Audio Engineering Society

Convention Paper

Presented at the 144th Convention
2018 May 23–26, Milan, Italy

Virtual Reality for Subjective Assessment of Sound Quality in cars

Angelo Farina, Daniel Pinardi, Marco Binelli, Michele Ebri and Lorenzo Ebri

University of Parma, Dipartimento di Ingegneria e Architettura, Parco Area delle Scienze 181/A, 43124, Parma, ITALY

Correspondence should be addressed to Angelo Farina (farina@unipr.it)

ABSTRACT

Binaural recording and playback has been used for decades in automotive industry for performing subjective assessment of sound quality in cars, avoiding expensive and difficult tests on the road. Despite the success of this technology, several drawbacks are inherent in this approach. The playback on headphones does not benefit of head-tracking, so the localization is poor. The HRTFs embedded in the binaural rendering are those of the dummy head employed for recording the sound inside the car, and finally there is no visual feedback, so the listener gets a mismatch between visual and aural stimulations. The new Virtual Reality approach solves all these problems. The research focuses on obtaining a 360° panoramic video of the interior of vehicle, accompanied by audio processed in High Order Ambisonics format, ready for being rendered on a stereoscopic VR visor. It is also possible to superimpose onto the video a real-time colormap of noise levels, with iso-level curves and calibrated SPL values. Finally, both sound level colormap and spatial audio can be filtered by the coherence with one or multiple reference signals, making possible to listen and localize very precisely noise sources and excluding all the others. These results have been acquired employing a massive spherical microphone array, a 360° panoramic video recording system and accelerometers or microphones for the reference signals.

1 Introduction

The First applications of Acoustic Holography date back to the experiments for analysis of vibration patterns and radiation behavior of loudspeakers conducted by Hladky, Jan in 1974 [1] and Frankort, F. J. M. in 1978 [2], recently proposed with technology and signal processing advancements in 2017 by Pinardi, D. [3].

Theoretical basis of acoustic holography are well explained in [4], whilst Harris N. presented in 2004 a practical solution for an acoustic source located in a non-anechoic room, showing colormaps of SPL variations [5]. In the last ten years the interest for visualization of SPL distribution in an acoustic field has grown consistently, also thanks to the spreading of massive microphones' array, and several hardware and software solutions developed for sound sources location, nowadays well known as "acoustic camera": Delikaris-Manias S. in 2016 [6] and McCormack L. in 2017 [7].

In parallel, innovative VR reproduction systems came to the market, such as Samsung Gear VR, Microsoft HoloLens, Facebook Oculus Rift, HTC Vive and the new Samsung HMD Odyssey. All these took benefits from spatial audio rendering technique, most of all Ambisonics method [8] and binaural decoding [9].

Many potential applications come up from the combination of these technologies: from the identification and analysis of noise sources, for example in workplaces, to the complete immersive reproduction of a vehicle cockpit, such cars, tractors or planes.

The paper presents a recording/playback method (and its practical implementation) developed by mean of current state-of-the-art equipment and technology coming from the Virtual Reality video approach, providing a degree of realism during playback unattainable with traditional binaural recording technology.

2 Experimental measurements

The experimental measurements reported here were carried out in three different situation: laboratory of acoustic in the Department of Industrial Engineering at the University of Parma (Italy), LABEL – electroacoustic laboratory of Casa del Suono, Parma (Italy) and cockpit of vehicle Alfa Romeo Giulia, property of ASK Industries Group, Reggio Emilia (Italy).

2.1 Experimental setup

The recording equipment consists in an Eigenmike™, featuring a spherical array of 32 microphones and an array of 8 GoPro cameras, mounted on a human torso for being placed on the seat: both the acoustical center of the microphone array and the optical center of the camera array are on the z-axis. Passenger’s head shares the same coincident reference on z-axis.

The support for the array of GoPro cameras (fig. 1) has been produced in our laboratory, with 3D printing technology, material ABS. It is designed for housing 8 GoPro Session cameras: their form factor, which is a cube, permits to mount the cameras tilted by 90 degrees, so that each frame has an aspect ratio of 3:4 instead of 4:3; this improves the vertical coverage angle. Moreover, the small size permit to mount them along a circle of small diameter, with an overlap that gives a satisfying stitching quality. The radius has been chosen in order to have a distance of 65 mm between the centers of each lens, which is the average distance between human pupils: this means that the 8 video recordings can be processed into a 3D stereoscopic panoramic video stream.

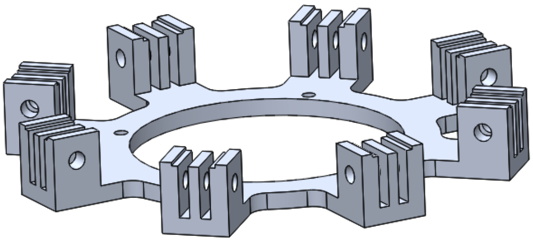

Figure 1. Mounting support for GoPro Cameras

The support for the Eigenmike has also been manufactured in our laboratory, and permits to mount the video array perfectly on axis with the acoustic center of the microphone (fig. 2). The small vertical offset can be compensated via software. The system is designed for being mounted both onto a dummy torso (fig. 3), to be placed on a seat, or on a standard microphone stand.

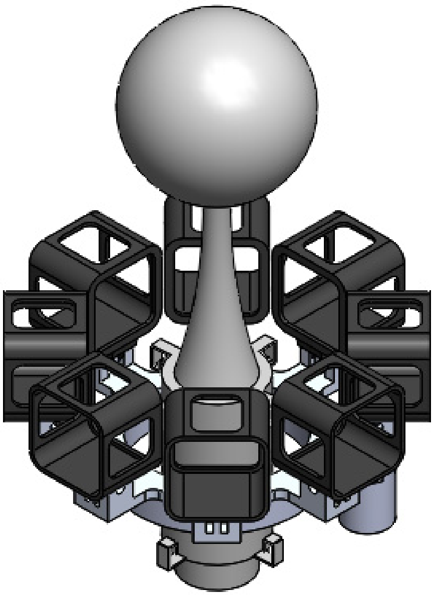

Figure 2. The audio plus video recording arrays

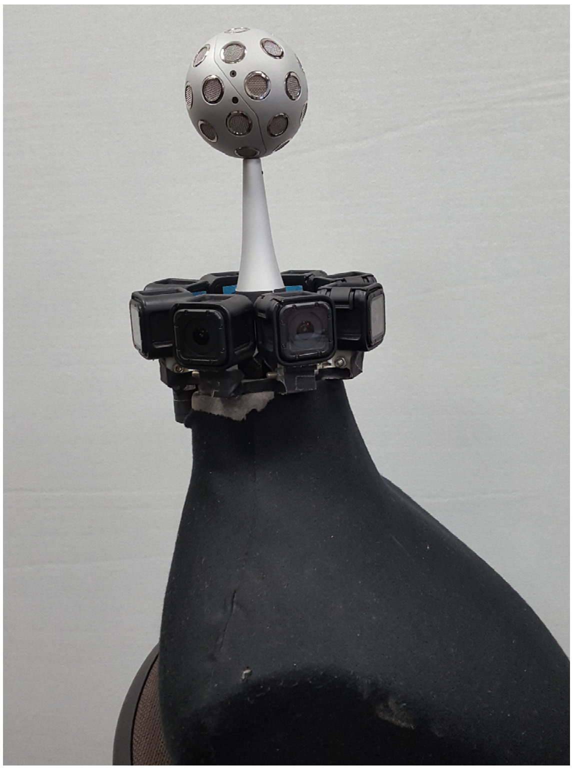

Figure 3. System on a dummy torso

If a high video quality is not mandatory, the array for video recording can be replaced by a simpler solution: a compact panoramic camera with two hemispherical lenses, such as the Samsung Gear 360. Reproduction with VR visors needs the highest possible resolution, typically 3840x1920 or even 4096x2048, but for noise analysis on laptop screen 1920x960 is in most cases even too much. Eventually, when the panoramic video scene is static and sound sources are not moving, for example for measuring noise engine or the car sound system with the car stopped, just a panoramic picture can be taken once before the measurements. This gives an additional advantage: if the panoramic background is taken in the exact position of the acoustic center of microphone array, there will not be need to compensate any offset between video and acoustic center.

To record reference signals from microphones or accelerometers at least one additional soundcard must be used and the same reference clock must be shared between it and the Eigenmike. The final recording system included three sound cards: one Emib for Eigenmike, a MetricHalo 2882 with 8 microphone inputs and a Roland Studio Capture with 12 microphone inputs and 4 line inputs (fig. 4). The master clock was imposed by the Roland Studio Capture, shared via digital S/PDIF to the Metric Halo and via Word Clock from the Metric Halo to the Emib. Recording unit was an Apple laptop and connections were via USB for Roland Studio Capture and FireWire for the others two, in daisy chain. Recording software employed was Plogue Bidule: thanks to Core Audio driver provided with macOS operating system, it is possible to create an aggregate device, which in practice permits to view all three soundcards as a single one, with a recording capability of 32 + 24 channels, 24 bits @ 48 kHz.

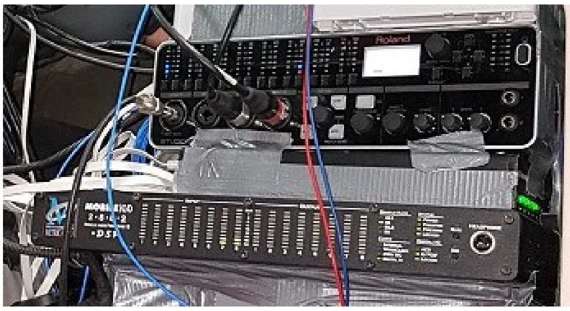

Figure 4. Recording system

2.2 Measurements

Two different situations have been created in laboratory of acoustic in the Department of Industrial Engineering at the University of Parma. The first consists in a static acoustic scene with two loudspeakers placed at the same height and playing together two uncorrelated white noises (fig. 5). The experiment has been repeated three times, with different distances of the two sources (30cm, 90cm and 180cm). The Eigenmike was placed at a distance of two meters in front of them, and one omnidirectional B&K microphone was placed in front of each loudspeaker at short distance (10 cm) for taking the corresponding reference signal.

Figure 5. Static acoustic scene of two uncorrelated sources

The second situation consists of the same two sources playing an uncorrelated white noise, with an omnidirectional microphone at short distance: this time one source was static while the other was above a moving trolley (fig. 6).

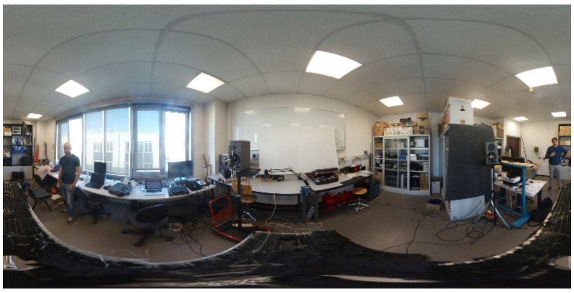

Figure 6. Moving acoustic scene of two uncorrelated sources

At LABEL – ELectroacoustic LABoratory of Casa del Suono, Parma (Italy), the reproduction system consists in a ring of eight studio monitor loudspeakers (fig. 7). The Eigenmike was placed in the center. In a first experiment 10s signal of white noise has been played sequentially from each loudspeaker. In a second, more tricky experiment, two signals have been played together: an anechoic recording of male voice from the front monitor and an anechoic recording of female voice from the back monitor.

Figure 7. Ring of 8 studio monitors at LABEL (Parma)

Inside the cockpit of a vehicle (Alfa Romeo Giulia, fig. 8), many experiments have been conducted: with the car stopped and engine off, the sound system have been tested both with exponential sine sweep excitation technique [10] and music playback; with the car running on road several maneuvers have been recorded, such as WOT – Wide Open Throttle, constant speed, or a ramp of descending speed with gear in neutral. During recording sessions on road, several microphone and accelerometers were placed for getting reference signals, with the aim of characterizing principal sources of noise, such as engine, rolling noise, exhaust or engine fan, for an ANC – Active Noise Cancelling system,.

Figure 8. Panoramic picture of Alfa Romeo Giulia cockpit

3 Signal processing

The 32 signals from the microphones are processed employing our own 3DVMS technology: a set of virtual microphone signals is derived employing a matrix of FIR filters, computed by inverting a set of anechoic impulse response measurements performed on the microphone array with a turntable, as explained in [11]. Typically two sets of virtual microphone signals are derived: a 16-channels third-order Ambisonics soundtrack is computed, to be used for rendering on an head-mounted VR visor during binaural playback with head tracking, and a 122-channels SPS (Spatial PCM Sampling) soundtrack which is employed for creating the colormap of sound level spatial distribution, as explained in [12].

The user can choose to apply an A-weighting filter: this is fundamental to keep the match between visual and aural stimulus, because it permits to compute a colormap that presents the energy distribution as it is perceived by human earing system. It is also possible to correct for an optical misalignment (vertical offset and horizontal angle) between the Eigenmike and the video recording system, if necessary.
The video colormap can be filtered with a pass-band filter, while static colormap can be computed with low-pass, high-pass, band-pass and octave band filters, and all of them are customizable.
The difference between the maximum and minimum SPL level to plot on colormap (which is in practice a noise gate) can be chosen: the lower the difference, the better will be the sound source localization, but reflections will be excluded.
Buffer size and output frame rate for generation of video colormap can be changed, iso-level lines and their value can be activated or not.

The software controlling the anechoic measurement procedure on turntable, the generation of FIR filters and of the colormap is coded in Matlab.

3.1 Colormap level calibration

Showing real SPL value is mandatory for noise analysis application, so colormap must be calibrated first. A Genelec Studio Monitor 8351A has been used to playback a pink noise filtered in the 1 kHz centered octave band, with a level of 90dB measured on axis at 1m with a sound level meter. Than Eigenmike has been placed with acoustical center on axis at 1m from the source, with the Emib gain set at 0dB, and a 30s recording has been taken. To determine the calibration coefficient for colormap, the recording has been processed to SPS 122ch format and the total average RMS signal has been computed summing the RMS values of the 122 virtual microphone signals. The difference between the resulting dB value and the reference value (90 dB) is the calibration offset. The operation performed to get the total SPL value is known as energetic summation:

$$SPL = 10 \cdot \log_{10}\{\sum_{i=1}^{122}[p_R(f)^2 + p_I(f)^2]\} \quad (1)$$

Where $p_R(f)$ and $p_i(f)$ are respectively the real and the imaginary part of the autospectrum of each of the 122 signals. Filtering becomes in this way very efficient because it can be performed in the frequency domain, and there is no need to go back to time domain (except for Ambisonics audio generation).

3.2 Coherence analysis

If reference signals have been recorded, coherence filtering can be performed. Both colormap and audio can be filtered, keeping the match between visual and aural stimulus. Colormap coherence filtering is obtained evaluating the transfer function between reference (x) and each of 122 SPS signals (y), whereas audio coherence filtering is obtained evaluating the transfer function between reference (x) and each of 32 signals from Eigenmike (A-format, y), before Ambisonics conversion.
The operator chosen for transfer function estimation is H_1, as shown in (2):

$$H_1 = \frac{P_{xy}(f)}{P_{xx}(f)} \quad (2)$$

Where $P_{xy}(f)$ is the *Cross Power Spectral Density of x and y* (or *Cross Spectrum*) and $P_{xx}(f)$ is the *Power Spectral Density of x* (or *Auto Spectrum*).
The transfer function takes as inputs the signals in time domain, but the result comes in frequency domain, so the application of filtering is simply the multiplication between the FFT of SPS signals and the transfer function itself, before computing the complex sum of eq. 1.
It is possible to estimate also the H_1 operator for a MIMO – Multiple Inputs Multiple Outputs system (3):

$$H_1 = \frac{P_{XY}(f)}{P_{XX}(f)} \quad (3)$$

Where $P_{XY}(f)$ is the *Cross Power Spectral Density Matrix* of k inputs x and i outputs y and $P_{XX}(f)$ is the *Power Spectral Density Matrix* of k and i inputs x.

For static colormap or video colormap of static and time invariant sources playing stationary signals, such the white noise, the H_1 operator is evaluated once and used for all frames. For video colormap of sources that are moving, not time invariant or playing nonstationary signal, such the voice, the H_1 operator is evaluated for each video frame.

Finally, the coherence map for each frame of the video can be calculated. Knowing the distribution of magnitude-squared coherence (4) between a reference signal and the 122 signals of SPS format, permits to determine for each frame the position of the reference on the map, tracking its motion in the space.

$$C_{xy}(f) = \frac{\left|P_{xy}(f)\right|^2}{P_{xx}(f) \cdot P_{yy}(f)} \tag{4}$$

Then, from the position of each reference for each frame, it is possible to retrieve azimuth and elevation values and pan the sound in Ambisonics format, by the solution of spherical harmonics equations.

4 Results

Static colormaps are saved as JPG images, video colormap can be saved in two different format, depending on the purpose. For laptop reproduction, a medium quality MP4 with a mono audio track corresponding to an omnidirectional microphone (first channel of Ambisonics), whereas for VR reproduction a high quality MOV with Ambisonics 3rd order track, in Ambix format, 16 bits 48 kHz. This is the highest value currently supported by the player chosen as target, Jump Inspector, an app released by Google for high-end Android smartphones.

The player comes with a built-in set of HRTFs measured with Neumann KU100 binaural head, capable of rendering a cube of eight loudspeakers, which is poor even for a second order Ambisonics. We are now working the possibility to load our own set of HRTFs, capable of rendering 32 loudspeakers distributed on a sphere, for avoiding spatial aliasing. This means we are near to render a complete 3rd order Ambisonics, customized for each person.

The smartphone can be placed inside a VR visor (fig. 9) and used to give to the user the real sensation of being inside the virtual scene: from a visual point of view, thanks to the 360° panoramic video, from an aural point of view thanks to binaural rendering of a full 3rd order Ambisonics.

Figure 9. A VR visor with integrated headphones

Figure 10 shows the colormap computed in case of two stationary sources playing at the same time two uncorrelated white noises. The sound seems to come from the apparent source located in the middle of the two loudspeakers, with a prominence of the red lobe on the left, because that loudspeaker was 3 dB more powerful. On the rear of laboratory, the reflection is easily recognizable.

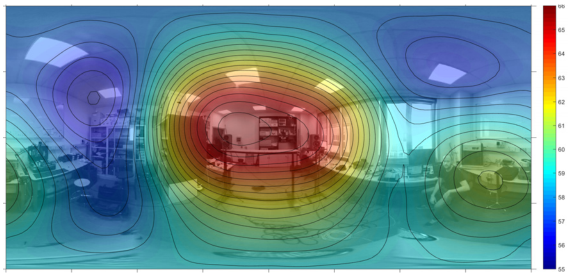

Figure 10. Colormap of two simultaneous sources

In figure 11 and 12 the same situation is analyzed introducing coherence filtering, respectively with left and right reference. Eventually, the colormap is computed with MIMO H_1 estimation, making clear the power of this analysis tool (fig. 13).

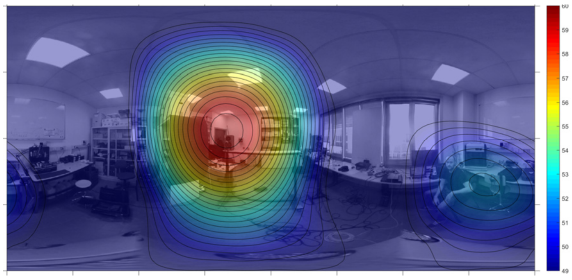

Figure 11. Colormap of two simultaneous sources, left coherence

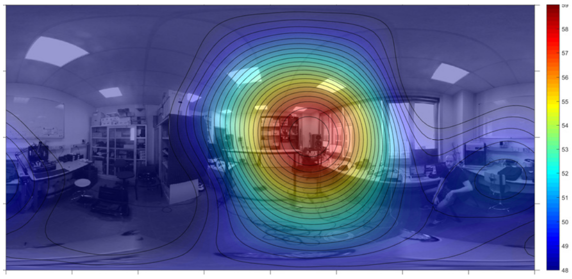

Figure 12. Colormap of two simultaneous sources, right coherence

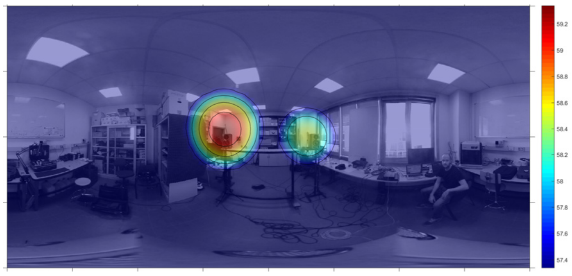

Figure 13. Colormap of two simultaneous sources, MIMO coherence

In the second experiment realized in our acoustic laboratory two loudspeakers were playing two uncorrelated white noise, with the first one stationary and the second one moving. In figure 14 the colormap of total energy is computed: note that the lobe in the middle has little meaning, because all the information is packed in a static image, but the loudspeaker was moving. In figure 15 instead the coherence filtering has been applied with the reference signal taken by the microphone facing the static source.

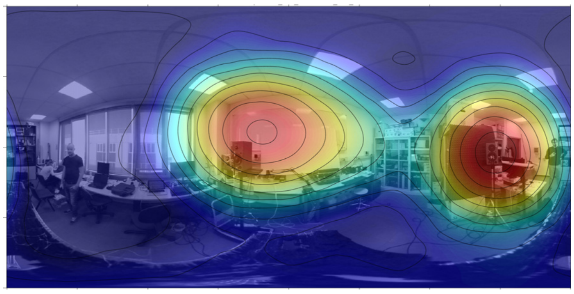

Figure 14. Colormap of two sources, with one moving (left), one static (right)

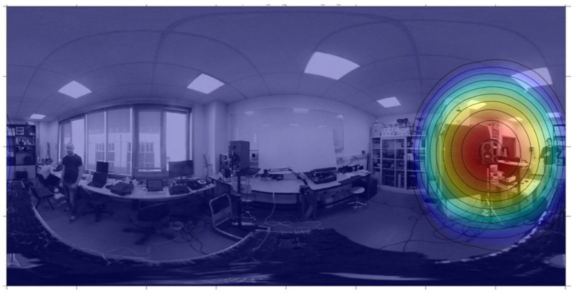

Figure 15. Colormap of two sources, coherence filtering with the static source

Here instead the situation created at LABEL is shown, where two opposite loudspeakers were playing not stationary sounds: human voices. A male voice was coming from the front while a female voice was coming from the back. In the figure 16 the total colormap is computed, while in figure 17 coherence filtering with female voice has been applied. In this case, the reference has not been taken with a microphone, but directly using the played signal. Note that in the second colormap, the orange lobe near the front loudspeaker is the reflection of the wall behind the loudspeaker.

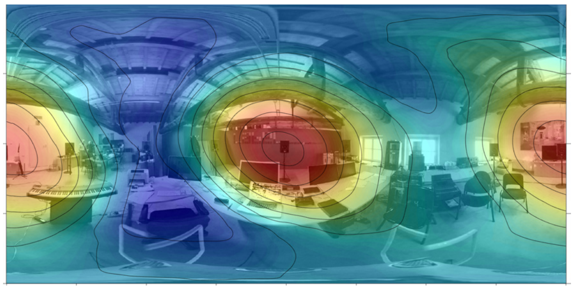

Figure 16. Colormap of LABEL with two sources

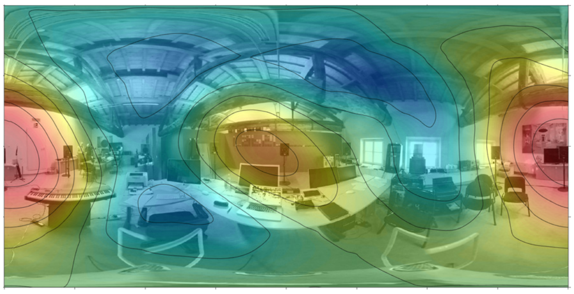

Figure 17. Colormap of LABEL with coherence filtering

The last case shown consists in the cockpit of Alfa Romeo Giulia. In figure 18, a colormap corresponding to WOT maneuver has been superimposed to the panoramic background. The noise comes out from the most solicited engine paw, on which the torque is discharged. In figure 19, the car was stopped with engine off, and a soundtrack was played by front left and front right loudspeakers. With flat setting of balance and fade and no equalization, the apparent sound source is in the middle of the sources and can be perfectly localized.

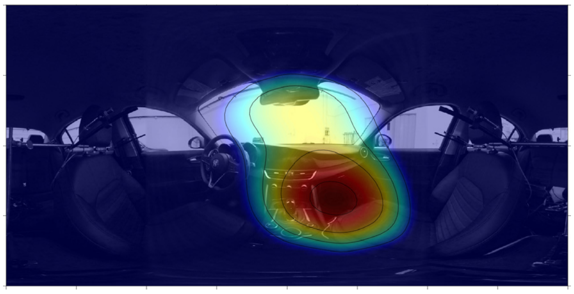

Figure 18. Colormap of Alfa Romeo Giulia in a WOT – wide open throttle

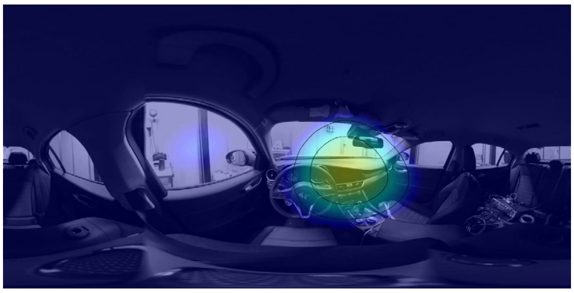

Figure 19. Colormap of Alfa Romeo Giulia – music playback with front left and front right loudspeakers

Acknowledgments

The authors would like to thank ASK industries S.P.A. for financial support and technical assistance.

This work is supported by the Italian Ministry of Economic Development (MISE)'s FUND FOR THE SUSTAINABLE GROWTH (F.C.S) under grant agreement (CUP) B48I15000130008, project VASM ("Vehicle Active Sound Management").

References

[1] J. Hladky, "The Application of Holography in the Analysis of Vibrations of Loudspeaker Diaphragms" *Journal of the Audio Engineering Society* vol. 22, Issue 4, pp. 247 – 250 (1974)

[2] F. J. M. Frankort, "Vibration Patterns and Radiation Behaviour of Loudspeaker Cones" *Journal of the Audio Engineering Society* vol. 26, Issue 9, pp. 609 – 622 (1978)

[3] D. Pinardi, A. Farina, M. C. Bellini, K. Riabova, L. Collini, "Measurement of Loudspeakers with a Laser Doppler Vibrometer and the Exponential Sine Sweep Excitation Technique" *Journal of the Audio Engineering Society* vol. 65, Issue 7/8, pp. 600 – 612 (2012)

[4] E. G. Williams, *Fourier Acoustics: sound radiation and nearfield acoustical holography* (1999)

[5] N. Harris, "Modelling Acoustic Room Interaction for Pistonic and Distributed-Mode Loudspeakers in the Correlation Domain" *Audio Engineering Society Convention 117,* paper 6180 (2004)

[6] S. Delikaris-Manias, D. Pavlidi, V. Pulkki, A. Mouchtaris, "3D Localization of Multiple Audio Sources utilizing 2D DOA Histograms" *24th European Signal Processing Conference*, pp. 1473 – 1477 (2016)

[7] L. McCormack, S. Delikaris-Manias, V. Pulkki, "Parametric Acoustic Camera for Real-time Sound Capture, analysis and tracking" *Proceedings of the 20th International Conference on Digital Audio Effects,* pp. 412 – 419 (2017)

[8] M. A. Gerzon, "Periphony: With-heigth sound reproduction" *Journal of the Audio Engineering Society* vol. 21, Issue 1, pp. 2 – 10 (1973)

[9] G. Kearny, T. Doyle, "Heigth Perception in Ambisonic Based Binaural Decoding" *Audio Engineering Society Convention 139,* paper 9423 (2015)

[10] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique" *Audio Engineering Society Convention 108,* paper 5093 (2000)

[11] A. Farina, A. Capra, L. Chiesi, L. Scopece, "A Spherical Microphone Array for Synthesizing Virtual Directive Microphones in Live Broadcasting and an Post Production" *40th Audio Engineering Society Conference* (2010)

[12] M. Binelli, A. Venturi, A. Amendola, A. Farina, "Experimental Analysis of Spatial Properties of the Sound Field inside a Car employing a Spherical Microphone Array" *130th Audio Engineering Society Conference* (2011)