A Study on an Efficient Spatialisation Technique for Near-Field Sound in Video Games

Manuel López Ibáñez, Nahum Álvarez, and Federico Peinado

Department of Software Engineering and Artificial Intelligence
Complutense University of Madrid
Profesor José García Santesmases, 9. 28040, Madrid (Spain)
manuel.lopez.ibanez@ucm.es, email@federicopeinado.com
`http://nil.fdi.ucm.es`
Artificial Intelligence Research Center
National Institute of Advanced Industrial Science and Technology (AIST)
2-3-26 Aomi, Koto-ku, Tokyo 135-0064 Japan
nahum.alvarez@aist.go.jp

Abstract This article presents a simple and efficient method for spatialising sound in virtual environments by adding low pass filters (LPF) to the already widespread panning and attenuation techniques. Through two different experiments, variations in subject performance when locating sounds in a virtual environment between regular 3D audio from popular game engines (Unreal Engine and Unity) and our proposed sound system were evaluated. The first experiment consists of an audio-only test via an online survey, whereas the second experiment employs a minimalistic 3D video game which allows for user interaction guided by sound. Results of both experiments suggest better performance and accuracy when using LPFs, the second one finding a significant difference when comparing both techniques. We conclude that the LPF technique, as a mean for spatialisation of sounds coming from behind the subject, could be applied to complement current audio systems due to their performance-oriented nature and their good results with real users.

Keywords: Acoustics · 3D Audio · Directional Sound · Spatial Attention · Entertainment Technology

1 Introduction

3D sound for video games has not been a particularly fertile field in the past years due to the continued use of traditional spatialisation techniques [1] and a relative lack of attention from both players and developers of virtual environments. However, the popularization of Virtual Reality (VR) has brought an increasing interest in improving sound systems for video games [2], so as to achieve levels of realism and presence [3] that were previously out of reach. This new wave of sound technologies for video games has generated interesting initiatives, such as Steam Audio[1], which try to go beyond Head Related Transfer

---
[1] https://valvesoftware.github.io/steam-audio/

Manuel López Ibáñez, Nahum Álvarez, and Federico Peinado

Functions (HRTFs) and take into account in-game geometry and materials to simulate auditory spaces. Yet, these systems focus mainly on realism, not necessarily on usability. Our intention with this study is the opposite: to focus on gameplay, improving player orientation and task performance, even if that means sacrificing realism.

A good example of how the exclusive use of a complex and realistic sound system can create gameplay problems is the recent addition of HRTFs to Valve's first person shooter (FPS) game *Counter Strike: Global Offensive*[2]. In this game, players are able to move their avatars' heads at a much higher speed than in the real world, which, together with a more realistic sound system that adds a delay to audio propagation, can generate confusion when trying to quickly locate sound sources, as the perceived delay is not consistent with HRTF generation [4]. That is: in this situation audio changes are slower than player movements. Our take on this problem is to propose an audio system that is not completely faithful to reality, but gives enough clues to allow players to quickly learn where sounds are located in virtual space. Besides, it works more efficiently than an HRTF-based system (in terms of computational efficiency), as it only needs to track if an object is not being rendered by the in-game camera, so as to decide when to apply an LPF to the sounds emitted by it, as it will be explained later.

Through this paper, we will propose a simple sound technique, based on LPFs, which tries to balance realism and usability, while aiming to achieve accurate sound source identification for all users. Our main intention is to allow players to have better performance when identifying sounds coming from behind, which constitutes one of the currently most important challenges in 3D and surround sound generation [5]. This is achieved through a very brief learning process. Using our method, users compare sound sources they can see with sound sources they cannot see, both being applied a different audio processing. Subjects easily identify what type of sounds are meant to come from behind, and which come from the front, just by comparison or by a process of elimination, and after just a few seconds of training.

The structure of the present article is as follows: First, we will review the state of the art in 3D sound spatialisation and HRTFs; in the next two sections, we will state our goals and the experiments we designed to reach them; after that, we will compare our initial hypothesis to the results achieved, interpreting and discussing the data; finally, we will end with some brief conclusions about applicability of our system and future lines of work.

Goals. The goals of the present research are the following:

- To explore how sound spatialisation techniques work by default in two of the most commonly used game engines (Unreal Engine and Unity Engine), and try to improve them.
- To identify an efficient and simple method for spatialising sound, which could be used together with other, more complex approaches.

---

[2] www.counter-strike.net

- To study possible differences in accuracy when users try to identify a sound coming from the rear with and without LPF applied to it.
- To test the performance of our proposal with real users locating sounds in a virtual environment, using a pragmatic approach, instead of focusing on the level of realism achieved.

2   3D Sound Spatialisation

It is well known that the key component in the process of spatialising sound in a 3D virtual environment is being able to simulate *sound direction* and *sound distance* [6]. In this subsection we focus on two different techniques: Head-Related Transfer Functions and LPFs.

2.1   Head-Related Transfer Functions

Currently, the most used technique to capture sounds that include information about their direction is through Head-Related Transfer Functions (HRTFs) [1]. HRTF capture consists, essentially, of recording sound as it would have been heard by an individual. To achieve that, a set of microphones is placed in front of both ears, in an attempt to capture sounds from all relevant directions. The subject used for capturing can either be a human with microphones attached to the head or a dummy specifically designed for that purpose. For example, the KEMAR HRTF database [7] used a synthetic head to recreate the hearing capacities of a human. Other commonly used databases include: LISTEN HRTF [8], CIPIC HRTF [9], FIU DSP Lab HRTF [10] and ARI HRTF [11].

The information captured during the recording process can later be used to create impulse responses, which can be attached to game audio by using a convolution reverberation plugin. The result is a processed sound that reflects the physical properties of the environment in which the impulse was recorded, and contains information about sound direction. The combination of HRTFs and reverberation effects allows for accurate 3D sound placement in a multichannel environment, and is currently the most commonly used technique in realistic audio content creation.

2.2   Low Pass Filters

Another common method for spatialising sound, and the one we chose for developing our technique, is the use of LPFs. These audio filters are linear and time-invariant; they emulate, by cutting off frequencies above a chosen number of hertzs (Hz), a common phenomenon of real-life sound: the dissipation of high frequencies in a way that depends on the substance the waves travel through and the distance from the listener at which they originate. A bigger portion of high frequencies are cut if sound travels through a more *dissipating* environment or from far away, and emulating this can give virtual environment's sounds depth and credibility.

Manuel López Ibáñez, Nahum Álvarez, and Federico Peinado

To mimic the mentioned effect, sound designers need to attenuate the correct range of frequencies of a sound attending to its position and the properties of the virtual environment. A very common way to do this in Digital Signal Processing (DSP) is through a Chebyshev (type 1) filter, which is the one being used in our proposal. Its attenuation (A, in decibels), according to Williams [12], can be represented as follows:

$$A_{dB} = 10log[1 + \epsilon^2 C_n^2(\omega)]$$

$C_n(\omega)$ is a Chebyshev polynomial of the nth order which oscillates between ± 1 for $\omega \leq 1$.

$$\epsilon = \sqrt{10^{\frac{R_{dB}}{10}} - 1}$$

And R_{dB} is the ripple in decibels (dB).

As stated by Smith, type 1 Chebyshev filters grant a "faster roll-off by allowing ripple in the passband" [13], which means undesired frequencies are quickly and precisely cut-off. This produces a more clear effect, which is the reason why we chose this variant over the rest.

In video games, high frequency attenuation using an LPF is useful when trying to communicate the distance at which an object is, the kind of materials that surround the user or whether a sound comes from behind or not.
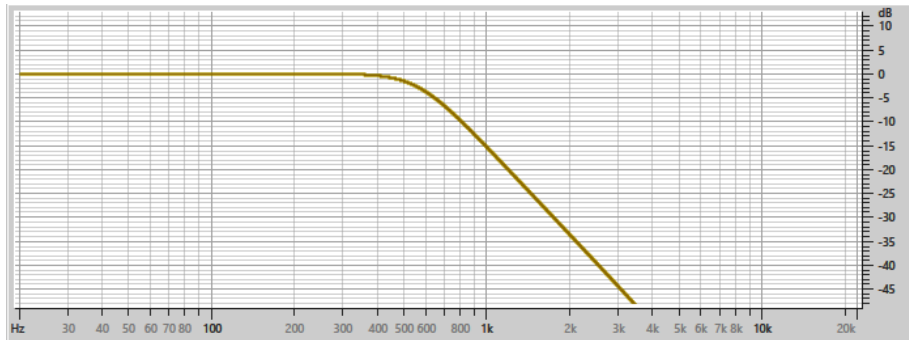
Figure 1. Graph of a Chebyshev filter in Adobe Audition (limit: 200 Hz; transition bandwidth: 300 Hz; ripple: 0,01 dB)

3   Experiment Design

We designed two different experiments (which were called Experiment 1 and Experiment 2 throughout this text) to achieve the results that will be revealed later.
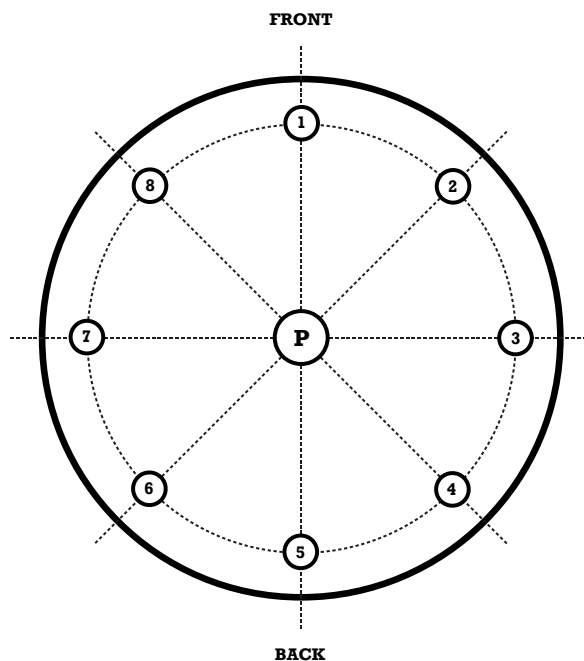
Figure 2. Position diagram used in the first pilot. P represents the player, while numbers indicate all possible sound positions.

Experiment 1. The first experiment was conceived as a mere pilot. It was an online survey, distributed through social networks and completed remotely, in which subjects had to express their opinion on the position of a sound in a complex audio environment while using a pair of headphones. The sound was a single, clear, high pitched alarm tone, which came from four different places in space (from a total of 8 possible directions), sequentially. Users did not have any visual references: only sound. There were two audio tracks playing simultaneously: on one hand, our alarm sound, normalized at -0,45 dB; on the other, the complex audio of an action scene, extracted from a technical demo by Epic Games called *Showdown VR*[3], normalized at -3 dB. The intention pursued when

---

[3] https://www.unrealengine.com/marketplace/showdown-demo

including these two tracks was to help users differentiate between our spatialised alarm sounds and a game-like 3D sound environment, which was used as a reference. The complex audio environment included sounds of guns firing, explosions, cries, etc., coming from a variety of directions.

There were two separated groups of 29 people each, which took two different tests ($1A$ and $1B$). The first one ($1A$) included the original sound of the mentioned demo, along with alarm sounds coming from four positions. Everything was recorded in Unreal Engine, using its default audio system and 3D spatialisation for each sound. The second one ($1B$) included the same background track, recorded in Unreal Engine, but the alarm sounds were processed separately, using an LPF when they came from behind the player. We utilised Adobe Audition[4] as a tool for designing audio with Chebyshev LPFs, achieving attenuation graphs similar to the ones in Figure 1.

Excluding the differently processed alarm sounds, both surveys had the same structure, which was as follows:

- First, the users were asked to listen to an isolated, non-spatialised sample of the alarm sound.
- Next, they were explained how to position sounds in a diagram like the one in Figure 2.
- Then, subjects were presented with a sound track which included four consecutive alarm sounds on top of the background noise described above. They were asked to identify their positions and mark them in the diagram. For example: first sound in position 4, second sound in position 1, etc.
- Lastly, a series of questions related to the demographic profile of each subject were asked. They included: age, sex, level of education, diagnosed audition (hearing) problems, perceived performance during the experiment, frequency with which the subject plays video games and opinion on the importance of video game audio.

All questions related to subjects' opinions were posed by using a Likert scale [14].

Experiment 2. Our second experiment (2) was an on-site test in which users had to play a minimalistic 3D video game made in Unity[5], using a consumer-quality pair of headphones (JVC HA-X570). There were also two versions of this test ($2A$ and $2B$), and each one was taken by 13 different subjects.

This game consisted on an empty room, except for eight spheres that floated around the player, as the ones shown in Figure 3. Subjects had to complete a brief interactive sequence during which they had to point and click (with the help of mouse rotation, buttons and a crosshair on the center of the screen) the sphere that they thought was emitting the looping alarm sound. If the position was correct, the sphere would stop playing its sound, start emitting a blue light,

---

[4] http://www.adobe.com/es/products/audition.html
[5] https://unity3d.com/es

and finally the next sphere of the sequence would start playing the same alarm sound from a different position. When all eight spheres had been turned on, the game ended.

Experiment 2 was thought as a way to increase feedback and allow trial and error, so that every user would end up having information about their general performance. Also, in this experiment time was a significant measure of how well a subject did, as even people with many incorrect answers could finish the experiment, and we could register their delay.
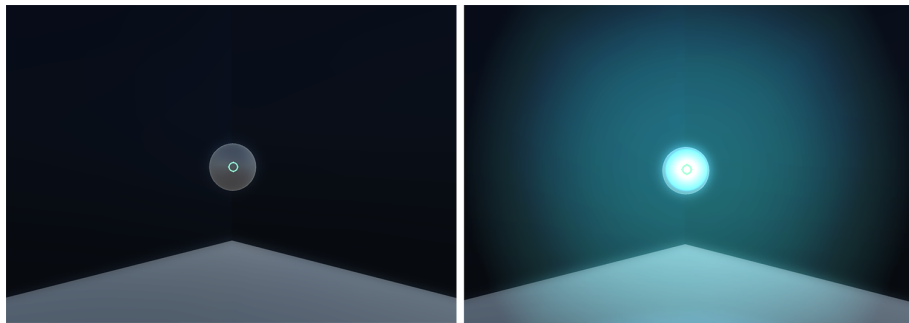
Figure 3. Two of the spheres which could be found in Experiment 2. The left one has not yet been activated, while the one on the right has already been correctly identified and clicked by the player.

Before starting the experiment, every subject was given the following guidelines:

– You will play a game from a first person perspective.
– You will be inside a small and dimly lit room.
– You will not be able to walk around. You will, however, be able to look around using the mouse.
– A crosshair is shown at the center of the screen. It indicates where you are looking at, and always follows the position of the mouse.
– Eight spheres will float around you. All will be at the same distance from you, and static. They will also be at the same distance from each other, forming a circumference around you.
– At the beginning of the game, one of the spheres will produce a looping alarm sound. Your task is to identify the sphere from which that sound is coming, point at it with the crosshair, and click the left mouse button.
– If you identified the position correctly, the sphere will turn blue and the same alarm will start playing from a different sphere. If you failed to identify the position, the alarm will keep sounding until you do.
– The game will finish when all eight spheres are blue and you do not hear any more sounds.
– You must complete the task as soon as possible.

A logger would save data from every user (e. g.: total time to complete the task and raycast hits from the crosshair), so as to be able to compare performance between the two different audio techniques utilised. Besides, everyone had to complete a small survey after finishing the experiment, in which they gave demographic information, as in Experiment 1, and expressed their level of agreement with the actual location of sounds. Relevant data will be detailed in the "Results" section.

Audio system for experiment 2. The only difference between $2A$ and $2B$ was the audio producing method used by each. While $2A$ used the original 3D sound system from Unity, $2B$ used a modified one which worked as follows. Each frame, all audio coming from visible spheres (that is, from spheres being rendered by the player's camera) was processed by using simple stereo panning and distance attenuation (the original methods used in Unity, by default), whereas sounds coming from objects not being currently rendered were applied a LPF. The parameters used for the filter were a cutoff frequency of 2456 Hz and a low pass resonance of 1. This effect is not consistent with how attenuation works in reality, but is clearly identifiable.

The field of view (FOV) of the in-game camera tried to mimic that of the frontal eye field (FEF) of a human eye (around 114 degrees [15]), so that everything outside on-screen space could be considered to be in the rear.

3.1 Hypothesis

Our main hypothesis is that our proposal, a low-latency spatialisation technique based on position-dependent LPF, can allow for more accurate sound position identification when compared to a 3D sound system based on panning, such as the default audio system present in video game engines like Unreal Engine[6] and Unity 3D.

Therefore, the null hypothesis (H_0) in this case is that performance of users when identifying sound positions does not improve when using an LPF-based system. The alternative hypothesis (H_1) is that it does improve only when using our system.

3.2 Demography

The first experiment (1) was passed to a sample of 58 people (41 men and 17 women), randomly distributed in groups of 29 for each version of the test ($1A$ and $1B$), with average ages of 33.03 and 31.28, respectively. The ages ranged between 22 and 51 in group $1A$ and between 20 and 43 in group $1B$.

48 of them had gone through college (17 degrees, 29 master degrees and 2 PhDs), whereas 9 had ended their education during high school.

The second experiment (2) had a smaller sample due to its face-to-face nature, with a total of 26 subjects (19 men and 7 women), randomly distributed

---

[6] https://www.unrealengine.com/

in groups of 13 for each version of the test ($2A$ and $2B$). Group $2A$ had an average age of 25 (18 to 38), whereas group $2B$ was 24.31 years old (18 to 37) on average. 12 subjects were studying a degree related to Computer Science, 9 of them had already finished it, 4 had a PhD and 1 had a master's degree in the field.

4   Results and Discussion

The goal of the first pilot (Experiment 1) was to study the possibility of a difference in accuracy, when users try to identify a sound coming from behind, with and without a spatialisation system based on LPF. Only one sound came from the rear in each version of the experiment, so we measured the number of subjects who got the position of the sound right in each case.

Table 1. Accuracy (right and wrong answers) when identifying sound direction for groups $1A$ and $1B$.

	Sound 1		Sound 2		Sound 3 (rear)		Sound 4	
Group 1A	Right	Wrong	Right	Wrong	Right	Wrong	Right	Wrong
	8	21	7	22	9	20	3	26
	Sound 1		Sound 2 (rear)		Sound 3		Sound 4	
Group 1B	Right	Wrong	Right	Wrong	Right	Wrong	Right	Wrong
	13	16	12	17	4	25	4	25

The first results were promising, though we could not consider them to be statistically significant due to the high rates of failure most subjects obtained in both cases. As can be seen in Table 1, the success rate (number of right answers divided by the total amount of subjects) of group $1A$ for sounds coming from behind was a mere 31.03%, while group $1B$ achieved a 41.38%. In spite of reaching a difference of more than 10 points, not having any set of answers with a success rate of more than 50% led us to the conclusion that the experiment was too difficult for a person with normal hearing.

The data collected during the second experiment (2) was more enlightening than the previous one, as is shown in Table 2 and Figure 4. The average time taken by subjects from group $2A$ to complete the task, 38.84 seconds, is far from the average time of 23.66 seconds achieved by group $2B$. This hints at an improvement due to the utilisation of the new spatialisation system in $2B$.

As Figure 4 shows, results were not regular in Experiment 2 due to the variation induced by the different levels of ability found in subjects. However, average times in $2A$ nearly duplicate those in $2B$, and the same happens with maximum completion times for each group: while $2A$ hits a maximum of 80.36, $2B$'s highest time was 41.4 seconds.

Besides, when asked about their own performance ("Was it easy for you to identify sound positions during the experiment?") in a 5-level Likert-type scale,

Table 2. t Student test for times achieved by users in $2A$ and $2B$.

Group	2A	2B
Subjects (N)	13	13
Average	38.8431	23.6600
Std. Deviation	19.1029	7.1668
SEM	5.2982	1.9877
P-value	0.0130	

11 users in group $2B$ answered in a positive way: "Strongly agree" (3) or "Agree" (8), and 2 gave neutral answers ("Neither agree nor disagree"). Users in group $2A$, on the other hand, gave a total of 8 positive answers (3: "Strongly agree"; 5: "Agree"), plus 5 neutral answers.

The results attained by the new system negate the null hypothesis (H_0) and confirm the alternative (H_1), as there exists a significant difference in performance between the two prototypes ($2A$ and $2B$). If we assume a normal distribution of subject auditive skills in both groups —during the survey, none of the subjects we used to collect data said to have any hearing problems—, the only difference between the two systems is the addition of LPF to sounds coming from behind the player, which makes this variable seemingly responsible for the above-mentioned changes in performance.
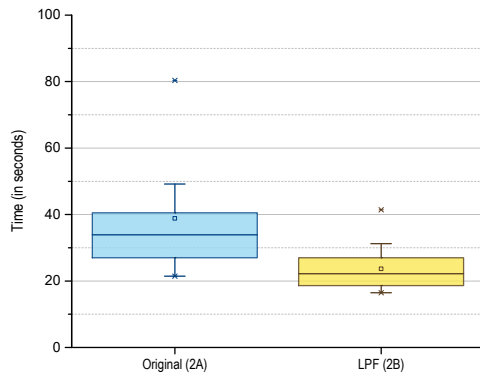
Figure 4. Comparison between the average times achieved by users in both the original ($2A$) and improved with an LPF ($2B$) sound systems.

Though the goals of this research were accomplished, our results would have been more consistent with a larger set of subjects. The second test (Experiment 2) being in-person, and due to time and space constraints, we had to limit the total amount of subjects taking the experiment to a set of 26, though having a greater sample of measures would have been highly beneficial.

Additionally, it would have been useful to integrate our tecnhique in real video games to test its capacities in real-world situations. This was not done due to the lack of popular open source video games which depend heavily on sound spatialisation, and the high development cost it would have had to change their audio systems. Besides, there were far more men than women among volunteers during our tests. Though women were distributed evenly between groups, their small numbers could have influenced the results, as differences in hearing between men and women have been previously discovered [16]. Moreover, the set of subjects utilised does not represent a particular statistical population, and therefore the results cannot be extrapolated to a more generic set of people.

5   Conclusions and future work

Due to the fact that A and B tests contained a single difference in sound processing, and considering the results are better in B groups for experiments 1 and 2, we can extract the conclusion that the addition of LPFs to rear sounds seems to improve —notwithstanding the lack of realism of this technique— recognition of those sounds in the game-like environments already tested. A probable reason for the quicker identification of sound location when applying LPFs is the nature of our implementation: as LPFs are only applied when sounds come from a place that is off camera, the usual reaction for most players was to quickly turn around every time the LPF effect was detected.

Perception of self-performance was evaluated generously by both groups in experiment 2 (stated in the "Results" section), as even users with the highest times gave a neutral answer to the question on this matter, which leads us to think there is no conscious advantage for subjects in group $2B$. However, their results were indeed better.

As for future research, we think it would be interesting to test our system when building near-field interfaces for first person or VR games, as it can be used in addition to a more complex and realistic audio engine, and can improve performance when locating interactive objects in virtual environments. It would be desirable to check if our system works the same way in a full-fledged video game, in which the player can usually find many more auditive stimuli. It would also be useful to build a different experiment in which all frontal sounds would have an LPF applied, and all rear sounds would be left untouched, so as to be able to judge if LPFs are automatically associated to places in the rear, or if they simply induce the observed behaviour by producing contrast (and thus, pattern learning) between two sound categories: *frontal* and *rear*.

Another interesting addition to this work would be to compare the performance of our technique to that of an engine which uses HRTFs, newer approaches such as physics-based sound (e.g.: Steam Audio), or a combination of these two. A possibly useful experiment for this research goal would be to test user performance when identifying sound location in three different audio environments: one which uses only our method, a second one which uses Steam Audio out-of-the-box, and a third one which uses a combination of both, so that LPFs are

only applied to near-field objects not being currently rendered, and the rest of the sounds are spatialised normally.

Acknowledgements

References

1. M. Morimoto and Y. Ando, "On the Simulation of Sound Localization," *Journal of the Acoustical Society of Japan*, vol. 1, pp. 167—-174, 1980.
2. D. Hong, T.-H. Lee, Y. Joo, and W.-C. Park, "Real-time Sound Propagation Hardware Accelerator for Immersive Virtual Reality 3D Audio," *Proceedings of the 21st ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, 2017.
3. J. Lessiter, J. Freeman, E. Keogh, and J. Davidoff, "A cross-media presence questionnaire: The ITC-Sense of Presence Inventory," *Presence*, 2001.
4. C. Phillip Brown and R. O. Duda, "A structural model for binaural sound synthesis," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 5, pp. 476–488, 1998.
5. Z. Cai, S. Makino, and T. M. Rutkowski, "Brain Evoked Potential Latencies Optimization for Spatial Auditory Brain-Computer Interface," *Cognitive Computation*, vol. 7, pp. 34–43, feb 2013.
6. D. R. Begault, "3-D Sound for Virtual Reality and Multimedia," *Computer Music Journal*, vol. 19, no. April, p. 99, 1995.
7. W. G. Gardner and K. D. Martin, "HRTF measurements of a KEMAR," *The Journal of the Acoustical Society of America*, vol. 97, no. 6, pp. 3907–3908, 1995.
8. O. Warusfel, "LISTEN HRTF database," 2002.
9. V. Algazi, R. Duda, D. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, pp. 99–102, 2001.
10. J. C. Gupta, N., Barreto, A., Joshi, M., & Agudelo, "HRTF Database at FIU DSP Lab," in *2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, (Dallas), pp. 169–172, 2010.
11. P. Balazs, "ARI HRTF Database," 2014.
12. A. Williams and F. J. Taylor, *Electronic Filter Design Handbook*. McGraw-Hill, 1995.
13. S. W. Smith, *The Scientist and Engineer's Guide to Digital Signal Processing*. 1999.
14. R. Likert, "A technique for the measurement of attittudes," *Archives of Psychology*, vol. 22, no. 140, pp. 1–55, 1932.
15. I. P. Howard and B. J. Rogers, *Binocular vision and stereopsis (extraits)*, vol. 29. Oxford University Press, 1995.
16. M. Don, C. W. Ponton, J. J. Eggermont, and A. Masuda, "Gender differences in cochlear response time: an explanation for gender amplitude differences in the unmasked auditory brain-stem response," *The Journal of the Acoustical Society of America*, vol. 94, pp. 2135–48, oct 1993.