Investigations on the Influence of a Dynamic Binaural Synthesis on Speech Intelligibility in Communication Applications

Nils Poschadel, Mahdi Alyasin, Stephan Preihs, Jürgen Peissig
Leibniz Universität Hannover, Institut für Kommunikationstechnik
Appelstr. 9A, 30167 Hannover, Email: poschadel@ikt.uni-hannover.de

Abstract

Within the project VIA^{2}mobiL, we developed signal processing algorithms and methods for a dynamic headphone based binaural synthesis, with a special focus on radio communication in mobile control centers. Our aim is to achieve a better speech intelligibility in radio communication through the binaural presentation of a conversation scene with several separately locatable interlocutors. In our investigations, a method for determining word recognition rates and 50 % speech intelligibility thresholds was developed on the basis of the methodology of the Oldenburg sentence test (OLSA). By means of conducted listening experiments, we examined whether the application of a dynamic binaural synthesis results in a gain in speech intelligibility if compared to monophonic or stereophonic reproduction. In this paper we present the study design and the results of the first part of our experiment.

Introduction

Binaural synthesis is a well known technology for 3D sound generation which has already found its way into a variety of applications [5]. The advantage of 3D audio if compared to conventional playback modes and with regard to different psychoacoustic aspects has already been proven in other studies. For example, Drullman and Bronkhorst [1] showed that a headphone based 3D presentation yields better speech intelligibility with two or more competing talkers compared to conventional monaural and binaural presentation, in particular for sentence intelligibility. For this reason, we implemented and tested this technology in a radio communication system of our project partner ATS Elektronik GmbH. A study was then designed to answer the question: How does a dynamic binaural synthesis affect speech intelligibility in this target application of radio communication?

3D Audio Dispatcher System

An overview of the entire system for mobile control centers, developed within the project VIA^{2}mobiL, is given in Figure 1. The TETRA audio data, together with some metadata such as the GPS positioning of the different radio operators, the Individual Short Subscriber Identity (ISSI) or call groups are transmitted to a dispatcher software via network. The desired (virtual) positions of the different sound sources/ interlocutors can now be derived from this metadata or just freely defined. The software of the audio application receives this absolute positioning information and uses the information about the dispatcher's viewing direction provided by a headtracker mounted on top of the headphones to calculate the position relative to this direction.
The dynamic binaural rendering system used in this study was implemented using the Max/MSP environment following [3]. For us, this approach represented the best compromise between flexibility, quality and computational complexity. The *multiconvolve~* object from the HISSTools impulse response toolbox [4] is used for the real-time convolution of the audio signals with the corresponding BRIRs. After a filter change, e. g. triggered by a head movement, a cross-fade is performed to eliminate audible artifacts. The BRIRs used were measured in the Immersive Media Lab (IML) of the Institute of Communications Technology with a KEMAR 45BC-12 dummy head and a resolution of 5° in the azimuth plane [3]. Afterwards, they were truncated to 46.4 ms (2048 Samples) and interpolated to a resolution of 1°.

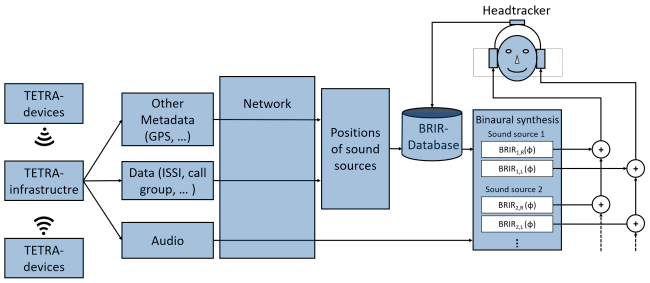

Figure 1: Overview of the 3D audio dispatcher system developed within the project VIA^{2}mobiL.

Experimental design

The listening experiment designed to answer our research question consisted of two parts. In this paper we describe the first part, in which the test subjects listened to the simultaneous playback of up to four speech signals (s. [1]). One speaker was the target talker (TT) and one to three other speakers were the competing talkers (CTs). The task of the study participants was to reproduce the sentences spoken by the TT by clicking the respective words in a GUI showing five columns of ten possible words each. The TT's voice was constant for one participant during the whole test. Six subjects each were randomly assigned a strong (male) voice (TT 1) or a thin (female) voice (TT 2). The CTs were randomly selected for each sentence. The identification of relatively well or badly understood voices and other hyperparameters was carried out in preliminary studies.

Figure 2: Screenshot of the graphical user interface of the study software. The subjects have to choose the words they hear from all the choices that can occur in the OLSA.

The audio was played back in the presentation modes monaural (mon), binaural (bin), 3D audio with TT in 0° (3D-0°) and 3D audio with TT in 60° (3D-60°). In the cases of 3D audio, the speech signals of the interlocutors were convolved with the respective BRIRs according to the procedure explained before.
For each combination of presentation mode and number of CTs, the positions of the interlocutors were defined as they would most likely be configured in our radio communication application. For example, in the 3D-60° case with two CTs, the TT is located at 60° azimuth, the first CT at -60° azimuth and the second CT at 0° azimuth. An overview over all scenarios is shown in Figure 3. The TT's position and voice were presented to the study participants before each change in scenario.
The speech intelligibility in the first part of the study was evaluated using the Word Recognition Score (WRS) defined by $\mathrm{WRS} = \frac{\#\{\text{correctly understood words}\}}{\#\{\text{all words}\}} =: \frac{N_{\text{correct}}}{N_{\text{all}}}$ with N_{correct} being the number of correctly understood words by the participant and N_{all} the number of all words played back. For each combination of number of CTs and presentation mode, 15 sentences were played to the subjects and then the mean value of the WRS of all subjects was calculated.

Table 1: Adaptation scheme of the voice level depending on the number of correctly understood words in the previous sentence (s. [2]).

N_{correct}	Level change in dB	
	Sentences 1-3	Sentences 4-28
5	-3	-2
4	-2	-1
3	-1	±0
2	+1	±0
1	+2	+1
0	+3	+2

Speech material

The speech material used for this study is based on the sentence components of the Oldenburg sentence test (OLSA) [2]. In the OLSA, every sentence consists of five words and has an identical structure: name – verb – numeral – adjective – object.

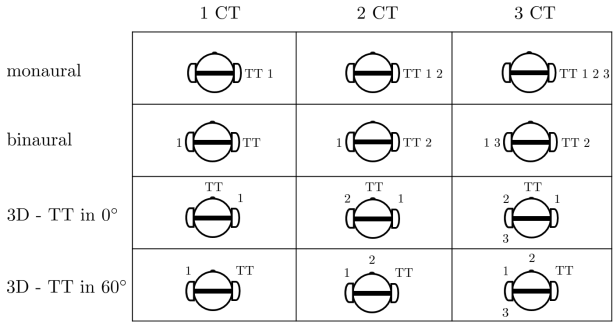

Figure 3: Positions of the TT and the CTs (numbers 1-3) for different number of CTs and presentation modes. Possible positions are at 0°, ± 60° and -120°.

Due to the need for different speakers in this study, the original recordings of the OLSA could not be used. For this reason, we let three men and three women record the OLSA's speech material. To take coarticulation effects into account, whole sentences were recorded, with each possible word transition appearing at least once. These sentences were then semi-automatically cut into the individual words. In total, $10^5 = 100{,}000$ different sentences from six different speakers could be generated. For every test subject, a different set of random sentences was created.

Procedure

12 test subjects (nine men and three women) aged from 18 to 55 took part in the experiment. They were German native speakers and stated that they had normal hearing. The study was also conducted in the IML of the Institute of Communications Technology. The audio was played on a Sennheiser HD 280 Pro with a level of 60 dB SPL. The test subjects sat in the IML in front of a PC that ran the study software. The subjects wore headphones on which a headtracker was attached. For the reproduction of noise in the second part of the study, four loudspeakers were located in the corners of the room.

The setup of the study is shown in Figure 4.

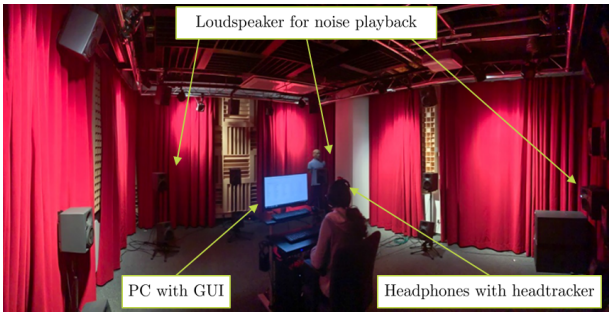

Figure 4: The setup of the study. The test person sits in the IML in front of a PC that runs the study software. The person wears headphones on which a headtracker is attached. For the reproduction of noise in the second part of the study, four loudspeakers are located in the corners of the room.

Results

Overall from the first part of the experiment we can state an increase in speech intelligibility for 3D audio playback by about

- 22 % compared to monaural playback with two CT,
- 44 % compared to monaural playback with three CT,
- 9 % compared to binaural playback with two CT and
- 5 % compared to binaural playback with three CT.

Furthermore, the variation of the results with monaural playback is considerably higher compared to the other presentation modes, in particular 3D audio (s. Figure 5). This indicates a higher dependence on the sentence-speaker-combination for the monaural case.

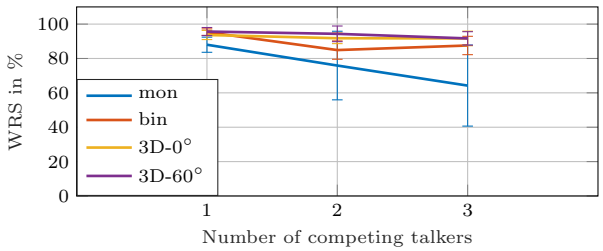

Figure 5: Average WRS depending on the number of CTs for the different presentation modes with error bars representing the standard deviations.

As shown in Figure 6, speech intelligibility in the conducted tests was also strongly dependent on the TT's voice (especially for monaural audio).

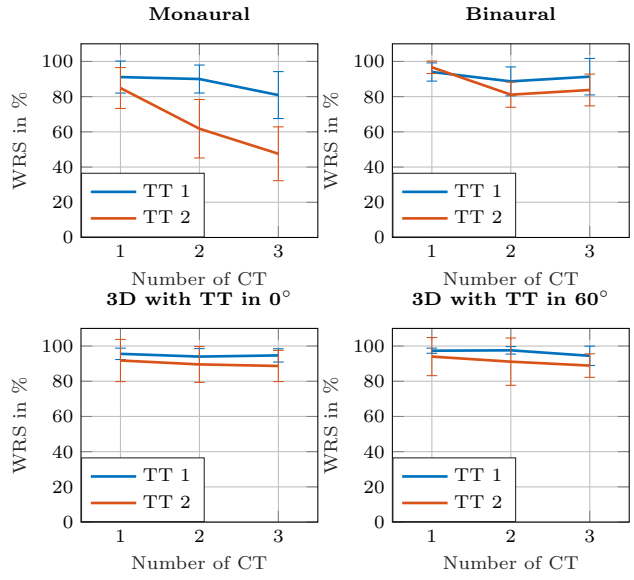

Figure 6: Average WRS depending on the number of CTs for different presentation modes and the two different TTs with error bars representing the standard deviations of the average WRS of each participant.

Conclusion and future work

In this paper, the results of the first part of a study on speech intelligibility using a dynamic binaural synthesis in the application of radio communication were presented. It could be shown that a binaural synthesis has a measurable advantage in the presence of noise compared to conventional monaural or binaural playback.
In detail our results show that up to 44 % and 9 % higher word recognition rates can be achieved with a binaural synthesis compared to a monophonic and stereophonic reproduction at the same level.
Our next steps will include an examination of the results of the second part of the study as well as the user feedback regarding the solvability of the task given.
In the future, we would also like to evaluate the time spent to complete the various tasks, which was recorded during the execution of the study. In addition, we would like to confirm our hypothesis of the reduction in cognitive load by also recording and evaluating physiological parameters.

Acknowledgement

The project VIA^{2}mobiL was funded by the German Federal Ministry for Economic Affairs and Energy (Grant ZF4298802LF7). We want to thank our project partner ATS Elektronik GmbH for the possibility of a cooperation and the fruitful discussions within the project.

References

[1] Drullman, R., 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 (2000), 2224–2235.

[2] Kuehnel, V., Kollmeier, B., Wagener, K.: Entwicklung und Evaluation eines Satztests für die deutsche Sprache I: Design des Oldenburger Satztests. Zeitschrift für Audiologie 38 (1999), 4–15 .

[3] Li, S., Schlieper, R., Peissig, J.: The Impact of Head Movement on Perceived Externalization of a Virtual Sound Source with Different BRIR Lengths. AES International Conference on Immersive and Interactive Audio (2019).

[4] Tremblay, P. A., Harker, A.: The HISSTools impulse response toolbox: Convolution for the masses. ICMC 2012: Non-Cochlear Sound - Proceedings of the International Computer Music Conference (2012), 148–155.

[5] Zhang, W., Samarasinghe, P., Chen, H., Abhayapala, T.: Surround by Sound: A Review of Spatial Audio Recording and Reproduction. Applied Sciences vol. 7 no. 5 (2017), 532.