time an audio signal containing five speech utterances was played back via loudspeakers, which were aligned at the dummy head. After reading the text, we asked the person to write down the utterances she understood. The more utterances the subject could reproduce, the higher the impact of the auditive interference on the cognitive load.

System	*#Utt(mean)*
Presented system	1.792
Reference system	2.792

Table 1: Results of the subjective interference tests. Average number of correctly recognized utterances is shown.

We compared the performance of the subjects within the proposed system against their performance when exposed to a reference system. The reference system was realized by using standard studio loudspeakers together with a band-pass filter imitating the frequency characteristics of a radio handset. Hence, we expected the presented system to exhibit significantly less cognitive load than the reference system.

Table 1 shows the results of the experiment. It can be seen that with the presented system significantly less utterances were recognized than with the reference system (paired t-test, $p < 0.003$). Here, participants could reproduce around 55% more utterances than in the system applying the parametric arrays. This confirms our hypothesis that the presented system supports the cognitive relief of the user, since significant less audio information was transferred to the test subject.

4.2 User acceptance

To evaluate the overall system we performed a user study in a working area where users both are confronted with a lot of acoustical information and work in an office environment. We opted for the traffic monitoring field where users receive audio information via telephone, radio handset, synthetic voice alarms, and other acoustical alerts. Typically, three to five operators work together in specially equipped control rooms, where each operator has a work station desk.

We were able to interview six operators in total (45 mean aged and male), four of them had more than 15 years of work experience. The experimental procedure was arranged into three parts; first, the main functionalities of the presented acoustic interface were demonstrated. We prepared a typical traffic monitoring scenario, where acoustical alerts, telephone calls and radio handset interaction were integrated in a fully automatic sequence of events. The scenario started with an acoustic alarm to grab the attention of the test subject. Then, the operator could accept an incoming call, talk to the calling person, and give a phone call to resolve the situation. In the second part, the operator could explore the acoustic interface in detail. The test person listened to several test signals (radio streams, telephone and radio handset calls), experiencing the immersion properties of the system. Moreover, we asked the operator to move around in the room while listening to the test

signals, judging the quality of the automatic alignment. We the interviewed the operator related to positive and negative aspects of the presented system and the respective components, as well as possible improvements and applications in the traffic monitoring area.

In the interviews all subjects commented positively on the overall impression of the presented system. We were able to transmit the idea of the device-less communication to the operators, who emphasized the integrating and centralizing aspects of the acoustic interface (i.e., all information is transferred over a single user interface). The combination of automatic aligned parametric arrays and the microphone array adds a lot of flexibility and mobility to the daily working routines. All operators were satisfied with the speed of the alignment as well with its mechanical noise properties (very quiet). They were impressed by the narrow sound beam emitted by the parametric arrays; the audio quality was judged between good and excellent.

5 DISCUSSION

The results presented in Section 4 suggest that the acoustic user interface concept – visually guided sound zoning – can be translated into a system implementation. The parametric arrays allow personalized audio playback for multiple users, since the emitted audio signal is only perceived by the target user and all other audible signals in the room originate solely from reflections. This is also shown in the experiments involving the acoustic crosstalk, which can be highly reduced with parametric arrays compared to regular loudspeaker systems. In combination with the microphone array, a device-less communication system is created offering maximum mobility and ergonomy.

The presented system offers a great potential for new applications in both home consumer and office environments. Its mobility property frees the user from any wearable device, while the zoning properties immerse the user. Moreover, for multi-user applications, the overall noise level is reduced which directly influences the cognitive load of the involved people.

6 CONCLUSION

We presented a new audio user interface based on a sound zoning approach. We control highly directional sound reproduction and capturing units by the output of a vision system, tracking the user's position and head movement. This enables a maximum mobility together with a maximal immersion of the user. The applications of such a system range from home usage to typical office environments. In our studies we could show that users are highly impressed by the capability of such a system and would greatly benefit from the inherent properties.

7 ACKNOWLEDGEMENTS

This work was funded by Austrian Ministry for Transport, Innovation and Technology in the fusInC[3] project.

8 REFERENCES

[Amo14] Amon, C., Fuhrmann, F., and Graf, F. Evaluation of the spatial resolution accuracy of the face tracking system for kinect for windows v1 and v2, 6th Congress of the Alps Adria Acoustics Association, 2014.

[Asa04] Asano, F., Yamamoto, K., Hara, I., Ogata, J., Yoshimura, T., Motomura, Y., Ichimura, N. and Asoh, H. Detection and separation of speech event using audio and video information fusion and its application to robust speech interface. EURASIP Journal on Applied Signal Processing, 2004.

[Bet15] Betlehem, T., Zhang, W., Poletti, M. , and Abhayapala, T. Personal Sound Zones: Delivering interface-free audio to multiple listeners, Signal Processing Magazine, IEEE, 32(2), 2015.

[Fuh16] Fuhrmann, F., Amon, C., Leitner, C., Maly, A. and Graf, F. Personalized sound zoning for communication means - design and evaluation of a bidirectional beamforming system, submitted to DAFx, 2016.

[Gan11] Gan, W., Tan, E. and Kuo, S. Audio projection. IEEE Signal Processing Magazine, 28(1), 2011.

[Gan12] Gan, W., Yang, J. and Kamakura, T. A review of parametric acoustic array in air. Applied Acoustics, 73(12), 2012.

[Gul09] Guldenschuh, M. Transaural beamforming. M.Sc. thesis, Graz University of Technology, 2009.

[Hua11] Huang, Y., Chen, J. and Benesty, J. Immersive audio schemes. IEEE Signal Processing Magazine, 28(1), 2011.

[Kai14] Kaiser, R. and Fuhrmann, F. Multimodal interaction for future control centers: interaction concept and implementation, Workshop on Roadmapping the Future of Multimodal Interaction Research, ACM, 2014.

[Kas14] Kashiwase, S. and Kondo, K. Towards a parametric speaker system with human head tracking beam control, 3rd IEEE Global Conference on Consumer Electronics (GCCE), 2014.

[Pom09] Pompei, J. Fundamental Limitations of Loudspeaker Directivity, `https://www.holosonics.com/tech_directivity.html`, 2009.

[Sha03] Sharma, R., Yeasin, M., Krahnstoever, N., Rauschert, I., Cai, G., Brewer, I., MacEachren, A. and Sengupta, K. Speech-gesture driven multimodal interfaces for crisis management, Proc. of the IEEE, 91(9), 2003.

[Str03] Strauss, M., Sontacchi, A., Noisternig, M. and Höldrich, R. A spatial audio interface for desktop applications, 24th International Audio Engineering Society Conference, 2003.

[3] `http://www.joanneum.at/digital/referenzprojekte/fusinc.html`