Immersion Control with Loudspeaker Reproduction of Stereo Recordings

JOHN G. BEERENDS[1], *AES Fellow*

RICHARD VAN EVERDINGEN[2]

EELCO GRIMM[3], *AES Fellow*

ANTAL VAN NIE[3]

JASPER OUDEJANS[3],

AMAN NUNDA[3],

BOSSE NEIJMAN[3],

MELLE BLITS[3],

ULMT VAN DER LINDEN[4]

HANS VAN MAANEN[5]

[1] TNO, The Hague, The Netherlands

[2] Delta Sigma Consultancy, Nunspeet, The Netherlands

[3] HKU University of the Arts, Utrecht, The Netherlands

[4] Helios Pro Audio Solutions, Haarlem, The Netherlands

[5] Temporal Coherence, Huizen, The Netherlands

The last decades have seen an ever-increasing number of audio channels for recording and playback, ranging from 5.1 towards 22.2 and more, mostly focused on improved surround localization and/or immersion. This paper shows that reproduction of stereo recordings can be improved significantly without complex upscaling algorithms and with no need for more than two additional surround loudspeakers. The basic insight is that for an optimal sense of immersion, a diffuse field should be created that has a minimal contribution to the direct field and should not cause localization problems. Subjective experiments were carried out that show that subjects highly appreciate the adaptable diffuse field approach as presented in this paper. There is however a

large variation in the preferred diffuse field level; some subjects prefer volumes close to the just noticeable difference, while others choose levels that are louder than the direct field.

0. INTRODUCTION

According to Wikipedia, 'Hi-Fi' [1] is high-quality reproduction of sound without audible noise and distortion, based on a flat (neutral, uncolored) frequency response within the human hearing range. Technically, these requirements can be easily met today, even with moderately priced consumer equipment. Note that this definition does not mention the number of channels required in the recording and playback system. For artificially generated sound, one should have some idea of the rendering and the number of audio channels necessary for a Hi-Fi playback experience. For recorded sound, we should have some guidance for the number of audio channels required in both the recording and playback system.

One could argue that having only two ears leads to the conclusion that we just need two audio channels. However, reconstructing the two-ear-signals that we experience in a live event is extremely difficult; we need to take into account individual HRTFs (Head Related Transfer Functions) and head movements. One could also argue that when only a single object is to be recorded, just one channel is sufficient. Take for example a single voice: Can we make a recording that allows for a natural reproduction? Yes, just make a mono recording of the voice in an anechoic room and play it back through a single loudspeaker with the same directional properties as the voice. If we listen to that recording in a room, we have a full Six-Degrees-of-Freedom (6DoF) [2] reproduction. The whole chain of capture and reproduction is transparent.

If we try to apply the idea of transparency between recording and playback to a set of musical instruments, we run into trouble. Firstly, musical instruments can have wild directivity patterns, so recording in an anechoic room will result in a spectral imbalance. This inequity will also propagate in the artificial reverberation that is often added in recordings that have a "dry" overall sound quality. Secondly, with large orchestras, we need a large number of anechoic mono recordings that have to be played back over at least the same number of correctly placed loudspeakers. And finally, we

should ask the question whether this way of thinking is the correct Hi-Fi approach for large orchestra's that do not fit in our living room. For the recording of music events, one should try to capture a similar feeling as one would have experienced at the live event, with the acoustics of the recording room providing proper acoustic integration across all instruments, including their directional patterns. In playback, we should then try to reproduce the sound field as would have been experienced in the live event, especially taking into account the feeling of immersion. This shows that there are two approaches in the Hi-Fi recording/playback chain: One focused on the transparency 'Here and Now', the augmented reality approach, and one on the transparency 'There and Then', the virtual reality approach. Both require completely different, incompatible, recording/playback techniques.

If we aim for the illusion 'There and Then', what is the minimum number of audio channels required for Hi-Fi quality loudspeaker reproduction? The introduction of stereo significantly improved the perceived loudspeaker reproduction quality of music events compared to mono. For headphone reproduction we do not need to extend the number of channels, although a high quality, transparent recording-reproduction chain requires personalized HRTF corrections, including compensations for head movements. In loudspeaker reproduction, personalized HRTF corrections are unnecessary, while the quality is determined by other characteristics of which the acoustics of the listening room is a dominating factor. For a high-quality experience, room modes should be damped and we should aim for a low reverberation time, preferably less than 0.5 seconds, allowing to hear the longer echoes of recording rooms. If the reverberation time of the listening room is extremely low, such as in an anechoic room, we will experience a lack of feeling immersed by the sound and to compensate for that effect, an extremely large number of audio channels would be required.

The number of audio channels needed for Hi-Fi quality loudspeaker reproduction of music events in a living room environment is not clear. While the expansion of the number of recording/playback channels from 1 to 2 – mono to stereo – was a great improvement, the extension from 2 to 4 – stereo to quadrophony – was a failure. Further expansion to more than a dozen channels (5.1 up to 22.2 or even higher) as in many surround systems, seems inconsistent

with the characteristics that improve music reproduction. Advanced systems like Dolby Atmos, Sony RA360, Auro3D, DTS-X, high order Ambisonics, Wave Field Synthesis or Object-Based Audio Coding ([3], [4], [5], [6], [7]), are mainly useful in films and games, where sound effects require a more exact localization. To ensure correct binaural and monaural decolorization in these environments, specific recording techniques are required that often forces one to use unconventional recording techniques, e.g. as used by foley artists.

While for movies and games sound localization is important, the reproduction of music is seldom improved by adding more reproduction channels. The feeling of being immersed by a natural sounding diffuse field is much more important than an improved localization. In fact, adding more reproduction channels can lead to uncontrolled degradations that can be characterized as “hearing things jumping around”.

While the number of audio channels on consumer equipment has been growing, recording of music has remained mainly in stereo. Although many upscaling algorithms have been developed, it turns out to be difficult to create a high-quality, immersive diffuse field. Complex algorithms can be formulated (see e.g. [8], [9]), possibly extended with a center channel, but in general, two-to-five up mixing algorithms provide a poorer front image quality and only a marginal improvement in immersion. In most cases, the original stereo reproduction is preferred [10].

This paper describes a simple method that optimizes the sense of immersion, while using ordinary stereo recordings. The design philosophy is based on the understanding that immersion does not require localization. In fact, it is even a distracting factor when listening to recordings of medium to large orchestras. In our opinion, this preference is based on evolution. Nearby objects have low levels of diffuse field and can be exactly localized, often indicating a dangerous situation that requires action. For objects further away, this is the opposite and these are therefore perceived as safe, leading to a higher sense of comfort.

The diffuse field approach we will present allows to achieve controllable levels of immersion that sound better than advanced multi-channel recording/playback systems and upmixing algorithms. The strength of the approach is that it allows to control the sense of immersion in such

a way that it can be easily adapted to the content being played and to one's own personal preference.

1. HISTORIC BACKGROUND

Immersion is a highly cognitive term that was only recently introduced in the world of sound reproduction over loudspeakers. The main problem is that it is related to sound quality and therefore difficult to define [11]. In general, quality has two different dimensions: function and beauty. Quality optimization usually starts with the former. An excellent car, for example, should never fail in its function of transportation; it must fulfill it with high reliability. Once that is achieved, the focus shifts towards the beauty dimension. But because that is in the eye of the beholder, it is difficult to quantize and optimize.

In sound quality research, the focus has thus been more on the functional quality aspect of localization and to a lesser extent on immersion. The first studies related to the latter were carried out in the context of speech perception, where being immersed in a (cocktail) party leads to functional difficulties in the understanding of a single voice [12], [13]. This aspect of immersion is thus related to the functional localization aspect, where one tries to optimize speech intelligibility. The same could be said about the increased number of audio channels in the aforementioned advanced audio systems that focus on localization accuracy.

The beauty aspect of immersion has only been studied more recently. Eaton and Lee [14] asked experts to rate ten aspects of sound quality in relation to immersion. Horizontal envelopment was found to be more important than vertical. What remains unclear, however, is to what extent subjects prefer to be immersed by a sound. Take the example we gave in the introduction: the reproduction of an anechoic recorded single voice, reproduced by a single loudspeaker, leading to full Six-Degrees-of-Freedom (6DoF) [2] reproduction. If we play that recording over a standard stereo setup, or over four loudspeakers using quadrophony (the big mono approach), the feeling of immersion improves with the number of loudspeakers used, but not the perceived sound quality.

An increase in immersion is expected to be observed if a natural surrounding diffuse field is created. Four direct-radiating loudspeakers are, due to their properties, not very suitable for this purpose.

Many designers recognize the importance of widespread directivity of loudspeakers for high-quality music reproduction. In general, the sense of immersion is less with classic direct radiating speakers than with omnidirectional designs. To improve the feeling of immersion, designers have used additional drivers that do not radiate directly towards the listener. The best-known is probably Amar Bose who designed an enclosure that has additional drivers in the back panel to produce reflections against the walls, thereby improving the balance between direct and diffuse field [15]. Kantor and De Koster from Acoustic Research extended this idea and developed an enclosure that use extra backwards radiating drivers to equalize the diffuse field room response independently from the direct field [16]. The first listed author of this paper demonstrated the quality improvement of this approach in 1988 for manufacturer BNS, at the Dutch Firato exhibition [17]. An extra set of back radiating loudspeakers that can be added to any regular stereo setup was used to equalize the diffuse field response. However, the weakness of all these setups is that they primarily create a frontally localized diffuse field.

In our view, the sound quality of a stereo loudspeaker setup that is focused on the ‘There and Then’ illusion should try to produce a natural surround sounding diffuse field focused on the reproduction of the ambient information of the recording. A very early attempt to do so from a stereo recording was formulated by Madsen [18], who based his ideas on the work of Lauridsen [19] and Damaske [20]. Madsen performed experiments under anechoic conditions and found that in a standard stereo setup, the recorded ambiance information is partly masked by the recorded direct sound. That loss can be countered by simultaneously playing the same stereophonic signal through an extra pair of loudspeakers behind the listener, with a short delay. Madsen concludes: “As long as the delay is shorter than the shortest time for which the discrimination of sound pulses is possible, yet long enough to obtain the necessary incoherence, the reverberation will have the desired spatial character. Moreover, the localization of the direct sound image will not change, but a reconstruction of the acoustical character of the original recording site is obtained.”

In his famous book “Sound Reproduction: Loudspeakers and Rooms”, Floyd Toole [21] states that delayed versions of sounds originating from the front soundstage create a feeling of immersion, but he does not provide insight on how to achieve high quality immersion with loudspeakers. Tomlinson Holman [22] does some suggestions. He writes: “An alternative to the use of an array of monopole loudspeakers facing the listening area and attempting to produce a diffuse field from multiple uncorrelated sources, is to use purpose-built loudspeakers which emphasize the diffuse-field component of their output to reproduce the reverberant field, as reflected sound in the reproducing space.” This proposal steers away from the (semi) anechoic approach of Madsen and instead uses reflections on the listening room walls to produce a more diffuse sound field that envelops the listener. Damaske [20] already noted that for a pleasant spatial impression, an incoherent field is preferred over a coherent one, so maximizing the amount of reflections via the listening room walls seems like a good strategy.

While new standards focus on an ever increasing number of reproduction channels [23], allowing improved localization, we are convinced that excellent quality of immersion can be achieved using ordinary stereo recordings that are reproduced by a regular stereo loudspeaker setup, complemented by only two additional surround loudspeakers of omnidirectional design, that project most of their sound energy towards the walls. In the next section, we investigate a setup based on this proposal, aiming to extract ambience information from the original recording, while allowing to adjust the amount of experienced immersion to the properties of the recording and to one’s own personal preference.

2. EXPERIMENTAL SETUP

The Immersion Control design is initiated by the observation that the sense of immersion of stereo recordings can be improved by the addition of two specifically constructed surround loudspeakers that simply reproduce a slightly lower volume of the Left and Right front loudspeakers [24]. These speakers should be designed to ideally contribute only to the diffuse field, so that the degree of immersion can be easily controlled without introducing localization errors. This also

reduces undesired comb filtering effects between front and rear. A simple, but effective way to create a diffuse radiating surround speaker is to use a cone-shaped diffuser that creates, for a substantial part, a 360 degrees horizontal radiation pattern. In the optimal construction, the speaker is designed to minimize its contribution to the direct field, for example by limiting the radiation to about 300 degrees (see Figure 1).

Figure 2 shows the layout of the complete speaker setup. Note that the Left and Right Diffuse Surround speaker mainly radiate towards the walls of the listening room, opposed to the standard surround setups where the surround speakers radiate towards the listener [10], [25]. This approach prevents localization degradations that can be characterized as “hearing things jumping around”. The setup is limited in its capability to create a perfect diffuse field due to its focus on the horizontal plane, but in general the feeling of immersion is dominated by sound in that direction, compared to vertical [14].

2.1. Timing

In the initial system design, care should be taken that the surround loudspeakers have the same propagation time as the front speakers. In our setup, the distance between all four speakers and the listener was the same. For other arrangements, one can use electronic delays to compensate unequal distances. To keep the front stereo image stable, and avoid localisation of the direct field in the rear loudspeakers, the audio signal for the surround speakers should be given an extra delay. As Madsen noted, this also enlarges the incoherence between front and back, which results in a more diffused sound field. According to the well-known ‘Haas effect’ [26], the recommended additional delay time is between 10 and 20 ms. Damaske [27] found an optimum for noise pulses of varying length at approximately 15 ms. Informal assessment of different delays showed that the optimal value is dependent on the acoustic properties of the room where the recording was made. Roughly speaking, more delay could be allowed for recordings in large concert halls than for dry pop recordings. In order to keep the task of the subjects simple, they were

only asked to set the preferred level of the diffuse field loudspeakers. The delay was set to the fixed estimated global optimal value of 17 ms.

Starting point in our experiment is a standard stereo setup with an angle between 40 and 60 degrees between the two, equidistant, front loudspeakers. The placement of the rear surround speakers is less critical and informal assessment of different placements showed that angles up to 120 degrees are acceptable, although the optimal value is expected to be lower than 100 degrees.

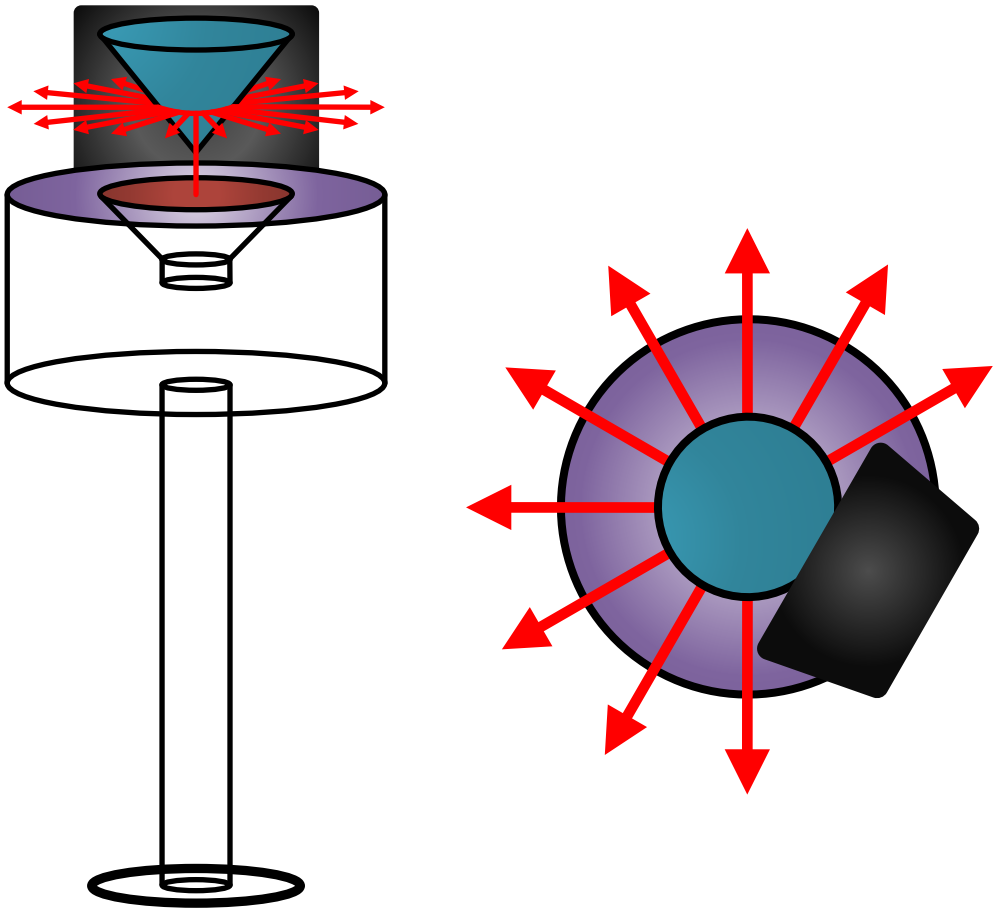

Figure 1. Side and top view of the surround loudspeaker used to create a diffuse field behind the listener. The grey block is attached to the mounting of the cone and is covered with absorbent material that provides shielding of the direct sound, so that dispersion is mainly limited to around 300 degrees omnidirectional.

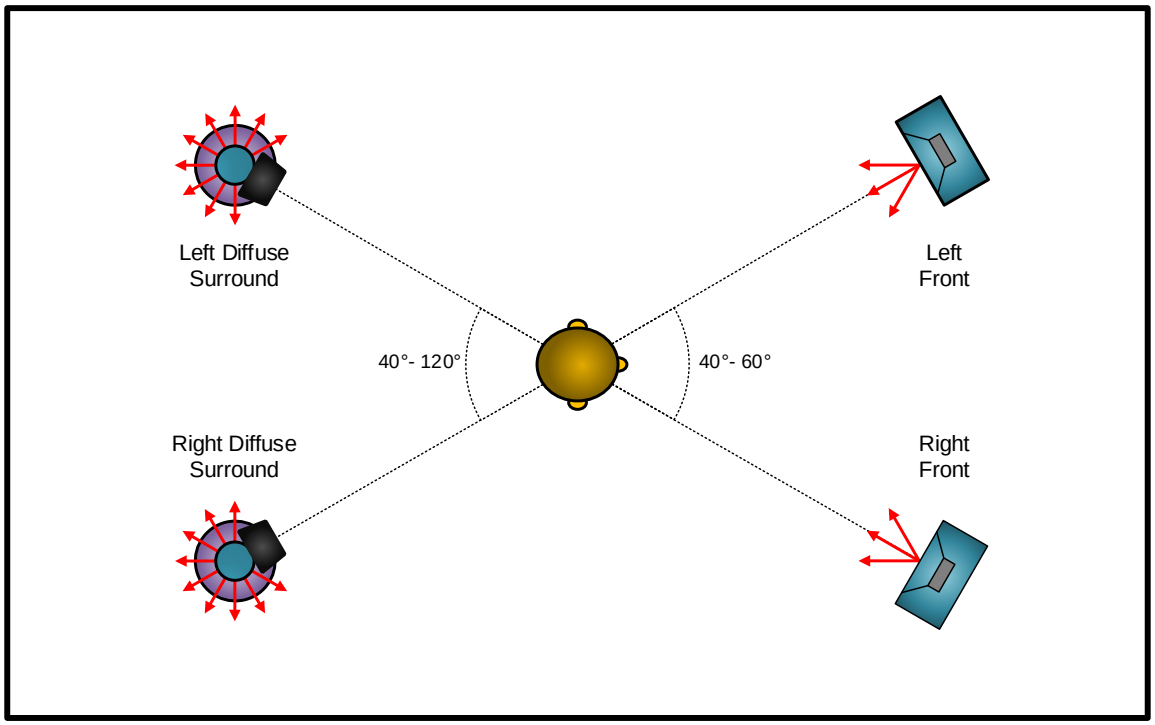

Figure 2. Loudspeaker setup that allows for Immersion Control by use of the cone-shaped diffuser of Figure 1, that mainly contributes to the surround diffuse field.

3. SUBJECTIVE TEST

3.1. Experimental Procedure

The goal of the subjective test procedure is to find diffuse field settings that lead to the best perceived overall sound quality of an ordinary stereo recording. As explained in section 1, the relation between immersion and quality is unclear, while maximizing one does not necessarily mean optimizing the other. During the test, subjects became confused when asked to rate both aspects. Therefore, they were asked to give their opinion only on the overall perceived sound quality.

Two subjective tests were carried out in two different, low reverberation acoustic environments, each with its own equipment. The first test used a total of 13 subjects, 3 female and 9 male, with ages between 18 and 68 year, consisting of naïve subjects, professional audio engineers and HiFi enthusiasts. A total of 11 subjects were used for the second test; the six authors from the HKU plus five of its students. Ages between 18 and 57 years, all male.

Each subject who participated was tested individually using seven different audio fragments, ranging from dry pop to lively classic concert hall recordings. Three fragments were taken from the EBU Subjective Quality Assessment Material [28], developed for evaluating stereo recordings (see Table 1). Each audio fragment was between 10 and 60 seconds long and was played in a loop for as long as it took the subject to form an opinion on the overall sound quality improvement with the diffuse field filler. The audio files were normalized to a level of –23 LUFS (Loudness Units Full Scale). When played through the stereo loudspeaker set, the maximum level difference between the loudest parts of each of the eight fragments was about 3 dBA. The range that could be set for the diffuse field fillers was between –20 and +3 dBA. The former was the estimated level of the just noticeable difference, the latter was configured for safe play back. At each setting, a mute switch allowed subjects to turn off the diffuse field fillers completely to form a well-founded opinion about their preference.

Table 1. Audio fragments used in the subjective evaluation.

	COMPOSER/PERFORMER	**TRACK**	**SOURCE**
1	Abba	Head over heels	EBU SQAM track 69
2	Beethoven	3e Symfonie Berlin Staatskapelle / Suitner	HiFi news and record review track 10
3	Haydn	Trumpet Concerto in E-flat Major Hob. VIIe:1:III Allegro	EBU SQAM track 55
4	Sara K.	Stars	Single track download
5	Stevie Ray Vaughan	Chitlins con carne	Single track download
6	Vivaldi	Concerto in Re maggiore, RV93	BIS CD290 track 1
7	Vocal	Quartet	EBU SQAM track 48

Each subjective test began by playing all seven audio fragments, only through the two stereo front loudspeakers. Subjects were asked to set the overall playback volume to their preferred level and whether they would like to adjust the loudness of any individual track. One subject requested an individual level adaptation of –2 dB for the Abba fragment. In the next step of the experiment, each individual fragment was continuously repeated (loop playback) and subjects were asked to adjust the level of the diffuse surround speakers for maximum perceived overall sound quality. If the volume of the diffuse surround speakers were set to a level that significantly increased the overall loudness (>2 dBA), subjects were asked to re-evaluate the direct field level setting to check whether this rise was related to their preference. This procedure prevents the loudness level from affecting the preference score.

When subjects reached their preferred volume setting for the diffuse field speakers, they were asked to rate their opinion on the improvement in perceived overall sound quality using a 5-point evaluation scale, where 0 represented turning off the rear diffuse surround speakers completely (see Table 2).

Table 2. Scale values used in the evaluation of the improvement in perceived overall sound quality.

SCALE VALUE	QUALITY IMPROVEMENT RATING
0	None
1	Very small
2	Small
3	Fair
4	Big
5	Very big

3.2. Equipment

Figure 3 illustrates the equipment used in the first experiment. The transparent blocks show the existing reference system of the first listening room, normally in use for high-end audio demonstrations. The source was a Personal Computer running as a Digital Audio Workstation. No requirements or conditions had been set from the project on the composition of this primary

equipment, other than to use the best that the facility could make available. In this experiment, both loudspeaker sets were passive and connected to their dedicated power amplifier.

Intentionally, the gear of the diffuse field loudspeakers – shown as grey blocks – was configured as a free-standing addition, thereby eliminating any possible influence on the original setup, while keeping track of the volume of the main system. The input signal therefore was connected to the outputs of the power amplifier, in parallel with the front loudspeakers, via a balanced attenuator network. This signal was converted to 24-bits digital audio with a sampling rate of 192 kHz and fed to an AV-receiver. The task of this device was fourfold: fixed delay (17 ms), high-pass filter (80 Hz), volume control and amplification. Figure 4 shows a sketch of the setup and dimensions of the first listening room.

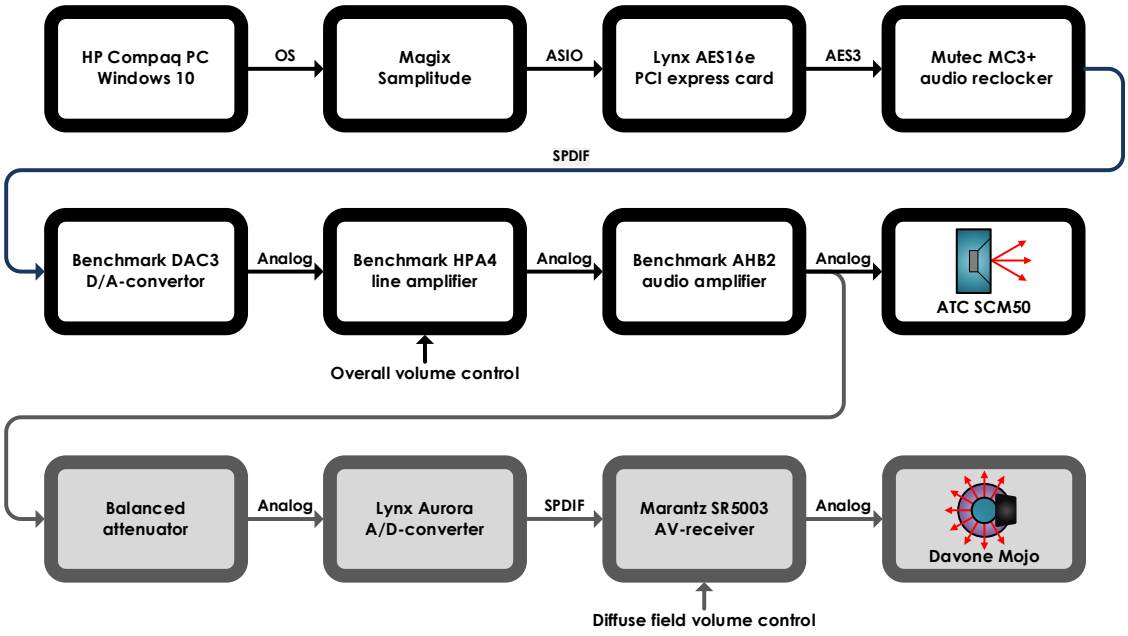

Figure 3. Equipment setup used in the first experiment.

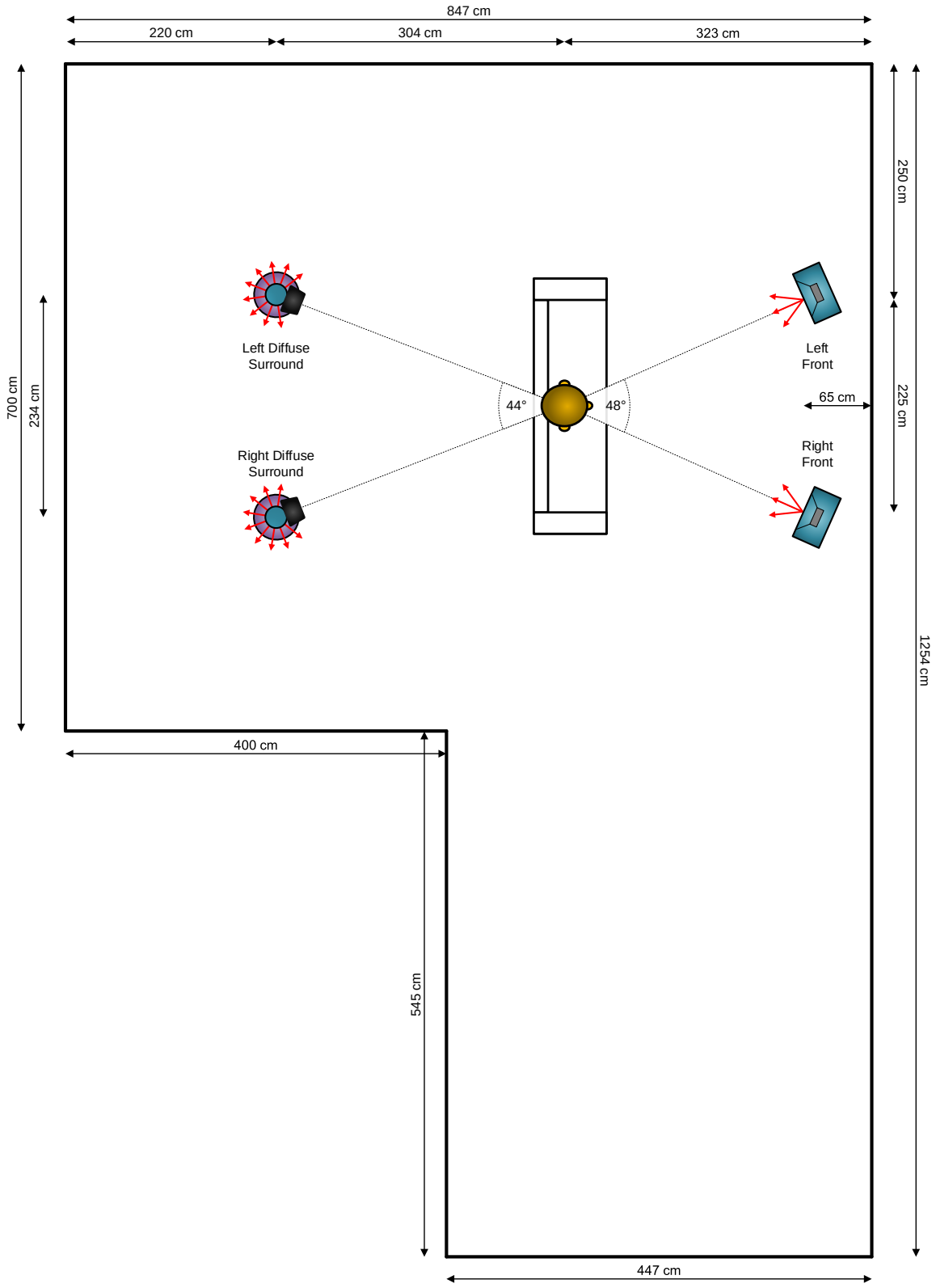

Figure 4. The loudspeaker setup and dimensions of the first listening room. Reverberation time above 200 Hz is around 0.25 seconds, with a slow rise to circa 0.45 seconds at 60 Hz.

Figure 5 illustrates the equipment used in the second experiment. The transparent blocks again show the existing reference system of the second listening area, normally in use as a control room. The source was a Personal Computer running MAX 8 as a digital controller of the audio streams of both the front and rear loudspeakers. In this setup, the front speakers were active and digitally controlled. At the rear, the same passive speakers were used as in the first experiment. Completed with a power amplifier, the addition to the main system is displayed by the grey blocks. Figure 6 shows a sketch of the setup and dimensions of the second listening room. The high-pass filter at 80 Hz, latency compensation for the digitally controlled speakers and the fixed delay of 17 ms between front and rear were programmed into MAX 8. All signals used a 44.1 kHz sample rate.

In both experiments, subjects were given a simple remote control with just three buttons: Volume Up, Down and Mute. The former enabled the level of the diffuse field fillers to be adjusted in 1 dB increments. The mute button allowed switching between the original set and the supplement at the touch of a button. This allowed a direct comparison to determine individual preference, with or without the added immersion system.

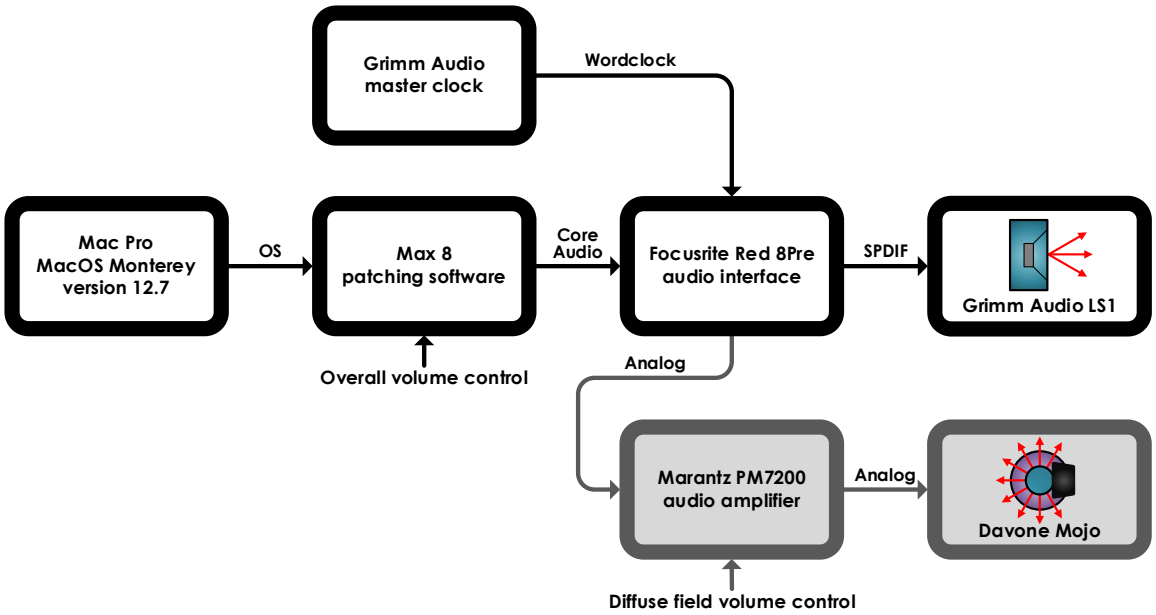

Figure 5. Equipment setup used in the second experiment.

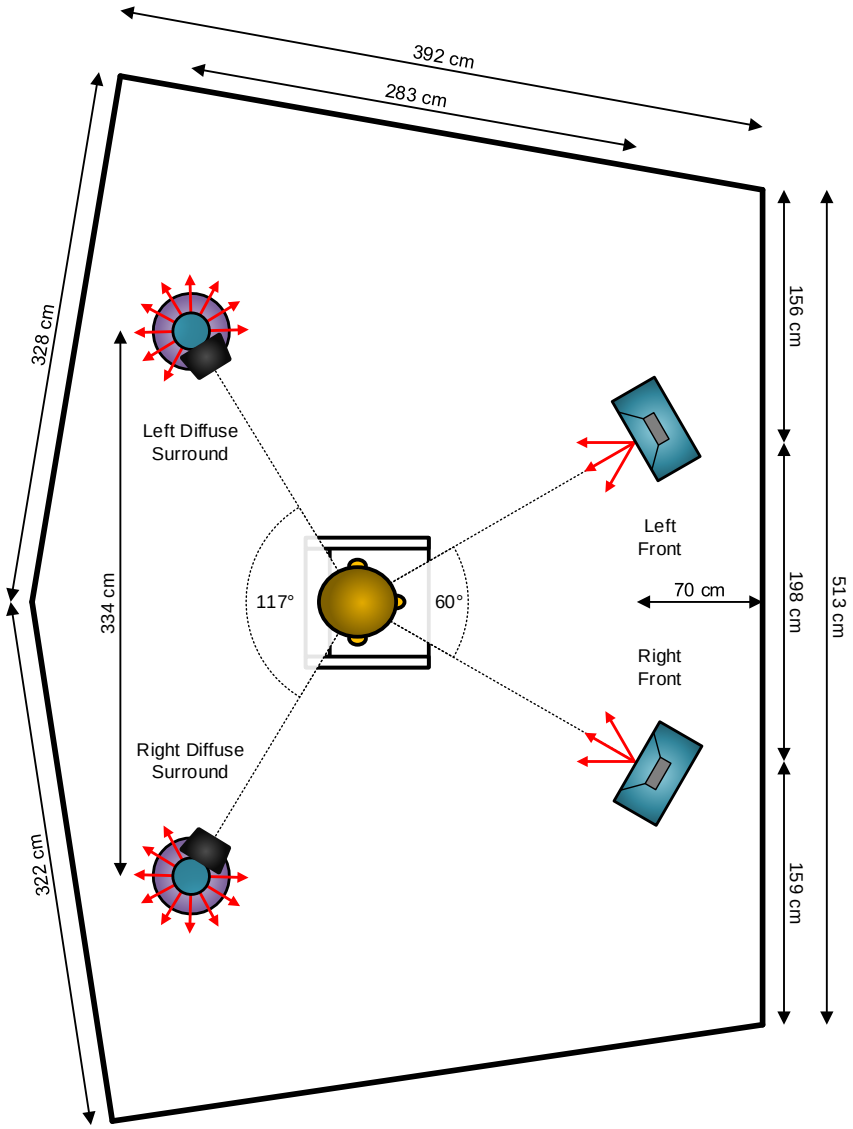

Figure 6. The loudspeaker setup and dimensions of the second listening room. Reverberation time above 100 Hz is around 0.2 seconds, with a slow rise to circa 0.4 seconds at 60 Hz.

3.3. Experimental Results

All subjects gave seven preferred level settings – one for each audio fragment – for the diffuse field speakers, including a score for the improvement in perceived overall sound quality. A first simple analysis of the results clearly shows that the additional diffuse field is highly appreciated; from the total of 168 observations, obtained with 24 subjects, 155 were in favor of adding it. It is

also clear from the average results as given in Tables 3 and 4, that the average improvement in sound quality is significant in both experiments. Over 155 preferred observations, it is 3.4 on a 5-point scale. This result is obtained with a level setting of the diffuse field loudspeakers that is on average about 4 to 6 dBA lower than the front loudspeakers. However, large differences are seen between subjects and audio fragments.

Analysis of the individual preference level settings for the diffuse field loudspeakers reveals an extremely wide variation between subjects, see Tables 5 and 6. Some prefer to set the diffuse field level even higher than the direct field. This result was not expected in the design of the experiment. For safety reasons, the diffuse field level was limited to 3 dB louder than the direct one.

Comparing the results of the two different experiments shows that adding the diffuse field gives a higher overall quality improvement in the first test that only used subjects not involved in the creation of the project. The rating in the second experiment, which included six of the authors of this paper, is 0.7 less on the 5-point scale. Also, the percentage of subjects that preferred to switch on the diffuse field for all seven audio fragments dropped from 77% in the first, to 55% in the second experiment. A reason for this lower appreciation may be the significantly smaller size of the listening room, which degrades the diffuse field. The wider angle of 117 instead of 44 degrees between the two rear loudspeakers may also have had a negative influence. Another possible explanation is that engineers assemble their mix without the extra diffuse field channels. They would likely have made different mix decisions when hearing the extra field. Some (professional) listeners may recognize that the original intent of the mix is altered by the addition of the diffuse field, and prefer to hear the original sound to stay closer to the original work.

The fact that in most cases subjects prefer to switch on the diffuse field speakers, and some of them even prefer to set them louder than the direct field speakers, might be explained by the fact that for a classical acoustic concert, performed in a good sounding hall, the loudness of the direct sound is lower than that of the indirect/diffuse field. Also note that some subjects prefer a diffuse level setting that was close to the estimated just noticeable level difference, which was estimated to be around –20dBA (see Table 3).

Table 3. First experiment – Averages over the 13 individually preferred levels of the diffuse field in dB relative to the direct field, with minimum and maximum settings and the associated quality improvements.

	FRAGMENT	**DIFFUSE LEVEL RELATIVE TO DIRECT in dB**				**QUALITY IMPROVEMENT**		
		AVERAGE	MIN	MAX		AVERAGE	MIN	MAX
1	Abba	–6.3	OFF	3		3.4	0	5
2	Beethoven	–5.5	–17	3		3.8	2	5
3	Haydn	–4.3	–17	3		3.5	1	5
4	Sara K.	–5.2	–16	0		4.2	3	5
5	Stevie Ray Vaughan	–6.1	OFF	3		3.9	0	5
6	Vivaldi	–5.7	OFF	3		3.2	0	5
7	Vocal quartet	–6.3	OFF	3		3.4	0	5
	OVERALL AVERAGE DIFFUSE LEVEL	**–5.6**				**3.7**		
	DIRECT LEVEL in dBA	**75**	**71**	**80**				

Table 4. Second experiment – Averages over the 11 individually preferred levels of the diffuse field in dB relative to the direct field, with minimum and maximum settings and the associated quality improvements.

	FRAGMENT	**DIFFUSE LEVEL RELATIVE TO DIRECT in dB**				**QUALITY IMPROVEMENT**		
		AVERAGE	MIN	MAX		AVERAGE	MIN	MAX
1	Abba	–4.5	–11	-1		3.0	2	4
2	Beethoven	–0.9	–5	3		3.9	2	5
3	Haydn	–3.5	–12	1		3.2	1	4
4	Sara K.	–5.5	OFF	-1		2.2	0	4
5	Stevie Ray Vaughan	–3.8	OFF	1		2.7	0	4
6	Vivaldi	–3.5	OFF	0		2.5	0	5
7	Vocal quartet	–3.6	OFF	3		3.5	0	5
	OVERALL AVERAGE DIFFUSE LEVEL	**–3.8**				**3.0**		
	DIRECT LEVEL in dBA	**78**	**72**	**80**				

Table 5. First experiment – Average diffuse field level over seven audio fragments of the individual settings in dB relative to the direct field, including average quality improvements and their variation. The average diffuse level and its variation are calculated using only the audio fragments for which the diffuse field speakers were switched on. Also, the percentage of sequences for which the subject perceived an improved overall sound quality is given (7 out of 7 is 100%).

SUBJECT NUMBER	AVERAGE DIFFUSE LEVEL	DIFFUSE LEVEL VARIATION	AVERAGE QUALITY IMPROVEMENT	QUALITY IMPROVEMENT VARIATION	IMPROVED PERCENTAGE
1	–4.6	4	3.4	4	71
2	+2.0	4	4.4	1	100
3	–2.9	3	4.7	1	100
4	–4.7	13	3.7	3	100
5	–5.6	4	5.0	0	100
6	–16.9	3	2.7	1	100
7	–6.8	3	2.0	3	86
8	–5.6	7	3.7	2	100
9	–5.0	11	3.6	3	100
10	–10.9	1	2.9	1	100
11	–7.3	13	4.0	3	86
12	–2.1	13	4.1	2	100
13	–2.9	9	4.4	2	100

Table 6. Same as table 5, but now for the second experiment.

SUBJECT NUMBER	AVERAGE DIFFUSE LEVEL	DIFFUSE LEVEL VARIATION	AVERAGE QUALITY IMPROVEMENT	QUALITY IMPROVEMENT VARIATION	IMPROVED PERCENTAGE
14	–0.6	6	4.3	2	100
15	–4.5	9	2.9	4	86
16	–1.9	4	3.3	4	100
17	–0.4	4	3.9	3	100
18	–2.0	8	4.0	3	100
19	–7.5	8	2.4	4	86
20	–4.6	5	2.3	4	71
21	–7.0	3	1.0	3	43
22	–2.4	7	3.7	3	100
23	–2.5	5	2.6	4	86
24	–8.4	14	2.9	2	100

4. CONCLUSIONS AND DISCUSSION

The most important conclusion that can be drawn from the experiments is that subjects highly appreciate the described diffuse field approach that can be adapted to their personal preference and the audio content. The results of the subjective tests show that there is a large variation in the preferred level of the extra diffuse field loudspeakers. Some subjects set the level close to the just noticeable difference, about 20 dB below the level of the direct field loudspeaker, while others choose to set it above the level of the direct field loudspeakers. Furthermore, the level settings and improvement in overall sound quality are strongly dependent on the audio fragment.

The advantage of the diffuse field extension given in this paper is that only two small additional surround speakers are needed for a significant increase in overall perceived sound quality, without introducing degrading localization errors as found in many surround setups. From the total of 168 observations, 155 were in favor of adding it, with an average improvement rating of 3.4 on a 5-point scale. Furthermore, the setup allows for an ‘Immersion Control’ that can be adapted to the characteristics of the recording, and to one’s own personal immersion preferences, through simple playback optimization.

The extreme dependence of the optimal perceived immersion on personal preferences makes it difficult to design an objective measurement system that allows to assess the overall sound quality of a system. For mono speech and music, and to some extent stereo music, perceptual models have been developed that show good correlation between objective measurement and subjectively perceived speech and music quality [29] [30] [31]. Models for spatial audio quality have been developed, but do not take into account the immersion optimization as given in this paper [32] [33] [34].

A possible way forward to allow for objective measurements, that take into account the immersion control results of this paper, is to introduce an individual ideal reproduction setup using an approach given in [35].

5. REFERENCES

[1] High Fidelity, Wikipedia.

[2] Six-Degrees-of-Freedom, Wikipedia.

[3] M. A. Gerzon, “Periphony: With-Height Sound Reproduction”, J. Audio Eng. Soc., vol. 21, 2-10 (1973 Feb.).

[4] M. A. Gerzon, “Ambisonics in Multichannel Broadcasting and Video”, J. Audio Eng. Soc., vol. 33, 859-871 (1985 Nov.).

[5] Furness, Roger K. “Ambisonics - an overview,” Audio Engineering Society Conference: 8th International Conference: The Sound of Audio, pp. 181-190 (UK, May 1990).

[6] A. J. Berkhout, D. de Vries and P. Vogel, “Acoustic control by wave field synthesis,” J. Acoust. Soc. Am., vol. 93, pp. 2764-2778 (1993 May).

[7] S. Spors, H. Wierstorf, A. Raake, F. Melchior, M. Frank, & F. Zotter, “Spatial Sound With Loudspeakers and Its Perception: A Review of the Current State,” Proceedings of the IEEE, 101(9), pp. 1920-1938 (2013 Sep.).

[8] R. Irwan and R. M. Aarts, “Two-to-Five Channel Sound Processing," J. Audio Eng. Soc., vol. 50, pp. 914-926 (2002 Nov.).

[9] C. Faller, “Multiple-Loudspeaker Playback of Stereo Signals*," J. Audio Eng. Soc., vol. 54, pp. 1051-1064 (2006 Nov.).

[10] F. Rumsey, “Controlled Subjective Assessment of Two-to-Five Channel Surround Sound Processing Algorithms," J. Audio Eng. Soc., vol. 47, pp. 563-582 (1999 July/Aug.).

[11] A. Raake and J. Blauert, "Comprehensive modeling of the formation process of sound-quality," *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*, Klagenfurt am Wörthersee, Austria, 2013, pp. 76-81, doi: 10.1109/QoMEX.2013.6603214.

[12] E. C. Cherry: Some experiments on the recognition of speech, with one and with two ears. J. Acoust. Soc. Am. 25, pp 975-979 (1953).

[13] A. W. Bronkhorst, “The Cocktail Party Phenomenon: A Review of Research on Speech Intelligibility in Multiple-Talker Conditions,” Acta Acustica united with Acustica, Volume 86, Number 1, pp. 117-128 (2000).

[14] C. Eaton and H. Lee, “Quantifying Factors of Auditory Immersion in Virtual Reality”, Audio Engineering Society International Conference on Immersive and Interactive Audio (UK, March 2019).

[15] A. G. Bose, “On The Design, Measurement, and Evaluation of Loudspeakers”, presented at the 35th Convention of the Audio Engineering Society (USA, Oct. 1968).

[16] K. L. Kantor and A. P. Koster, “A Psychoacoustically Optimized Loudspeaker,” J. Audio Eng. Soc., vol. 34, pp. 990-996, (1986 Dec.).

[17] J. G. Beerends, “BNS Energetic Diffuse Field System”, presented at the FIRATO (NL, Aug. 1988).

[18] E. R. Madsen, “Extraction of Ambiance Information from Ordinary Recordings," J. Audio Eng. Soc., vol. 18, pp. 490-496 (1970 Oct.).

[19]. H. Lauridsen, "Nogle forsøg med forskellige former for rumakustisk gengivelse" ("Experiments Concerning Different Kinds of Room-Acoustic Recordings"), Ingeniøren, No. 47, 906 (1954).

[20] P. Damaske, “Subjective Investigation of Sound Fields,” Acta Acustica united with Acustica, Volume 19, Number 4, pp. 199-213 (1967).

[21] F. E. Toole, “Sound Reproduction: Loudspeakers and Rooms,” Elsevier 2008.

[22] T. Holman, “The Number of Audio Channels”, Audio Eng. Soc, preprint 4292 (1996 May).

[23] J. Herre, J. Hilpert, A. Kuntz and J. Plogsties, "MPEG-H 3D Audio—The New Standard for Coding of Immersive Spatial Audio," in *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 770-779, Aug. 2015, doi: 10.1109/JSTSP.2015.2411578.

[24] M. Tohyama and A Suzuki, “Interaural cross-correlation coefficients in stereo-reproduced sound fields,” J. Acoust. Soc. Am., vol. 85, pp. 780-786 (1989 Feb.).

[25] ITU-R BS.1116, “Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems”, International Telecommunication Union, Geneva, Switzerland (1997).

[26]. H. Haas, Über den Einfluss eines Einfachechos auf die Hörsamkeit von Sprache [On the influence of a single echo on the intelligibility of speech]. Acustica 1, 49-58 (1951).

[27]. P. Damaske, Die psychologische Auswertung akustischer Phänomene [The psychological interpretation of acoustical phenomena]. Proceedings, 7th Int. Congr. On Acoustics, Budapest, 21 G2 (1971).

[28] EBU SQAM, “Sound Quality Assessment Material Recordings For Subjective Tests”, European Broadcasting Union, Geneva, Switzerland (2008).

[29] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullman, J. Pomy and M. Keyhl, “Perceptual Objective Listening Quality Assessment (POLQA), The Third Generation ITU-T Standard for End-to-End Speech Quality Measurement Part I – Temporal Alignment,” J. Audio Eng. Soc., vol. 61, pp. 366-384 (2013 June).

[30] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullman, J. Pomy and M. Keyhl, “Perceptual Objective Listening Quality Assessment (POLQA), The Third Generation ITU-T Standard for End-to-End Speech Quality Measurement Part II – Perceptual Model,” J. Audio Eng. Soc., vol. 61, pp. 385-402 (2013 June).

[31] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, C. Colomes, M. Keyhl, G. Stoll, K. Brandenburg, B. Feiten, “PEAQ - The ITU-Standard for Objective Measurement of Perceived Audio Quality,” J. Audio Eng. Soc., vol. 48, pp. 3-29 (2000 Jan./Feb.).

[32] R. Conetta, T. Brookes, F. Rumsey, S. Zielinski, M. Dewhirst, P. Jackson, S. Bech, D. Meares, S. George, “Spatial Audio Quality Perception (Part 1): Impact of Commonly Encountered Processes”, J. Audio Eng. Soc., vol. 62, pp. 831-846 (2014 Dec).

[33] R. Conetta, T. Brookes, F. Rumsey, S. Zielinski, M. Dewhirst, P. Jackson, S. Bech, D. Meares, S. George, “Spatial Audio Quality Perception (Part 2): A Linear Regression Model”, J. Audio Eng. Soc., vol. 62, pp. 847-860 (2014 Dec).

[34] P. M. Delgado and J. Herre, "Objective Assessment of Spatial Audio Quality Using Directional Loudness Maps," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019, pp. 621-625, doi: 10.1109/ICASSP.2019.8683810.

[35] J. G. Beerends, K. van Nieuwenhuizen, and E. vd Broek, “Quantifying Sound Quality in Loudspeaker Reproduction”, J. Audio Eng. Soc., vol. 64, pp. 784-799 (2016 Oct.).