Acoustic spatial encoding in a portable navigation aid for the blind

Alexis Guibourgé[1], Viviane Ghaderi[1], Jörg Conradt[1] and Bernhard U. Seeber[2]
[1]*Neuroscientific System Theory, 80333 München, E-mail: office@nst.ei.tum.de*
[2]*Audio Information Processing, , 80333 München, E-mail: aip@ei.tum.de*
Technische Universität München

***Abstract*— Independent navigation and obstacle avoidance in an unknown environment is highly challenging for visually impaired people. A portable device, translating visual information into 3D sound, can assist and provide the user with a spatial impression of the surroundings. To successfully localize obstacles represented through virtual sound-sources in space, sounds need to be adapted to individual listeners. Three virtual sound-source creation methods, allowing fast individualization at low-cost, were implemented and compared. They were all based on a two-step selection of head-related transfer functions (HRTF) from a catalogue. In the pre-selection step, a subset of HRTFs was found, from which the final HRTF was subjectively selected. The pre-selection relied on pinna-dimensions in the first virtual sound-source creation method, and on head-dimension in the second and third methods. The individualization took less than fifteen minutes for the first and three minutes for the second and third methods. To improve the elevation perception of the virtual sound-sources, two different elevation coding techniques, using low-pass filtering and band boosting respectively, were implemented. These coding techniques improved precision and accuracy of sound localization without increasing the duration of the personalization process. The two different coding techniques were implemented in the second and third virtual sound-source creation methods. The three sound creation methods were validated in a 2D localization test. The results of the ear-dimension based selection indicated an overestimation of lateral positions, which was not observed in the head-size based selection. A better elevation estimation performance was observed with the methods using the elevation coding techniques.**

I. INTRODUCTION

In order to assist visually impaired people during mobility tasks via sound, the perceived location of the virtual sound-source has to be precisely controlled. The difficulty with precise sound localization is that everyone relies on individual features for localization; ear shape, head and torso size are parameters that influence the sound reaching the eardrum of the listener. Therefore, using a generic virtual sound for every listener would lead to suboptimal results for sound localization performance. One method to address this issue is to measure each user's individual head-related transfer functions (HRTF).

The long-term goal of this work is to develop a widely adopted mobility device for the blind. To achieve this, the navigation aid has to be low-cost and easily adaptable to the user. Since measuring individual HRTFs is either expensive or time consuming, this method is not an option.

Consequently, two HRTF individualization processes, based on HRTF selection, were developed considering the following criteria:

- Sound localization performance
- Cost and time consumption

Moreover, two different elevation-coding techniques were used to improve the elevation estimation performance. Finally, three different virtual sound-source creation methods, using HRTF individualization and elevation coding, were designed and tested for validation. The individual accuracy, precision and group scattering of the sound-position estimates were the main criteria used to evaluate the achievable performance with these methods.

II. SOUND CREATION METHODS

To achieve precise localization of sound sources, it is necessary to individualize the sound; this means the generated sound should be personalized with respect to the user's anthropometric characteristics. Moreover, in order to improve the localization performance, elevation coding can be used. The following presents three different methods implemented and tested in this work, which generate sounds based on individualization and elevation coding. For each method, a pre-selection of HRTFs from the CIPIC database was performed based on anthropometric parameters followed by a subjective selection process for the final HRTF [CIP, SF03].

A. Pinna-Based HRTF Individualization

The first method uses only sound-individualization without any position encoding. A two-step procedure is used for the HRTF selection; first a pre-selection is performed which is then followed by a final selection. The pre-selection was based on the ear dimensions of the user. For this the CIPIC database was used as it contains a large set of HRTFs and an expanded set of additional data such as anthropometric parameters [CIP]. In order to have an efficient selection, the most important dimensions had to be chosen out of the ten different pinna-dimensions that are available in the CIPIC database. Based on a statistical analysis performed by Zhang et al. [MZA11] four pinna dimensions were chosen. The relative distance of the ears was computed, and the five HRTFs with the lowest relative error were selected.

Once the pre-selection was completed, a subjective selection was used to find the final HRTF for the testing [SF03]. During this process, the subject listened to virtual sounds produced by the different pre-selected HRTFs. The displayed sound was a virtual source moving stepwise from the right (-40°) to the left (+40°). The user could listen to the signal as many times as needed, was allowed to take notes, and to choose the order of display, so that back-to-back comparison could be performed.

The user was asked to focus on three criteria while comparing the different stimuli:

- The range of display; the sound source was displayed in the frontal area, between -45° and +45°. If a stimulus gave the impression that the sound source was outside this range, this pre-selected HRTF should be discarded.

- The second criterion was accuracy; the sound source moved from the right to the left side step by step, with an equal step size. If the user had the impression of speed up or speed down of the sound source, the pre-selected HRTF was to be discarded.
- Finally, the subject was asked to discard those pre-selected HRTFs for which the sound source appeared to be located behind instead of in front.

To allow the user to compare these criteria easily, a Matlab® GUI was developed.

B. Head-size based Individualization and elevation coding

The second method was a modification of the first one. For this method, the pre-selection was not based on pinna dimensions but on head size. Based on Zhang et al., the head width is the head parameter showing the highest correlation with the head-related impulse response and was therefore the chosen parameter for HRTF pre-selection [MZA11].

Additionally, in this method, sound elevation was encoded in the stimulus characteristics to help the listener estimating the elevation of virtual sources. Inspired by Susnik et al., elevation was coded by lowpass filtering the pink noise pulses with different cutoff frequencies [RST05]. Each cutoff frequency was mapped to a particular elevation. The choice of the cutoff frequencies was made based on the critical bands following cutoff frequencies were therefore chosen: 0.7 kHz, 1.17 kHz, 1.850 kHz, 2.9 kHz, 4.8 kHz and 8.5 kHz to encode -33.75°, -22.5°, -11.25°, 0°, +11.25°, +22.5°. No filtering was used for the +33.75° elevation.

C. Band Boosting for Elevation Encoding

The third method differs from the second one in the way the elevation of the virtual source was coded. Here, instead of filtering the signal, a specific frequency band in the stimulus was boosted depending on the elevation of the source. For high sources, high frequencies were boosted.

III. METHOD EVALUATION

To evaluate the performance of each of the three sound creation methods, a testing process was developed and tests were conducted with normal hearing subjects having no previous experience with sound experiments. The test procedure is described in the following section.

A. Experimental Setup

Binaural hearing relies on interaural and spectral cues and the sound stimuli were generated to emphasis these cues. A pink-noise signal was used since it is broadband. The sound was modulated with pulses to assist the interaural time difference cue. The pulse width was equal to 100ms and the time-space between two pulses was equal to 150ms. The sampling rate of the sound was 44100 Hz. The virtual positions were generated at -30°, -20°, -10°, 0°, +10°, +20°, +30° for the azimuth and -33.75°, -22.5°, -11.25°, 0°, +11.25°, +22.5°, + 33.75° for the elevation, where azimuth and elevation refer to the interaural polar coordinate system.

B. Localization Paradigm

For the subject to communicate the perceived positions of the virtual sound-sources, a laser pointer was used [See02]. The laser pointer was statically connected to two servos and controlled with the help of a computer. The subject was positioned in front of a white wall at a distance of 1.4 meters. After listening to the sound, the subject could change the laser position with the keyboard to point at the virtual sound-source. The estimated direction could then be derived from the position of the pointer. The laser could be pointed at directions in a range of [-80°, +80°] and [-60°, +60°] for azimuth and elevation, respectively. The subject was positioned right below the laser pointer, and the elevation of the chair was adjusted so that the subject's eyes were exactly at 1.3 meters elevation. Markers were used to verify that the head wouldn't move. This paradigm was estimated to be 1° precise. In this work, the paradigm was assumed to be intuitive and no training was used.

C. Experiment

To evaluate the performance of the three different methods a two-step procedure was followed; the first step was the HRTF selection based on anthropometric and subjective selection, and the second step was the localization test. For the localization test, sound sources were displayed and the subject had to localize them by moving the laser pointer. 45, 70 and 70 sound sources were localized per subject for the first, second and third experiment, respectively.

D. Results

1) First Method

12 subjects participated in the experiment. The duration of the individualization was 10 minutes on average, and the subsequent test lasted 15 minutes. The results are shown in Figure 1.

For the azimuth, a bias was observed for the mean estimations of each subject. This bias was equal to 11.4° on average. The mean unsigned error was equal to 16.5° and varied strongly across azimuth. Indeed, this value was equal to 8.6° for a target azimuth equal to 0° and to 18.2° for a target azimuth equal to 30°. A linear regression of the mean estimated values further showed this overestimation, as the slope of the linear approximation was equal to 1.6. This indicated that the estimates were on average 60% too high. The averaged individual standard deviation was 10.4 degrees. A difference appeared between the standard deviations for the different azimuth angles. The standard deviation was the lowest for the center and for the extreme positions, that is respectively 0°, -30° and 30°.

For the elevation performance, listeners were poor at estimating it as the 17.8° unsigned error and the 0.2 linear regression slope

Azimuth	**Exp 1**	**Exp 2**	**Exp 3**
Signed error	$11.4° \pm 1.7°$	$-3.0° \pm 1.3°$	$1.7° \pm 1.8°$
Standard deviation	$10.4° \pm 1.8°$	$8.2° \pm 1.1°$	$7.2° \pm 1.7°$
Between mean standard deviation	$13.5° \pm 3.8°$	$6.7° \pm 1.8°$	$8.0° \pm 2.5°$
Elevation			
Signed error	$-12.6° \pm 2.1°$	$-2.8° \pm 1.4°$	$-3.1° \pm 2.4°$
Standard deviation	$9.8° \pm 1.7°$	$9° \pm 1.8°$	$12.0° \pm 2.6°$
Between mean standard deviation	$7.9° \pm 2.3°$	$7.8° \pm 1.6°$	$6.0° \pm 2.3°$

Figure 1: Obtained results for the three different experiments. Exp 1, 2 and 3 refer to the experiments 1, 2 and 3 respectively. The signed error and the standard deviation are averaged over all the subjects. The between mean standard deviation is the standard deviation of the mean estimations of the subjects.

revealed.

2) Second and Third Methods

12 and 5 subjects, respectively, participated in the tests of the second and third virtual sound creation methods. The duration of the individualization was on average 4 minutes, and the test duration was 20 minutes.

Regarding the azimuth, linear regression performed on the experiment results revealed a bias equal to -18% and +12% using the low-pass and band-boosting techniques, respectively. The signed error was equal to -3.0° and +1.7° for the second and third experiments, respectively, which represents a 70% and 83% improvement compared to the results obtained with the first method. The between mean standard deviation also deceased compared to the first experiment, of 50% for the second experiment and of 45% for the third experiment. These values were equal to 6.8° and 8.0° for the second and third experiments, respectively. The standard deviation of the azimuth estimations also decreased from 21% and 24% for the second and third experiment, respectively.

Regarding the elevation, both methods showed that the achievable localization performance for coded elevation sound was very high compared to the results obtained with the first method. The accuracy improved by 85%, with a mean error of -2.8° compared to -17.6° for the first experiment. The low-pass and band-boost coding both showed that a higher signed error occurred for virtual sound-sources at a +33.75° elevation.

E. Discussion

1) Comparison With Other Individualization Methods

The first selection led to a global unsigned error equal to 34.3°, the second and third methods showed a global unsigned error equal to 21.5° and 22.7°, respectively. The head-size based HRTF selection combined with elevation coding achieved a lower global unsigned error than the 34.5° and 35.8° global unsigned error, obtained by Gardner and Rothbucher et al., respectively, for virtual sound-sources produced via manikin HRTF [Gar97, PP14]. The second and third experiment also showed a lower global unsigned error than the 24.0° and 32.1° error found by Rothbucher et al. and Wenzel et al., respectively, for virtual sound-sources produced through individual HRTF [Gar97, KWW93]. However, the global unsigned error found here was higher than the 20.1° error found by Wightman et al. using real sound sources [WK89b]. Finally, compared to the 3.4° global error obtained by Seeber and Fastl using HRTF selection to produce the virtual sound-sources, the global errors found here were much higher [SF03]. Unfortunately, the localization paradigms used in the above mentioned experiences were not the same, and this prevents us to go further in the localization performance comparison, since the localization paradigm influences the results.

2) Head Size Consideration and Elevation Coding

From the experimental results of this work, it is clear that the second and third methods provided better localization results than the first one. Indeed, improvements of 20% up to 70% were made for azimuth and elevation localization for accuracy and precision. Improvement for the mean standard deviation was also observed. Indeed, this value decreased from 13.5° for the first method to 6.7° and 8.0° for the second and third methods, respectively. In these last methods, the regression slope could be reduced to 0.82 and 1.12 respectively, suggesting that selecting the HRTF based on the head-size influenced the azimuth overestimation.

As mentioned before, a higher signed error occurred for virtual sound-sources located at the +33.75° while using the elevation coding method. This may be explained with the directional bands. Indeed, Blauert stated that sounds with frequencies beyond 8 kHz are perceived to be coming from behind and not above [Bla96]. Therefore, it seems that the initial instructions were not sufficient to remap these high frequencies to high position. Consequently, two options are available to counter this effect: a longer training stage can be implemented to force the remapping, or the elevation code has to be adjusted to use cutoff frequencies constrained to the 500Hz - 8kHz range. The drawback of the first solution is that it would be time consuming; the drawback of the second solution is that the resolution of the possibly encoded positions would be lower.

IV. CONCLUSION AND FUTURE WORK

This work aimed to design sound individualization methods with respect to the three following criteria: cost, time consumption and performance. These criteria reflected the desire of creating a device intuitive and affordable for blind people. The sound sources were created through three different methods. The methods using elevation encoding combined with head-size-based sound individualization lead to better performance regarding time consumption and sound localization than the performance obtained with the pinna-based individualization method. This paper therefore suggests these methods for implementation in the assistive device.

The next important topic to investigate is the amount of sound sources a listener is able to detect simultaneously. Indeed, to fully replace vision with sound, a large amount of information has to be transmitted to the ears. Consequently, sound interpretation is indisputably the limiting factor for the achievable information rate.

ACKNOWLEDGEMENTS

BS is supported by BMBF 01 GQ 1004B

REFERENCES

[Bla96] Jens Blauert. Spatial hearing, revisited edition, the psychophysics of human sound localization, The MIT Press, October 1996.

[CIP] http://interface.cipic.ucdavis.edu/sound/hrtf.html

[KWW93] Doris J. Kistler, Elizabeth M. Wenzel and Frederic L. Wightman. Localization using nonindividualized head-related transfer functions. Journal of the Acoustical Society of America, Vol. 94(1): 11-123, July 1993.

[Gar97] William G. Gardner. 3-D Audio Using Loudspeakers. PhD thesis, Massaschusetts Institute of Technology, 1997.

[MZA11] R. A. Kennedy, M. Zhang and T. D. Abhayapala. Statistical method to identify key anthropometric parameters in HRTF individualization, Hands-free Speech Communication and Microphone Arrays (HSCMA), pp. 213-218, 2011.

[RST05] Jaka Sodnik, Rudolf Susnik and Saso Tomazic. Coding of elevation in acoustic image of space, Proceedings of Acoustics 2005, Western Australia

[SF03] Seeber, B. U., and Fastl, H. "Subjective Selection of Non-Individual Head-Related Transfer Functions," in Proc. 9th Int. Conf. on Aud. Display, edited by E. Brazil, and B. Shinn-Cunningham (Boston University Publications Prod. Dept., Boston, USA), pp. 259-262. 2003.

[See03] Seeber, B. U., (2003) “Untersuchung der auditiven Lokalisation mit einer Lichtzeigermethode”. PhD thesis, Technische Universität München, 2003.

[PP14] Klaus Diepold, Philipp Paukner and Martin Rothbucher. “Sound Localization Performance Comparison of Different HRTF-Individualization Methods”, Technische Universität München, 2014.

[WK89b] Frederic L. Wightman and Doris J. Kistler. Headphone simulation of free field listening and psychophysical validation, Journal of the Acoustical Society of America, Vol. 85(2): 858-867, February 1989.