INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2024, VOL. 70, NO. 2, PP. 355-360
Manuscript received February 20, 2024; revised June, 2024. doi: 10.24425/ijet.2024.149552

# Sound localisation definition in parametrically and non-parametrically decoded first-order ambisonic systems

Jacek Majer

***Abstract*—This study assessed sound localisation definition in ambisonic systems using two-non-parametric and three parametric decoders, in a two-dimensional format. The sound samples were played back through eight loudspeakers arranged in a circle. The participants compared pairs of sound samples to determine which sample offered a more precise perception of the sound source's location. The data analysis, using a Bradley-Terry probability mode, revealed that parametric decoders were preferred with a 60–83% probability. Among the parametric decoders, the COMPASS method, which utilizes the Multiple Signal Classification algorithm for sound source direction estimation, received the highest scores for sound localisation judgements.**

***Keywords*—ambisonics; spatial audio; parametric decoding; psychoacoustics**

## I. Introduction

AMBISONICS, primarily developed in 1970s by Michael Gerzon [1], is a method of 3D sound spatialisation, using directional components. The components are derived by means of decomposition into a series of spherical harmonics. The number of terms in the series can be truncated to a finite order *N*, providing a flexible representation in terms of bandwidth and fidelity of reproduction. The objective of playback is to faithfully reproduce the original sound field, whether in a one-dimensional (two-dimensional variant) or multi-plane (three dimensional variant) configuration. This requires the use of at least $2N+1$ or $(N+1)^2$ speakers accordingly.

Common ambisonic systems are typically limited by low spatial resolutions, as they often utilize small speaker arrays and first-order ambisonic channel formats. However, a subset of parametric methods is available [2]-[4] that can significantly enhance perceptual performance in such systems. Instead of static reproduction, only most essential auditory information (termed as spatial parameters) is rendered, based on time varying soundfield analysis and synthesis stages. This allows for bandwidth reduction during transmission, without compromising spatial quality. Parametric systems also offer higher perceptual quality during playback and greater flexibility with respect to the speaker system in use. This flexibility is achieved through the synthesis of the extracted parameters.

Although non-parametric decoding methods do not take spatial parameters into account, their design is deeply grounded in sound localisation theory. To enable a comparison with the parametric approach, the following section provides a detailed explanation of this theory. In the subsequent section, parametric methods and their current implementations are introduced. Following this, a brief review of published listening experiments is presented. Finally, the article discusses the results and offers the author's conclusions based on his conducted experiment.

### *A. Sound Reproduction in Ambisonic Systems*

Spherical microphone arrays are used to capture the 3D sound field during the process of recording (encoding) ambisonic channels [5], [6]. The transducers and their directional characteristics are usually closely related to spherical harmonics of particular order. For example, the first order of ambisonics requires $N = 4$ harmonics: one omnidirectional and three figure-of-eight patterns, which are orthogonal to each other. This configuration enables the storage of information about the sound's position within the available channels. Different sound directions and distances correspond to different relative phases and amplitudes in these channels.

The encoding process can also be generated synthetically [6]. Synthetic coding involves the transformation of existing recordings, usually monophonic and anechoic, into ambisonic format. This procedure requires a virtual representation of the array in the form of a virtual microphone and information about the direction of sound arrival. This technique is mainly used for sound scene presentation with single or multiple virtual sounds sources.

The process of converting ambisonic channels into speaker feeds for playback, also referred to as the decoding stage, relies on sound localisation theory [7]. The aim is to recreate, as accurately as possible, binaural localisation cues [8] for listeners, while taking into consideration the size of the listening area, the number of speakers, and their placement. Interaural intensity difference (IID) and interaural time difference (ITD), which are frequency dependent cues, determine the position of the sound source. When wavelengths are shorter than the diameter of the listener's head, an acoustic shadow begins to form, resulting in differences in sound intensity between the ears. For longer wavelengths, the time difference of arrival at the listener's ears becomes more significant.

To enhance directional sound reproduction through loudspeakers, this theory was reformulated by Gerzon resulting

The Author is with Chopin University of Music, Department of Sound Engineering, Chair of Musical Acoustics and Multimedia, Warszawa, Poland (e-mail: jacek.majer@chopin.edu.pl).

in what is known as the Makita theory [9] and Energy Vector theory [7]. The apparent direction of sound for low frequencies is defined as the directional vector $\widehat{r_V}$, referred to as Makita localisation, and is based on interaural time difference. The directional vector ($\widehat{r_V}$) can be derived by summing complex signals representing sound pressure and velocity from each speaker.

For high frequencies, where IID is significant, the directional vector $\widehat{r_E}$, known as energy vector localisation, represents the perceived source position. The directional vector ($\widehat{r_E}$) is calculated by summing signals representing energy from each speaker. Additionally, both $\widehat{r_V}$ and $\widehat{r_E}$ have associate magnitude quantities, denoted as $r_V$ and $r_E$, and their values depend on the number of sources. In the presence of a single natural source, $r_E$ is always equal to 1. However, if there are two or more sources, such as in sound reproduction systems with multiple loudspeakers, the values will always be less than 1.

Based on this information, the main decoder design guidelines can be summarized as follows. 1) To accurately reproduce sound from a specific direction, the $\widehat{r_V}$ and $\widehat{r_E}$ should conicide; 2) The $r_V$ should be equal to 1 and $r_E$ should be as close to 1 as possible. The first guideline implies a regular speaker placement scheme, where all speakers are equally distant from the centre and are placed in diametrically opposite pairs. This applies to both horizontal placement and multiple planes. The $r_V$ and $r_E$ values can also be influenced by changing the proportions between velocity channels and pressure channels. The norm of the energy vector $\|r_E\|$ can also be used to objectively describe the blur width of the reproduced sound source [10], as an angle $\alpha_E \ = \mathrm{acos}(r_E)$. For every finite order *N* and listening area size, there exists a theoretical upper limit frequency for the exact reproduction of the soundfield. As the order increases, the constraints on the listening area and limit frequency decrease. This, in turn, leads to energy vector values $r_E$ that are closer to 1 and, consequently, lower values of the blur width $\alpha_E$.

In terms of signal processing, the reproduction method is both linear and time-invariant. During the reproduction phase, sound pressure and velocity are encoded into ambisonic channels by summing weighted signals from a microphone array. These weights are specific to the array design. Similarly, during the decoding process, each speaker feed results from a weighted sum of all ambisonic channels. The weights are derived from a described decoder design. The values of these weights remain constant throughout playback. There are no parameters that would influence the behaviour of the encoding or the decoding processes because they do not depend on information about the sound field itself. Therefore, this method is also known as non-parametric.

### B. *Parametric decoding*

The main principle of ambisonic systems employing parametric methods is described by the Directional Audio Coding (DirAC) method [2]. First-order ambisonic channels, containing recorded or artificially generated sound fields, undergo analysis in both the time and frequency domains, followed by the extraction of discrete directional parameters. This processing is guided by a sound field model that describes two spatial components: the sound source (a single plane wave) and diffuse sound. These components are rendered for an arbitrary sound system in a manner that preserves the appropriate spatial cues for the listener.

The method has proven to be more effective compared to non-parametric decoding, although it can produce audible sound distortion, mainly due to non-ideal methods for sound source estimation and the precision of time-frequency analysis. Incorrect parameter assignments occur when analysing complex sound fields. This can happen, for instance, when significantly early reflections are present or when multiple sources are being registered.

Solutions for this problem have been formulated in works related to Higher-order Directional Coding (HO-Dirac) [3] and Coding and Multidirectional Parametrization of Ambisonic Sound Scenes [4] (COMPASS) methods for parametric decoding. The HO-Dirac method employs the same estimation methods as DirAC but supports higher-order ambisonic channels, providing more detailed directional information. This method is less error-prone and can estimate a greater number of spatial parameters, by dividing the analysed sound field into non-overlapping sectors. The COMPASS method estimates spatial parameters based on the general subspace principle of array processing, derived from speech enhancement methods. Unlike the direct microphone signals, this method supports variable-order ambisonic signals.

Subjective comparisons were made between HO-Dirac and COMPASS, and non-parametric first and higher-order ambisonic systems [3], [4]. The conducted listening experiments evaluated reproduction accuracy, using a Multiple-Stimulus with Hidden Reference and Anchor (MUSHRA) method [11]. Synthetic coding was used to generate ambisonic input of first and higher orders, containing sound scenes with variable number of virtual sound sources. Anechoic and reverberant conditions were simulated. The results revealed that when using the same order of ambisonic input, both parametric methods enhanced the overall perceived quality compared to non-parametric methods.

## II. EXPERIMENT

### A. *Objective of the experiment*

The objective of the experiment was to compare the localisation definition of sound sources rendered by non-parametric and parametric decoding methods. The term *localisation definition* is used because the participants were explicitly comparing the uncertainty in sound source localisation between two sources. Instead of employing methods for determining source directions, the experiment sought participant's preferences regarding sound direction ambiguity for the sake of simplicity and experiment robustness.

### B. *Preliminary listening session*

Before the experiment, a listening session was conducted with the aim of collecting spatial sound attributes that best described the perceived differences between non-parametric and parametric systems. The goal was to identify attributes that would be easy to assess in a comparison of both decoding methods. Two listening experts participated in the session.

The attributes that showed the most pronounced differences were:

- Localisation definition of sound sources

- Stability of sound source positions during head movements
- Perceived depth/distance of sound sources

It was decided that the attribute to be evaluated in the experiment will be the localisation definition of sound sources.

### C. *Reproduction system*

A two-dimensional ambisonic system was used, consisting of eight PSI M14 broadband studio monitors evenly placed in a ring configuration ($r$ = 2m, $\phi$ = 45°), in an acoustically treated listening room. The room had an average reverberation time, T30, of 0.2 s. No additional processing, such as time delay alignment or speaker equalization, was applied.

Five types of decoding methods were evaluated, as presented in Table 1, using First Order Ambisonics (FOA) as an input. Non-parametric methods, based on the design principles described in Section I, utilized the Sampling Ambisonic Decoder (SAD) [6], with both the original and psychoacoustically improved weightings, referred to as basic and max $r_E$ weighting. Parametric decoding employed the original HO-DirAC method and COMPASS method. Both algorithms for Direction of Arrival (DoA) estimation implemented in the COMPASS decoder were tested: Multiple Signal Classification (MUSIC) and Estimation of Signal Parameters via Rotational Invariance Techniques (ESPIRIT).

The decoding and reverb simulation were performed using the REAPER Digital Audio Workstation with the SPARTA and COMPASS audio plugin suite [12]. All the processing related to reproduction was conducted in real time. The experiment was controlled by a custom script, written in the ReaScript environment [13] which adjusted playback parameters and received listeners responses as MIDI messages via a controller.

TABLE I
EVALUATED DECODING METHODS

| Name | Decoding method | Method type |
|---|---|---|
| basic | Sampling Ambisonic Decoder (SAD) | Non-parametric |
| maxre | Sampling Ambisonic Decoder with energy preserving weighting for high band | Non-parametric |
| HO_Dir | HO-Dirac | Parametric |
| CompA, CompB | Compass | Parametric, two DoA estimators, ESPIRIT (CompA) and MUSIC (CompB) |

### D. *Stimuli*

Since only playback methods were essential for the presentation, synthetic encoding was preferred over encoding real sound sources. In synthetic encoding, real ambisonic recordings were replaced by monophonic recordings processed by a virtual microphone. Ambisonic channels of arbitrary order could be generated based on given sound directions and the ideal directivity patterns of the virtual microphone. In the experiment, a virtual microphone generating FOA channels was used. Virtual sound sources were created at six uniformly spread positions (0°, ±60°, ±120°, 180°). These sources were presented in two diffusion modes: with and without simulated early reflections.

Three anechoic, monophonic recordings were used. These were sourced from the Bang & Olufsen *Music for Archimedes* sound library [14] and were selected based on their proportion of transient and steady-state characteristics: 1) Speech (Danish female speaker, featuring a mix of transient and steady-state characteristics). 2) Stringed instrument (acoustic guitar, primarily steady-state). 3) Percussion instrument (conga, primarily transient).

An additional listening session with an experienced expert was conducted to equalise the loudness levels for all combinations of sound type, position, reverberation, and decoder type. The procedure involved matching stimuli against one arbitrary chosen reference that represented comfortable loudness level for listening.

### E. *Listeners*

The experiment was run on nine listeners, comprising both males and females with normal hearing. The participants were sound engineering students from the Chopin University of Music in Warsaw and possessed extensive prior experience in critical sound evaluation.

### F. *Procedure*

The listening sessions were conducted using a within-pair comparison sound evaluation procedure. Each trial consisted of two stimuli played in sequence. Listeners were instructed to provide one of two possible responses, indicating which of the samples in the pair more precisely defined the localisation of virtual sound sources. They could respond with either “A is better” or “B is better.” Ties, such as “A is the same as B,” were not permitted in the experiment. In cases of uncertainty, listeners were instructed to provide their guess. There was no time limit for providing a response, and the next trial began after receiving the answer. The total trial presentation had a duration of 7 s, with 3 s allocated to each sound in the pair and 1 s for a break between the sounds.

The set of stimuli included 360 test items: (3 sound types × 6 sound positions × 2 diffusion modes × 10 combinations of decoded pairs). To account for the possibility of constant error, both AB and BA pairs were presented, resulting in a total of 720 items. The decoding method for each item was randomized during presentation, while other parameters were randomized for each trial. This ensured that the virtual sound source direction, diffusion mode, and sound type were consistent within compared sounds. The listening position was in the centre of the system. During the session, the listeners were instructed not to move and to maintain their orientation toward the front of the system (0°direction)

Each listener participated in a short training session and then completed three series of judgements of the subset of 240 items, one for each sound type. The listeners were free to take short breaks between the trials. The duration of one session was approximately 40 to 50 minutes.

## III. RESULTS

The Bradley-Terry probability model [15], [16] was used to create ranking of decoding methods, based on the results of paired comparison judgements. According to this model, the probability of A being preferred to B is determined by equation (1):

$$P(A > B) = \frac{\pi_A}{\pi_A + \pi_B}, A \neq B \qquad (1)$$

The $\pi_n$ values, where $n$ is the number of ranked objects, are known as scale parameters, estimated by maximizing the log-likelihood function, based on the empirical probability of preference.

Figure 1 shows the compiled results for the Maximum Likelihood Estimation (MLE) of $\pi$ parameters for all listeners. It includes subsets of stimuli with and without reflections.

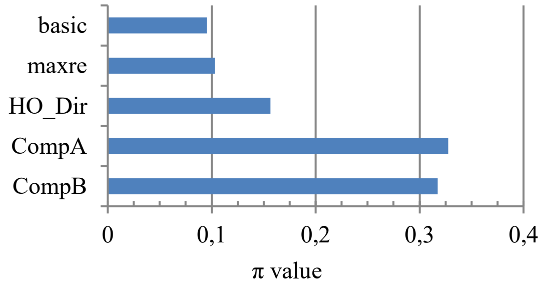

Fig. 1. Compiled values of π parameter for all listeners

The estimated $\pi$ values for non-parametric methods were approximately of 0.1. For the HO_Dir decoder the range of values fell between 0.1 and 0.23, and for COMPASS it ranged from 0.25 to 0.4. The largest differences in $\pi$ values were observed in the subset with reflections, whereas the smallest differences were found in the subset without reflections.

The same results, presented as probabilities of preference, are shown in Table 2. Matching decoder A from a row to a decoder B from a column reveals the probability of preference of A over B. The probability values were as follows: 1) between 67% and 83% for COMPASS compared to basic and max $r_E$ comparisons, 2) between 53% and 69% for HO-Dirac compared to basic and max $r_E$ comparisons, 3) between 52% and 81%. for COMPASS comparing to HO-Dirac, and 4) between 47% and 53% for basic and max $r_E$ comparisons.

TABLE II
PROBABILITIES OF DECODING METHOD PREFERENCE (ALL STIMULI)

| P (A > B) | basic | maxre | CompA | CompB | HO_Dir |
|---|---|---|---|---|---|
| basic | - | 48% | 23% | 23% | 38% |
| maxre | 52% | - | 24% | 25% | 40% |
| CompA | 77% | 76% | - | 51% | 68% |
| CompB | 77% | 75% | 49% | - | 67% |
| HO_Dir | 62% | 60% | 32% | 33% | - |

Figure 2 shows the compiled results for MLE of $\pi$ parameters for all listeners. The results are shown for a subset of stimuli without reflections, and for a subset with reflections. The corresponding probabilities of preference are shown in Table 3. In both cases, non-parametric methods yielded similar values. However, the HO-Dirac decoder had considerably lower scores when reflections were present, while the opposite trend was observed for the COMPASS decoder.

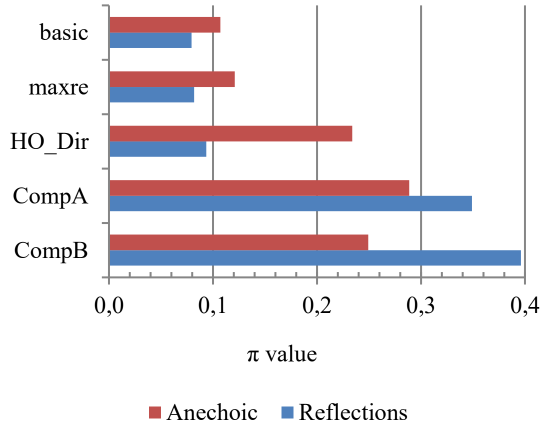

Fig. 2. Compiled values of π parameter for all listeners presented separately for anechoic and reverberant stimuli

Figure 3 shows the results for MLE of $\pi$ parameters for all listeners, categorized by the type of sound. All methods showed relatively similar values. Notably, the most substantial differences in $\pi$ values were observed for the HO-Dirac and Conga sound types.

Normalized $\pi$ parameter values for different sound source directions are shown in Figures 4-6, both for subsets of stimuli without reflections and for subsets with reflections. In each graph, the parameter value is presented relative to the distance from the centre. Across all decoding methods, the values exhibit a uniform distribution with minor to moderate deviations. The least variability in $\pi$ values regarding the direction was observed when employing the COMPASS B method with reflections present. In Figure 5, more consistent judgements between decoders are evident for anechoic stimuli originating from a direction of 120°.

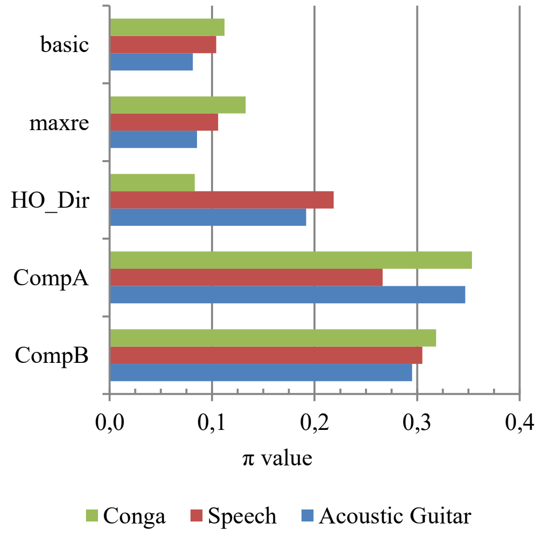

Fig. 3. Compiled values of π parameter for all listeners, presented separately for different types of sounds

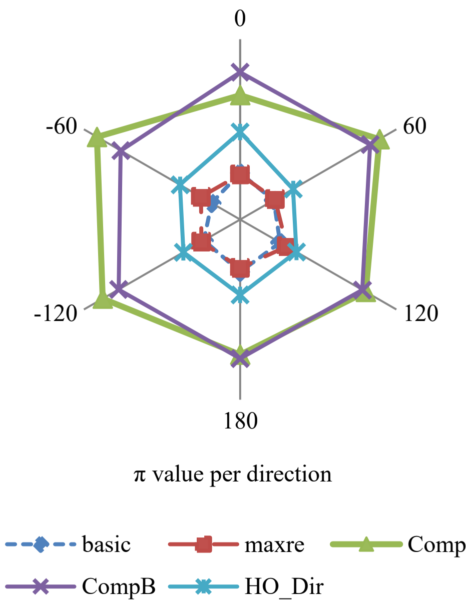

Fig. 4. Normalized π parameter values for different sound source directions, all results

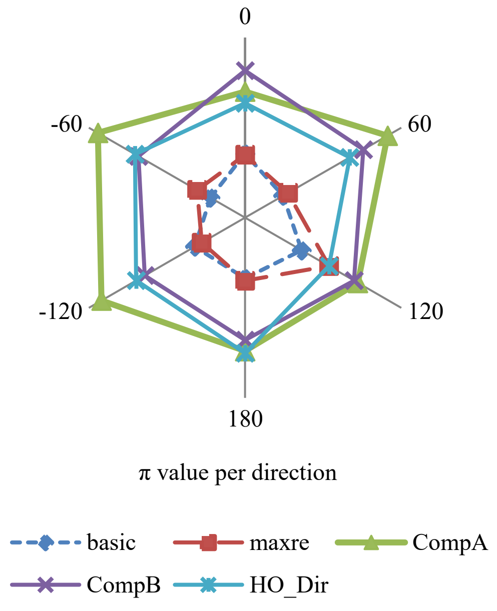

Fig. 5. Normalized π parameter values for different sound source directions, for stimuli without reflections

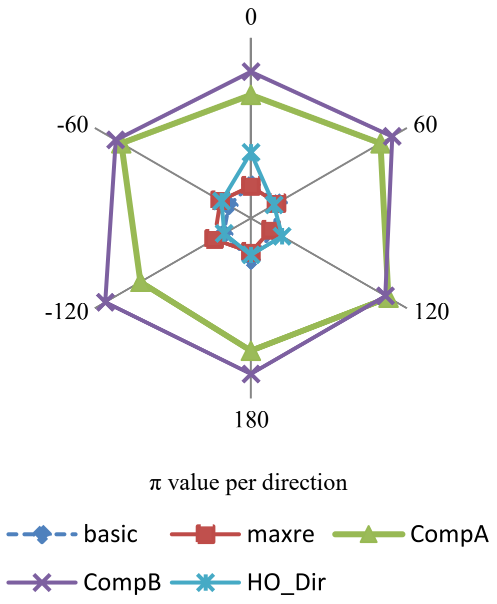

Fig. 6. Normalized π parameter values for different sound source directions, for stimuli with reflections

## IV. Discussion

When $\sum \pi_n = 1$, the data are statistically uniform, and no significant preferences occur between objects if $\pi_1 = ... = \pi_n = 1/n$. This hypothesis was tested against the alternative $\pi_i \neq \pi_j, i \neq j, i, j = 1, ..., n$ ,using the chi-squared test statistic ($df$ = 4, α = 0.05) and all the results were above critical value.

The results demonstrate a moderate to substantial preference for parametric methods over non-parametric methods when evaluating the localisation definition of virtual sound sources. This is in agreement with the findings of the experiment conducted for HO-Dirac and confirms the subjective COMPASS results for the loudspeaker system, which were previously obtained solely from a binaural system.

The subset of sources with present reflections exhibits noticeably more consistent judgements, while the opposite is true for sources without reverberation. This is likely attributed to variations in the quality of Direction of Arrival (DoA) estimates, which differ among parametric decoding methods and results in audible differences. This differences are considerably more pronounced when compared to stimuli without reflections. The preference is also associated with audible alterations in timbre, as reported by listeners during the

Table III
Probabilities of decoding method preference (anechoic and reflections stimuli separated)

| P (A > B) | Anechoic | | | | | Reflections | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | basic | maxre | CompA | CompB | HO_Dir | basic | maxre | CompA | CompB | HO_Dir |
| basic | - | 47% | 27% | 30% | 31% | - | 49% | 19% | 17% | 46% |
| maxre | 53% | - | 30% | 33% | 34% | 51% | - | 19% | 17% | 47% |
| CompA | 73% | 70% | - | 54% | 55% | 81% | 81% | - | 47% | 79% |
| CompB | 70% | 67% | 46% | - | 52% | 83% | 83% | 53% | - | 81% |
| HO_Dir | 69% | 66% | 45% | 48% | - | 54% | 53% | 21% | 19% | - |

experiment. Stimuli with a high proportion of transients to steady-state sound, such as conga, were reported to have the most pronounced audible distortions.

No significant difference in preference was observed when comparing non-parametric methods. According to Zotter's research [17], the max $r_E$ decoder demonstrated the best overall performance. However, in this case, the perceptual differences may be negligible as this experiment employed only first order Ambisonics, while the publication evaluated performance with 5th order.

Similar results in decoder preference were obtained across all sound directions. Unlike the accuracy of localisation judgements, which decreases for lateral directions, the ability to judge localisation definition appeared to remain consistent. When asked, listeners did not report experiencing fatigue when evaluating localisation definition from directions behind them.

The consistent judgements observed at 120° directions in Fig. 4, are probably related natural reflections, that were present in the listening room.

## V. Conclusions

Based on the conducted experiment, the following conclusions can be drawn regarding evaluation of non-parametric and parametric decoding methods:

In general, parametric methods were preferred over non-parametric ones in the horizontal loudspeaker rendering of virtual sound scenes, containing single sound sources. However, when early reflections were simulated, similar judgements in the localisation definition of sources were observed for the HO-Dirac method and non-parametric methods. When comparing HO-Dirac and COMPASS, the latter method was consistently preferred, regardless of sound type, diffusion and the evaluated directions in the experiment. A significant difference between DoA estimators in COMPASS was noted when reflections were present.

The preferences for decoding methods remained consistent across sound source directions, except for the anechoic variant and a single direction where natural room reflections might have an influence.

## Acknowledgements

The author wishes to thank Barbara Okoń-Makowska and Andrzej Miśkiewicz for their valuable contributions in shaping the final design of the listening experiment. Special thanks are also extended to Joanna Szczepańska-Antosik for her assistance during the preliminary listening sessions.

## References

[1] M. A. Gerzon, “Periphony: With-Height Sound Reproduction,” J. Audio Eng. Soc., vol. 21, pp. 2–10, 1973.

[2] V. Pulkki, “Spatial Sound Reproduction with Directional Audio Coding,” J. Audio Eng. Soc., vol. 55, pp. 503-516, 2007.

[3] A. Politis, J. Vilkamo and V. Pulkki, "Sector-Based Parametric Sound Field Reproduction in the Spherical Harmonic Domain," in IEEE Journal of Selected Topics in Signal Processing, vol. 9, no. 5, pp. 852-866, 2015, https://doi.org/10.1109/JSTSP.2015.2415762

[4] A. Politis, S. Tervo and V. Pulkki, "COMPASS: Coding and Multidirectional Parameterization of Ambisonic Sound Scenes," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 2018, pp. 6802-6806, https://doi.org/10.1109/ICASSP.2018.8462608

[5] S. Moreau, J. Daniel, and S. Bertet, “3D Sound Field Recording With Higher Order Ambisonics–Objective Measurements and Validation of a 4th Order Spherical Microphone,” presented at the 120th Convention of the Audio Engineering Society (2006 May), paper 6857

[6] F. Zotter, M. Frank “Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality”, 2019. https://doi.org/10.1007/978-3-030-17207-7

[7] M. A. Gerzon, “General Metatheory of Auditory Localisation,” presented at the 92th Convention of the Audio Engineering Society (1992 March), paper 3306

[8] T. Letowski and S. Letowski, “Localisation Error: Accuracy and Precision of Auditory Localisation,” in Advances in Sound Localisation. InTech, Apr. 11, 2011. https://doi.org/10.5772/15652

[9] Makita, Y. “On the Directional Localisation of Sound in the Stereophonic Sound Field”, EBU Review, part A no. 73, pp.102-8, 1962

[10] Bertet, S., Daniel, J., Parizet, E., and Warusfel, O., “Investigation on localisation accuracy for first and higher order ambisonics reproduced sound sources,” Acta Acustica united with Acustica, 99(4), pp. 642–657, 2013, ISSN 1610-1928, https://doi.org/10.3813/AAA.918643

[11] ITU-R Recommendation BS.1534-1, “Method for the subjective assessment of intermediate quality levels of coding systems,” Tech. Rep., International Telecommunication Union (ITU), Geneva, Switzerland, 2003

[12] https://leomccormack.github.io/sparta-site/docs/plugins/compass-suite/

[13] https://www.reaper.fm/sdk/reascript/reascript.php

[14] Bang and Olufsen. Music for Archimedes. CD B&O 101, 1992

[15] R. A. Bradley and M. E. Terry, “Rank analysis of incomplete block designs: I. The method of paired comparisons,” Biometrika, vol. 39, pp.324–345, 1952. https://doi.org/10.2307/2334029

[16] J. S. Lee, "Paired comparison for subjective multimedia quality assessment: Theory and practice," 2013 IEEE International Symposium on Circuits and Systems (ISCAS), Beijing, China, 2013, pp. 1099-1102, https://doi.org/10.1109/ISCAS.2013.6572042

[17] M. Frank and F. Zotter, “Localisation experiments using different 2D Ambisonics decoders,” in 25th Tonmeistertagung, Leipzig, 2008.