Audio Engineering Society

Convention e-Brief

Presented at the 141st Convention
2016 September 29 – October 2, Los Angeles, CA, USA

STEAK: Backward-compatible Spatial Telephone Conferencing for Asterisk

Dennis Guse and Frank Haase

TU Berlin

Correspondence should be addressed to Dennis Guse (dennis.guse@alumni.tu-berlin.de)

ABSTRACT

In this paper, we present our implementation of a telephone conferencing system that renders a spatial representation via binaural synthesis. The implementation extends the open-source software Asterisk and complies with established Voice-over-IP standards. The implementation only requires clients to be capable of receiving and reproducing the rendered binaural signals (two channels). Furthermore, the implementation is backward-compatible as clients not fulfilling these requirements are provided with mono-rendered signals without additional spatial information. The implemented system is released as open-source software and will enable researchers to investigate the (dis-)advantages of spatial conferencing under real-world conditions.

The project name is *Spatial TelephonE conferencing for AsterisK* (STEAK)

Website: *http://www.SteakConferencing.de*

Introduction

Telephone conferences are an ubiquitous communication tool, for example, to coordinate distributed teams or communicate with remote business partners. A telephone conference provides a shared virtual space in which participants can communicate using audio in real time. As multiple participants interact via a shared virtual space, recognition of individual talkers and thus attribution of statements as well as listening to individual talkers in case of double talk are issues. These issues are especially problematic if the individual acoustical representation of the virtual shared space is rendered and transmitted as a mono signal (i. e., downmixing the signals of all remote, speaking participants into one mono signal). In fact, the established telephony infrastructure is still mainly limited to the transmission of mono signals. Using this infrastructure for connectivity to a telephone conferencing system has the important advantages that *(a)* participants only need to have access to standardized, ubiquitous hardware and *(b)* can participate from almost anywhere. A mono representation might be sufficient, but it prevents exploiting the spatial hearing capability of humans (cf. [1]). Advantages of exploiting spatial information has been shown to improve talker identification, talker recognition, and intelligibility (e. g., [2, 3, 4, 5]). Carefully positioning the individual participants in a spatial acoustical rendering of the virtual space, enables participants to experience the virtual space similar to a face-to-face meeting. One option to add spatial information is rendering the virtual space using binaural synthesis. Here, the rendered signals are presented via a pair of headphones.

In this paper, we present a production-ready telephone conferencing system with binaural synthesis based upon the open-source software Asterisk. Due to the implementation's compliance to established *Voice-over-IP* (VoIP) standards, it does not require proprietary technology and can be integrated into the existing telephony infrastructure. The system aims at closing the gap in research on spatial telephone conferencing. Research in this direction was so far limited as no production-ready system for centralized spatial conferencing was available as open-source software. Using proprietary systems has the disadvantages that the complete signal processing is often not publicly documented and in general cannot be modified if desired. This paper is structured as follows. First, we describe options for implementing a production-ready system with binaural synthesis as well as their (dis-)advantages. Second, we describe the implementation of the *Spatial TelephonE conferencing for AsterisK* (STEAK) in detail. Afterwards, the results of a performance evaluation with regard to delay of the implemented system is presented. This paper concludes with a summary and an outlook for future research on spatial telephone conferencing.

Considerations for Spatial Conferencing

In the following, first the method of binaural synthesis is introduced briefly. Afterwards, the two contrasting topologies for rendering a telephone conference are presented and discussed.

Binaural Synthesis

A spatial representation of a virtual space can be created using binaural synthesis. In binaural synthesis, the signal of every sound source is rendered depending on their position in the acoustical space to be simulated and then all rendered signals are downmixed into one signal. This step is conducted for the left ear and right ear individually. The applied signal modifications might include *Interaural Time Difference*, *Interaural Level Difference*, spectral changes, and also simulation of a room (i. e., reverberation). These modifications are usually applied by convolving each source signal with the *Head Related Transfer Functions* (HRTFs) for the left ear and right ear. The rendered signals can then be presented using a pair of headphones and provide the impression of a virtual acoustical space. Often binaural synthesis is applied using HRTFs created with a dummy head (i. e., a human-like head with artificial ears). Although it has been shown that individualized HRTFs (e. g., accounting for individual shapes of the outer ear) improve the spatial representation slightly, it is often sufficient to use general HRTFs (e. g., [6]). The spatial hearing capability of humans has two limitations. First, the angular resolution of a human observer is not uniform (i. e., highest resolution on the horizontal plane is achieved in front of the listener). Second, front-back confusion is problematic (cf. [1]). Human listeners overcome these issues by actively exploring their acoustical environment (i. e., sampling the acoustical space over time by moving their head). For binaural synthesis, this requires that (a) the movement of the head can be measured precisely enough and that (b) the delay between changes in head position including head orientation and the update of the virtual acoustic space is low enough (i. e., below 75 ms [7]).

Telephone Conferencing Topologies

Telephone conferences can be created using either a decentralized topology or a centralized topology (cf. [8]). A visualization of both topologies is shown in Figure 1. In a decentralized topology, the signals of all clients are transmitted to all other clients while each client renders his representation of the virtual space himself. The major drawback of this topology is that increasing the number of clients results in an increase of the required network bandwidth. However, this topology allows to achieve a low rendering delay. In a centralized topology, rendering of the virtual space for each client is conducted by one central instance and the rendered signals are transmitted to each client. Thus, the central instance requires the network bandwidth, computational resources, and electric power for rendering while each client only needs to receive and reproduce his rendered signal. From the perspective of a client, participating in a centralized telephone conference is, in fact, not different to a one-to-one call. The central instance can also operate as a translator between different, incompatible technologies, such as signaling technologies (e. g., *Session Initiation Protocol* (SIP) and *Web Real-Time Communication)* (WebRTC)) and media transmission technologies (e. g., used codecs). Due to these advantages, telephone conferencing systems are usually implemented using a centralized topology.

With regard to the application of binaural synthesis, the centralized topology exhibits two inherent limitations.

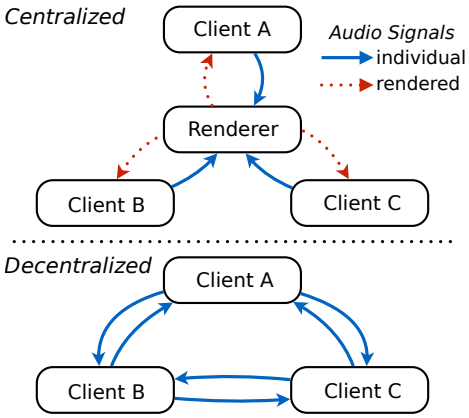

Fig. 1: Topologies for rendering a telephone conference including signal flow (adapted from [9]).

First, the central rendering increases the rendering delay due the required transmission of the rendered signals. Thus, using head tracking is hardly feasible as the upper boundary of the rendering delay (i. e., 75 ms) is hard to fulfill. The second limitation is related to the transmission of the rendered signals. In addition to requiring the transmission of *two* signals from the central instance to each client, the rendered signals need to be compressed to be suitable for transmission. In fact, compression might introduce artifacts in the binaurally rendered signals. This might negatively influence the spatial representation and also the perceived quality. Beside these limitations, a practical issue occurs if multiple participants use the same microphone (e. g., a conference phone). For spatial separation, the speech signal of these participants would need to be unmixed before applying binaural synthesis (e. g., [10]). However, the impact of the resulting artifacts in the unmixed source signals on binaural synthesis has been unknown.

Implementation

In the following, the design aspects and implementation details of the here presented telephone conferencing system are explained. The major goal of the implementation was to create a spatial conferencing system that can be integrated into the established ecosystem of telephony technology while not requiring specialized (i. e., standard-incompatible) clients. This system uses a centralized topology to avoid the network bandwidth issue and can thus act as a translator between different technologies. This allows users to select their desired connecting technology depending on their current situation. An issue for the implementation is the transmission of binaurally rendered signals as the established telephony infrastructure is limited to mono. This can be overcome by *(a)* using two separate transmission channels in parallel or *(b)* using a single binaural-capable transmission channel. While (a) is technically feasible it requires to carefully synchronize both transmission channels. In addition, this might introduce undesired effects due to differences in degradations on each transmission channel (e. g., packet-loss [9]). Using a binaural-capable transmission channel avoids these complex issues while it requires the availability of such a channel. Although not widely used, the *Real-Time Transport Protocol* (RTP), which is the common basis for media transmission in VoIP, allows the transmission of binaural signals (i. e., two channels) over one transmission channel. However, this functionality is only implemented for a limited number of codecs. One of these codecs is Opus ([11]), which is a relatively new, versatility, low-latency codec that is royalty-free and open-source. Opus is the de facto standard for audio transmission using WebRTC.

As basis for the implementation, the open-source software *Asterisk* (*http://www.asterisk.org*) was selected. Asterisk forms the perfect platform for the implementation, as it *(a)* provides connectivity to a wide range of telephony technologies, *(b)* can translate between different technologies, and *(c)* provides out-of-the-box the functionality to create *non*-spatial telephone conferences (i. e., downmixing to mono). For spatial conferencing via binaural synthesis, Asterisk was extended with four features. First, the internal signal processing was enhanced by the capability to process two-channel signals. Second, the default conference renderer was modified, so it can handle mono signals and two-channel signals while applying binaural synthesis for binaural-capable transmission channels. Binaural synthesis was implemented using the open-source library *FFTW* (*http://www.fftw.org*). This library implements Fourier transformation and is compatible to the software license of Asterisk (i. e., *GNU General Public License*). Third, the codec Opus including two-channel support was added to Asterisk. Fourth, the two-channel support has been enabled for SIP including SIP-over-WebRTC. In fact, the signaling remains in full compliance to SIP. To receive a binaural representation, a client needs to connect to the enhanced Asterisk and to announce support for two-channel-capable codecs via *Session Description Protocol* (SDP). Therefore, any SIP client implementing Opus with two-channel sup-

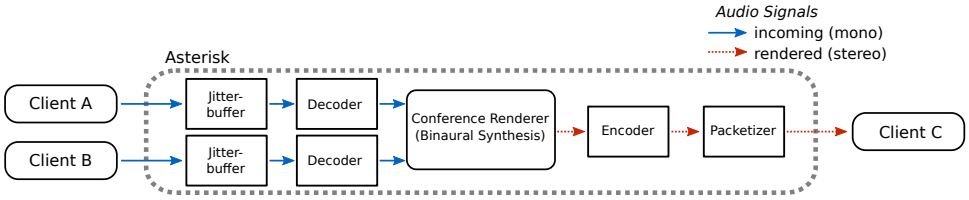

Fig. 2: Exemplary overview of the internal signal flow within Asterisk for the binaural-capable conference renderer. Only signal flow from client A and client B towards the binaural-capable client C are depicted.

port is supported. Note that the conference renderer is decoupled from the actual transmission technology. An exemplary overview of the signal flow for spatial conferencing is shown in Figure 2. A client can interact with the conference renderer by using *dual-tone multi-frequency signaling* (DTMF). This allows clients to enable and disable binaural synthesis as well as change the layout of the virtual conference space.

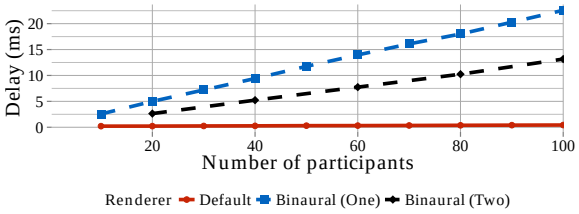

Fig. 3: Rendering delay for the default renderer and the binaural renderer by number of participants.

Performance Evaluation

The performance of the implementation, based upon Asterisk 13.6.0 and FFTW 3.3.4, was evaluated on a computer (Intel Core i7-4790 at 4 GHz) running Ubuntu Linux 16.04 (64bit). This evaluation focused on the processing delay of the binaural synthesis while omitting coding and transmission. Binaural synthesis was conducted at 48 kHz with a block size of 20 ms while rendering one virtual acoustical space for each telephone conference (i. e., one representation for all participants). Figure 3 shows the results for the default implementation as well as one binaural renderer and two binaural renderers (1000 iterations; 2 s of babble noise). Note that Asterisk executes each conference renderer instance in one thread. Thus, parallelism can only be exploited using multiple renderers (i. e., multiple conferences in parallel). While the default renderer performs nearly constant, the binaural renderer adds delay approximately in a linear manner. For the binaural renderer, the processing delay is satisfying up to 80 *actively speaking* participants in one conference (18 ms). After that, it might become problematic as the processing delay adds to the end-to-end delay and thus might negatively affect the perceived quality.

Conclusion & Future Work

In this paper, we presented our enhancement (i. e., spatial conferencing via binaural synthesis) to Asterisk. Although, the implementation as a central renderer has drawbacks (e. g., accounting for head movements is not timely feasible), it is advantageous as clients only need to be capable of receiving and reproducing binaurally-rendered signals. Most important, the implementation is backward-compatible as mono-capable clients (without binaural synthesis) and binaural-capable clients can participate in the same telephone conference. In addition, the presented implementation complies to VoIP standards. We hope that the open-source release of our implementation will stimulate future research on spatial conferencing especially with regard to the (dis-)advantages and limitations under real-world conditions. Two important aspects are not yet completely solved in this domain. First, how to position individual participants in the virtual space and how to adjust positions if participants join or leave an ongoing telephone conference. Initial work has been conducted for positioning (e. g., [12]) and also algorithms for repositioning have been investigated (e. g., [13]). However, these investigations were limited and not conducted under real-world conditions. Second, the quality of the recording at the client-side might be problematic. Degradations resulting from the recording equipment as well as recording environment are very likely to occur under real-world use. This might include low-budget microphones, background noise, and reverberation. In fact, binaural synthesis is applied in general using signals acquired under anechoic conditions while often using high-end equipment. The presence of recording-

related degradations might negatively influence the result of the binaural synthesis with regard to spatial representation and also perceived quality. This might, in fact, be problematic for the user acceptance. The impact of recording-related degradations and their impact has to the authors knowledge not been investigated. In addition, also degradations due to transmission of the signals must be considered. Beside packet-loss, where initial work has been conducted [9], also the impact of applying lossy compression on binaurally rendered signals has only received limited attention (e. g., [14]).

The system presented here provides an excellent platform for further research on spatial conferencing in laboratory environments and field studies. Moreover, the system is ready for productive use and thus allows for *really* experiencing spatial conferencing in practice.

Acknowledgements

We would like to thank Prof. Dr.-Ing. Alexander Raake and Janto Skowronek (both TU Illmenau) for introducing us to the topic of spatial telephone conferencing as well as Prof. Dr.-Ing. Jens Ahrens (Chalmers University of Technology) and Dr. Hagen Wierstorf (TU Berlin) for providing us practical, hands-on knowledge on binaural synthesis. The project received funding by the German Federal Ministry of Education and Research in context of the SoftwareCampus program (grant number 01IS11556).

References

[1] Blauert, J., *Spatial Hearing - Revised Edition: The Psychophysics of Human Sound Localization*, The MIT Press, 1996, ISBN 0-262-02413-6.

[2] MacKeith, N. W. and Coles, R. R. A., "Binaural advantages in hearing of speech," *J Laryngol Otol*, 85(3), pp. 213–232, 1971.

[3] Ihlefeld, A. and Shinn-Cunningham, B., "Spatial release from energetic and informational masking in a selective speech identification task," *The Journal of the Acoustical Society of America*, 123(6), p. 4369, 2008, ISSN 00014966, doi: 10.1121/1.2904826.

[4] Drullman, R. and Bronkhorst, A. W., "Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation," *The Journal of the Acoustical Society of America*, 107(4), pp. 2224–2235, 2000.

[5] Kilgore, R., Chignell, M., and Smith, P., "Spatialized audioconferencing: what are the benefits?" in *Proceedings of the 2003 conference of the Centre for Advanced Studies on Collaborative research*, pp. 135–144, IBM Press, 2003.

[6] Møller, H., "Fundamentals of binaural technology," *Applied acoustics*, 36(3-4), pp. 171–218, 1992.

[7] Lindau, A., "The perception of system latency in dynamic binaural synthesis," *Proc. of 35th DAGA*, pp. 1063–1066, 2009.

[8] Singh, K., Nair, G., and Schulzrinne, H., "Centralized conferencing using SIP," in *Internet Telephony Workshop*, volume 7, pp. 57–63, 2001.

[9] Spur, M., Guse, D., and Skowronek, J., "Influence of Packet Loss and Double-Talk on the Perceived Quality of Multi-party Telephone Conferencing with Binaurally Presented Spatial Audio Reproduction," in *Fortschritte der Akustik: Tagungsband d. 42. DAGA*, Aachen, Germany, 2016.

[10] Raake, A., Spors, S., Ahrens, J., and Ajmera, J., "Concept and evaluation of a downward-compatible system for spatial teleconferencing using automatic speaker clustering." in *INTERSPEECH*, pp. 1693–1696, 2007.

[11] Spittka, J. and Vos, K., "RTP Payload Format for the Opus Speech and Audio Codec," *Internet Engineering Task Force*, 7587, 2015, ISSN 2070-1721.

[12] Hyder, M., Haun, M., and Hoene, C., "Placing the participants of a spatial audio conference call," in *Consumer Communications and Networking Conference (CCNC), 2010 7th IEEE*, pp. 1–7, IEEE, 2010.

[13] Roegiers, D., *Dynamical Aspects of Spatial Audio in Multi-participant*, Master Thesis, Ghent University, 2012.

[14] Katz, B. F. and Prezat, F., "The Effect of Audio Compression Techniques on Binaural Audio Rendering," in *Audio Engineering Society Convention 120*, 2006.