A Perceptual Evaluation of an Immersive System for Automotive Environment

Loris Grossi
Inst. of Inf. Systems and Networking
Univ. of Applied Sciences and Arts
Southern Switzerland
loris.grossi@supsi.ch

Andrea Quattrini
Inst. of Inf. Systems and Networking
University of Applied Sciences and Arts
Southern Switzerland
andrea.quattrini@supsi.ch

Alberto Vancheri
Inst. of Inf. Systems and Networking
University of Applied Sciences and Arts
Southern Switzerland
alberto.vancheri@supsi.ch

Tiziano Leidi
Inst. of Inf. Systems and Networking
University of Applied Sciences and Arts
Southern Switzerland
tiziano.leidi@supsi.ch

Valeria Bruschi
Dept. of Information Engineering
Università Politecnica delle Marche
Ancona, Italy
v.bruschi@staff.univpm.it

Nefeli Aikaterini Dourou
Dept. of Information Engineering
Università Politecnica delle Marche
Ancona, Italy
n.a.dourou@pm.univpm.it

Stefania Cecchi
Dept. of Information Engineering
Università Politecnica delle Marche
Ancona, Italy
s.cecchi@staff.univpm.it

***Abstract*—Immersive audio systems enhance the listening experience by creating spatially rich and realistic soundscapes. The increasing interest in these systems has paved the way for their adoption across various fields, ranging from gaming and virtual reality to automotive environments. While headphones are traditionally used for binaural signal reproduction, loudspeaker-based systems are emerging as alternatives for specific scenarios. However, the use of loudspeakers introduces the phenomenon of crosstalk, and crosstalk cancellation (CTC) techniques are specifically required when loudspeakers are used for binaural reproduction. In this paper, a four-channel loudspeaker system, arranged in a car simulator, is employed for the binaural reproduction, applying a multiband crosstalk cancellation algorithm. The system has been perceptually evaluated through listening tests, in comparison with other existing CTC approaches. The results demonstrate the strong performance of the implemented system.**

***Index Terms*—immersive audio, crosstalk cancellation, perceptual evaluation**

I. Introduction

Over the past decades, spatial audio rendering has attracted significant attention from both the scientific and industrial communities due to its potential to enhance sound and music perception across various applications. These include video conferencing [1], gaming [2], virtual and augmented reality [3], [4], as well as advancements in the therapeutic field [5]. These systems rely on headphones to reproduce binaural signals, which include spatial information from virtual sound sources specifically designed for each ear. As demonstrated in [6], advanced digital signal processing frameworks can support spatial audio synthesis through both headphone and loudspeaker playback, including time-varying filtering, dynamic source positioning, and room acoustics modeling. However, headphones are not practical for certain applications, such as shared listening environments or situations requiring long-term comfort, making loudspeaker-based reproduction a preferred alternative. One such method is the stereo dipole configuration, where closely spaced loudspeakers are used to generate spatially localized virtual sound sources around the listener, as explored in [7]. This, however, introduces the crosstalk effect, as each ear receives signals from both channels, contrary to the intended design of binaural signals for individual ears. To address this issue, a crosstalk cancellation (CTC) algorithm becomes necessary [8], [9]. Recent work has explored adapting CTC techniques to conventional stereo and surround systems for reproducing spatial audio content originally intended for multichannel 3D formats [10].

In parallel, emerging paradigms such as the Internet of Sound (IoS) are reshaping how spatial audio technologies are designed, deployed, and experienced across distributed and interactive environments [11], [12]. IoS frameworks promote interoperability, low-latency communication, and edge computing architectures, important features for real-time immersive audio applications in mobile or constrained contexts, including automotive settings [13], [14].

With the growing amount of time spent in transportation, binaural headrest systems have the potential to shape the future of immersive listening environments. These systems consist of loudspeakers mounted on a seat behind the listener's head and

Part of this work was supported by the financial program DM MiSE 3370 27 October 2023, project "Epignosis" F/340008/01/X59.

can be utilized in various seated scenarios, including cars [15], [16], helicopters or aircraft [17], [18], and yacht settings [19]. In the literature, audio headrest applications are mostly related to active noise cancellation [18], [19] or the creation of personal sound zones [20]–[22]. However, only a limited number of approaches emphasize binaural audio. A crosstalk cancellation algorithm tailored for a binaural audio headrest was developed and tested in [16]. The study evaluated three loudspeaker positions and found that, for certain frequencies and when the loudspeakers are placed close to the head, it is possible to achieve effective crosstalk cancellation, resulting in high-quality binaural reproduction. Additionally, they observed that at specific frequencies, the positioning of the loudspeakers is closely linked to the issue of an ill-conditioned matrix [23]. Furthermore, the loudspeaker arrangement naturally enhances channel separation due to head shadowing effects. In [24], the feasibility of binaural rendering using drive units integrated into an automotive headrest, without employing crosstalk cancellation, was explored. The study specifically analyzed natural channel separation in configurations where the rear of the skull directly contacted the loudspeakers. They discovered that, under anechoic conditions and with no gap between the head and the headrest, sufficient natural channel separation can be achieved for broadband binaural audio reproduction without the need for crosstalk cancellation. This study overlooked two important aspects. Firstly, in practical scenarios, there is likely to be a gap between the head and the headrest, which may vary during playback. Secondly, environmental reverberation can significantly impact the effectiveness of binaural reproduction.

An alternative method was examined in [17], where loudspeakers were positioned laterally to the listener's ears on a test rig attached to an aircraft seat. The objective was to expand the sweet spot of the crosstalk canceller by utilizing an optimized filter designed to minimize error within a small spatial area around the listener, termed space-average. While this approach reduces the impact of listener movement, the lateral placement of the loudspeakers is impractical for real-world applications, such as in automotive environments.

The simplest crosstalk cancellation procedure consists of the inversion of the matrix containing the four head-related transfer functions (HRTFs), which describe the acoustic paths between the loudspeakers and the listener's ears. It was introduced by Bauer in [25]. Despite the simplicity of this approach, the transfer function inversion is not always possible due to the non-minimum-phase characteristics of most systems. For this reason, more efficient CTC algorithms have been studied over the years. A widespread solution was proposed by Kirkeby in [23], consisting of a fast deconvolution (FD) method with regularization. Another approach used in CTC algorithms is the least mean square (LMS), employed thanks to its simplicity and robustness [26]. However, the CTC techniques discussed above necessitate the knowledge of the HRTFs and are sensitive to listener head movements. In this context, Glasgal [27] proposed the recursive ambiophonic crosstalk elimination (RACE) algorithm that is based on the inversion and attenuation of unwanted signals and does not require the HRTFs knowledge. A variation of the RACE was introduced in [28], which relies on modeling and controlling the propagation of acoustic waves from the sources to the listener's ears. This approach, closely resembling ray tracing techniques used in computer graphics, enhances robustness by regularizing the order of the cancellation signal cues. Leveraging these concepts, a solution that incorporates a multiband approach to crosstalk cancellation was proposed in [29]. This frequency-band-dependent CTC allows for an improvement in the cancellation, a reduction of the coloration, and an extension of the bandwidth, maintaining a good system robustness.

In this paper, the CTC approach of [29] is applied to a four-channel immersive system arranged in a car simulator to perform a perceptual evaluation. The algorithm has been implemented through a demo application, widely described in the paper. This approach is compared to the fast deconvolution method of [23] and to the RACE [27] using different soundtracks.

The paper is organized as follows. Section II describes the multiband crosstalk cancellation algorithm. Section III illustrates the application that implements the CTC algorithm. Section IV explains the experimental setup and how the tests are performed. Section V reports the experimental results of the perceptual evaluation. Finally, Section VI concludes the paper.

II. Multiband crosstalk cancellation

At the core of the method introduced in [28] lies an approach based on modeling CTC using geometrical acoustics principles, which extends the well-known Recursive Ambiophonic Crosstalk Elimination (RACE) approach, but instead of relying strictly on recursive filters, it employs geometrical modeling for computing the acoustical propagation paths. As part of this method, the concept of a cancellation complex is introduced. A cancellation complex consists of the following components:

- A virtual source S_0 called the primary source (see below for the definition of virtual source).
- Two virtual sources S_1 and S_2 called the cancellation sources.
- Two receivers E_1 and E_2, to be identified with the ears of the user, called target and non-target ear.

A virtual source is a sound source used in computation that is physically identified with a real loudspeaker when the sound is outputted. This definition makes the algorithm more flexible and allows us to localize, with some technical constraints, more than one virtual source on the same real loudspeaker.

The cancellation complex works as follows. The source audio signal is emitted from the primary source S_0 directed to the target ear E_1. The cancellation sources S_1 and S_2 generate signals to cancel the crosstalks: S_1 cancels the crosstalks received at E_1 and S_2 the crosstalks received at the E_2. With these definitions, a cancellation complex is used to impose a single audio channel to a target ear and can be implemented with two real loudspeakers if S_0 and S_1 are co-localized on the same real loudspeaker.

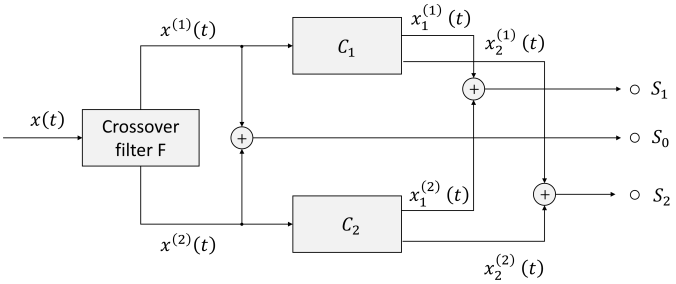


Fig. 1. Scheme of the multiband crosstalk cancellation algorithm for two bands. The input signal is split into $n = 2$ bands, obtaining $x^{(j)}(t)$, with $j = 1, ..., n$. Each output is processed by C_j, producing the signals $x_1^{(j)}(t)$ and $x_2^{(j)}(t)$, sent to the sources S_1 and S_2, respectively. S_0 is computed after the crossover filter F to preserve phase coherence, when using IIR crossover filters.

The impulse responses (IRs) related to the cancellation sources are calculated as infinite successions of pulses with timing and amplitudes derived from the HRTF and the relative position of the head with respect to the loudspeakers. It is possible to prove that the amplitudes of the pulses in the IRs are geometrically decaying with a common ratio G determined by the HRTF and the head position and rotation in accordance with the following definition:

$$G = \frac{g_{12}g_{21}}{g_{11}g_{22}} \tag{1}$$

where $g_{ij} = |H_{ij}|$ is the gain of the HRTF from the source S_i to the ear E_j.

The IRs can be practically truncated at a length N, called the cancellation order. The truncation process leaves a non-canceled residual at the ears. Two truncation approaches can be implemented: the counterlateral, when the truncation is at the same length N for both cancellation sources, and the ipsilateral, when the truncation is at length $N-1$ at the loudspeaker S_1. The counterlateral method leaves the residual at the non target ear E_2, producing a loss of cancellation effectiveness, whereas the ipsilateral methods leaves the residual at the target ear E_1, producing coloration. The truncation order is chosen to balance the negative effect of the non-cancelled residual, that decreases with N, with the negative effects, in real emission conditions, of an infinite recursion.

Stability and robustness are analyzed via the parameter G defined above. The cancellation complex is stable if $G < 1$ and robustness is greater for smaller values of G.

This geometrical acoustics-based approach simplifies implementation and provides effective, robust, and flexible real-time cancellation of crosstalk. Laboratory experiments confirm a robust cancellation effectiveness.

The evolution of this method, provided in [29], introduces a multiband time-domain strategy called MB CTC (Multiband CrossTalk Cancellation) to enhance the performances of the geometrical acoustics method. In the MB CTC approach, the frequency range is initially segmented into n overlapping bands, and 4th-order Linkwitz-Riley filters are subsequently designed to target each band. Each output signal of the crossover filter is processed separately, using gain and delays derived from the HRTF at the center of the corresponding band. The scheme of the algorithm is shown in Figure 1, where the case $n = 2$ is considered for simplicity (in real implementations n is usually equal to or larger than 5). With reference to the figure, the input signal $x(t)$ is processed as follows:

- The filter bank F splits the input signal $x(t)$ into n components $x^{(j)}(t)$, $1 \leq n$.
- The j-th component $x^{(j)}(t)$ is processed by the unit C_j that generates signals $x_1^{(j)}(t)$, $x_2^{(j)}(t)$ to be sent to the sources S_1 and S_2, respectively.
- The output of F are directly sent to S_0 after being mixed together again. The crossover filter F is applied to preserve phase coherence, when using IIR crossover filters.

The key benefits of this multiband extension include improved CTC effectiveness and reduced residual coloration. In experiments, the MB CTC approach exhibited significantly improved cancellation effectiveness compared to the single band approach. The MB CTC approach efficiently addresses the limitations inherent to the geometrical acoustics single-band model, enabling more robust, accurate, and immersive rendering of binaural audio content over loudspeakers, and presents a promising avenue for dynamic and adaptive implementations, as further studied in [30].

III. System Implementation

The MB CTC algorithm has been implemented through a demo application shown in Figure 2. The application comprises two panels: the pre-process and the process panel. The parameters of the pre-process panel, shown in Figure 2(a), are the following:

- **InputFileName:** specifies the audio file to play. A 4-channel track with a sampling frequency of 48 kHz is used. This track is constructed by duplicating the original stereo channels.
- **HeadTableJSONFile:** JSON file containing HRTF information, the sampling frequency, set to 48 kHz, and the frequency band limits. In the experiments, five bands have been considered with the following limits: [0, 150, 500, 1500, 5500, 24000] Hz.
- **ConfigJSONFile:** JSON file containing all the information required to configure the CTC algorithm, including the position of the four loudspeakers, expressed in meters following the measurement coordinate system shown in blue in Figure 3.
- **HeadPosition:** specifies the coordinates of the midpoint of the ears. According to the setup of Figure 3, the following coordinates are set: `x=0 y=0 z=1`. The head position is assumed to be fixed.
- **HeadRotation:** specifies the rotation of the head. The rotation is specified using a string formatted as follows: `n_th=0.0 n_phi=0.0 rot=-90`. A rotation angle of -90 degrees means that the listener is looking straight

ahead, as shown in Figure 3. Head rotation is also considered to be fixed.

- **CancellationOrder:** the recursion depth of the crosstalk cancellation process. An order of 7 is chosen for the experiments.
- **ResidualType:** specifies the method used to characterize the residual crosstalk after the truncation of the cancellation process at a finite order. It is set to `IPSILATERAL RESIDUAL`, so the non cancelled residual will reach the ipsilateral ear (with respect to the direct sound).
- **StoreToFile:** specifies if the output signal elaborated by the CTC algorithm is saved or not.
- **OutputFileName:** the name of the output file, if the storing is enabled, can be specified.

The parameters of the process panel, shown in Figure 2(b), are the following:

- **MuteCancellation:** specifies whether the CTC algorithm is enabled.
- **MainGain:** global gain of the entire application.

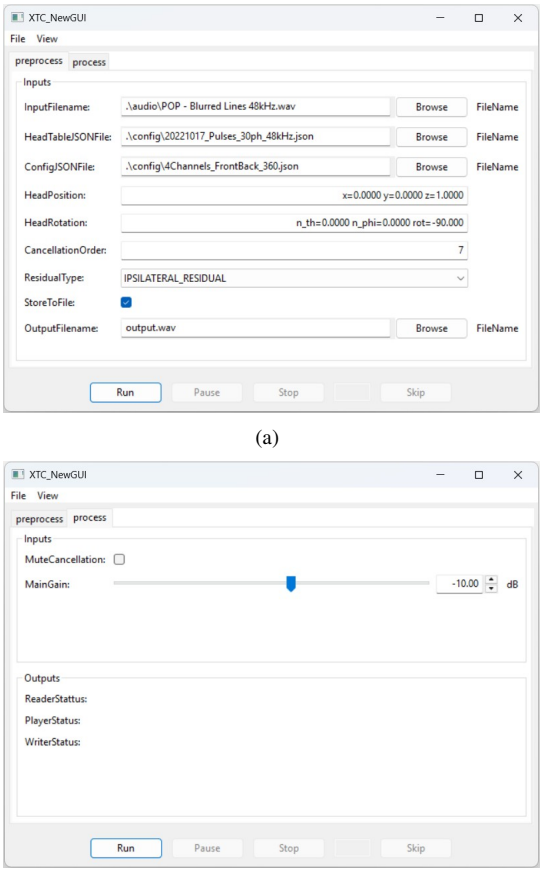
Moreover, in the process panel, the following application status labels are shown:

- **ReaderStatus:** shows possible error messages related to the file reader.
- **PlayerStatus:** shows possible error messages related to the audio player.
- **WriterStatus:** shows possible error messages related to the output audio writer.

IV. Experimental Setup

The presented system has been evaluated through subjective listening tests using the setup shown in Figure 3. The PC was connected to a Focusrite Scarlett 18i20 sound card, which managed four Genelec 6010A loudspeakers. The listener was sitting in a vehicle seat surrounded by the four loudspeakers, as shown in Figure 4, with their head remaining relatively static. The frontal loudspeakers (FL and FR) were positioned one meter from the listener and were spaced 1.12 meters apart. The rear loudspeakers (RL and RR) were placed close to the listener, 10 cm behind the center of the head, and 71 cm apart, to recreate a headrest setup.

The spatial quality perception of the crosstalk cancellation filters was evaluated according to the protocol outlined in Recommendation ITU-R BS.1284 (general methods for the subjective assessment of sound quality) [31]. Specifically, the spatial impression attribute was assessed. According to ITU-R BS.1284, spatial impression refers to the extent to which the auditory scene gives the listener a convincing sense of space, including envelopment and the perceived spatial environment of the sound source. Listeners were asked to rate the spatial impression based on the definition provided in the recommendation: “The performance appears to take place in an appropriate spatial environment”. In stereo recordings, spatial impression is associated with the perceived degree of stereo widening, that is, how spread out or expansive the sound feels across the left–right field. Following the guidelines of



(a)

(b)

Fig. 2. Screen of the (a) pre-process and (b) process panel of the XTC demo application.

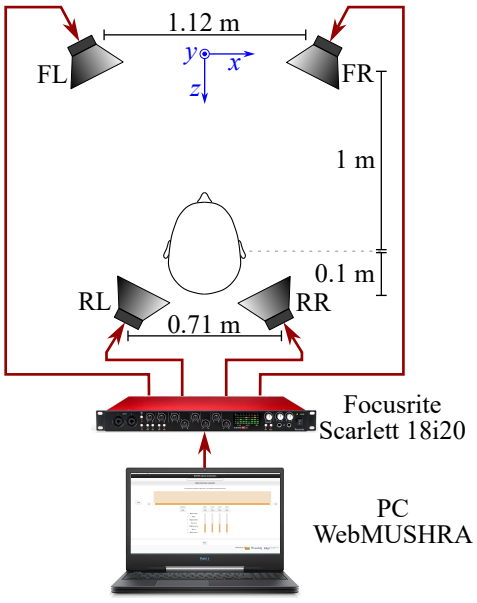


Fig. 3. Setup of the experiments.

[31], expert listeners were preferred, as they can provide a better and quicker indication of the likely results in the long

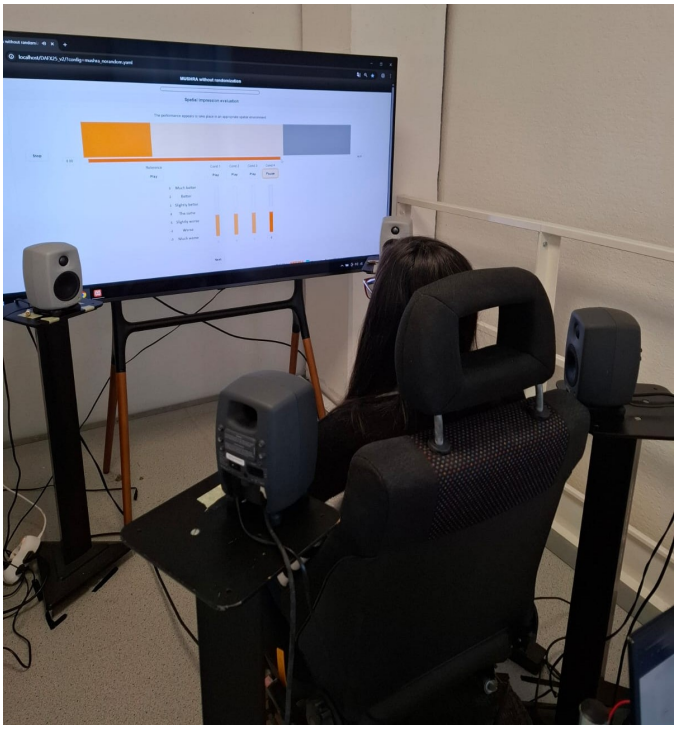


Fig. 4. Photo of the experimental setup.

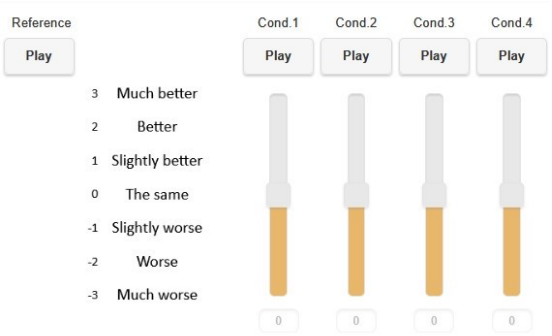


Fig. 5. Graphical User Interface (GUI) for subjective listening tests.

term. However, to minimize bias in the listening test outcomes, the listeners were not informed about the study objectives or the anticipated results.

The experimental graphical user interface (GUI) is shown in Figure 5. The GUI was implemented using webMUSHRA, a web audio API based experiment software [32]. A multiple comparison test with a hidden reference was used. Specifically, the following four conditions were evaluated in relation to the known reference within the same trial. All evaluations were conducted using a bipolar discrete seven-grade scale, ranging from -3 (Much worse) to 3 (Much better):

- 4-channel sound filtered to the 4 crosstalk filters calculated through the multiband CTC (MB CTC) algorithm, explained in Section II.
- 4-channel sound filtered to the 4 crosstalk filters calculated through the RACE algorithm of [27].
- 4-channel sound filtered to the 4 crosstalk filters calculated through the fast deconvolution (FD) method with regularization of [23].
- 2-channel hidden reference (i.e., original stereo track reproduced solely by the frontal loudspeakers).

It is worth noting that the hidden reference does not represent the best-case scenario. In fact, the other conditions can be evaluated either higher or lower than the reference. Ideally, the best-case reference would be headphone reproduction. However, this was not included in the experiments, as it would require changing the reproduction setup within the same session, which is not feasible. For all 4-channel conditions, crosstalk cancellation is applied simultaneously to both the front and rear loudspeakers. The fast deconvolution approach has been applied to HRTFs measured with a Brüel & Kjær - Type 4128C binaural mannequin, after a frequency warping [33]. The FD has been employed considering a regularization factor of 10^{-6}, and a frequency threshold ± 20 dB has been applied to the final FD crosstalk filters. These values have been chosen through experimentation.

The experimental procedure used seven tracks, the details of which are presented in Table I. In the table, the "Type" refers to the input audio track. The four music genres are originally stereo tracks that were adapted to our 4-channel system with crosstalk cancellation. The two speech samples are spatially panned to the left and right channels, respectively. Finally, the binaural track includes pre-binauralized content and features natural sounds arriving from various directions and moving dynamically within the scene. All stimuli were approximately 20 seconds long, ensuring that sentences or phrases were not cut off. The stimuli were presented in a randomized order throughout the trials. Participants had the option to switch freely between the reference signal and any of the conditions being evaluated, allowing them to make direct comparisons between the algorithms and the reference stereo signal. The ability to identify hidden references facilitates the post-screening of subjects. The listening level was maintained consistently throughout the entire experimental procedure for all participants.

V. Experimental Results

Twelve participants took part in the experiment; however, two participants who failed to identify the hidden reference

TABLE I
EXPERIMENTAL TRACKS USED IN THE SUBJECTIVE TESTS.

Label	Title	Type
Classical	The Four Seasons Spring op.8 no.1, Vivaldi	Stereo
Pop	Blurred Lines, Robin Thicke	Stereo
Rock	The Chain, Fleetwood Mac	Stereo
Soul	Feeling Good, Nina Simone	Stereo
Female Speech	Female Speech	Stereo, left panning
Male Speech	Male Speech	Stereo, right panning
Binaural	Sound of nature	Binaural track

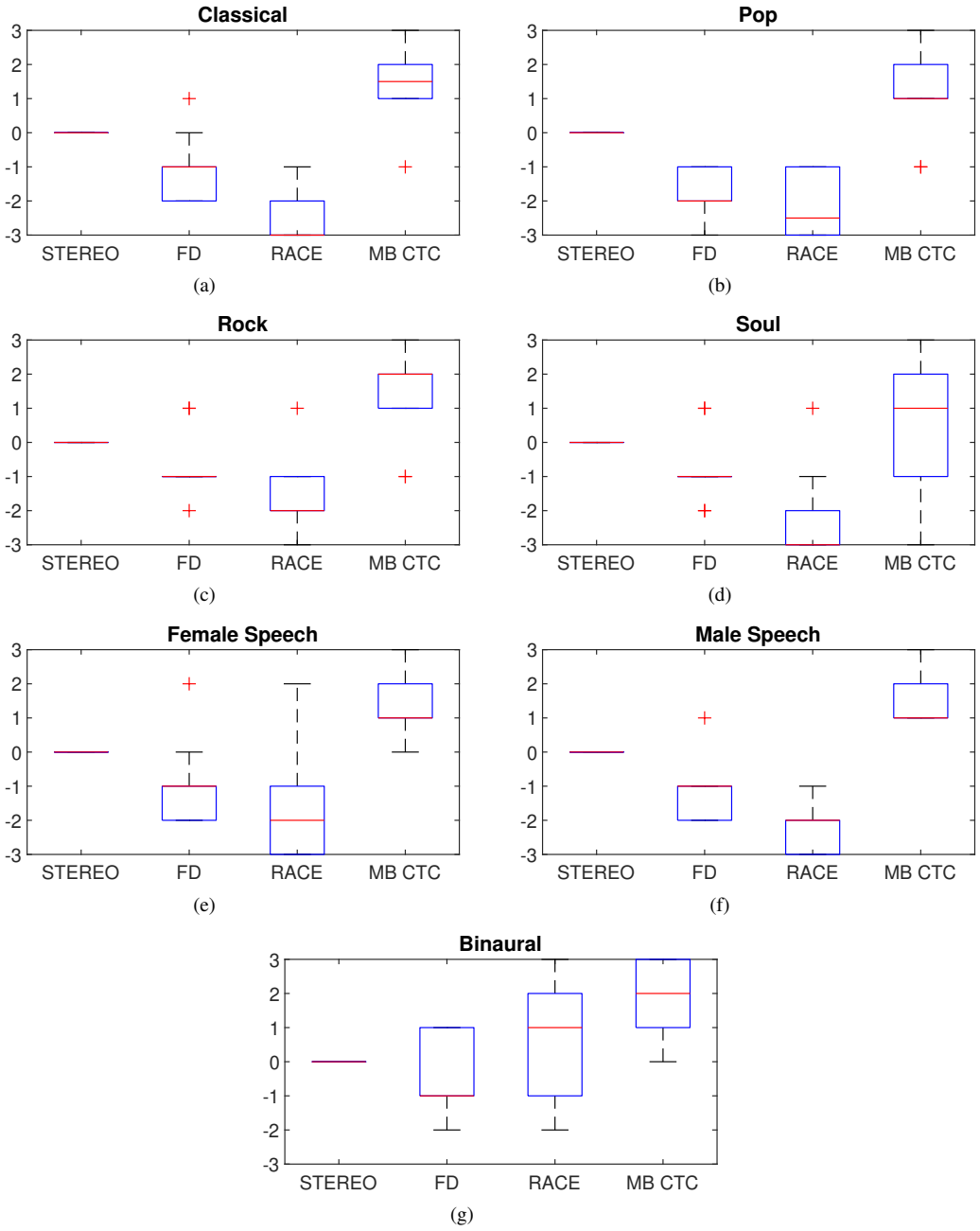


Fig. 6. Spatial impression ratings for the (a) Classical, (b) Pop, (c) Rock, (d) Soul, (e) Female speech, (f) Male speech, and (g) Binaural track, where STEREO is the reproduction without CTC using only the front loudspeakers, FD is the Fast deconvolution method based on the FFT [23], RACE is the Recursive Ambiophonic Crosstalk Elimination algorithm [27], and MB CTC is the multiband crosstalk cancellation algorithm.

were excluded. The final panel of subjects consisted of 10 expert listeners, including 7 males and 3 females, aged between 23 and 30 years (mean age of 26.2, and standard deviation of 3 years). All participants possessed a technical background in acoustics and prior experience with binaural audio listening. In Section V-A, descriptive statistics summarize the ratings provided by the 10 participants, while Section V-B employs inferential statistics to analyze samples and make predictions about larger populations.

A. *Descriptive statistics*

The evaluations of spatial impression for the seven experimental tracks are presented in Figure 6. In each box plot, the central red mark represents the median, while the lower and upper edges of the blue box indicate the 25th and 75th percentiles, respectively, defining the interquartile range (IQR). The whiskers extend to the most extreme data points that are not classified as outliers, while outliers are displayed separately using the "+" marker symbol.

The spatial impression is primarily evaluated based on the

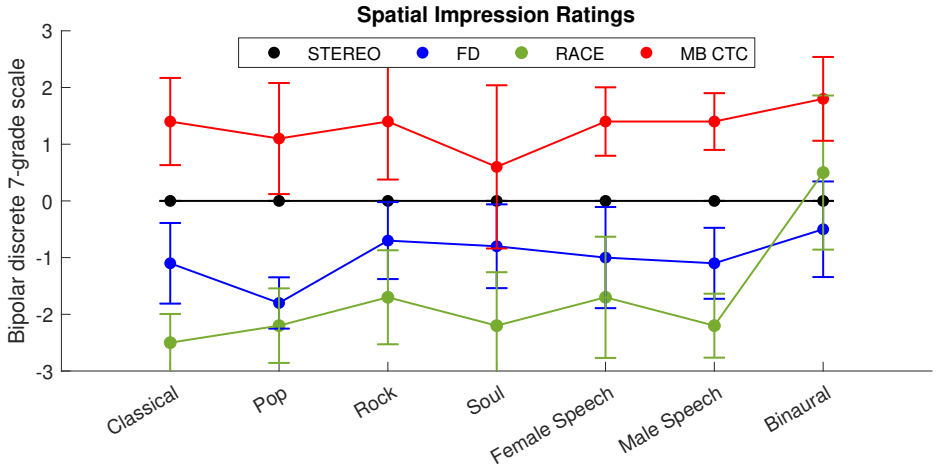


Fig. 7. Estimated marginal means and 95% confidence intervals for the spatial impression ratings for the seven experimental tracks, where STEREO is the reproduction without CTC using only the front loudspeakers, FD is the Fast deconvolution method based on the FFT [23], RACE is the Recursive Ambiophonic Crosstalk Elimination algorithm [27], and MB CTC is the multiband crosstalk cancellation algorithm.

median value of the ratings. The median is preferred over the mean due to the relatively small sample size, the non-normal distribution, and the presence of significant outliers [34]. The stereo version received a score of zero for all participants and tracks, indicating that the participants included in the analysis were able to identify the hidden reference. The fast deconvolution and RACE algorithms performed worse compared to the stereo version. Specifically, the fast deconvolution algorithm demonstrated a slight deterioration in spatial impression, with an average median value of -1.1 (SD = 0.38) relative to the reference (i.e., stereo) version, while the RACE algorithm exhibited a more significant decline, with an average median value of -1.9 (SD = 1.37) in spatial impression across the seven tracks. Notably, the RACE algorithm outperformed the reference version in the single case of the binaural track, as shown in Figure 6(g). The multiband crosstalk cancellation algorithm consistently demonstrated the highest spatial impression among all the algorithms, including stereo, with an average rating of 1.4 (SD = 0.48). Remarkably, the MB CTC algorithm outperformed RACE even in the case of the binaural track, where RACE exhibited a relatively high spatial impression.

TABLE II
STATISTICAL SIGNIFICANCE OF THE DIFFERENCE BETWEEN EACH CTC ALGORITHM AND THE EVALUATION FOR THE STEREO VERSION (BASED ON ESTIMATED MARGINAL MEANS). SIGNIFICANT CASES ($p < 0.05$) ARE IN BOLD.

Track	FD	RACE	MB CTC
Classical	**0.040**	**< 0.001**	**0.016**
Pop	**< 0.001**	**< 0.001**	0.191
Rock	0.267	**0.007**	0.077
Soul	0.221	**0.003**	1.000
Female Speech	0.192	**0.035**	**0.003**
Male Speech	**0.019**	**< 0.001**	**0.001**
Binaural	1.000	1.000	**0.002**

From another perspective, it is interesting to observe the diversity of responses given for each algorithm and track, by analyzing the interquartile range of them. The least diversity is noted for the fast deconvolution crosstalk cancellation, with an average IQR of 0.9, followed by the RACE and the MB CTC algorithms, which have an average IQR of around 1.5.

The poor performance of fast deconvolution and RACE in crosstalk cancellation systems is due to limitations in their assumptions and robustness. Fast deconvolution relies on the assumption that HRTFs are minimum-phase, which simplifies the inversion but discards important spatial and phase information [35]. Conversely, RACE employs a recursive algorithm in the time domain [27]. When multiple orders of cancellation remain perceptible, the overlapping delayed and inverted signals can interact destructively, resulting in perceptual artifacts such as spectral coloration or timbral distortion.

B. *Inferential statistics*

A two-factorial repeated measurements analysis of variance (rmANOVA) model is employed to generalize the results obtained in 10 participants. The problem can be seen as one for which "two within-subjects factors (crosstalk cancellation algorithm and audio material) are completely crossed, and at least one rating is obtained for each combination of listener, audio material, and condition" [34]. Specifically, the analysis evaluates the effect of the crosstalk cancellation algorithm on the spatial impression of each of the seven experimental tracks, using the stereo reproduction as a reference.

The analysis in Figure 7 illustrates the estimated marginal means (EMMs) along with the 95% confidence intervals (CIs) for each CTC algorithm (including the stereo version) and track, as calculated from the ANOVA modeling. Estimated marginal means represent predicted population means based on a statistical model rather than raw data. Following, in Table II, post-hoc tests evaluate the statistical significance of the differences in EMMs ratings between the CTC algorithms

and the stereo signal presented in the form of the hidden reference. Since the hidden reference was correctly identified in all cases, this comparison is equivalent to analyzing the effect with the reference of zero on the bipolar evaluation scale. A value below 0.05 signifies a statistically significant difference in the spatial impression rating, but does not indicate whether it is an improvement or a decline. In this context, a small value of statistical significance is interpreted as clear evidence of the algorithm's noticeable effect. Proper adjustments using Bonferroni corrections have been considered.

Focusing on Figure 7 and Table II, we again confirm that the EMM for the stereo is zero for all instances. A slight decrease in spatial impression is observed with the fast deconvolution algorithm, which is clear in three stimuli: classical ($p = 0.040$), pop ($p < 0.001$), and male speech ($p = 0.019$). The RACE algorithm shows a more pronounced reduction, statistically significant across all tracks ($p < 0.050$) except for the binaural ($p = 1.000$), where the RACE seems to improve spatial impression compared to stereo. Lastly, for all tracks, the EMMs of the MB CTC are consistently higher than those of the stereo. Applying the MB CTC algorithm, statistical significance is achieved for the majority of the tracks, though one track (i.e., soul) shows a less evident improvement ($p = 1.000$).

VI. Conclusions

The paper presents a perceptual evaluation of a four-channel immersive system based on a multiband crosstalk cancellation algorithm. The system setup consists of two frontal loudspeakers and two rear loudspeakers, arranged in a headrest configuration on a vehicle seat. The presented multiband CTC approach has been compared to the fast deconvolution method and the RACE algorithm, thorough a multiple comparison test with a hidden reference (i.e., the original stereo sound), considering seven soundtracks and twelve listeners. Descriptive and inferential statistics were employed to evaluate spatial impression, demonstrating a clear superiority of the MB CTC algorithm over stereo reproduction across all experimental tracks. In contrast, the fast deconvolution algorithm exhibited a slight deterioration across all tracks, while the RACE algorithm displayed a more pronounced decline. However, an exception was noted for one track, where the RACE algorithm improved spatial impression, achieving a performance level closer to that of the MB CTC.

The current evaluation of the presented system focuses exclusively on the driver's listening experience, assuming a fixed head position. Future research could extend this analysis to more realistic scenarios by incorporating head-tracking systems for adaptive crosstalk cancellation and by considering the auditory perception of passengers.

References

[1] M. Wong and R. Duraiswami, "Shared-space: Spatial audio and video layouts for videoconferencing in a virtual room," in *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*. IEEE, 2021, pp. 1–6.

[2] J. Broderick, J. Duggan, and S. Redfern, "The importance of spatial audio in modern games and virtual environments," in *2018 IEEE Games, Entertainment, Media Conference (GEM)*. IEEE, 2018, pp. 1–9.

[3] T. Potter, Z. Cvetković, and E. De Sena, "On the relative importance of visual and spatial audio rendering on VR immersion," *Frontiers in Signal Processing*, vol. 2, pp. 904866, 2022.

[4] T. Langlotz, H. Regenbrecht, S. Zollmann, and D. Schmalstieg, "Audio stickies: visually-guided spatial audio annotations on a mobile augmented reality platform," in *Proceedings of the 25th Australian computer-human interaction conference: augmentation, application, innovation, collaboration*, 2013, pp. 545–554.

[5] D. Johnston, H. Egermann, and G. Kearney, "The use of binaural based spatial audio in the reduction of auditory hypersensitivity in autistic young people," *International journal of environmental research and public health*, vol. 19, no. 19, pp. 12474, 2022.

[6] J.M. Jot, V. Larcher, and O. Warusfel, "Digital signal processing issues in the context of binaural and transaural stereophony," in *98th Audio Engineering Society Convention*, Paris, France, Feb. 1995.

[7] O. Kirkeby, P.A. Nelson, and H. Hamada, "The Stereo Dipole - Binaural Sound Reproduction Using Two Closely Spaced Loudspeakers," in *102nd Audio Engineering Society Convention*, Munich, Germany, Mar. 1997.

[8] D. H. Cooper and J. L. Bauck, "Prospects for Transaural Recording," *J. Audio Eng. Soc.*, vol. 37, no. 1/2, pp. 3–19, Feb. 1989.

[9] William Grant Gardner, *Transaural 3-D audio*, Citeseer, 1995.

[10] A. Baskind, T. Carpentier, J.M. Lyzwa, and O. Warusfel, "Surround and 3D-audio production on two-channel and 2D-multichannel loudspeaker setups," in *3rd International Conference on Spatial Audio (ICSA)*, 2015.

[11] Luca Turchet, Mathieu Lagrange, Cristina Rottondi, György Fazekas, Nils Peters, Jan Østergaard, Frederic Font, Tom Bäckström, and Carlo Fischione, "The internet of sounds: Convergent trends, insights, and future directions," *IEEE Internet of Things Journal*, vol. 10, no. 13, pp. 11264–11292, 2023.

[12] Luca Turchet, Claudia Rinaldi, Carlo Centofanti, Luca Vignati, and Cristina Rottondi, "5G-enabled internet of musical things architectures for remote immersive musical practices," *IEEE Open Journal of the Communications Society*, vol. 5, pp. 4691–4709, 2024.

[13] Claudia Rinaldi, Fabio Franchi, Andrea Marotta, Fabio Graziosi, and Carlo Centofanti, "On the exploitation of 5G multi-access edge computing for spatial audio in cultural heritage applications," *IEEE Access*, vol. 9, pp. 155197–155206, 2021.

[14] Stefano Giacomelli, Carlo Centofanti, José Santos, Mauro Galbiati, Tiziano Salvi, Fabio Graziosi, and Claudia Rinaldi, "Remote immersive audio production: State of the art implementation, challenges, and improvements," in *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*. IEEE, 2024, pp. 1–10.

[15] J. Kovačević, N. Kaprocki, and A. Popović, "Review of automotive audio technologies: Immersive audio case study," in *2019 Zooming Innovation in Consumer Technologies Conference (ZINC)*. IEEE, 2019, pp. 98–99.

[16] A. Lundkvist, A. Nykänen, and R. Johnsson, "3D-sound in car compartments based on loudspeaker reproduction using crosstalk cancellation," in *130th Convention of Audio Engineering Society*. Audio Engineering Society, 2011.

[17] S. Elliott, C. House, J. Cheer, and M. Simon-Galvez, "Cross-talk cancellation for headrest sound reproduction," in *Audio Engineering Society Conference: 2016 AES International Conference on Sound Field Control*. Audio Engineering Society, 2016.

[18] J. Buck and D. Sachau, "Active headrests with selective delayless subband adaptive filters in an aircraft cabin," *Mechanical Systems and Signal Processing*, vol. 148, pp. 107164, 2021.

[19] H. Chen, D. Long, H. Zou, S. Wang, J. Tao, and X. Qiu, "Improving near-field spatial uniformity of secondary sources for active noise control headrests," *Applied Acoustics*, vol. 230, pp. 110437, 2025.

[20] S. Goose, L. Riddle, C. Fuller, T. Gupta, and A. Marcus, "Paz: In-vehicle personalized audio zones," *IEEE MultiMedia*, vol. 23, no. 4, pp. 32–41, 2015.

[21] J. Linjama and V. Välimäki, "Immersive personal sound using a surface nearfield source," in *153rd Convention of Audio Engineering Society*. Audio Engineering Society, 2022.

[22] H. Oppermann and S. Checa, "Sonic opportunities presented by personalized sound zones," in *Audio Engineering Society Conference: AES 2022 International Automotive Audio Conference*. Audio Engineering Society, 2022.

[23] O. Kirkeby, P. A. Nelson, H. Hamada, and F. Orduna-Bustamante, "Fast Deconvolution of Multichannel Systems using Regularization," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 189–194, Mar. 1998.

[24] E. Stanhope, L.J. Hobden, and S.G. Oxnard, "Near-field binaural rendering: Evaluating the natural channel separation of loudspeakers mounted in a headrest," in *Audio Engineering Society Conference: AES 2023 International Conference on Spatial and Immersive Audio*. Audio Engineering Society, 2023.

[25] B. B. Bauer, "Stereophonic Earphones and Binaural Loudspeakers," *J. Audio Eng. Soc.*, vol. 9, no. 2, pp. 148–151, Apr. 1961.

[26] L. Lim and C. Kyriakakis, "Multirate Adaptive Filtering for Immersive Audio," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Salt Lake City, UT, USA, May 2001, vol. 5, pp. 3357–3360.

[27] R. Glasgal, "360° localization via 4.x RACE processing," in *Proc. of 123rd Audio Engineering Society Convention*, New York, USA, Oct. 2007.

[28] A. Vancheri, T. Leidi, T. Heeb, L. Grossi, N. Spagnoli, and D. Weiss, "Geometrical acoustics approach to cross talk cancellation," in *Audio Engineering Society Convention 152*. Audio Engineering Society, 2022.

[29] A. Vancheri, T. Leidi, T. Heeb, L. Grossi, and N. Spagoli, "Multiband time-domain crosstalk cancellation," in *153rd Audio Engineering Society Convention*. Audio Engineering Society, 2022.

[30] A. Vancheri, T. Leidi, T. Heeb, L. Grossi, and N. Spagnoli, "Dynamic adaptation in geometrical acoustic CTC," in *Proceedings of the 154th Audio Engineering Society Convention*. Audio Engineering Society, 2023.

[31] ITU-R BS. 1284-2, "General methods for the subjective assessment of sound quality ," 2019.

[32] M. Schoeffler, S. Bartoschek, F. R. Stöter, M. Roess, S. Westphal, B. Edler, and J. Herre, "webMUSHRA—A comprehensive framework for web-based listening tests," *Journal of Open Research Software*, vol. 6, no. 1, 2018.

[33] A. Farina, A. Bellini, E. Armelloni, et al., "Implementation of cross-talk canceling filters with warped structures-subjective evaluation of the loudspeaker reproduction of stereo recordings," *Proc. of SHARC2000, Boston*, pp. 11–13, 2000.

[34] ITU-R BS. 1534, "Method for subjective listening tests of intermediate audio quality," 2001.

[35] Juhan Nam, Miriam A Kolar, and Jonathan S Abel, "On the minimum-phase nature of head-related transfer functions," in *125th Convention of the Audio Engineering Society*. Audio Engineering Society, 2008.