Real-time Spatialized Sound Generator on Embedded Wearable Platform

Mihai Daraban
Applied Electronics Department
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
mihai.daraban@ael.utcluj.ro

Karoly Lengyel
DOTLUMEN
Cluj-Napoca, Romania
karoly@dotlumen.com

Agota Faluvegi
DOTLUMEN
Cluj-Napoca, Romania
agota@dotlumen.com

Cornel Amariei
DOTLUMEN
Cluj-Napoca, Romania
cornel@dotlumen.com

2021 IEEE 27th International Symposium for Design and Technology in Electronic Packaging (SIITME) | 978-1-6654-2110-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/SIITME53254.2021.9663671

***Abstract*— In this paper a novel 3D real-time sound generation and spatialization system, optimized for wearable computers, and enhancements specific for transferring spatialized information to the blind is presented. For visually impaired people, non-visual information sources are paramount to attain accessibility and inclusion. Blind people use their hearing as the main source of information. This research presents a sound engine capable of generating binaural sounds in real-time. Besides real-time binaural sounds, the engine can also enhance real-time spatialization, which results in an improved sound cue perception. The engine performance in providing sound cue feedback was tested through different trials, having visually impaired individuals as test subjects.**

Keywords—binaural sounds, HRIR, HRTF, component, formatting, style, styling, insert *(key words)*

I. Introduction

Eyesight is the primary stimulus used by people for reliable information for spatial tasks and navigation. During evolution, our eyes have developed in providing high spatial resolution to protect us from predators and allowing to gather or catching food. As a result, vision started to dominate other senses and became a calibration tool for sensory cues during spatial tasks [1]. Most objects built today are specifically designed to be used by sight-enabled people. Because of this, the general environment poses multiple challenges to the visually impaired.

When it comes to navigating through an environment, there are two perspectives from which the navigation process can be seen: allocentric (obstacles are relative to each other) and egocentric (spatial representation build on the subject position in the environment). Depending on when an individual's vision is affected, they can understand and represent spatial information allocentricaly. However, in the case of congenitally blind persons, egocentric perception is the only one they can understand.

To help visually impaired or blind persons navigate through large spaces, sensory substitution devices (SDDs) have been developed. The paper proposes a headset capable of providing audio and haptic feedback to guide visually impaired individuals in real-life scenarios. Such a device can help over 36M blind individuals today, a number expected to increase to over 100M by 2050 [2].

For such a device to be feasible, it must be able to perform auditory feedback in real-time. This is a challenge from both a hardware and a software perspective. In the following sections of the paper, details about the implementation are given in order to achieve a sensory substitution embedded device.

II. Binaural Sounds

A. Sound Spatialization

The key point of the proposed system is to transfer the presence of obstacle and other relevant information into spatialized sound cues. The generated sound image should have the following proprieties:

- to be as little intrusive as possible, minimizing the interference with other sounds from the environment,
- the sound frequency and amplitude field should be at the range at which humans are most sensitive for sound spatialization.

When it comes to sound spatialization, the accuracy of sound source localization depends on [3]:

- position in the horizontal plane (better precision in front then to the side),
- maximum number of simultaneous spatially separated sources that can be distinguished is around 3 for tonal stimuli,
- signal bandwidth can influence the accuracy (wider bands make for better accuracy).

For horizontal plane sound spatialization two parameters are important [4]:

- interaural time differences (IDT) – difference between the times sounds reach the two ears,
- interaural level differences (ILD) – differences in sound pressure level reaching the two ears.

The ITD and the ILD are very important parameters for pinpointing the sound origin, but their value changes not only with position but also with the signal frequency. As a result, based on the frequency range, the sound position is detected relying on one principle (Table I). Besides ITD and ILD, there are also studies showing that in the sound frequency domain direction bands exist, such that some frequencies are better to use for front cues, rear or above (elevation plane) [3]. From one person to another there can be differences in the directional band regarding frequency, however, no significant differences were observed between the 1/3 and 1/6 octave band noise signals.

TABLE I. AZIMUTHAL SOUND SOURCE LOCALIZATION AS A FUNCTION OF ITD AND ILD PARAMETERS AND FREQUENCY BAND

Parameter	**Localization accuracy vs. Frequency domain**		
	Bellow 1 kHz	***Between 1 and 3 kHz***	***Over 3 kHz***
ITD	Good	Mediocre	Impossible
ILD	Impossible	Mediocre	Good

The previous studies are important in helping to define a sound signal that will feel as less intrusive as possible and still provide the best experience for spatial localization on the horizontal and vertical plane.

B. Binaural Sounds

To reproduce the same natural sensation through headphones, when generating a sound cue, it is important to also consider the reflections and the head level and propagation of the sound through the ears.

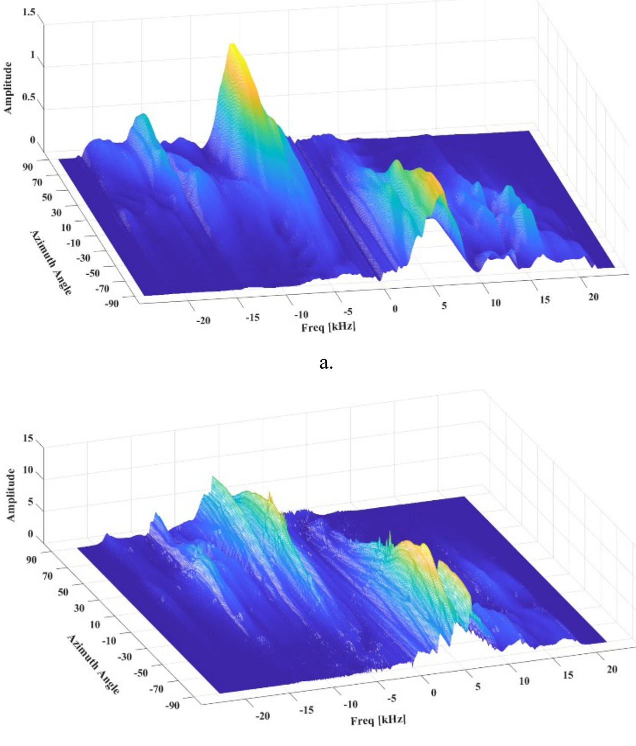
A solution is to use head related transfer functions (HRTF), that incorporate the frequency representation of the changes in sound pressure before reaching the listener's eardrums, which are variable with the shapes of the head, pinna, and torso. HRTF is influenced by the following parameters [3]:

- HRTF varies with distance in the one meter range. Over this threshold no significant impact is observed,
- the absolute phase of HRTF can be ignored, as it does not influence the perception of sound direction,
- there is a tight relationship between HRTF variation with the direction of the sound source, because of asymmetrical shape of head and pinna,
- because of differences in head and pinna shapes, HRTFs vary from individual to individual.

The parameter that has the largest impact over binaural sound generation is the variation of HRTF with human shape and size [5,6]. Using the HRTF measured on someone else to generate the sound cues will cause sounds to be perceived at erroneous azimuth angles or elevation [3].

To produce virtual sound cues that mimic what an individual would perceive in real-life, in terms of spatialization, is very complex. First, a tailored HRTF is need, which can only be obtained through complex sound measurement setup in an anechoic room [7]. A solution is to use either HRTF or HRIR (head related impulse response) measured on an artificial head-and-torso system such as the Kemar (GRAS Sound & Vibration A/S, Holte, Denmark). The Kemar dummy head is created as an average shape of multiple anthropometric head measurements, especially in the ear areas. An advantage of using dummy heads is that the measurements can be more precise as there would be no movements caused by fatigue or imbalance specific to human tailored HRTF setups.

For the development of the embedded binaural sound generator, the Aachen Kemar database was used [8]. In Fig.1 the HRTF of the Kemar data vs a human head can be observed. Compared to the human HRTF, the Kemar data seems to have a more monotonic evolution across the front band frequencies.



a.

b.

Fig. 1. HRTF of Kemar dummy (a) vs HRTF of human head (b)

Using the human HRTF (Fig. 1 b) and applying a sound with a band frequency between four and six kHz that changes position from left to right, as it reaches -20 degree on azimuth the amplitude of the sound will be much higher than at the previous angles creating the sensation that the sound comes from the left ear direction (-90 degree on azimuth). For the sound cues to be perceived at correct angles, it is important to have a monotonic relationship between the angle position and the sound intensity on the left and right ear.

III. BINAURAL SOUNDS PROPERTY

A. Sound properties

To define the sound properties for the audio feedback system, an analysis over the Aachen database with respect to ITD and ILD evolution as a function of angle and frequency bandwidth was done. During this test the following sound bandwidths were used: 3000 – 5000, 3000 – 4000 and 4000 – 5000.

What we observed was that the ILD is having a narrow interval in which the intensity of the sound presents with a monotonic evolution as the azimuth angle changes. Also, the ITD has shown situations where the delay vs azimuth angle was not always monotonic as it would have been expected. In such situations the sound would present with a higher intensity at side angles compared to the ipsilateral ear, Fig. 2.

The current results show that using single tone sounds, the range of sound localization is greatly reduced as there will be difficulties in pinpointing the correct sagittal plane in which a sound source is located. Using white noise allows to have a monotonic evolution for ITD and ILD which is closer to what humans are expecting when a sound cue moves back and forth from left to right in the median plane, Fig. 3.

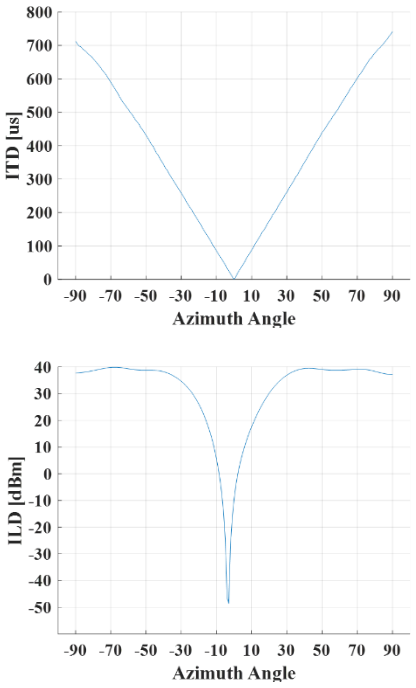


Fig. 2. ITD and ILD of Kemar dummy in 3000 – 5000 Hz bandwidth

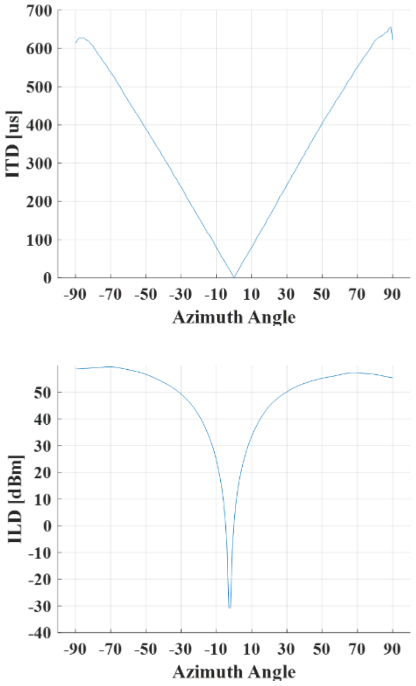


Fig. 3. ITD and ILD of Kemar dummy for full bandwidth (white noise)

Another parameter that is important, besides the frequency band, is the length of the sound cue. Long signal cues will interfere with the human perception and expected feedback. The signal cue should be short such that a new cue can be generated as the person changes the head orientation, but long enough so that the position of the sound cue can be understood. For that, sound lengths between 200 ms and 500 ms were tested. As the results will show both settings yield good results, however the shortest time length was preferred when it comes to sending sound guidance.

Another important sound property is its simultaneous perception. In real life scenarios, multiple parallel sound cues are needed in other to distinguish multiple obstacles or guiding instructions.

It is a known fact that humans can track two simultaneous conversations or having a discussion while there is music in background. However, this ability was developed when hearing known sounds such as instruments or human speech. It was necessary to test the simultaneous spatialization perception of multiple arbitrary sound cues, which don't resemble typically known sounds from the environment. For this, test tones and noise signals were used, which are less likely to be heard in real-life scenarios:

- Sin1: low frequency multiple sinus signal (500 Hz + 600 Hz + 700 Hz + 1000 Hz)
- Sin2: middle frequency multiple sinus signal (3000 Hz + 4000 Hz + 4500 Hz + 5000 Hz)
- Pink1: pink11/6 octave Pink noise signal with a nominal center frequency at 1000 Hz
- Pink2: 1/6 octave Pink noise signal with a nominal center frequency at 3500 Hz
- Pink3: 1/6 octave Pink noise signal with a nominal center frequency at 7500 Hz.

The signals are paired during cue generation such that the mono sounds do not have common frequencies. For example, the low frequency multiple sinus signal will never be paired with 1/6 octave Pink noise signal with a nominal center frequency at 1000 Hz.

Another purpose of the sound generation engine is to give navigation cues which help blind individuals avoid obstacles and be guided in real-life scenarios. While the sound spatialization engine can provide such cues as well, their specifics are trade secret and beyond the purpose of this research paper.

IV. Embedded Wearable Platform for Guiding Visually Impaired People

The real-time spatialized sound generator requires an embedded platform which can be wearable, while providing enough performance for this real-time application. The hardware used for running the custom sound engine is made from a high-performance portable computer, an odometry acquisition device and a set of over the ear headphones.

Multiple high-performance portable computers are available in the industry. Due to the increased interest in robotics, several solutions such as the NVIDIA Jetson series or the Intel NUC devices exist.

The odometry acquisition device must understand the translation and rotation of the human head, at a high-enough rate and good enough precision such that the generated sounds are well spatialized. These are paramount requirements for a spatialization performance which can convince humans of correct spatialization. Multiple technologies can provide this information, such as Inertial Measurement Units (IMUs), Inertial Navigation Systems (INSs), or even some modern Visual odometry sensors such as the Intel Realsense T265.

The portable computer obtains the odometry data from the odometry sensor which must be placed on the human head such that the movement of the head is correctly detected in the odometry. Then, the odometry is used as input for the audio engine, Fig. 4, in generation spatialized sounds that are intuitively perceived in term of direction, azimuth, elevation and distance. As the final device needs to be portable, the sound engine was developed having as target real-time feedback and low-power embedded platforms suitable for wearable devices.

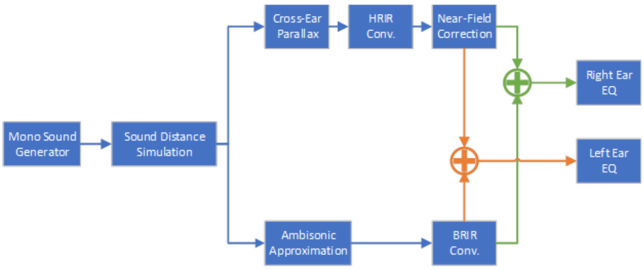


Fig. 4. Sound engine module diagram

A. Mono Sound generator

Based on a few parameters, one can generate a large range of distinct sound types. The first and the most important parameter refers to the shape of the sound wave. At this point the sound can include components like sinusoidal, triangular, square signals or white or pink noise or any combination of these. The second parameter is the frequency or frequency range of the sound, that can take values from the human audible spectrum (20 – 20kHz). Another important parameter is the amplitude of the signal, in other words, the volume of the sound. These parameters refer to each component of the sound, and each sound cue can contain multiple such sound components.

Further parameters include the number of samples or sampling frequency, which are two different representations of the same aspect, namely how many sound samples are in one second of sound. Another parameter is the global amplitude, which refers to the general loudness of the sound, compared to each separate component that can have their own volume.

There are two more parameters, that are related to the cyclicality of the sound: the sound period and the duty cycle. The first one indicates the time required for the sound wave to make a complete oscillation, while the second one refers to the amount of time the sound is audible in a period. A duty cycle of 100% means a continuous sound, while 0% means complete silence.

B. Sound player

The sound player is a custom software implementation that can recreate real-time sounds.

At creation, a block size of 1024 samples is specified, and when the block is consumed, a callback function requests the next one. Since this is a real time system, the callback must contain only simple and fast operations, to avoid buffer underruns. More specifically, the execution time should not exceed $1/f_s$ seconds, where f_s is the sampling frequency.

Considering this criterion, a sound player engine was developed, consisting of a player, which contains a mixer, which contains multiple channels. The external application sends sound arrays to their corresponding channels, which can be the length of a block size, or longer. If the sound array is longer than the specified block size, at input phase, the data is preprocessed and separated in chunks which have the length of the specified block size and are fed in a first in first out (FIFO) queue. If the sent sound array has exactly the length of block size, it is fed directly in the queue. At callback time, the mixer requests a chunk from the FIFO queues of all the channels, mixes them based on their relative amplitude, and sends the result to the player.

This structure enables real-time sound processing in the application, since only individual chunks are processed at a time. Furthermore, it enables parallel computations.

C. Sound spatializer engine

The sound spatializer contains the sound distance simulation, cross ear parallax [9], HRIR convolution, near-field correction, ambisonic approximation [10] and BRIR convolution modules, as presented in Fig. 2.

When spatialized sound is required, a mono sound is requested from the Mono Sound generator. When this is enqueued, the spatializer engine starts listening to the system odometry, which is published by the odometry sensor at an externally specified rate. When the data is available, a callback function registers the odometry, and based on the point which is the spatialization target, it calculates the azimuth, elevation, and distance. The signal path at this point is divided in two ways: the anechoic path and the echoic path.

The first stage in the anechoic path is the cross-ear parallax calculus, which is applied based on recommendations from [9]. This stage selects the appropriate HRIR for each ear, since there is an offset between the center of the head and the left and right ears, and a different azimuth is required for each.

The HRIR convolution is performed using the Overlap Save convolution (OLS) technique [11]. Since the spatialization azimuth and elevation can vary over time, the convolution filter must be changed dynamically. This leads to sudden changes in signal amplitude because of signal phase change, which are perceived as sound glitches. To overcome this, a linear crossfade technique was applied when filter change was imposed:

$$y_{out}(n) = \begin{cases} y_0(n) \cdot f_{out}(n) + y_1(n) \cdot f_{in}(n), & n < L \\ y_1(n), & otherwise \end{cases} \quad (1)$$

where, y_{out} – output signal,

y_0 – input signal before changing the convolution filter or spatialization angle,

y_1 – input signal after changing the convolution filter or spatialization angle,

f_{out} – fade out envelope described by $f_{out} = 1 - \frac{n}{L}$,

f_{in} – fade in envelope described by $f_{out} = \frac{n}{L}$,

n – the crossfade filter current sample index,

L – length of the crossfade filter, application dependent.

The following stage is the near field correction stage, a stage seen in previous research such as [12]. This stage also implies dynamic filter replacement, however, the above mentioned crossfade solves this problem.

The echoic path runs in parallel with the anechoic path. At core it is a BRIR filter, which contains a binaural room impulse response. In our case, it contains four filters, at four different azimuths: front, back, left, right. To obtain responses at azimuths in between, ambisonic approximation is applied.

The ambisonic approximation stage consists of B-format ambisonic encoding and decoding. This calculates the gains for each filter. At convolution stage, the signal is split in 4 parallel ways, multiplied with its corresponding gain calculated in the previous stage, and convolved with its corresponding filter. Since room impulse responses are very long, Uniformly Partitioned Overlap Save convolution (UPOLS) was applied. In this path, crossfade was not required since the filters are static, only the gains are changing.

V. Test Results

A. Simultaneous sound discrimination

One of the tests performed was to analyze simultaneous sound perception. As described in section III, pairs of mono sounds were used to perform this test. The generated cues were placed on different azimuth angels and at different distances using the sound engine. This test was performed on 13 blind and visual impaired persons. It was a static test, with no changes in the azimuth direction during the test.

The generated simultaneous sound cues were as follows:

- at the same azimuth angle (i.e. 10º or 340º), but at different distance in space. For example, one cue was set at 0.2 m and the second one at 1 m.
- at the same distance i.e. (0.2 m or 1 m), but at different azimuth angles. During the tests the following pairs of angles were used (Sound 1 azimuth angle, Sound 2 azimuth angle): (10º, 20º), (20º, 45º), (60º, 30º), (320º, 340º).
- neither the azimuth angle or distance was the same.

In all three scenarios the mono sounds, used as input for the binaural engine, did not have common frequency components.

For all three scenarios it could be observed that if the two sound cues were similar in content (both multiple sinus components or both pink noise bands) more than 70% of the test participants would perceived the sound cues as one. The distance or the angle difference between the cues did not help in perceiving to sound images.

Using a 1/6 octave Pink noise signal and a multi sin tone as input for the sound cues had a greater impact in creating the sensation of two sound images at different points in space. Combining the • 1/6 octave Pink noise signal with a nominal center frequency at 1000 Hz with middle frequency multiple sinus signal resulted in over 75% of the test subjects reporting to perceive two sound cues. This result was accomplished even if the sound images were at the same angle and different distances in space (i.e. 0.2 m and 1 m). However, if the same mono sound were used for a scenario in which the distance was set at 0.2 m and the sound cues were generated at different angles, only 50 % of the test participants perceived two sound cues. The previous test results showed that because of the loudness associated with close sound images (at 0.2 m) the brain interpreted the generated sounds as one single source.

From the previous tests the most important parameter in distinguish between multiple simultaneous sound cues it seams to be the content of the mono sources. Even if the angle between the cues was set as clos as 10º, the test subjects were able to distinguish them as long as the content was different.

B. Sound tracking

The next step had the purpose to test if a person can be guided using the hardware setup described in section IV. The test setup consisted in providing a sound cue at a specific angle.

In this test, the purpose is to reproduce how humans are used to pinpoint a sound source during daily activities. This usually involves a few changes of head orientation to what seems to be the origin from which the sound comes. By changing the head orientation with respect to the sound source the human brain can compare the sound origin to front of the head which acts as reference. During the test procedure, the taking part subject will have to move its head towards the direction from which it is perceives the binaural cue. Based on IMU device input, the Sound engine module will change binaural cue angle as the subject changes the head orientation to localize the sound image origin, **Error! Reference source not found.** 5.

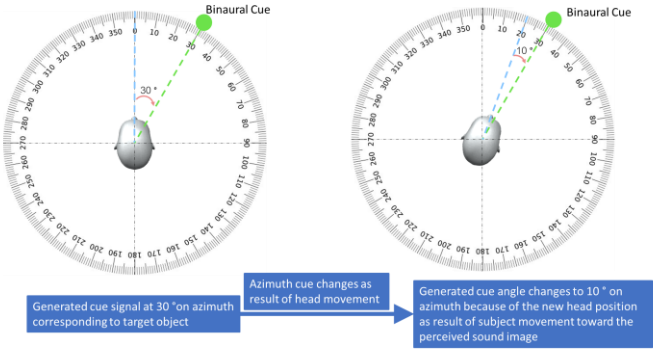

Binaural Cue
30 °
Binaural Cue
10 °
Generated cue signal at 30 °on azimuth corresponding to target object
Azimuth cue changes as result of head movement
Generated cue angle changes to 10 ° on azimuth because of the new head position as result of subject movement toward the perceived sound image

Fig. 5. Binaural cue generation vs head movement

For this test the same set of sounds described in section III were used. However, there was used also white noise because of the much better ITD and ILD performances as described in section III Fig. 3. During a test the blind person will hear a sound cue, and it needs to orientate the front of the head in the direction in which perceives the sound cue. Once it believes that the sound cue is centered (same intensity on both ears) it pushes a button to stop the test. At first the system records the system initial state, and only after 2000 samples will start playing the sound cue, Fig. 6.

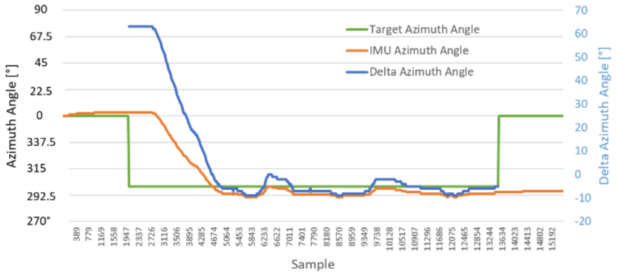


Fig. 6. Angle tracking during head turner test

Testing the three pink noise signals revealed that Pink1 and Pink3 allowed a much better orientation for the participating test subjects. For these two signals the orientation error towards the target angle was below 10º azimuth angle error in 85% (Pink1) and 70% (Pink3) of the performed tests. In case of Pink2 only 55% of the tests were bellow the 10º error margin.

The results obtained with signal Pink1 were also confirmed when Sin1 signal was used as input for the ssound engine module. When this signal was used 80% of the tests were under the ±10 º azimuth error. The middle frequency multiple sinus signal managed to get only 60% of the tests under the ±10 º azimuth error. White noise provided the same performance as Pink1 and Sin1 signals. However, when considering the sensation of hearing a noisy background, white noise was the least preferred by the participating test subjects.

Beside analyzing the azimuth orientation, tests involving azimuth and elevation were also performed. When trying to reproduce the elevation sensation through binaural sounds there is a chance that the azimuth perception to be affected. Using Pink1 only 35% of the tests reported a precision bellow ±10º on both azimuth and elevation. Pink3 signal was close, providing 40% of the tests in the ±10º error margin. Using Sin1, Sin2 or white noise did not improve to much the perception when it comes to elevation. The current results targeted an error margin imposed on horizontal and median planes. However, in majority of the cases the median plane was affected by a higher error then the horizontal plane.

VI. Conclusions

This research presents a sound engine capable of generating binaural sounds in real-time. Besides real-time binaural sounds, the engine can also enhance real-time spatialization, which results in an improved sound cue perception. The engine performance in providing sound cue feedback was tested through different trials, having visually impaired individuals as test subjects. During the test setups two main features were under evaluation: multiple sounds discrimination and sound cue orientation on traverse and sagittal planes. These tests allowed to identify sound types and frequency domains that could be used in a navigation system for visual impaired or blind people.

The developed device proved that horizontal orientation can be obtained with high precision, while elevation can still be improved. In a future research, a solution might be to use dynamic sound cues to suggest the direction or to combine the current hardware with another feedback technique to improve elevation orientation.

Acknowledgment

This work is a small part of the research and development done at DOTLUMEN (.lumen), a research start-up that builds glasses that empower the blind to live a better life. .lumen is committed to solve the discrepancy between the lack of assistive mobility solutions, and the high number of people with visual disabilities.

References

[1] C. Jicol, T. Lloyd-Esenkaya, M. J. Proulx, S. Lange-Smith, M. Scheller, E. O'Neill and K. Petrini, " Efficiency of Sensory Substitution Devices Alone and in Combination With Self-Motion for Spatial Navigation in Sighted and Visually Impaired," Frontiers in Psychology, vol. 11, pp. 1-17, 2020.

[2] R. R. A. Bourne et al., "Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis", Lancet Glob Health, Vol. 5, No. 9, pp. 888-897, September 01, 2017;

[3] K Iida, Head-Related Transfer Function and Acoustic Virtual Reality, Springer Nature Singapore Pte Ltd., 2019, pp.7–24.

[4] M. Risound, J. N. Hanson, F. Gauvrit, C. Renard, P. E. Mesre, N. X. Bonne and C. Vincent, "Sound Source Localization, European Annals of Otorhinolaryngology", Head and Neck Diseases, Vol. 135, No. 4, pp. 259-264, 2018.

[5] B. E. Treeby, J. Pan and R. M. Paurobally, "The effect of hair on auditory localization cues", The Journal of the Acoustical Society of America, Vol. 122, No. 6, pp. 3586-3597, 2007.

[6] X. Zhong, X. Xu and J. Zhang, "Influence of small and large pinnae on virtual auditory perception", 21st International Congress on Sound and Vibration 2014, July 2014, Beijing, China

[7] S. Li and J. Peissing, "Measurement of Head-Related Transfer Functions: A Review", Applied Sciences, Vol. 10, No. 14, pp. 1-40, 2020.

[8] H. Braren and J. Fels, "A High-Resolution Head-Related Transfer Function Data Set and 3D-Scan of Kemar", RWTH Aachen University, pp. 1-6, 2020, DOI: 10.18154/RWTH-2020-11307

[9] D. Romblom and B. Cook, "Near-Field Compensation for HRTF Processing", Journal of The Audio Engineering Society, October, 2008.

[10] P. Mahé, S. Ragot and S. Marchand, "First-order ambisonic coding with quaternion-based interpolation of PCA rotation matrices". EAA Spatial Audio Signal Processing Symposium, Sep 2019,Paris, France, pp.7-12.

[11] F. Wefers and J. Berg, "High-Performance Real Time FIR Filtering Using Fast Convolution on Graphics Hardware", Proc. of the 13th International Conference on Digital Audio Effects (DAFx-10), Graz, Austria , September 6-10, 2010.

[12] M. Cuevas-Rodríguez, L. Picinali, D. González-Toledo, C. Garre, E. Rubia-Cuestas, L. Molina-Tanco, A. Reyes-Lecuona, "3D Tune-In Toolkit: An open-source library for real-time binaural spatialisation", PLOS ONE, Vol. 14, No. 3, 2019.