A NEW VALIDATED METHOD FOR IMPROVING THE AUDIOVISUAL SPATIAL CONGRUENCE IN THE CASE OF STEREOSCOPIC-3D VIDEO AND WAVE FIELD SYNTHESIS

Cédric R. André[*,†], *Étienne Corteel*[‡], *Jean-Jacques Embrechts*[*], *Jacques G. Verly*[*], *Brian F.G. Katz*[†]

[*] INTELSIG Laboratory, University of Liège, Liège, Belgium.
[†] LIMSI-CNRS, Orsay, France.
[‡] sonic emotion labs, Paris, France.
Email: [*] C.Andre@ulg.ac.be [†] brian.katz@limsi.fr [‡] etienne.corteel@sonicemotion.com

ABSTRACT

While 3D cinema is becoming increasingly established, little effort has focused on the general problem of producing a 3D sound scene spatially coherent with the visual content of a stereoscopic-3D (s-3D) movie. The perceptual relevance of such spatial audiovisual coherence is of significant interest. In this paper, we explain why the combination of accurate sound positioning and stereoscopic-3D images can lead to an incongruence between the sound and the image for multiple spectators. Then, we adapt to s-3D viewing a method originally proposed for 2D images in the literature to reduce this error. Finally, a subjective experiment is carried out to prove the efficiency of the method.

Index Terms— Auditory-visual integration, stereoscopic video, Wave Field Synthesis, 3D cinema, three-dimensional television, auditory displays, audio-visual systems

1. INTRODUCTION

This article addresses the question of the perceptual congruence between the sound and the image when the spectator in a cinema is presented with a 3D sound scene spatially coherent with the stereoscopic 3D (s-3D) scene.

In essence, the depth perception in s-3D is created by presenting a different image to the two eyes. Both images in an s-3D pair are displayed on the cinema screen and all spectators thus look at the same pair of images. When one compares the visual perception of two spectators seated at different locations in the room, one finds, both geometrically and experimentally, that the objects of the scene displayed on the screen are rendered at different locations in the room [3].

The present paper considers the potential error in the angle between the sound and the image when presenting precise spatial sound through Wave Field Synthesis (WFS) in combination with s-3D video to spectators seated at different locations. Particularly, we wish to determine the angular disparity range between the auditory and visual stimuli that provides the same feeling of congruence as compared to no angular disparity. Our first contribution is the adaptation to s-3D of an existing method to reduce this angular disparity. Our second contribution is the perceptual validation of the new proposed method.

2. BACKGROUND

2.1. Subjective evaluation of audiovisual congruence

When people are presented with a time-synchronous but spatially mismatched auditory-visual stimulus, they tend to perceive the sound coming from closer to the location of the visual stimulus, the so-called "ventriloquism" effect [29]. This effect decreases with increasing angular difference between the positions of the sources [20].

Vision and audition give us information about the same objects as those we find in our surroundings. This information is integrated in the brain to form the percept of a single audiovisual object. Experiments previously conducted in laboratory conditions used an audiovisual stimulus consisting of a simultaneous pair of brief, simple, and arbitrary stimuli, such as an auditory beep, and a visual flash. For such experiments, a statistically optimal model approximates well the mechanism of bimodal integration [1].

The magnitude of the auditory-visual integration has been found to depend on both spatial relations and temporal relations of the unimodal stimuli. The auditory-visual window of integration of arbitrary stimuli extends up to about 100 ms in time and 3° in azimuth angle [21]. It is centered around azimuth 0° in space (directly in front), when the stimuli are co-located, and about 50 ms in time, when the auditory signal arrives after the visual signal [28, 22]. The effect of a temporal disparity on the spatial localization acuity is greatest when the spatial error between the sound and the image is below the (spatial) threshold of integration [28]. However, this effect is not significant below a 50 ms time delay.

When the stimuli are more natural, i.e. carry meaningful information, such as for a speaking character, then the "unity assumption" must be taken into account. The unity assumption arises from properties shared by the unimodal stimuli (here sound and image) such as spatial location, temporal rate, size, shape, ... [32]. The more numerous the common properties, the stronger the association of the stimuli. Conversely, the more numerous the conflicting cues, the weaker the integration.

Therefore, when more natural stimuli are used, such that the unity assumption holds, the multimodal integration is maintained at much higher angles of discrepancy than those obtained with arbitrary stimuli. Simply by letting the participants assume that the arbitrary stimuli had a common cause, the spatial window can be increased to about 12° [22]. The temporal window, also, can be enlarged. Using a speech stimulus, a 200 ms time window can be obtained [30].

2.2. Off-axis s-3D viewing

In a cinema theater, not all spectators sit exactly in front of the middle of the screen (some sit off-axis), and the distance from the seat to the screen varies also. We review here the potential consequences of viewing an image from an unintended point of view.

First, the regular 2D case is considered. The 2D camera allows to capture on a plane, its imaging sensor, a planar projection of a scene. Mathematically, this is called a linear perspective. When the viewer's eyes are at the correct viewpoint, the picture duplicates the original scene on the viewer's retina [19].

When the viewer's eyes are not at the correct viewpoint, however, the retinal image suggests a scene with a different layout. Still, the viewer experiences the scene in the same way. In particular, judgments of the spatial layout seem relatively constant over the viewing angle [18]. However, judgments on the orientation of lines in space vary systematically with the viewing angle [18]. Directions that point to the sides of the pictures remain constant up to about 20° away from the viewpoint. On the contrary, directions that point out of the picture seem to "follow" the viewer. A famous example of this is Uncle Sam's finger in the "Uncle Sam wants you" poster.

Second, we consider the s-3D case. The question of the perception of the spatial layout in s-3D viewing was investigated in [5]. It was argued that the purely geometrical approach given in [33] to explain s-3D visual localization makes a strong assumption by considering that the visual percept is not corrected for the viewpoint. This assumption had not yet been evaluated. In [5], participants were asked to watch in s-3D a scene consisting of a static hinge with a 90° angle on a display which could be rotated, so that they were not always at the correct viewpoint. The geometrical approach [33] predicts that the hinge angle can be perceived as lower or higher, depending on the position of the viewer with respect to the correct viewpoint. Participants had to judge whether they perceived the hinge with a 90° angle or not. Following a psychophysical procedure, the details of which were not provided, the results showed that viewers did not compensate for their incorrect viewpoint. Instead, the geometrical approach accounted fairly well for the results.

2.3. Off-axis s-3D viewing and precise spatial sound

Visual objects can only be placed by the stereographer of an s-3D movie in our field of view. It is roughly a truncated cone with the apex at the viewer and extending towards infinity behind the screen. Since all the viewers look at the same s-3D image pair, this truncated cone thus follows each spectator and, when the visual perception of two spectators seated at different locations in the room are compared, one concludes that each object in the scene is not rendered at the same physical location in the room. In fact, only visual objects with a zero parallax, such that they are perceived as located at the screen plane, are consistently perceived among spectators. All other positions are not consistently perceived within the room. However, they are all perceived at directions that cross on the screen, at the intersection with the line between the spectator and the perceived object location.

In combination with 3D sound, this property of an s-3D image can lead to an audiovisual error for spectators seated off-axis. Indeed, Wave Field Synthesis, an example of modern audio spatialization techniques, can reproduce audio sources in a large listening area that are consistently perceived by all listeners as coming from the same location. Therefore, sound does not follow the same geometrical distortions as s-3D images. We now investigate this problem geometrically.

In Fig. 1, two spectators at S_1 and S_2 look at the same s-3D pair of images displayed on a screen. We assume that the pair of images contain one object and that the images are such that the spectator at S_1, the ideal (correct) viewpoint, perceives the object as being located at V_1 (behind the screen). For the spectator seated at S_2, the visual object appears at V_2, resulting in an angular error δ between the sound and the image if the sound is positioned at $A = V_1$, the ideal viewpoint.

3. PROBLEM ADDRESSED

3.1. Improving the spatial congruence

We introduce a method to reduce the angular error described in the previous section. This method already exists in the 2D case [14]. Here, we extend it to s-3D. This constitutes one contribution of the present paper.

First, we describe the method which was developed to combine spatially accurate sound rendering, by means of Wave Field Synthesis (WFS [7]), with regular 2D video to build a teleconferencing system [14]. The researchers faced a problem related to linear perspective. A user of their system,

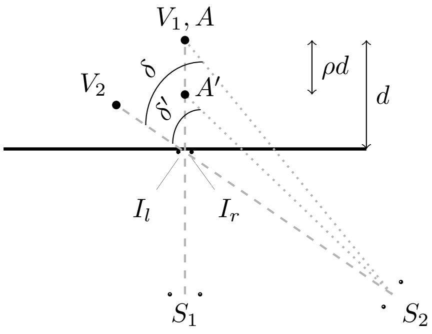

Fig. 1. Illustration of the method of reduction of the angular error between sound and image as a function of seating position. The spectators' eyes are symbolized by two dots. The spectator at S_2 watches the same point-like s-3D object (I_l and I_r) as the spectator at S_1, the ideal viewpoint, with an angular disparity in perceived position of δ with respect to V_1, the "correct" location of the s-3D object as perceived by S_1. The dashed lines S_1V_1 and S_2V_2 are the cyclopean lines of sight. The intersection of S_1V_1 and the screen is termed I, so that the distance d is $\overline{V_1I}$. The compression of the audio depth ($\overline{A'I}$ instead of $\overline{AI}$) allows one to reduce the angular error between the sound and the image ($\delta' < \delta$).

not sitting at the ideal viewpoint, would experience a discrepancy between the sound of the voice and the image of the face of his/her interlocutor.

The paradox between the judgments of spatial layout and orientation of lines makes it difficult to compute the location of the viewer's visual percept when watching a 2D picture from an off-axis location. The researchers placed the sound sources at the exact positions specified by the true 3D layout. Participants to the experiment then graded the perceived discrepancy between the sound and the image according to the ITU five-point impairment scale defined as follows: (1) very annoying, (2) annoying, (3) slightly annoying, (4) perceptible, but not annoying, and (5) imperceptible. The experiment revealed that annoying effects did occur when viewers shifted away laterally from the ideal viewpoint. A shift in depth seemed less disconcerting.

As also suggested in [14], it is possible to reduce the angular discrepancy between the sound and the image by pulling the audio sources towards the screen along the line between the visual object and the ideal viewpoint. At the same time, the audio gain is adjusted to produce the same sound level as the original source at the ideal viewpoint.

Second, the proposed method is adapted to s-3D video. We consider again the geometry in Fig. 1. Given the positions of the visual object V_1, the ideal viewpoint S_1, and the screen, one can compute the positions of the two points I_l and I_r in the left and right images on the screen corresponding to the visual object [17]. The sound can be placed at a point A' anywhere along the cyclopean line of sight S_1V_1, say according to a real parameter ρ defined by

$$A' - V_1 = \rho(I - V_1). \qquad (1)$$

where I is the intersection of S_1V_1 and the screen. Therefore, $\rho = 0$ yields $A' = V_1$ and $\rho = 1$ yields $A' = I$.

For a spectator seated at S_2, the visual object appears at V_2, resulting in an angular error δ between the sound and the image if the sound is positioned at $A = V_1$. When the sound is pulled closer to the screen, say at A', the angular error decreases for the spectator at S_2, i.e. $\delta' < \delta$. Note that this remains true when the line S_1V_1 is not perpendicular to the screen, as will be the case in this experiment. Provided that the sound level at A' is adjusted to match the volume it would have produced from A, the audiovisual congruence should be maintained at S_1. It should however be noted that a single adjustment will not be correct for all seating positions, as the acoustic attenuation is a function of the distance squared, such that the error in level adjustment will be greater for seating positions closer to the screen than S_1. The limit of audiovisual integration found here is also a measure of the sweet spot for accurate audiovisual reproduction [4].

3.2. Objectives

In the present article, an experiment is conducted with naive spectators to prove the efficiency of the method introduced in Section 3.1 in reducing the angular error between sound and image in the case of Wave Field Synthesis (WFS) and stereoscopic 3D (s-3D) video. A virtual scene consisting of a character standing in an apartment is chosen to simulate a cinema context. Audiovisual rendering is provided via the SMART-I^2 platform (Spatial Multi-user Audio-visual Real-Time Interactive Interface) [26, 27, 25], described below, using passive s-3D video and acoustic WFS. This virtual reality system provides its users with stable auditory and visual cues in a large rendering area. The paradigm is a yes/no experiment with the method of constant stimuli. The objective is to verify that the compression of the audio space towards the screen reduces the perception of inconsistency between sound and image when viewing s-3D contents combined with spatially accurate sound. This work is part of a larger study. Elements of this work, including the experimental protocol and preliminary results analysis, have been published in [2].

4. METHOD

4.1. Experimental design

The general layout of the experiment is shown in Fig. 2. Three possible seating positions (S_1, S_2, S_3) were situated 2 m from the right panel of the SMART-I^2, which was the active projection screen. Seating position S_1 faced the middle of the panel, with S_2 and S_3 situated 0.6 m and 1.2 m to the right of S_1, respectively. A virtual character, the visual stimulus, was

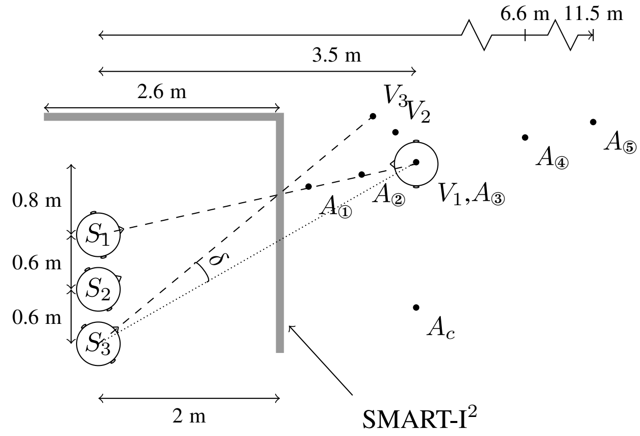

Fig. 2. Layout of the experimental setup with respect to the SMART-I^2 audio-visual panels (in gray). The S_i's are the positions of the subjects. V_i is the perceived position of the virtual character seen from S_i. The $A_{\textcircled{i}}$'s are the audio positions of the rendered speech. The angle δ illustrates the angular separation between the perceived location of the character and a position of the rendered speech.

rendered 1.5 m behind the screen, at 0.8 m to the left of S_1. A speech signal, the auditory stimulus, was rendered at five different positions along the line going through S_1 and the virtual character position, V_1. These positions are labelled $A_{\textcircled{1}}$ (closest to the screen) to $A_{\textcircled{5}}$ (farthest from the screen). $A_{\textcircled{3}}$ corresponds to the position of the virtual character (V_1), i.e. there is no audiovisual discrepancy for this sound position if the spectator is at S_1. In addition, a control position A_c is defined as the mirror image of $A_{\textcircled{3}}$ with respect to the perpendicular to the screen passing through S_1. The different subscripts used to denote the audio and visual object positions underline that these are independent.

A total of 17 subjects took part in the experiment (14 men, 3 women, age 19 to 30 years old, Mean = 23.5, Std. Dev. = 3.2). They all worked at the LIMSI. They were naive as to the experiment and they were not financially compensated. All but one participant had previously seen at least one s-3D movie in a cinema. Twelve participants played 3D video games (but not necessarily in s-3D) at most once a month. Only five participants used spatialized audio systems more than once a month, and three of them were the only ones to use virtual reality systems. The subjects can therefore be considered as being naive with respect to the combination of audio and video technologies used here.

The chosen experimental design was a within-subjects design with three factors: seat position (three levels), sound position (six levels), and repetition (four levels) (see Section 4.4). Together, the seat position and the sound position define the angular error `AVangle` (14 levels) between the sound and the image, in degrees [deg].

Table 1. Chosen values of ρ and the corresponding angles of error `AVangle` [deg] for each position S_i in the layout of Fig. 2.

	$A_{\textcircled{1}}$	$A_{\textcircled{2}}$	$A_{\textcircled{3}}$	$A_{\textcircled{4}}$	$A_{\textcircled{5}}$	A_c
ρ	0.79	0.40	0.01	-2.07	-5.32	control
S_1	0.0	0.0	0.0	0.0	0.0	26
S_2	1.9	4.3	6.0	10.2	12.2	31
S_3	2.9	6.9	9.9	17.4	21.2	34

The chosen values of ρ (where ρ is defined by Eq. 1) and the corresponding values of `AVangle` are given in Tab. 1. These values were chosen as a compromise between sampling the whole range of angles from complete congruence to complete incongruence, and ensuring that the SMART-I^2 was able to reproduce exactly the sound source at the chosen location. The values were verified in a pilot experiment with six subjects [2]. The values of `AVangle` corresponding to each value of ρ and each position S_i can be obtained from the geometry in Fig. 2. Given the coordinates S_1 and V_1 and the coordinates of the eyes of the viewer at S_1, the projections of V_1 in the left and right images can be obtained. Then, the coordinates of V_2 and V_3 can be computed [17]. It is assumed that each viewer is facing the direction of the point I, the midpoint between I_l and I_r.

4.2. Experimental setup

The present study was carried out using an existing system for virtual reality, called the SMART-I^2 [26, 27, 25], which combines s-3D video with spatial audio rendering based on WFS [7].

The SMART-I^2 system (Fig. 3) is a high-quality 3D audiovisual interactive rendering system developed at the LIMSI-CNRS in collaboration with *sonic emotion* [16]. The 3D audio and video technologies are brought together using two Large Multi-Actuator Panels (LaMAPs [8]), each of size 2.6 m × 2 m, forming a "corner", with the panels acting both as a pair of orthogonal projection screens and as a 24-channel loudspeaker array. The s-3D video is presented to the user using passive polarized technology, and 24 actuators attached to the back of each LaMAP allow for a WFS reproduction in a horizontal acoustic window corresponding to the s-3D video window. WFS is a sound field reproduction techniques that synthesizes the physical radiation properties of sound sources within an extended listening area [26]. The 20 cm spacing between the actuators corresponds to an aliasing frequency of about 1.5 kHz, the upper frequency limit for a spatially correct wavefront synthesis, accounting for the size of the loudspeaker array, and the extension of the listening area [9]. The implementation of WFS used here is restricted to the synthesis of sound sources located in the horizontal plane [10].

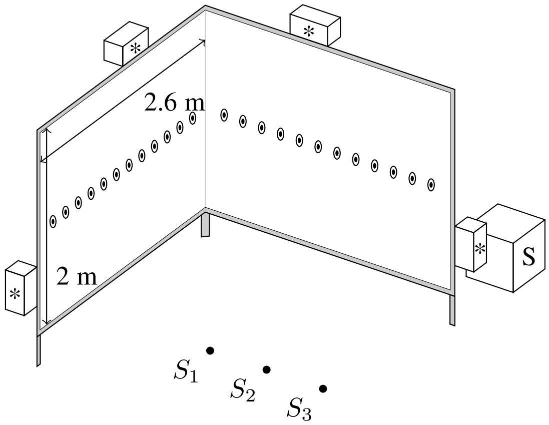

Fig. 3. Schematic view of the SMART-I^2. ⨀: WFS actuators, **S**: subwoofer, $*$: surround speakers. The WFS actuators and the screens are co-located in depth. Three dots on the ground plane indicate the positions S_1, S_2, and S_3.

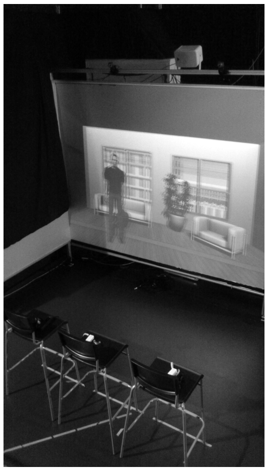

Fig. 4. Photo of the experimental setup showing the three bar stools and a projected s-3D image. The left screen is hidden behind a piece of dark cloth.

Azimuth and distance localization accuracies of sound events in the SMART-I^2 are globally consistent with corresponding real life localization accuracies. The azimuth localization accuracy of the WFS system in the SMART-I^2 was evaluated in [26]. Participants had to determine the origin of a 150 ms white noise burst coming from 17 possible virtual targets, each separated by 3°. The median of the angular error was always less than 3°, with a variability between 3 and 4°. These results are in line with the literature [31]. The distance perception in the SMART-I^2 was evaluated in [26, 25]. Participants estimated the distance to virtual sources in the auditory, visual, and auditory-visual modalities. Using two methods, visual target selection and blind-walking triangulation, results were in line with the literature on real auditory source distance perception [34]. The perceived distance d_p to the auditory targets was modeled by the curve $d_p = kd_s^a$ where d_s is the simulated distance, and k and a are parameters of the model. The median values of k and a were 1.72 ± 0.09 and 0.33 ± 0.03, respectively. This is also in line with the literature [34].

The software used to render the visual part of the experiment is MARC (Multimodal Affective and Reactive Characters [11]), a framework for real-time affective interaction with multiple characters [12]. The MARC architecture also provides a lip-synch functionality, for the characters, based on a pre-analysis of the speech wave file. The integration of MARC in the SMART-I^2 is described in [13].

4.3. Audiovisual material

The visual material consisted of one MARC character (Simon) in a scene depicting an apartment (Fig. 4). The point of view was chosen so that the character's mouth was at the height of the SMART-I^2's actuators, to avoid any vertical discrepancy. The scene was rendered at a 1:1 scale, i.e. life-size. The visual content was played continuously throughout the trial sessions.

The audio signal was the speech pronounced by the virtual character. There were two different five-second long sentences from two tales selected from a corpus [15]. For all sound positions, the level of the stimuli was adjusted at 52 dBA RMS at S_1. The ambient noise in the room was 33 dBA RMS at the seating positions.

4.4. Experimental task

In order to make efficient use of the installation and minimize total experiment time, up to three participants took part together in each experiment session. Each participant sat successively at the three positions S_1, S_2, and S_3 (not necessarily in this order). The participants were first provided with written instructions regarding the experiment. They wore passive linear polarizing s-3D glasses and received a Wiimote. There was no physical restriction on their head movement.

Each session consisted of three consecutive blocks to allow for each participant to sit at the three different positions. Each block consisted of 24 trials for data collection, corresponding to six sound positions, repeated four times. The first block started with a training session to make sure that the participants understood the task. This training alternated between two situations: sound at the correct position ($A_{③}$) and sound at the control position (A_c). The order of the stimuli was randomized in each block. Each value of the repetition factor was associated with one of two different speech sentences, alternating between the two. This was done to avoid monotony during the experiment. Each trial started with a five-second stimulus followed by a five-second period during which subjects answered the question "Is the voice coherent

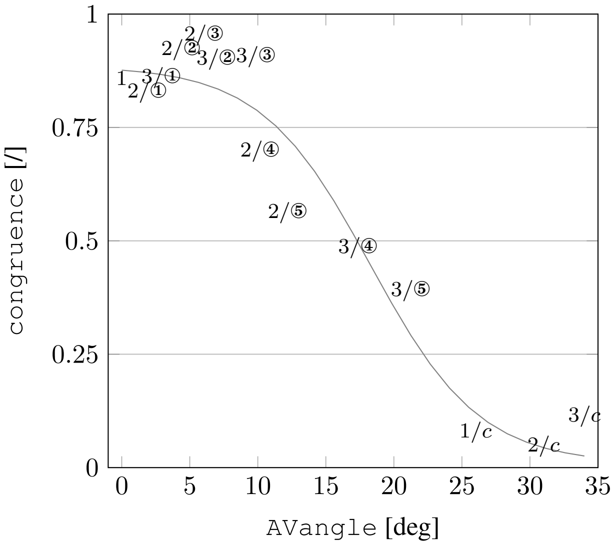

Fig. 5. Mean responses over all participants and corresponding fit, discarding the outliers. The data points are marked as "seat position/sound position".

with the character's position?" by pressing a button of the Wimote. The number of repetitions was chosen to keep the experiment short (about 8 minutes per block, and 30 minutes in total). The stimuli in each block were played in an automated way, with the subjects being observed remotely. Five-minute rest periods were offered between each block.

5. RESULTS

We present the results of the statistical analysis carried out on the answers from the participants. The answers yes and no are coded as 1 and 0, respectively. All statistics are reported at the 0.05 significance level.

We started by analyzing the panel performance and searching for potential outliers in the data. One participant's results obtained a `congruence` score almost constant irrespective of the angle of error `AVangle`. This behavior was clearly different from that of the other participants. A numerical analysis, based on the Median Absolute Deviation [23], confirmed this result. Therefore, the data from this participant was excluded from the subsequent analysis.

After removing the outlier from the data, the proportion of "yes" answers (the `congruence` score), was computed for each value of `AVangle` over all participants, as shown in Fig. 5. To understand how the mean score values relate to the positions of the sound and participants, the number of yes and no answers are reported in Tab. 2, summed over all participants.

We can see that, for participants at S_3 and the sound at $A_{\textcircled{4}}$ (equivalent to an angle of error of 17.4°), the numbers of answers yes and no are approximately equal. This point, where participants are equally likely to answer yes or no, is called the point of subjective equivalence (PSE).

Table 2. Counts of answers yes/no summed over all participants, discarding the outliers, for each sound position ($A_{\textcircled{1}}$ to $A_{\textcircled{5}}$ and A_c) and each participant position (S_1 to S_3). Due to minor technical glitches, 17 values were not recorded (out of 1152 trials).

	$A_{\textcircled{1}}$	$A_{\textcircled{2}}$	$A_{\textcircled{3}}$	$A_{\textcircled{4}}$	$A_{\textcircled{5}}$	A_c
S_1	56/8	56/8	54/10	59/4	49/15	5/59
S_2	53/11	58/5	61/3	44/19	36/27	3/59
S_3	53/9	54/6	58/6	31/32	25/37	7/55

Table 3. Results of the χ^2 test comparing the data collected at each participant position for each sound position (first line), and the corresponding p-values (second line).

	$A_{\textcircled{1}}$	$A_{\textcircled{2}}$	$A_{\textcircled{3}}$	$A_{\textcircled{4}}$	$A_{\textcircled{5}}$	A_c
χ^2	0.56	0.73	4.3	30.2	17.0	140.8
p	0.75	0.70	0.12	$< 10^{-6}$	0.0002	$< 10^{-6}$

In order to determine the values of `AVangle` at which the perception of the `congruence` is statistically different from that at $\texttt{AVangle} = 0$, a χ^2 test was performed for each value of ρ (see Appendix A for information on the statistical procedure). At each sound position, except at the control position A_c, one sample corresponds to S_1 and serves as a reference for the congruence ($\texttt{AVangle} = 0$). Since there is no reference sample at A_c (all the samples are incongruent by design), we include in this particular χ^2 test the values obtained at S_1 with the sound source at $A_{\textcircled{3}}$. Therefore, each χ^2 test was performed on three populations (df $= 2$), corresponding to three different values of `AVangle`, except at the control position, where four populations were compared (df $= 3$).

The results of the tests are given in Tab. 3. The table shows that the value of the χ^2 statistic increases almost monotonically with the distance from S_1 to the sound position (decreasing value of ρ). The lower value of the χ^2 statistic when the sound is located at $A_{\textcircled{5}}$ is a result of the lower proportion of yes answers at S_1 (the reference sample) for this sound position. When the sound is placed too far away, the assumption that only adjusting the sound level is enough to maintain the congruence at S_1 ceases to be valid.

At the three closest sound positions ($\texttt{AVangle} < 10°$), the proportions obtained in each test are statistically identical, irrespective of `AVangle`. An overall estimate $\bar{p}$ can be computed for each group of tested samples by collapsing all the corresponding counts in Tab. 2. The resulting mean proportion $\bar{p}$ is 0.85, when the sound is at $A_{\textcircled{1}}$, and 0.90 when the sound is at $A_{\textcircled{2}}$ and $A_{\textcircled{3}}$.

At $A_{\textcircled{4}}$, $A_{\textcircled{5}}$, and the control position A_c, the χ^2 test reaches significance at the 0.05 level, and therefore at least one proportion is different from the others. The Marascuilo procedure is applied to compare all pairs of proportions [6].

At each sound position, all comparisons between the proportion at S_1 and the other seats are significant, except for the comparison between the proportions at S_1 and S_2 when the sound is at $A_{⑤}$. Note, however, that the significance is obtained if the sample at S_1 is replaced by another reference sample, indicating once again that this result is obtained because of the lower count of yes at S_1 when the sound is at $A_{⑤}$.

6. DISCUSSION

An increasing angle of error `AVangle` between the sound position and the perceived character position decreased the reported `congruence`, i.e. the proportion of "yes" answers to the judgement of spatial congruence between the sound and the image of the character.

When the angle of error between the sound and the image was superior to 10°, the congruence statistically significantly decreased. The reported `congruence` continued to decrease with increasing angular discrepancy.

When the angle of error was less than 10°, the reported feeling of congruence was statistically independent of the angle of error, and the congruence score was maximal, between 0.85 and 0.9. These values of the angle of error (below 10°) also correspond to the cases where the sound was located nearest to the screen. This indicates that the method consisting in pulling the audio sources close to the screen with respect to an "ideal" viewer helps to improve the audiovisual congruence when accurate spatial sound is used in combination with s-3D images. For memory, all experiments were carried out for sound sources located in the horizontal plane.

7. CONCLUSION

We presented here a study of the auditory-visual spatial integration of 3D multimedia content by naive subjects. The audiovisual rendering was provided by a combination of passive stereoscopic 3D (s-3D) imaging and acoustic Wave Field Synthesis (WFS).

A subjective experiment was carried out where a situation corresponding to an angle of error between an s-3D video and a spatially accurate sound reproduced through WFS was presented to naive subjects. Motivated by a cinema application, we chose a stimulus consisting of a talking character in an apartment scene. After a five-second speech stimulus, subjects gave their answer to the question "Is the voice coherent with the character's position?"

A method, originally proposed for reducing the error between 2D images and spatially accurate sound reproduction, was adapted to s-3D viewing. When the sound was brought closer to the screen plane with respect to an ideal seating position, we found that the auditory-visual spatial integration was maintained to a high level for all tested seating positions. Conversely, when the sound was pulled away from the screen, we found that the integration was progressively reduced, an effect which was more pronounced at more off-axis seating positions.

8. REFERENCES

[1] D. Alais and D. Burr. The ventriloquist effect results from near-optimal bimodal integration. *Curr. Biol.*, 14(3):257–262, February 2004.

[2] C. R. André, É. Corteel, J.-J. Embrechts, J. G. Verly, and B. F.G. Katz. Subjective evaluation of the audiovisual spatial congruence in the case of stereoscopic-3D video and Wave Field Synthesis. *Int. J. Hum.-Comput. St.*, 72(1):23–32, January 2014.

[3] C. R. André, J.-J. Embrechts, and J. G. Verly. Adding 3D sound to 3D cinema: Identification and evaluation of different reproduction techniques. In *Proc. 2nd Int. Conf. on Audio Language and Image Processing (ICALIP 2010)*, pages 130–137, 2010.

[4] C. R. André, M. Rébillat, J.-J. Embrechts, J. G. Verly, and B. F. G. Katz. Sound for 3D cinema and the sense of presence. In *Proc. of the 18th Int. Conf. on Auditory Display (ICAD 2012)*, pages 14–21, Atlanta, GA, 2012.

[5] M. S. Banks, R. T. Held, and A. R. Girshick. Perception of 3-D layout in stereo displays. *Inform. Display*, 25(1):12–16, January 2009.

[6] M. L. Berenson, D. Levine, and T. C. Krehbiel. *Basic Business Statistics*. Prentice Hall, 12th edition, 2012.

[7] A. J. Berkhout, D. de Vries, and P. Vogel. Acoustic control by Wave Field Synthesis. *J. Acoust. Soc. Am.*, 93(5):2764–2778, 1993.

[8] M. M. Boone. Multi-Actuator Panels (MAPs) as loudspeaker arrays for Wave Field Synthesis. *J. Audio Eng. Soc.*, 52(7/8):712–723, 2004.

[9] É. Corteel. On the use of irregularly spaced loudspeaker arrays for Wave Field Synthesis, potential impact on spatial aliasing frequency. In *Proc. 9th Int. Conf. on Digital Audio Effects (DAFx'06)*, Montréal, Canada, 2006.

[10] É. Corteel, L. Rohr, X. Falourd, K.-V. NGuyen, and H. Lissek. Practical 3-dimensional sound reproduction using Wave Field Synthesis, theory and perceptual validation. In *Proc. of the 11th French Congr. of Acoustics and 2012 Annu. IOA Meeting*, pages 895–900, Nantes, France, 2012.

[11] M. Courgeon. Multimodal affective and reactive characters. marc.limsi.fr, October 2013. (Last accessed: 2013/11/18).

[12] M. Courgeon and C. Clavel. MARC: a framework that features emotion models for facial animation during human–computer interaction. *J. Multimodal User Interfaces*, pages 1–9, 2013.

[13] M. Courgeon, M. Rébillat, B. F.G. Katz, C. Clavel, and J.-C. Martin. Life-sized audiovisual spatial social scenes with multiple characters: MARC & SMART-I^2. In *Proc. of the 5èmes Journées de l'AFRV*, Orsay, France, December 2010.

[14] W. P. J. de Bruijn and M. M. Boone. Application of Wave Field Synthesis in life-size videoconferencing. In *Audio Eng. Soc. Conv. 114*, 2003.

[15] D. Doukhan, A. Rilliard, S. Rosset, M. Adda-Decker, and C. d'Alessandro. Prosodic analysis of a corpus of tales. In *INTERSPEECH–2011*, pages 3129–3132, 2011.

[16] Sonic Emotion. Sonic emotion absolute 3D sound. www.sonicemotion.com, October 2013. (Last accessed: 2013/11/18).

[17] M. Évrard, C. R. André, J. G. Verly, J.-J. Embrechts, and B. F. G. Katz. Object-based sound re-mix for spatially coherent audio rendering of an existing stereoscopic-3D animation movie. In *Audio Eng. Soc. Conv. 131*, 2011.

[18] E. B. Goldstein. Spatial layout, orientation relative to the observer, and perceived projection in pictures viewed at an angle. *J. Exp. Psychol. Human*, 13(2):256–266, May 1987.

[19] E. B. Goldstein. Pictorial perception and art. In *Blackwell Handbook of Sensation and Perception*. Blackwell Publishing, 2005.

[20] C. V. Jackson. Visual factors in auditory localization. *Q. J. Exp. Psychol.*, 5(2):52–65, 1953.

[21] J. Lewald, W. H. Ehrenstein, and R. Guski. Spatio-temporal constraints for auditory-visual integration. *Behav. Brain Res.*, 121(1-2):69–79, June 2001.

[22] J. Lewald and R. Guski. Cross-modal perceptual integration of spatially and temporally disparate auditory and visual stimuli. *Cognitive Brain Res.*, 16(3):468–478, May 2003.

[23] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *J. Exp. Soc. Psychol.*, 49(4):764–766, July 2013.

[24] L. A. Marascuilo. Large-sample multiple comparisons. *Psychol. Bull.*, 65(5):280–290, 1966.

[25] M. Rébillat, X. Boutillon, É. Corteel, and B. F. G. Katz. Audio, visual, and audio-visual egocentric distance perception by moving subjects in virtual environments. *ACM Trans. Appl. Percept.*, 9(4):19:1–19:17, October 2012.

[26] M. Rébillat, É. Corteel, and B. F. G. Katz. SMART-I^2: Spatial Multi-User Audio-Visual Real Time Interactive Interface. In *Audio Eng. Soc. Conv. 125*, 2008.

[27] M. Rébillat, B. F. G. Katz, and É. Corteel. SMART-I^2: "Spatial multi-user audio-visual real-time interactive interface", A broadcast application context. In *Proc. of the 3DTV Conf.*, Potsdam, Germany, 2009.

[28] D. A. Slutsky and G. H. Recanzone. Temporal and spatial dependency of the ventriloquism effect. *Neuroreport*, 12(1):7–10, January 2001.

[29] W. R. Thurlow and C. E. Jack. Certain determinants of the "ventriloquism effect". *Percept. Motor Skill*, 36:1171–1184, 1973.

[30] V. van Wassenhove, K. W. Grant, and D. Poeppel. Temporal window of integration in auditory-visual speech perception. *Neuropsychologia*, 45(3):598–607, January 2007.

[31] E. N. G. Verheijen. *Sound reproduction by Wave Field Synthesis*. PhD thesis, TU Delft, 1998.

[32] R. B. Welch. Meaning, attention, and the "unity assumption" in the intersensory bias of spatial and temporal perceptions. In *Cognitive Contributions to the Perception of Spatial and Temporal Events*, pages 371–387. Elsevier Science, 1999.

[33] A. J. Woods. Image distortions in stereoscopic video systems. In *Proc. of SPIE 1915*, pages 36–48, San Jose, CA, 1993.

[34] P. Zahorik, D. S. Brungart, and A. W. Bronkhorst. Auditory distance perception in humans: A summary of past and present research. *Acta Acust. united with Acust.*, 91:409–420(12), May 2005.

A. THE CHI–SQUARED TEST FOR DIFFERENCES BETWEEN PROPORTIONS

We describe here the χ^2 test for homogeneity [6] for our dataset. The test compares c (two or more) different groups (corresponding to `AVangle`) on a binary outcome (yes or no).

The statistic to be computed can be found in [6]. In the statistical software R, it is computed with the function `chisq.test`. This statistic approximately follows a χ^2 distribution with $c-1$ degrees of freedom.

When the statistic is not significant, there is no statistical difference between the proportions of answers yes or no at each tested value of `AVangle`. This is the null hypothesis H_0: the proportions at each value of `AVangle` are the same. The alternative, H_1, is that the proportions at each value of `AVangle` are different.

Under H_0, the proportions in each group vary only by chance and can be collapsed into one global proportion $\bar{p}$ by summing all counts in the corresponding groups.

Under H_1, the Marascuilo procedure is used to determine which pairs of groups have statistically differing proportions [24].