INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2024, VOL. 70, NO. 2, PP. 307-313
Manuscript received February 14, 2024; revised June, 2024. doi: 10.24425/ijet.2024.149546

Ambisonics' setup quality assessment through measurements and computations based on ITD and ILD functions using subbands and a Gammatone Filter Bank

Marcin Dąbrowski, Jan Skorupa, Wojciech Raszewski, and Maciej Głowiak

***Abstract*—Poznan Supercomputing and Networking Center (PSNC) developed an ambisonic installation and workflow as part of audio-visual 8K VR 360° immersive media experiments. This work aimed to investigate the quality of performance of the PSNC setup through both subjective tests as well as simulations providing objective parameters of interaural characteristics in a real-life scenario of PSNC studio. For the objective part, an algorithm for angle estimation has been proposed and computations were performed.**

***Keywords*—immersive audio; sound localization; ambisonics**

I. Introduction

AMBISONICS is a spatial sound reproduction method based on sound field deconstruction (recording and processing) and reconstruction (audio field synthesis during playback). This technique uses the multipole expansion method with spherical harmonics. It is a more accurate method in terms of virtual sound source separation and direction of arrival estimation compared to other immersive audio techniques such as quadrophonics. Ambisonics is used for music playback with accurate instrument separation, subjective human hearing condition tests, ambient noise analysis as well as in the field of psychoacoustics.

First order ambisonics consists of four independent components, which allow spherical acoustic field pressure reproduction. It can be implemented by four channel microphone recordings including one omnidirectional microphone and three bipolar pattern microphones set in X, Y, Z axes for capturing sound directional properties. Higher order ambisonics uses more components according to the formula $N = (l + 1)^2$ where N is the number of ambisonic B format components and l is the order. For a given playback setup of N_s speakers, the number of components shall satisfy the inequality $N \leq N_s$.

In our experiment $N_s = 24$ and $N = 16$, $l = 3$. Order 4^{th} or higher would need more speakers than available for our measurements.

II. Aim and Scope of Work

Poznan Supercomputing and Networking Center (PSNC) developed an ambisonic installation as part of audio-visual 8K VR 360° workflow for immersive media experiments. As a participant in Immersia TV and Immersify research projects, PSNC produced audiovisual content including the creation of 360° videos with surround sound. The ambisonics technique was chosen in order to achieve immersive audio with good angular resolution without the need of wearing headphones. In the Immersive project we combined virtual reality video with ambisonic audio and real time head position tracking. Thanks to ambisonics, we could localize virtual sound sources synchronously to head movements.

Sound quality using the target installation has not been analyzed as part of above-mentioned projects. Therefore, it was decided to evaluate the quality of the PSNC setup through both objective measurements of interaural characteristics as well as through subjective angle estimation in a real-life scenario at the studio.

In this work we used ambisonics for test pulses playback. Sound analysis was performed in binaural domain by the test participants (subjective part) as well as binaural recordings using a dummy head and a postprocessing algorithm we developed (objective part).

It was intended that the tests would answer the question whether the sound quality of the sound installation allows the reproduction of ambisonic recordings in the full azimuthal range with acceptable angular accuracy not worse than 10°.

III. Ambisonic Setup

The ambisonic installation under test is located at the PSNC main TV studio, which is a rectangular room acoustically isolated from external noise and partially equipped with sound absorbing surfaces and wooden dividers. However, due to large windows, the recording conditions are far from anechoic chamber. The multichannel audio interfaces for ambisonic signal production and other noise sources were generating

This work was supported by research and innovation project called Immersify which has been financed by European Commission in Horizon 2020 program.

Marcin Dąbrowski, Jan Skorupa, Wojciech Raszewski, Maciej Głowiak are with Institute of Bioorganic Chemistry of the Polish Academy of Sciences Poznan Supercomputing and Networking Center, Poland (e-mail: mdabrowski@man.poznan.pl, jskorupa@man.poznan.pl, wraszewski@man.poznan.pl, mac@man.poznan.pl).

audible noise at the studio of the level not greater than 25 dB_{SPL}. The ambisonic auditory space consists of 24 active studio monitors Genelec 8010A and one subwoofer Genelec 7350A, which gives a 24.1 speaker setup. Speakers were located on three heights of elevations -33°, 0°, 33°. This setup provided full sphere studio monitoring space.

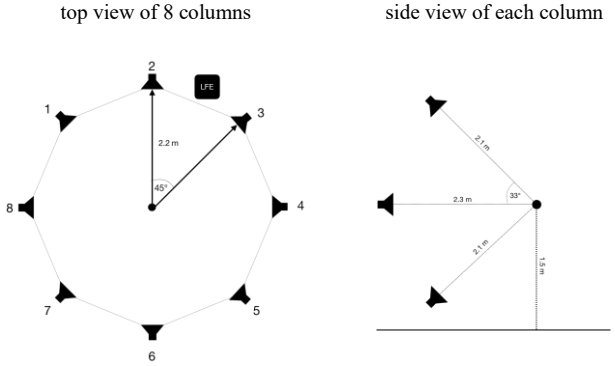

Fig. 1. Ambisonic 24.1 speaker setup of average radius 2.2 m

All speakers were directed towards the center of the sphere and the radius from the center to each speaker is around 2.2 m. For the measurements the 24.0 layout was used, the tested installation is shown in Figures 1 and 3A.

Signal workflow was based on Focusrite RedNet 3 audio interface, which was connected through ADAT protocol to four Behringer ADA8200 preamplifiers [1]. The interface was also equipped with Dante protocol, through which it was connected to the Focusrite RedNet PCIe card (see Figure 2). This connection provided up to 32 physical audio outputs with sample rate up to 48 kHz.

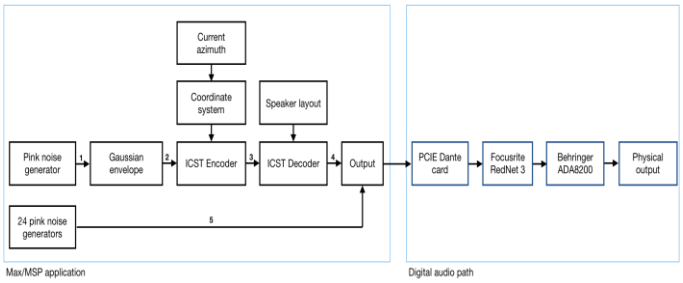

Fig. 2. Ambisonic audio path

For the purpose of measurements, we built an application using Max/MSP environment with ICST Ambisonics externals for Max/MSP [2, 3]. In order to get sufficient perceived spatial precision, pulses were encoded into 3rd order Ambix B-format with ACN channel ordering and SN3D weighting type. This format ensures the coefficients never exceed amplitude of 0th order component, which helps to avoid clipping effects and distortion of test signals [3]. The whole audio path was decoded into 24 channels.

IV. Experiment

As a test signal we used pink noise pulses (i.e. of spectral density $1/f$) with Gaussian envelope of duration 0.2 s. We used this kind of noise in order to get a wideband pulse. Pink noise was also chosen to imitate natural noise sources. Using sinusoidal pulses would produce too narrow bandwidth, which could cause directional ambiguities.

A virtual sound source was generated as 3rd order ambisonic ambix B-format. For simplicity, stimuli were located at elevation 0°. Adding the elevation degree of freedom will be the next step of our study. The test signals occupied frequencies in the range ca. 100 Hz (peak power) – 20 kHz (-80 dB), however, the mean signal frequency was around 1.5 kHz in terms of spectrum power density weighting.

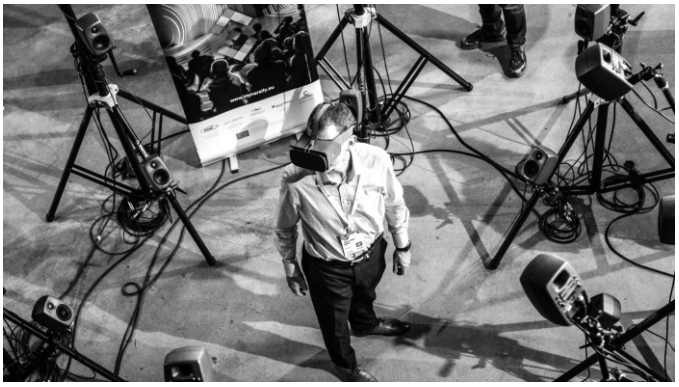

Fig. 3. PSNC immersive audiovisual demonstration with virtual reality mixed with ambisonic sound from 24 speakers, The Networking Conference, Tallin 2019

Our test participants were PSNC employees, 2 females and 8 males of the age in the range 24 – 40 years. In this group we had both experts in sound postproduction as well as representatives of professions unrelated to sound engineering. None of the participants reported any hearing erroneousness.

During the test, only the examinator (test manager) and one participant were present in the studio. Participants were sitting with the center of their head at so called sweet spot (sphere center). In front of the listener, we placed a table 0.65 m far from the sweet spot. On the table, we placed a clock face with a protractor (see Figure 4).

Participant was asked to turn the arrow in the perceived direction of sound every time a pulse was heard. A camera was located above the protractor and angles were read from the camera image. At the beginning of the test participants were trained how to point angles using an exercise series of not recorded ten random example angles.

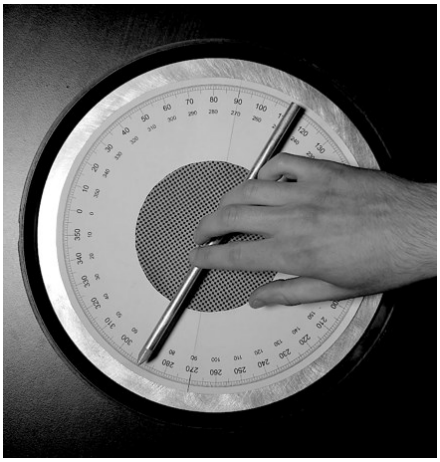

Fig. 4. Protractor during a subjective test

In order to minimize conscious direction analysis done by the participants and catch only the *low-level* hearing sense characteristics without overthinking, the interval between pulses was the shortest possible to allow protractor arrow positioning without discomfort. During the measurements the best value turned out to be 3 s between consecutive pulses.

One test series consisted of 32 pulses synthesized in a pseudorandom order; the pre-generated pseudorandom angle series is shown in Figure 5.

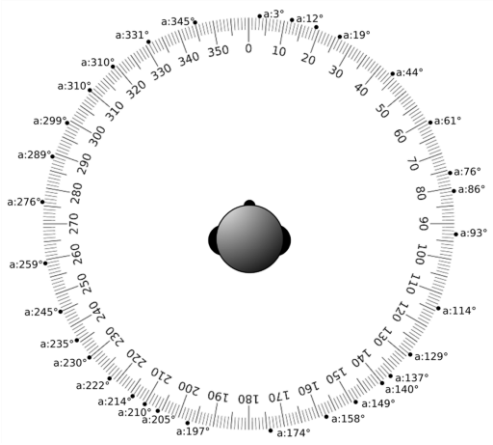

Fig. 5. Pseudorandom angle series of zero elevation

After each test signal set participants were given a hearing-relaxing break in the form of out-of-phase pink noise emitted independently from each speaker. This was because some participants reported irritability of long series of pulses. At the time of relaxation, the test manager was changing the position of the protractor table and camera in accordance to the proper reference frame position. In order to compensate the TV studio room characteristics, the reference frame was sequentially placed in four different positions (0°, 90°, 180°, 270°) relative to the room geometric axis. For each of these rotations, every 32-sample angle series was repeated three times: for head position 0° (a), -45° (b) and +45° (c), relative to the reference frame zero direction (Figure 6).

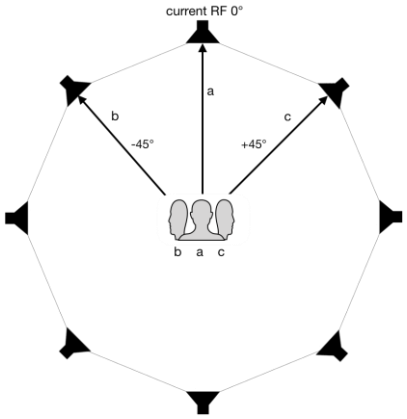

Fig. 6. Angle series (a, b, c) repeated for three head positions

This tri-angle procedure was needed as we noticed during preliminary experiments, that some participants could not properly estimate directions in the range 90° – 270° when their head was fixed at 0°. Better results were obtained when the head was kept in a number of fixed positions or was moving [4]. In order to simulate such movement, head was held in three fixed positions, i.e. the participants were instructed to direct their eyes to the proper pivots of the speaker stands.

Four reference frame positions and three hearing angles produce twelve repetitions of each angle from the pseudorandom series. Full test for a participant or dummy head consisted of $32 \cdot 12 = 384$ pulses, which took ca. 30 minutes including relaxing breaks.

We acquired the angles from the video files with a 1-5° precision, depending on how the participant was fluttering the arrow due to their uncertainty.

Due to the distance from the protractor to the participant's head center, a compensating formula was derived for the indicated-desired angle difference:

$$\delta = \tan^{-1}\frac{d\sin\alpha}{R - d\cos\alpha} \tag{1}$$

where α is the indicated angle, $d = 0.6$ m is the distance between the head center and the arrow center and $R = 2.2$ m is the radius of the ambisonic installation. The desired angle, which is the angle the participant intended to indicate is obtained from the formula:

$$\alpha_d = \alpha - \gamma \cdot \delta \tag{2}$$

where γ is an additional correction coefficient. The angle compensation was of considerable importance around 90° where the error due to lack of such compensation would reach 15.2°.

We noticed that some participants tried to compensate for the head-protractor distance effect intuitively and unconsciously. Sometimes better results were obtained when we used γ between 0.5 – 0.6 instead of the theoretical value 1. We also found that the optimum γ value minimizing the overall root mean square error for all participants was 0.55, however the shape of the angle-error function was affected. Ultimately for further calculations we fixed the gamma coefficient at the value 1 in order to get smoother error function shapes.

Neumann KU 100 dummy head was used for binaural recording using standard earlobes and standard capacitor microphones powered by phantom power. We use this dummy head with microphones placed inside it to record signals resembling those impinging human cochlea. Left and right microphones receive different acoustic pressure in terms of phase and amplitude due to the human head shape and distance between ears.

The high-pass filter inside the head was switched off and it was further assumed for simplicity, that the frequency and phase responses of the head microphones are flat. The head was motionless within a measurement series. The recorder connected to the head produced .wav files of sampling rate 48 kHz and bit-depth 24 (21 μs per sample), however during experiments we noticed that 96 kHz would be more sufficient rate for time difference estimation as the time difference of one sample corresponds to average human time difference precision of 10.4 μs.

V. ANGLE ESTIMATION ALGORITHM

The proposed algorithm was designed to resemble natural human hearing. The original left and right microphone signals were partitioned into subbands (audible spectrum segments) using 12-step ERB (equivalent rectangular bandwidth) space and a gammatone [5] filter bank using Octave scripts. Gammatone filters are audio linear filters with impulse responses modelling the operation of basilar membrane, which is part of ear cochlea. The subband center frequencies are shown in Figure 7. The ERB scale was proposed to reflect constant distances on the basilar membrane [6]. Partitioning of the audible spectrum into ERB subbands and using gammatone filter bank resembles the operation of human cochlea and auditory nerve.

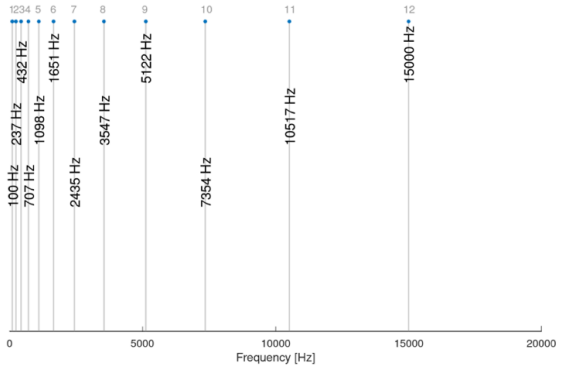

Fig. 7. Subband center frequencies using ERB scale

Two parameters can be obtained from binaural recordings in order to estimate the angle of arrival in a similar way the human hearing does: ITD (*Interaural Time Difference*) and ILD (*Interaural Level Difference*) [6]. Interaural time difference is the incoming sound arrival time difference between ears, whereas interaural level difference is the sound level difference between signals recorded from dummy head microphones. ITD and ILD enable incoming wave angle estimation, their values result from scattering and shadowing effects due to the shape and size of human head.

A number of theoretical and numerical models of ITD and ILD function shapes can be found. For ITD three formulas for low and high frequencies can be found in literature [8, 9, 10, 11]:

- Rayleigh, frequencies > 4 kHz:

$$\mathrm{ITD}(\theta) = 10^6 \cdot 2 \cdot {}^{a}/_{c} \cdot \sin\theta\ [\mu s], \quad \theta = 0 \dots 2\pi \tag{3}$$

- Kuhn, frequencies < 4 kHz:

$$\mathrm{ITD}(\theta) = 10^6 \cdot 3 \cdot {}^{a}/_{c} \cdot \sin\theta\ [\mu s], \quad \theta = 0 \dots 2\pi \tag{4}$$

- Woodworth:

$$\mathrm{ITD}(\theta) = 10^6 \cdot {}^{a}/_{c} \cdot (\theta + \sin\theta)\ [\mu s], \quad \theta = -\pi/2 \dots \pi/2 \tag{5}$$

where θ is the angle of arrival, $a = 0.085$ m is the average head radius and $c = 344$ m/s is the average speed of sound at 20°C.

There are also simplified models for non-zero elevations φ [12], however we focused on zero elevation problem here and left the elevation domain for further work.

Our preliminary measurement and calculation results showed that the best ITD fit was given by the Rayleigh formula for broadband signals. ILD was modelled by a formula:

$$\mathrm{ILD}(\theta) = 1 + (f[\mathrm{Hz}]/1000)^{0.8} \cdot \sin\theta, \quad \theta = 0 \dots 2\pi \tag{6}$$

where f[Hz] is the center frequency of a broadband signal or a subband. Our measurement results have shown the best performance of this approximation was for subbands 9-12 (5-15 kHz). According to [7], human brains use ILD with best results above 3 kHz in the range 0-30 dB with the resolution of 1 dB. In case of our dummy head, the ITD and ILD shapes are not ideally sinusoidal and symmetrical around 90° and 270° and are shown in Figure 8.

ITD was calculated using maximum cross correlation between the left and right channels of the broadband signal composed of all subbands. We also tried to use certain subsets of subbands, however during our experiments it turned out that ITD calculations for any subband subsets or even single subbands always gave less accurate results than for broadband signals.

ILD was calculated for each subband in order to have a number of samples to calculate further median values. During the experiment, we noticed that subbands 9-12 were the most useful for this purpose. Thus, for each sample pulse, we could determine 5 parameters: ITD and four ILDs coming from subbands 9-12. Using precalculated lookup tables (LUTs) composed of spline approximations, we could find angle estimations. ITD and ILD curves were approximated using periodic spline fitting (Figure 8) with the number of breaks in the range 10-12.

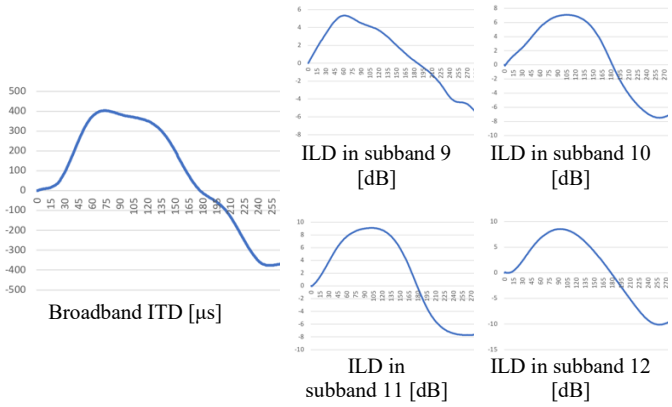

Fig. 8. ITD and ILDs approximation based on all measurements done with the dummy head, curves used further on as reference patterns for angle estimation

As ITD and ILD functions are not bijections for both spherical and a real head model, there are ambiguities in LUT value picking. To overcome this problem, we propose a method to distinguish between 90° - β, 90° + β, 270° - β and 270° + β angles. Two subsets of angle estimations, i.e. LUT readouts, are created:

- subset 1: scan LUT from 0° towards 90° for positive measured ITD/ILD values and scan from 360° towards 270° for negative,

- subset 2: scan LUT from 180° towards 90° for positive measured ITD/ILD values and scan from 180° towards 270° for negative.

Doing so we obtain readouts from both sides of sinusoidal shapes extrema. Now we have to make a decision which side of the extremum is more likely; we take the one for which we get the smaller value of the approximation standard deviation based on five samples (ITD and four subband ILDs). This method has proven to be correct as the quality of the algorithm is good compared to human accuracy.

VI. Measurement Results

We could observe considerable directional skills differences between participants in terms of angle error MAE (mean absolute error), RMSE (root mean square error) or mean standard deviation. Some participants had good MAE but high standard deviation whereas some the opposite.

In order to compare the results, we propose a measure of erroneousness in logarithmic scale (we name it $dB_{MAE\sigma}$):

$$1/Q = 10 \log \frac{\mathrm{MAE} \cdot \sigma}{\mathrm{MAE}_0 \cdot \sigma_0} \qquad (7)$$

where the reference values are $MAE_0 = 10°$ and $\sigma_0 = 10°$ and were taken arbitrarily by the authors.

The erroneousness ranking is presented in Table I. The worst participant was a male of the age 37 and he has shown an interesting directional erroneousness feature. For the majority of pulse samples, he could not properly distinguish between angles symmetrical to the 90° - 270° line (actual angle in the form 90° - β, 90° + β, 270° - β and 270° + β), i.e. he reported the actual angle 130° close to 50°, etc. This participant has not ever experienced any directional hearing erroneousness in his everyday life. This problem was unveiled using our specific workflow and pulse envelope and duration. We suppose that using a shorter pulse or a pulse of different characteristics, would mitigate the observed phenomenon.

TABLE I
ERRONEOUSNESS VALUES FOR HUMAN TEST PARTICIPANTS (P01-P10) AND DUMMY HEAD (DH)

Part. ID	$1/Q$ [$dB_{MAE\sigma}$]
P01	3.37
P02	4.11
DH	***6.09***
P03	6.63
P04	6.82
P05	6.90
P06	7.00
P07	8.08
P08	8.21
P09	10.13
P10	13.41

The same phenomenon was observed with dummy head data if we used symmetrical ITD and ILD approximations (e.g. sinus), the directional ambiguity in such case was of the same nature to the function symmetry around 90° and 270°. Human head breaks this symmetry thanks to shape of head, including earlobes, nose and eyes.

The dummy head result, due to relatively high standard deviation of angle estimations, achieved third position in terms of erroneousness (see Table I) although our algorithm gave least estimation error (see Table II). Participant IDs were assigned with respect to the 1/Q ranking position in Table I.

In Table II we present MAE and standard deviation of angle estimation. Results in Tables I and II were calculated using all 384 samples per participant.

Performance of human hearing in terms of directivity was the worst for all participants around 140°-160° and 200°-220°, that is also the region of the worst ambiguity effect for participant P10. It needs further investigation on how it is dependent on the setup or workflow we use. What is noticeable, although the error reaches a minimum around 0°, standard deviation reaches a minimum around 90° and 270°, thanks to ITD and ILD maxima.

TABLE II
MAE AND STANDARD DEVIATION VALUES FOR HUMAN TEST PARTICIPANTS (P01-P10) AND DUMMY HEAD (DH); RESULTS SORTED BY MAE; PARTICIPANT IDS AS IN TABLE I

Part. ID	Mean Absolute Error [°]	σ [°]
DH	***6.0***	***67.8***
P01	9.2	23.6
P02	12.1	21.3
P05	12.6	39.0
P03	13.0	35.5
P04	13.8	34.8
P06	14.5	34.6
P09	14.6	70.4
P07	19.7	32.7
P08	19.8	33.5
P10	32.6	67.2

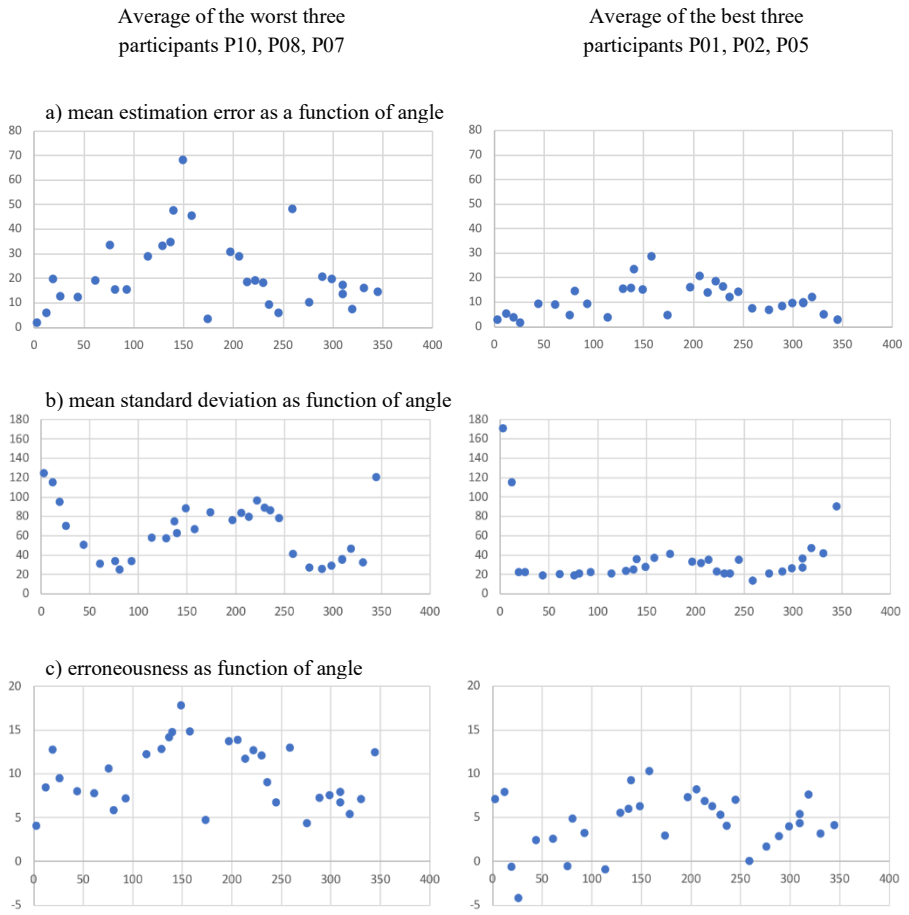

Fig. 9. Results for worst and best human participants as a function of angle

If we treat all participants' data as one dataset of one virtual participant, the total MAE is 9.7° with a standard deviation 49.3°. This gives us the erroneousness 6.8 dB. Compared to our algorithm (1/Q = 6.09 dB) we can say that the objective method using the dummy head is better than the human sense.

Mean error values fulfilled our expectations and literature values [7, 13, 8], however we are concerned about relatively high standard deviations, which needs more signal generation workflow analysis.

VII. Outcome and Further Research

Our research shows that the proposed setup is capable of reproducing of high-quality sound field with accurate directivity close to human natural hearing. In our setup, the perceived average angular accuracy was in the range of 5-10°, whereas the human hearing in conditions close to an anechoic chamber is approximately 7° [14]. However, ambisonic sound creators shall avoid angles between 150° and 270° due to high angle estimation errors. Additionally, due to the ambiguity of 0° and 180° angles for some listeners, such sources shall be avoided as well.

We are planning further investigation of perceived angle ambiguity errors, whether it came from the studio acoustic properties or erroneousness of the signal generation equipment or software. So far in our experiments, all virtual sources were placed in the horizontal plane at zero elevation. In further research, we will focus on investigating the impact of vertical virtual sources positions on the angle detection accuracy. We will also answer the question if using additional postprocessing effects imitating acoustic properties of various room types in terms of reverberance, can improve localization of virtual sources perception in the range of 150° and 270°.

References

[1] J. Skorupa, M. Głowiak, Content production guidelines: Ambisonics Recordings and Postproduction, Immersify project deliverable, https://immersify.eu/home/guidelines-reports/ambisonic-sound-production/, 2019

[2] J. C. Schacher, P. Kocher, Ambisonics Spatialization Tools for Max/MSP, 2006.

[3] J. C. Schacher, Seven Years of ICST Ambisonics Tools for MaxMSP – A Brief Report, Proc. Of the 2nd International Symposium on Ambisonics and Spherical Acoustics, 2010.

[4] H. Wallach, The Role of Head Movement and Vestibular and Visual Cues in Sound Localization, Journal of Experimental Psychology, vol. 27, no. 4, 1940. https://doi.org/10.1037/h0054629

[5] A.G. Katsiamis, E.M. Drakakis, Lyon R.F., Practical Gammatone-Like Filters for Auditory Processing, EURASIP Journal on Audio, Speech, and Music Processing, 2007.

[6] B. C. J. Moore, Development and Current Status of the "Cambridge" Loudness Models, Trends in Hearing, 2014. https://doi.org/10.1177%2F2331216516682698

[7] X. Zhong, Dynamic Spatial Hearing by Human and Robot Listeners, PhD Dissertation, Arizona State University, 2015.

[8] R. Sridhar, E. Y. Choueiri, Capturing the elevation dependence of interaural time difference with an extension of the spherical-head model, AES 139th Convention Papers, 2015.

[9] N. L. Aaronson, W. M. Hartmann, Testing, correcting, and extending the Woodworth model for interaural time difference, Journal of the Acoustical Society of America vol.135 no. 2, 2014. https://doi.org/10.1121/1.4861243

[10] J. Estrella, On the Extraction of Interaural Time Differences from Binaural Room Impulse Responses, TU Berlin 2010.

[11] J. Wall, Post-Cochlear Auditory Modelling for Sound Localisation using Bio-Inspired Techniques, Doctor of Philosophy thesis, pp. 98-99, 2010

[12] J. Huopaniemi, Virtual Acoustics and 3-D Sound in Multimedia Signal Processing, Doctor of Science degree dissertation, pp. 75-76, 1999

[13] B. C. J. Moore, A. Kolarik, M. A. Sonte, Evaluation of a method for enhancing interaural level differences at low frequencies, The Journal of the Acoustical Society of America, vol. 140 no.4, 2016. https://doi.org/10.1121/1.4965299

[14] J.C. Makous, J.C. Middlebrooks., Two-dimensional sound localization by human listeners, The Journal of the Acoustical Society of America, 1990. https://doi.org/10.1121/1.399186