Audio Engineering Society

Convention Paper 7519

Presented at the 125th Convention
2008 October 2–5 San Francisco, CA, USA

Absolute Threshold of Coherence of Position Perception between Auditory and Visual Sources for Dialog

Roberto Muñoz Soto[1], Manuel Recuero López[2], Diego Durán Blanc[1], and Manuel Gazzo[1,]

[1] Universidad Tecnológica de Chile INACAP, Santiago, Chile.
rmunozs@inacap.cl

[2] Grupo de Investigación en Instrumentación y Acústica Aplicada (I2A2), Universidad Politécnica de Madrid, España.
manuel.recuero@upm.es

ABSTRACT

Under certain conditions, auditory and visual information are integrated into a single unified perception, even when they originate from different locations in space. The main motivation for this study was to find the absolute perception threshold of position coherence between sound and image, when moving the image across the screen, and when panning the sound.

In this manner, it is possible to subjectively quantify, by means of the constant stimulus psychophysical method [1], the maximum difference of position between sound and image considered coherent by a viewer of audiovisual productions. This paper discusses the accuracy necessary to match the position of the sound and its image on the screen.

The results of this study could be used to develop sound mixing criteria for audiovisual productions.

1. INTRODUCTION

According to the Audiovisual Language, the person who receives the message is conditioned by its own characteristics, properties and essence. On the other hand, the person who sends the message must be concerned about making the message understandable.

The Audiovisual Language perceptive cycle begins its process with the following biological characteristics: the stimulus is perceived through the visual and aural physiological mechanisms, which determine the sensate interpretations of the diverse acoustic and luminous variations of the media presented. All this followed by a conditioned recognition, which is stored according to

biological and cultural characteristics of the subjects (memory); finally, there is a response from the person who receives the message. It is important to emphasize that it is not only a single stimulus, but a systematic group of stimuli which is organized by the subject according to its contextual situation. Therefore, the correlation between auditory and visual stimulus is critical in order to produce an enhancement of the subject's audiovisual experience in relation to the perceived reality of what is presented.

The evidence for the interdependence between the auditory and visual senses shows that this perceptual synergy depends on the coincidence or degree of coherence between visual and auditory information presented to the subjects.

The visual and auditory perceptions do not work as isolated processes; both modalities cooperate in the improvement of people's ability and efficiency in perceiving their surrounding environment. When the auditory information is supported by coherent visual information, or when the visual information is reinforced by a coherent auditory reference, the synergistic interaction between these two modalities reinforces stimulus comprehension [2].

The visual bias of auditory localization is generally known as the "ventriloquism effect" [3][4]. With temporally coincident presentation of auditory and visual stimuli, these stimuli can be integrated into one unified perception, even when they are spatially disparate.

There is a spatial-temporal window for auditory-visual integration of 100ms and 3º; i.e., when auditory and visual stimuli are within this window, they are always perceived as spatially coincident [5].

With sources distributed on the horizontal plane, the mean minimum audible angle (MAA) threshold is about 0.97º [6].

The possibilities of panoramic positioning offered for film sound mixing systems are very extensive. The sound mixer is able to locate any element of the soundtrack in any position of the horizontal plane. The technical limitations of a multichannel audio system which need to be considered are the precision level that is capable of recreating a satisfactory image in order to match the sound with the associated picture, and, with respect to aesthetic criteria, whether or not it would be advisable to move the sound source to different positions.

It is well known that dialog in audiovisual productions comes mainly from the front, because the image of the characters is on the screen most of the time. However, they are not always in the center of the screen, but constantly moving horizontally along the screen.

The question thus proposed is: What is biggest difference in position between auditory and visual sources which is not considered incoherent, in such a way that allows for adjustments which follow the movement of the characters, yet which doesn't annoy viewers or interfere with the "Suspension of disbelief" concept [7].

When stereo sound was first released there was a lot of experimentation with panoramic positioning. That new technology created restrictions at the moment of filming a take, since many takes are generally carried out for each scene. This led, after the editing process, to instantaneous changes in the position of characters on the screen. The "jump" of the sound from one location of the screen to another was considered distracting. Subsequently, the "all dialogue in the center" criterion was again applied. Nevertheless, there would still be scenes in which it would be possible to match the position of the sound to the image. Moreover, in some scenes, it is necessary to maintain the coherence between what is seen and what is heard. For this reason, it is important to know the greatest possible difference between the position of the image and its sound that doesn't induce a sensation of incoherence in the viewers.

Currently, audiovisual productions are distributed in standardized formats based mostly on ITU-Recomendation-775. The 5.1 channel system has been recommended as the standard for multichannel and stereophonic sound systems, both with and without accompanying pictures [8]. In addition, the cinema electro-acoustic system must be calibrated and equalized following a developed standard. That is why the following experiment was carried out in a standard dubbing stage, in order to reproduce theatrical presentation systems.

2. METHOD

2.1. Subjects

Participants were 10 undergraduate students of different degrees at Universidad Tecnológica de Chile INACAP (7 male and 3 female, mean age of 23 years). All subjects reported having normal hearing and normal or corrected to normal vision.

2.2. Procedures

In this study all the audiovisual sequences were recorded in digital format. The video sequences were created at a resolution of 720 x 480 pixels. The audio of the sequences was edited using Protools HD3 Audio System and Final Cut software installed on a G4 Macintosh computer.

The visual stimuli were presented on an acoustically transparent screen of 2.4 m X 4.27 m, positioned at 5.6 m from the mixing position. The image size was 4.18 m X 2.35 m with a 16:9 aspect ratio.

Auditory stimuli were presented over a speaker arrangement which follows the ITU-R BS. 775 recommendation [8], where the L C R speakers are positioned behind an acoustically transparent screen. The audio system was calibrated at 85 dBC (slow rate) at -20dBFS Pink noise signal, measured at the standard cinema listening position, and equalized following the "X" curve [9][10]. The experiment was carried out in a small cinema dubbing stage with dimensions of 8.26 m X 5.65 m X 3.45 m.

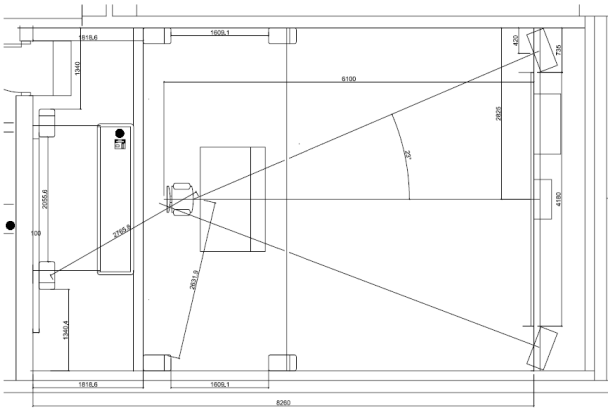

Fig. 1: Universidad Tecnológica de Chile INACAP's dubbing stage layout.

The text used was chosen taking into account the following criteria:

1. Credibility and sense of the communicative text: it is not desirable to include nonsensical words, sounds, or sentences, in order to avoid the effects of subject's comprehension and/or incomprehension on the final results of the research [8].

2. Length of the text: the sentence must be short to avoid subject fatigue.

3. Content: the text must be neutral so that it does not introduce any semantic bias to the subjects. The chosen text was "fue un ajuste de cuentas" ("it was an account adjustment"), which is a phonetically balanced sentence in Spanish.

2.3. Absolute Threshold of Coherence Position Perception

The method of Constant Stimuli was chosen in order to obtain the Absolute Threshold. The judgments were summarized in a table, where these values represent the percentage of the times that each comparison was judged "Coherent".
In order to get the psychometric function, the percentages of "Incoherent" judgments for each adjustment were represented on a graph (see figures 3 and 4).
Two experiments were carried out, using Long Shot (figure 2), one of them moving the image and the other panning the sound (phantom image).

Fig. 2. Long Shot

The sequence structures are the following:

The experiments consisted of 234 randomized trials, 9 for each of the 13 position adjustments, in steps of 2º.
The audiovisual test begins with 3s of a letter from "A" to "I" followed by a sound beep of 1 frame with blank video, then 14 black video frames. All these are followed by one and a half seconds of a number from "1" to "13" and a half second of black video. After that the sentence was presented with duration of 2 seconds. Four seconds were given to respond. Therefore, each of the 9 tests has a total duration of 2 minutes and 6 seconds. Each experiment took approximately 19 min to complete, including rehearsals and explanations.
Experiments were with one subject at a time, and each subject sat at the closest position recommended from the screen. The distance from the screen was 3.7 m.

3. RESULTS

For all experiments, the proportions of "coherent" responses were summarized in a table (see Tables 1 and 2), and its values represent the percentage of the times that each of the 13 positions adjustments (conditions) was judged as "coherent".

The observed distribution of responses was fitted to a curve of a mathematical function given in Equation 1, finding the least total error and the maximum R^2. The percentages of "coherent" judgments for each condition were represented on a graph (figure 3 and 4).

The observed distribution was compared with a normal distribution for each participant using the Kolmorogov-Smirnov goodness-of-fit-test. All observations could reasonably have come from the specified distribution ($p>.05$) for each experiment.

The data of nine tests for each experiment were analyzed by means of a one–way ANOVA. The results of test "A", for both experiments, were not included in the analysis. The result of the eight others test (B-I) were included because of the result of the one-way ANOVA revealed that the result of each one of them was not statistically different; when moving the image (F (7, 96) = 0.34, p =0.93), and when moving the sound (F (7, 96) = 0.052, p =0.999).
The Table 1 shows the results of the experiment when moving the image across the screen, and Table 2 shows the results of the experiment when moving the sound (phantom image). A t test revealed that both results were not significantly different ($t(13) = 1.713$, p = 0.112).

Image Position (°)	Probability "Incoherent" (%)	Numbers of Trials	Subjects Responses "Incoherent"
2	2.5	80	2
4	5.0	80	4
6	8.8	80	7
8	15.0	80	12
10	25.0	80	20
12	27.5	80	30
14	41.3	80	33
16	52.5	80	42
18	63.8	80	51
20	66.3	80	53
22	71.3	80	57
24	85.0	80	68
26	39.8	80	75

Table 1: Results of experiment when moving the image across the screen.

Image Position (°)	Probability "Incoherent" (%)	Numbers of Trials	Subjects Responses "Incoherent"
2	8.8	80	7
4	5.0	80	4
6	8.8	80	7
8	7.5	80	6
10	10.0	80	8
12	5.0	80	4
14	11.3	80	9
16	20.0	80	16
18	37.5	80	30
20	75.0	80	60
22	80.0	80	64
24	96.3	80	75
26	97.5	80	78

Table 2: Results of experiment when moving the sound across the screen.

To calculate the 50% threshold of preferences, the interpolation of values of the psychometric response was required. To achieve this, a mathematical model given in Equation 1 was used.

$$P_{(\alpha)} = \frac{P_{max}}{1+\left(\frac{P_{max}}{P_{min}}-1\right)e^{-KP_{max}\alpha}} \quad (1)$$

Eq. (2) was used to fit the curve to the observed data (Figures 3 and 4).

$$\alpha = \frac{1}{-KP_{max}}\left\{\ln\left(\frac{P_{max}}{P(\alpha)}-1\right)-\ln\left(\frac{P_{max}}{P_{min}}-1\right)\right\} \quad (2)$$

Using equation 1 the function with the best fit to the data in the table 1 is given by K=0.00256 which gives a R2 = 0.983 and a Total error of 156.7 %.
Calculating the degree for the 50% of "Incoherence" responses, an angle α = 15.55° was obtained.

The measured results and their mathematical approximation for both experiments can be observed in figures 3 and 4. The value obtained for the experiment when moving the sound is an angle α= 17.75°.

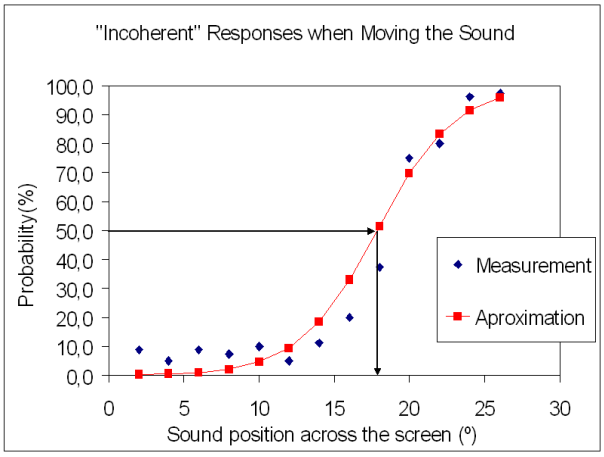

Fig. 3: Graph of measured values graph and their mathematical approximation for phantom image.

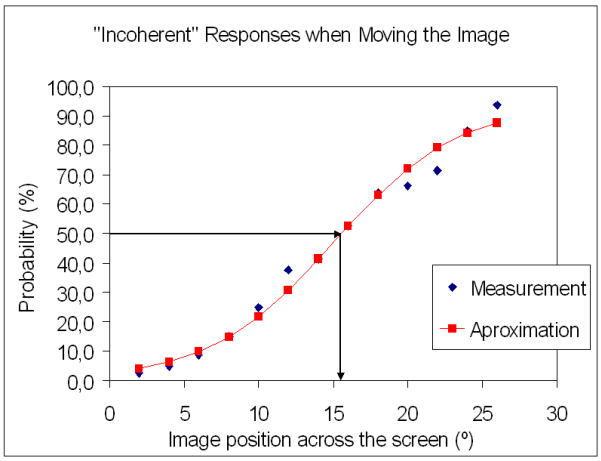

Fig. 4: Graph of measured values and their mathematical approximation when moving the image.

4. CONCLUSIONS

The results show that for angles between the sound and image of less than 15° (assuming the sound is in the center), there is no perception of positional incoherence.

The experiment was conducted with subjects seated in the first row of the theater. From this position, 15° of separation between the sound (in the center) and the image (to the right) corresponds to half the distance between the center and the far right side of the screen. For the mix engineer seated at the console, this location on the screen represents an angle of approx. 11°, and even less for people seated in the last row.

If the screen is divided in four equal parts (4 rectangles, one next to each other horizontally), it will only be necessary to make a panoramic sound adjustment when the visual source moves to or is located on the farthest left or farthest right area. And it will be necessary to move it in the same proportion to the visual source movement as it overpasses the two middle parts, in order for the sound not to be perceived as "incoherent" to the image.

On the other hand, the intelligibility of the sentences decreases when other sounds originate from the same position as the sentence [12]. Hence, it is recommended to keep the dialog in the center channel in a 5.1 system, and other sounds, such as discrete effects, close to 11° off the center, since they will be perceived as if they

come from the center. This will allow for better dialog comprehension.

5. ACKNOWLEDGMENTS

This work was developed thanks to the support of the Universidad Tecnológica de Chile INACAP and its Pérez Rosales Campus Vice-rector; Mrs. Karin Riedemann. The authors also thank Mark Berger, Adjunct Professor of Film at U.C. Berkeley and 4 time Oscar winner for best sound, for his help, useful advice and reviewing, as well to Victor Rojas and Jaime Delannoy for their efficient technical support.

6. REFERENCES

[1] S. Gelfand: "Hearing, an introduction to psychological and physiological acoustics". 3rd Ed. (Marcel Dekker Inc, New York, USA, 1990).
[2] W. Woszczyk, S. Bech, V. Hansen. "Interactions between audio-visual factors in a Home Theater System: Definition of Subjetives Attributes". Paper presented at the 99th AES Convention (New York, USA, 1995).
[3] Radeau M. (1994): Auditory-visual interaction and modularity. Curr. Psychol. Cong. 13: 3-51.
[4] R. Welch, D. Warren (1990): Immediate perceptual response to intersensory discrepancy. Psychol. Bull. 88:638-667.
[5] J. Lewald, W. Ehrenstein, R. Guski Lewald J., (2001): Spatio-Temporal constraints for auditory-visual integration. Behav. Brain. Res: 121, 69-79
[6] D. Perrot, K. Saberi (1990): Minimum audible angle threshold for sources varying both elevation and azimuth. Journal of the Acoustical Society of America 87:1728-1731.
[7] S. Taylor. "Biographia Literaria". H.J. Jackson, Ed. (Oxford, UK 1985).
[8] "Multichanel stereophonic sound system with and without accompanying picture". ITU-R BS. 775 (1994).
[9] T. Holman. "5.1 Surround Sound up and Running" (Focal Press, USA, 2000).
[10] "Cinematography – B - chain electro-acoustic response of motion-picture control rooms and indoor theatres -- Specifications and measurements". ISO 2969 (1987).
[11] T. Soto. "Influencia de la percepción visual del rostro del hablante en la credibilidad de su voz". Tesis doctoral departamento de comunicación audiovisual y publicidad (Universidad Andrés Bello, Chile, 2000).
[12] Bradley, John S. "From Speech Privacy to speech security". 19 International Congress On Acoustics. Madrid, España. 2-7 Septiembre 2007.