Impact of spatial audiovisual coherence on source unmasking

Julian Palacino, Mathieu Paquier, Vincent Koehl, Frédéric Changenet, Etienne Corteel

Buenos Aires – 5 to 9 September 2016
Acoustics for the 21st Century…

PROCEEDINGS of the 22nd International Congress on Acoustics

Free-Field Virtual Psychoacoustics and Hearing Impairment: Paper ICA2016-493

Impact of spatial audiovisual coherence on source unmasking

Julian Palacino(a), **Mathieu Paquier**(a), **Vincent Koehl**(a),**Frédéric Changenet**(b), **Etienne Corteel**(c)

(a)UBO - LabSTICC, France, julian.palacino@univ-brest.fr
(b)Radio France, France, Frederic.CHANGENET@radiofrance.com
(c)Sonic Emotion Labs, France, etienne.corteel@sonicemotion.com

Abstract

The influence of the spatial audiovisual coherence is evaluated in the context of a video recording of live music. In this context, audio engineers currently balance the audio spectrum to unmask each music instrument getting it intelligible inside the stereo mix. In contrast, sound engineers using spatial audio technologies have reported that sound source equalization is unnecessary in live music mixing when the sound sources are played at the same location of the physical instruments. The effects of spatial audiovisual coherence and sound spatialization have been assessed : expert subjects were asked to compare two mixes in audio only and in audiovisual mode. For this aim, music concerts were visually projected and audio rendered using WFS. Three sound engineers did the audio mixing for all pieces of music in the same room were the test have been carried out.

Keywords: audiovisual coherence, WFS

Impact of spatial audiovisual coherence on source unmasking

1 Introduction

Wave Field Synthesis (WFS) [1], based on acoustic field reconstruction, has been recently adopted for live music mixing. This technology allows to reproduce sound sources in different places for a wide sweet spot (in contrast to the stereophonic rendering). In conventional mixing (stereo), spectrum equalization is a classical technique to unmask concurrent sources. However, some sound engineers reported that this would be unnecessary for live music (as audio and visual information is available) when sound and image are spatially coherent for a given source. This is achievable with precision using WFS but not using stereo rendering. In addition, this technology ensures the parallax effect (distance of sources) providing coherent auditory impressions for any seat in the audience. It therefore ensures a consistent localization of sound sources on stage throughout the audience. This is a unique property of WFS that no other reproduction technique can offer within an extended listening area.

Our final goal is to determine if the interest of the WFS comes from the audiovisual coherence provided by sound sources spatialization (allowed by this system) or from the source spatialization itself (allowing spatial unmasking). For this purpose, spatialized mixes were done. One, sticking sound sources to the their visual position (denoted *Y-mix*) and another without audiovisual coherence (denoted *X-mix*). Different music styles were mixed by several sound engineers. As the general frame of this study is music mixing for live performances and in order to prevent all bias of a live concert (for example the repeatability of music performance), the experiment was done in a controlled environment using recorded audiovisual stimuli. The realism and immersion were increased using stereoscopic video as visual stimulus [2].

As a preliminary test confirmed that subjects were able to detect small differences between *X-mix* and *Y-mix* mixes [5], subjects were asked to evaluate two mixes between them in audio only mode and in audiovisual mode.

2 Experimental setup

The experiment took place in an acoustically treated room in the University of Brest (The background noise was < 30 dB(A) and $RT_{60} = 0.2$ s)

2.1 Experimental setup

30 Amadeus PMX 4 loudspeakers were placed on a supporting structure at the height of the ears of an average seated person (1.20 m). As illustrated on Figure 1, the distribution of loudspeakers was settled in order to increase the density in front of the listener. Frontal loudspeakers were behind an acoustically transparent projection screen. Low frequencies were rendered by a Genelec 7070A sub-woofer placed in the left corner of the room.

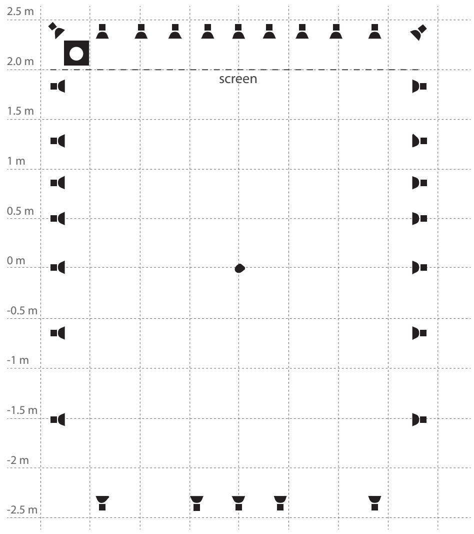


Figure 1: Loudspeaker configuration

Once the whole system installed, the frequency response of each loudspeaker was slightly corrected by equalization (± 2 dB max).

The WFS rendering was ensured by the Sonic Emotion Wave 1 processor [1] and the stereoscopic HD video was displayed using an Epson EHTW6000 projector (using active 3D shutter glasses).

2.2 Subjects

12 expert listeners with normal hearing took part in the experiment. The subjects were students at the Image and Sound Master's degree of the university of Brest. Their average age was 22.5 years old. All subjects were payed for their participation to the test.

2.3 Stimuli

Three concerts of different music styles were recorded (see Table 1). For mixing purposes, all microphone signals were recorded independently in a multi-track recorder. From each concert, a 30 seconds stimulus was extracted namely (*rock, jazz* and *classical*).

A wide static shot of each concert, as depicted in Figure 2 was obtained using a stereoscopic Panasonic AG-3DP15 camera in the aim of improving the immersion [2].

One should notice the importance of the sound engineer in the aesthetics of a mix (dynamics, spatial, spectral processing, etc.). In order to reduce the influence of the sound engineer,

[1] http://www2.sonicemotion.com/professional/

Table 1: **Excerpt description**

Excerpt	Piece	Instruments	Location
classical	baroque sonata	flute, oboe, cello and harpsichord	16c.Chuch
jazz	vocal Jazz	female voice, trumpet, sax, 2 keyboards, guitar and drums	Concert hall
Rock	Rock-funk	Male voice, choirs, guitar, bass guitar, trumpet, sax, keyboard, and drums	open-air concert

(a) Classic

(b) Jazz

(c) Rock

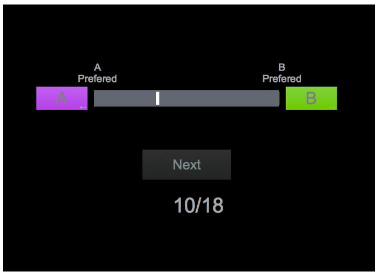
Figure 2: **Screen capture of video tracks of three excerpts.**

the mix has been done by three sound engineers. They mixed all pieces of music, for both modes (Audio only and Audiovisual). The sound engineers were asked to mix the whole song or movement. First, they were asked to do a spatial mix using only the audio material (none video material was available). The resulting mix is named *X-mix*. Once the first mix finished the video material was provided and they were asked to do an audiovisual coherent mix sticking the sound sources to their location on the screen named *Y-mix*.

All mixes were done in the test room and all stimuli were equalized in loudness (≈ 75 dB(A)).

3 Experimental protocol

This experiment is based on a pairwise protocol. In a trial, 2 *test* stimuli (*Y-mix* and *X-mix*) were proposed to the subject randomly. A trial for a given excerpt and a given sound engineer was repeated twice. In each trial, the subject had to listen to each stimulus (by freely switching between the 2 stimuli) and move a slider along a continuous scale to select his level of



(a) sessions 1 and 3

(b) sessions 2 and 4

Figure 3: **User interface screenshot**

preference. Once his choice was done, the subject was allowed to click on the *Next* button then a new randomly chosen trial was displayed. It should be noticed that no number or tick was displayed on the evaluation scale in order to prevent bias on the assessors choices. Only the labels A-preferred and B-preferred were displayed at the limits of the scale (figure 3). The evaluation values were stored in a continuous scale from -50 to 50. -50 means that the *X-mix* mix is preferred, 50 means that the *Y-mix* mix is preferred and 0 when no perceptual difference have been noticed between two mixes.

Each session was composed of 18 trials (2 repetitions $\times$ 3 excerpts $\times$ 3 sound engineers). The test was completed after four sessions, two for each presentation mode (audio only and audiovisual). In the first session of each mode (sessions 1 and 3), subjects were asked to select the level of general preference of each pair of stimuli (Figure 3a), then (sessions 2 and 4) subjects were asked to evaluate several attributes independently (*intelligibility, timbre, immersion, spatial precision* and *Realism*) (Figure 3b). The evaluated attributes were extracted from [3] and [4]. Subjects were asked to read a glossary containing the meaning of all attributes at the beginning of session 2 and 4 (table 2). In case of doubt subjects were able to consult the glossary at any moment during the test.

Subjects completed 2 sessions consecutively in each laboratory coming. It has been chosen to start the test with the audio only sessions (1 and 2) in order to ensure that subjects did not know the source positions. In the second coming (on another day) the audiovisual mode was tested (sessions 3 and 4). Session 1 and 3 lasted roughly 30 minutes and sessions 2 and 4 45 minutes. A 5 minutes break was imposed to the assessors between two consecutive sessions. The subjects sat in the center of the room where the mixes were made.

4 Results

Normality of distributions for the Audio only (A) and audiovisual (AV) presentation modes (for all excerpts and sound engineers) has been checked for *preference* and attributes with a

Table 2: **Glossary of attributes as given to assessors**

Intelligibility / Segregation :	Ability to separate each instrument within the mix.
Color/Timbre :	Sensation of timbral changes (richer/poorer) in high frequency, middle frequency or low frequency or sensation of a muffled or a metallic sound for an instrument or for the whole mix.
Immersion / Envelopment :	Impression of being inside a presented scene or to be spatially integrated into the scene.
Spatial Precision :	Sensation of being able to associate a precise position to each sound.
Credibility / Realism :	The sound seems to come from real sources around you.

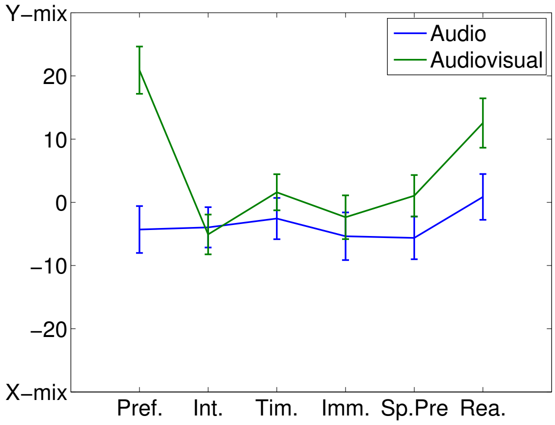


Figure 4: **Global mean and 95% confidence intervals for *preference* and attributes for each audio only and audiovisual session (For all 3 excerpts, 3 sound engineers, and subjects)**

Kolmogorov-Smirnov test. A *t*-test showed that *preference* ratings are significantly lower than 0 in A mode $(t(198) = -2.287; p < 0.05)$; *X-mix* was slightly preferred. On the other hand ratings are significantly higher than 0 in AV mode $(t(198) = 11.039; p < 0.05)$; *Y-mix* was largely preferred (figure 4). Results show that *X-mix* are preferred in A mode whereas in AV mode the *preference* of *Y-mix* mix is increased because it is coherent with the visual cues.

A *t*-test showed that *realism* results are not significantly different from 0 in the A mode. A *t*-test showed that *realism* results are significantly higher than 0 in AV mode, $(t(198) = 3.33; p < 0.05)$: in A mode *X-mix* and *Y-mix* have the same level of *realism*, however in AV mode the *Y-mix* is considered as more realistic. Furthermore a t-test showed that the impression of *realism* of the *Y-mix* was increased in the AV mode in contrast to the *X-mix* $(t(198) = 3.33; p < 0.05$; figure

4). *Realism* results show that the *realism* impression is the same for both mixes in A mode whereas in AV mode the audiovisual coherence of *Y-mix* mixes increases their *realism*.

About intelligibility, results are significantly lower than 0, but not significantly different in A mode $(t(198) = -2.448; p < 0.05)$ and in the AV $(t(198) = -3.186; p < 0.05)$ mode. *X-mix* has been considered as more intelligible whatever the presentation mode.

About *immersion*, results are significantly lower than 0 $(t(198) = -2.81; p < 0.05)$ in A mode only and a *t*-test do not indicate significant difference between the results obtained in the two modes.

About *spatial precision*, results are significantly lower than 0 $(t(198) = -3.28; p < 0.05)$ in A mode only and a *t*-test do not indicate significant difference between the results obtained in the two modes.

No significant correlation was observed between *preference* and any of the five different attributes or between two attributes themselves whatever the presentation method was.

5 Summary

This paper describes a test about the influence of the audiovisual coherence for live music mixing. This study aims to assessing the effect of the image on the perception of a 3D audio spatialized mix. Results indicate that *preference* and *realism* of a mix in audiovisual mode are principally increased by the spatial coherence of the mix to the position of the sound sources on the visual support. On the other hand, results do nor reveal any influence of the presentation mode on *immersion, timber, spatial precision* and *intelligibility*. However, further analysis as multidimensional scaling or Principal Components Analysis should be done to determine if there is any relation between preference and tested attributes.

This study confirms that 3D audio systems could increase the perceived quality of a mix by sticking sound sources to their position on the stage in live music mixing or in audiovisual music production.

Acknowledgements

This work has been supported by the ANR-13-CORD-0008 EDISON 3D project, funded by the French National Agency of Research (http://www.agence-nationale-recherche.fr/).

References

[1] D. de Vries. *Wave Field Synthesis*. AES monographs. Audio Engineering Society, 2009.

[2] W. IJsselsteijn, H. d. Ridder, J. Freeman, S. E. Avons, D. Bouwhuis, and others. Effects of stereoscopic presentation, image motion, and screen size on subjective and objective corroborative measures of presence. *Presence*, 10(3):298–311, 2001.

[3] A. Lindau, V. Erbes, S. Lepa, H.-J. Maempel, F. Brinkman, and S. Weinzierl. A Spatial Audio Quality Inventory (SAQI). *Acta Acustica united with Acustica*, 100(5):984–994, Sept. 2014.

[4] R. Nicol, L. Gros, C. Colomes, E. Roncière, and J.-C. Messonier. Etude comparative du rendu de différentes techniques de prise de son spatialisée après binauralisation. In *CFA/VISHNO 2016*, LE MANS, FRANCE, Apr. 2016. CFA.

[5] J. Palacino, M. Paquier, V. Koehl, F. Changenet, and É. Corteel. Assessment of the impact of spatial audiovisual coherence on source unmasking : preliminary listening test. Paris, June 2016. AES.