Impact of spatial audiovisual coherence on source unmasking

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

Proceedings of Meetings on Acoustics

Volume 28 http://acousticalsociety.org/

22nd International Congress on Acoustics
Acoustics for the 21st Century

Buenos Aires, Argentina
05-09 September 2016

Psychological and Physiological Acoustics: ICA2016 - 493

Impact of spatial audiovisual coherence on source unmasking

Julian Palacino, Mathieu Paquier and Vincent Koehl
Universite de Bretagne Occidentale, LabSTICC, Brest, France; julian.palacino@univ-brest.fr; mathieu.paquier@univ-brest.fr; vincent.Koehl@univ-brest.fr

Frédéric Changenet and Etienne Corteel
Radio France, Paris, France; Sonic Emotion Labs, Paris, France; Frederic.CHANGENET@radiofrance.com; etienne.corteel@sonicemotion.com

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 making 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 Wave Field Synthesis (WFS). Three sound engineers did the audio mixing for all pieces of music in the same room where the tests were carried out.

Published by the Acoustical Society of America

 [DOI: 10.1121/2.0000476]

1 Introduction

Wave Field Synthesis (WFS) [1], based on acoustic field reconstruction, has been recently adopted for live music mixing. This technology allows the reproduction of 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 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 in 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 subwoofer placed in the left corner of the room.

Once the whole system was 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 and the stereoscopic HD video was displayed using an Epson EHTW6000 projector (using active 3D shutter glasses).

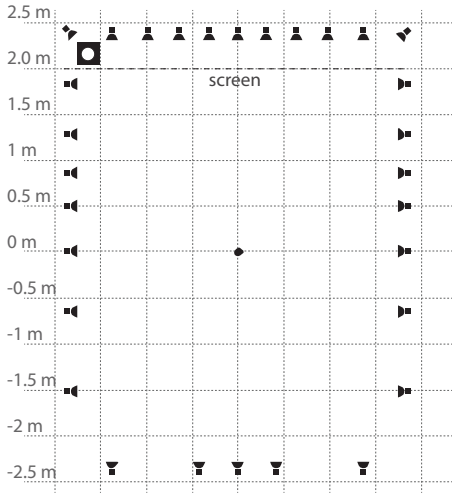


Figure 1: Loudspeaker configuration

2.2 Subjects

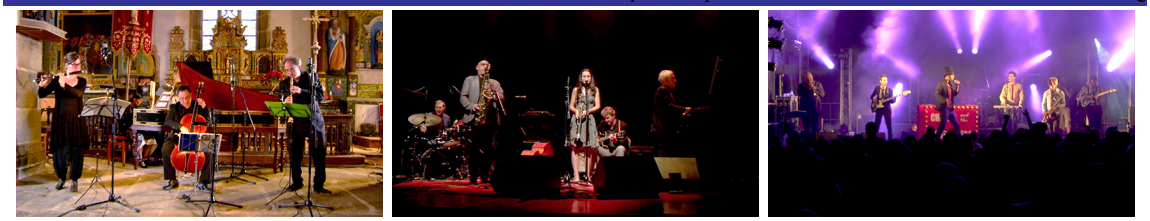
12 expert listeners with normal hearing took part in the experiment. The subjects were students at the Image and Sound Master's degree of the University of Brest. Their average age was 22.5 years. All subjects were paid for their participation in the test.

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*.

Table 1: **Excerpt description**

Excerpt	Piece	Instruments	Location
Classical	Baroque sonata	Flute, oboe, cello and harpsichord	16th c. Church
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.**

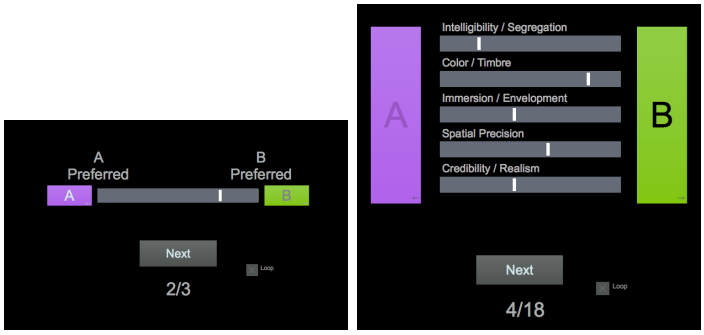
A wide static shot of each concert, as depicted in Figure 2, was obtained using a stereoscopic Panasonic AG-3DP15 camera with the aim of improving the immersion [2].

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, the mix was 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 (no 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 was based on a pairwise protocol. In a trial, 2 *test* stimuli (*Y-mix* and *X-mix*) were presented to the subject randomly. Each trial for a given excerpt and a given sound engineered stimulus were 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 preference. Following each selection, the subject was required to click on the *Next* button, then a new randomly chosen trial was displayed. It should be noted that no number or tick was displayed in the evaluation scale in order to prevent bias in the assessor's choices. Only the labels A-preferred and B-preferred



(a) sessions 1 and 3 (b) sessions 2 and 4

Figure 3: **User interface screenshot**

were displayed at the limits of the scale (Figure 3). The evaluation values were stored on a continuous scale from -50 to 50. -50 meant that the *X-mix* was preferred, 50 meant that the *Y-mix* was preferred and 0 meant no perceptual difference was noticed between two mixes.

Each session was composed of 18 trials (2 repetitions $\times$ 3 excerpts $\times$ 3 sound engineered stimuli). 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.

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.

Subjects completed 2 sessions consecutively in each laboratory visit. Testing began with the audio only sessions (1 and 2) in order to ensure that subjects did not know the source positions. In the second visit (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 lasted 45 minutes. Assessors were required to take a 5-minute break between two consecutive sessions. The subjects sat in the center of the room where the mixes were made.

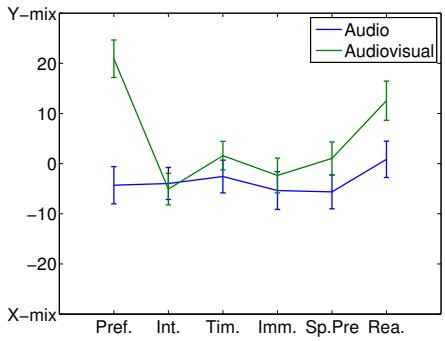


Figure 4: **Global mean and 95% confidence intervals for *preference* and attributes for each audio only and audiovisual session (For all 3 excerpts, 3 sound engineered stimuli, and subjects)**

4 Results

Normality of distributions for the audio only (A) and audiovisual (AV) presentation modes (for all excerpts and sound engineered stimuli) was examined for *preference* and attributes with a Kolmogorov-Smirnov test.

A *t*-test showed that *preference* ratings were significantly lower than 0 in A mode ($t(198) = -2.287; p < .05$); *X-mix* was slightly preferred. On the other hand, ratings were significantly higher than 0 in AV mode ($t(198) = 11.039; p < .05$); *Y-mix* was largely preferred (Figure 4). Results show that *X-mix* was preferred in A mode whereas in AV mode the *preference* of *Y-mix* was increased probably because *Y-mix* in AV mode was coherent with the visual cues.

A *t*-test showed that *realism* results were not significantly different from 0 in the A mode. A *t*-test showed that *realism* results were significantly higher than 0 in AV mode, ($t(198) = 3.33; p < .05$): in A mode *X-mix* and *Y-mix* had the same level of *realism*, however in AV mode the *Y-mix* was 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* ($t(198) = 3.33; p < .05$; Figure 4). *Realism* results showed that the *realism* impression was the same for both mixes in A mode whereas in AV mode the audiovisual coherence of *Y-mix* mixes increased their *realism*.

About intelligibility, results were significantly lower than 0, in A mode ($t(198) = -2.448; p < .05$) and in the AV ($t(198) = -3.186; p < .05$) mode and a *t*-test did not indicate significant difference between the results obtained in the two modes. *X-mix* has been considered as more intelligible whatever the presentation mode.

About *immersion*, results were significantly lower than 0 ($t(198) = -2.81; p < .05$) in A mode only and a *t*-test did not indicate significant difference between the results obtained in the two modes.

About *spatial precision*, results were significantly lower than 0 ($t(198) = -3.28; p < .05$) in A mode only and a *t*-test did 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 described a test about the influence of the audiovisual coherence for live music mixing. This study aimed to assess the effect of the image on the perception of a 3D audio spatialized mix. Results indicated that *preference* and *realism* of a mix in audiovisual mode were 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 did not reveal any influence of the presentation mode on *immersion, timber, spatial precision* and *intelligibility*. However, further analysis such as multidimensional scaling or Principal Components Analysis should be considered to determine if there is any relation between preference and tested attributes.

This study suggests 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. New York. Audio Engineering Society, 2009.

[2] W. IJsselsteijn, H. de Ridder, J. Freeman, S. E. Avons, and D. Bouwhuis. Effects of stereoscopic presentation, image motion, and screen size on subjective and objective corroborative measures of presence. *Presence: Teleoperators and virtual environments*, 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, 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, Société Française d'Acoustique*, Le Mans, France, 2016.

[5] J. Palacino, M. Paquier, V. Koehl, F. Changenet, and É. Corteel. Assessment of the impact of spatial audiovisual coherence on source unmasking: Preliminary discrimination test. *140th Audio Engineering Society International Convention*, 2016.