Audio Engineering Society

Convention Paper

Presented at the 140th Convention
2016 June 4–7, Paris, France

The Difference Between Stereophony and Wave Field Synthesis in the Context of Popular Music

Christoph Hold[1], Hagen Wierstorf[2], and Alexander Raake[2]

[1]*Assessment of IP-based Applications, Technische Universität Berlin, Berlin, Germany*
[2]*Audiovisual Technology Group, Technische Universität Ilmenau, Ilmenau, Germany*

Correspondence should be addressed to Christoph Hold (Christoph.Hold@Telekom.de)

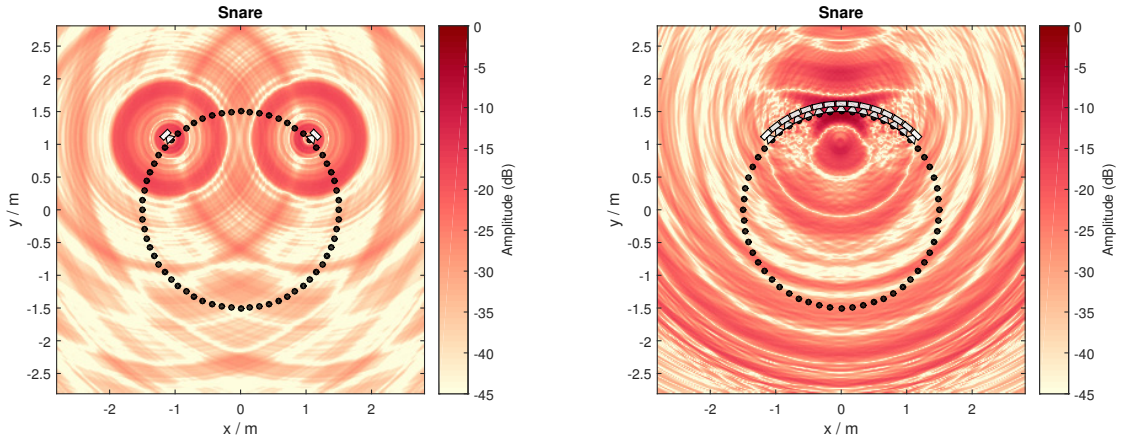
ABSTRACT

Stereophony and Wave Field Synthesis are capable of providing the listener with a rich spatial audio experience. They both come with different advantages and challenges. Due to different requirements during the music production stage, a meaningful direct comparison of both methods has rarely been carried out in previous research. As stereophony relies on a channel- and Wave Field Synthesis on a model-based approach, the same mix cannot be used for both systems. In this study, mixes of different popular-music recordings have been generated, each for two-channel stereophony, surround stereophony, and Wave Field Synthesis. The focus during the mixing process is on comparability between the reproduction systems in terms of the resulting sound quality. In a paired-comparison test listeners rated their preferred listening experience.

1 Introduction

An important aspect of popular music is sound quality, which largely influences the listening experience. Besides the overall tonal balance, the spatial impression and aspects such as the resulting desired excitement need to be considered in sound quality assessment. Studies have shown that a higher number of loudspeakers active during play-out leads to better results, for example in spaciousness and listener envelopment [1], which benefit high immersion of listeners. Furthermore, higher immersion is linked to stronger emotional reactions [2]. The present study investigates the benefits that spatial audio systems may provide to the listening experience of popular music.

Up until now, the music production industry almost exclusively covers spatial reproduction systems based on stereophony, like two-channel stereophony (Stereo) or 5.1 surround stereophony (Surround) [3]. They rely on applying level and time differences between the loudspeakers in order to position phantom sound sources. For this purpose, channel-based production techniques have been developed and perfected during the past decades. A different method for spatial music (re-)production is Wave Field Synthesis (WFS), which tries to synthesize a sound field in an extended listening area. WFS relies on a model-based mixing approach [4], wherefore traditional recording and mixing approaches are not suitable. Stereophonic mixes



(a) Stereo, phantom source at $(0, 1)$ m.

(b) WFS, focused source at $(0, 1)$ m.

Fig. 1: Simulated sound pressure of a snare hit signal, reproduced by different spatial audio systems, plotted at a given time sample for the listening area. The black dots indicate the positions of available loudspeakers, the white loudspeaker symbols highlight the active ones for the actual reproduction.

cannot easily be transformed so as to perform optimally in a WFS-reproduction context.

These problems have prevented direct comparisons between stereophonic methods and WFS for popular music in the past. Only single perceptual attributes for very basic sound scenes were investigated. Wierstorf et al. [5] showed that localization for a single source is better in WFS than in stereophony. Impacts of the sound mix are thereby bypassed, but the results cannot directly be transferred to more complex music scenes. It is not clear yet, how WFS affects listening preferences in the context of popular music.

Besides all the spatial benefits of WFS, it can add artifacts in the form of comb-filter like alterations of the frequency spectrum and smearing of the time signals [6]. Both are due to spatial aliasing which occurs for frequencies above 1.5 kHz in the utilized WFS system setup. Those artifacts are not always audible, but especially in the case of focused sources, they could alter considerably the hearing impression [7] and have to be kept in mind during the mixing process.

Since the sound pressure distribution for the same acoustic scene differs substantially (Fig. 1), the question arises, how the perceived scene differs. The goal of the present study is to compare Stereo, Surround, and WFS against each other in terms of which system is preferred for the reproduction of popular music. A listening test investigates the preference applying a paired comparison paradigm. The listening test includes different complex scenes that are created in a mixing step. To ensure that the preference of the listener is not dominated by the mixing steps, they were all performed by the same person with the mixing goal of comparability. The next section details this process.

2 Methods

The first part of this section describes in detail how the stimuli for the listening experiment were designed. Since we do not want to rely on up- or down-mixing algorithms, this included a mixing stage that had to deal with the two – sometimes contradicting – goals of achieving comparable mixes between the systems and achieve the best possible result for every observed system.

2.1 Reproduction Systems

The observed systems are set up in parallel by using corresponding speakers of a circular loudspeaker array shown in Fig. 2. This leads to a native Stereo and Surround setup according to ITU-R [8]. WFS is played out employing 56 available loudspeakers. All variants are using a subwoofer, integrated as a dedicated secondary source in WFS, via bass-management in both

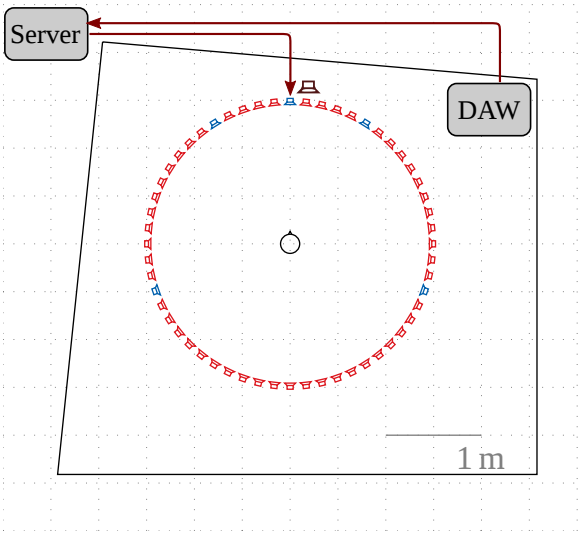


Fig. 2: Setup of the loudspeaker array used for all reproduction methods. In the case of WFS all loudspeakers were available to the renderer, in the case of Surround the loudspeakers marked in blue were used. For Stereo, only the two front loudspeakers of the Surround setup are active. For all methods a subwoofer was available in the front of the loudspeaker array. The listener was always placed at the center of the loudspeaker array.

stereophonic setups respectively. The chosen setup allowed a outsourced *Digital Audio Workstation (DAW)* to run from the monitoring position, which forwards all output channels via *MADI*. The following *audio server* routes all signals, including bass-management, and performs the WFS processing. We used a WFS implementation available as open source by the SoundScape Renderer [9]. We modified the source code of the renderer in order to disable the inherent amplitude decay for different virtual source positions. This leads to a more practical setup for music mixing, because it allows for amplification of single sources independently of their spatial position.

2.2 Stimuli and Mixing Process

All stimuli are generated from freely available or self-made recordings and are composed of mainly acoustic instruments [10]. Every song has to be available as multi-track recordings, which means that every instrument is picked up individually, often by multiple microphones and repetitions. Four different popular music recordings supply the source material. Two of them include stylistically different bridge parts and these are therefore presented additionally. The four recordings consist of a pop-rock song with live feeling and male vocals, presented complete ('Track A1') and from its guitar solo bridge ('Track A2'), a more sentimental pop song with deep male vocals, again complete ('Track B1') and with its very spatial guitar bridge ('Track B2'), a slightly heavier rock song with female vocals ('Track C') and a shorter hip-hop track ('Track D') with male rap vocals.

The mixing process of popular music involves stages that are more or less independent from the involved reproduction system. On the other hand, some stages like panning and reverberation have to be adjusted on each single systems in order to ensure they are used in the best possible way. Applying advanced mixing techniques demands for an adaptation to WFS, as most modern mixing techniques are based on channel-based stereophony and not on the model-based approach taken by WFS.

Figure 3 shows a block diagram of the basic layout we applied in order to create comparable mixes for different systems. Red arrows symbolize the path of audio signals, small blue boxes the sound processing stages and green shows major dependencies and influences of the mixing engineer. The upper large gray box describes all the system independent steps, the lower one all steps performed on the actual systems.

2.2.1 System Independent

The majority of the mixing process involving sound processing is system independent and was performed on a reference system. The listening test does not contain the reference system. This consists of a Stereo setup in a control room with studio loudspeakers including a subwoofer and is well known by the mixing engineer.

The mixing steps performed on the reference system include *level adjustments*, *equalizing*, and *dynamic range compression*. After listening to the source material individually, a basic level matching is applied. Equalizers are establishing a certain tonal balance and are also separating individual instruments. Since acoustic instruments and especially the human singing voice can drastically change their volume, the next step manipulates the signal's dynamic range. Dynamic range

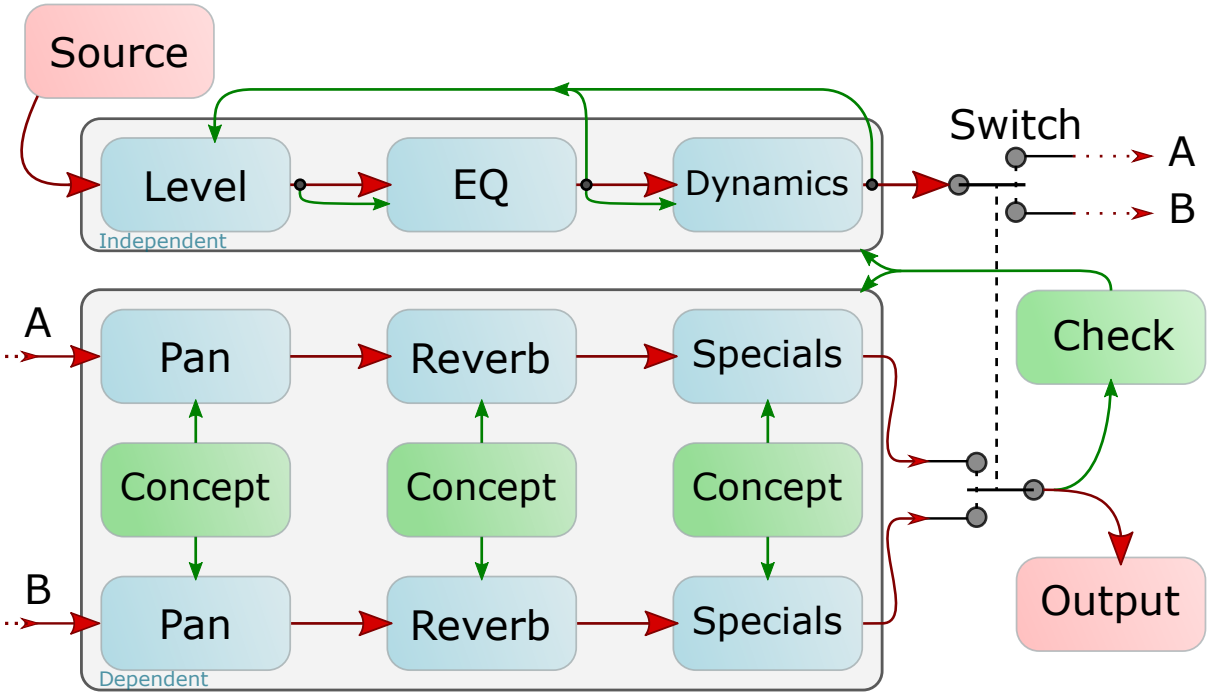


Fig. 3: Block diagram illustrating the basic multi system sound mixing procedure, from the source material to the finished output. Red corresponds to audio signals, blue to sound processing stages and green shows major dependencies.

compression is widely used in nowadays mixing and is therefore necessary to create the requested modern sound appearance. The green lines in Fig. 3 illustrate that each step affects previous decisions. For example, boosting high frequencies with an equalizer may lead to the demand of reducing the corresponding signal level. Compression obviously influences volume, but also alters the tonal character often. Proper adjustments on these first block layers result in a good overall balance, independent of the playback system. At this stage, the *Digital Audio Workstation (DAW)* project contains all non-spatial processing and specifies the fundamental sound character of each song. Finally, the created basic project enters an environment for instantaneous switching between the examined systems.

2.2.2 System Dependent

The final system dependent processing includes *positioning*, *reverb/delay*, and *special effects*. All system dependent processing is guided by an underlaying joint concept and is only adapted to the requirements of each system, avoiding fundamentally different results. Those adaptations, especially the positioning of the single objects in the music scene, followed a conservative handling, which implies avoiding extreme and rather unconventional settings as well as omitting moving sources. Hence, all main components of a song are positioned in the frontal scene. Lead vocal, snare drum and bass are always positioned in the center. Since further positioning is mainly driven by the available capabilities of each individual system, especially their spatial performance, elements that likely create envelopment are allocated to all directions. In particular, reverbs and delays should exhaust and thereby demonstrate the spatial capabilities.

All special processing, such as modulation based effects, are made as similar as possible between the systems. The final automation and correction/checking stage is again valid for all systems and therefore performed after the switching. The validation affects every stage. Note, that this stage closes a circle, which corresponds to the time-consuming mixing procedure.

Whenever certain processing is only applicable to stereophonic tracks in an adequate manner, these two

channels can be integrated in WFS by routing them to two corresponding virtual sources in the object based environment [11]. This technique based on virtual panning spots allows especially *bus-compression* or delay unit outputs to be integrated. Thus the combination or fusion of those two unequal methods, stereophony and sound field synthesis, can produce suitable results.

As the listeners are advised to rate the system they preferred in the test, it is required to guarantee equal loudness between the different systems at the listening position. This is important as preference ratings can easily be dominated by differences in loudness [12]. The loudness is adjusted between the systems by means of dummy-head recordings at the listener position. One system represents a reference and the other two systems are measured around the level of the reference system. For all recordings, the Zwicker and Fastl [13] model for temporally variable sounds estimates loudness and allowed for an accurate adjustment of the different systems. The model is available as part of the *Loudness Toolbox for Matlab* [14].

2.3 Participants

24 test listeners, including both man and woman, participated the test. Subjects were between 18 and 40 years old. There was no special selection regarding expertise. All listeners were financially compensated for their effort.

2.4 Procedure

In a paired comparison test the listeners rated their preferred listening experience, while switching instantaneously between the different systems with the associated mix. The attenders rated all six tracks on all systems in a randomized sequence, one track at a time. The participants received written instructions explaining their tasks. They were allowed to ask the examiner further questions at any time. The experiment started with a short training piece, which contained a looped 4 s song intro phrase of 'Track D'. During this training phase, the test conductor explained the GUI. By clicking on either button 'A' or 'B', the participants could switch between the stimuli as often as they like. After 45 s they were allowed to choose their preferred condition, a confirmation started the next pair of stimuli. The generated pop music mixes were presented in a randomized order and each playback system was rated with all six tracks. The switched conditions were Stereo, Surround and WFS, which got randomly assigned to the buttons 'A' and 'B', but not one system to both. The whole procedure lasted under 30 minutes in total.

2.5 Bradley-Terry-Luce Model

The *Bradley-Terry-Luce* model allows to create a continuous ranking of considered stimuli from paired comparison choices [15]. It can be shown that the Bradley-Terry-Luce model is a special case of the *elimination by aspects* models implemented by Wickelmaier and Schmid [16].

Besides estimating the systems' *ability* values, the model allows to determine corresponding 95% confidence intervals. Those ratio utility scale values sum to unity. For further details, the overall performances of the systems are split and analyzed for each track separately.

Feeding equation (1) with the ascertained *ability* values $u(x)$ from the estimated utility scale calculates the probability of choosing x from a set of Ψ alternatives [16].

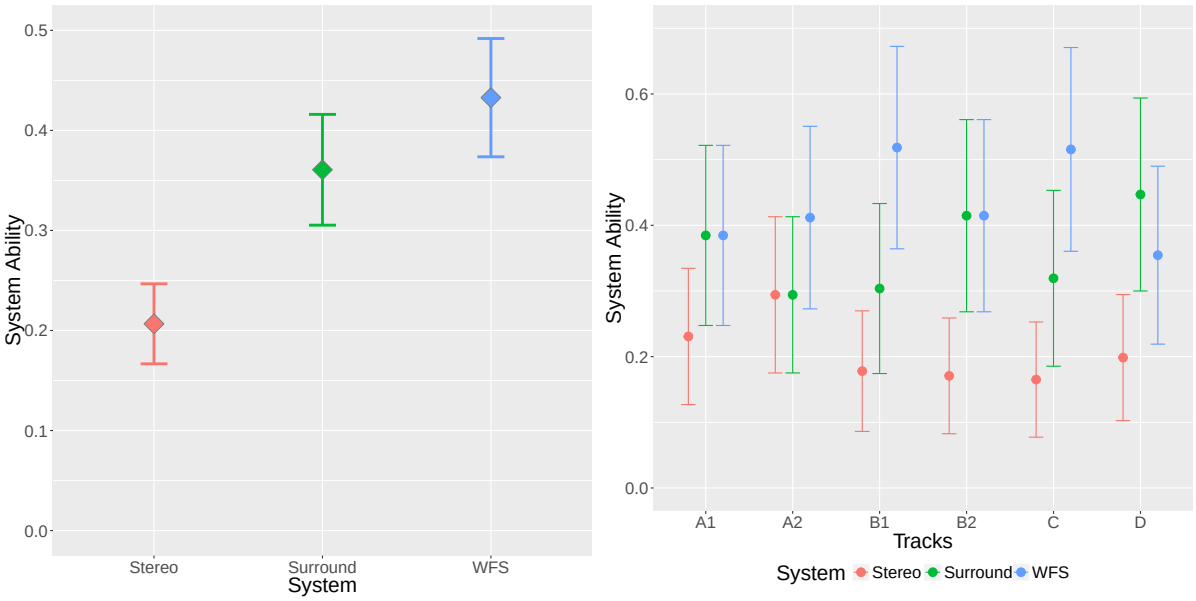
$$P(x,\Psi) = \frac{u(x)}{\sum_{y\in\Psi} u(y)} \tag{1}$$

3 Results

Table 1: Absolute frequency table of preferred reproduction systems in the paired comparison test as rated by 24 listeners for six different music mixes.

Systems \ Winner	Stereo	Surround	WFS
Stereo vs. Surround	51	93	–
Stereo vs. WFS	48	–	96
Surround vs. WFS	–	64	80

For the described listening test, Table 1 shows the absolute frequencies of all observations. Each row represents one pair of competing systems, the respective columns present the frequency of the listeners' preference. Listeners preferred Stereo the less, while against



(a) Total system abilities

(b) System abilities analyzed for each presented track

Fig. 4: Results of the paired comparison listening test regarding reproduction system preference. Analyzed with a Bradley-Terry-Luce model. The dots mark the generated, unity summing system abilities u(x), estimating the ability of each system to be preferred. The bars denote the corresponding 95% confidence intervals.

WFS lesser than against Surround. In this absolute perspective, WFS also beats Surround, but less conspicuously than against Stereo.

While consulting the frequency table, a first tendency becomes apparent. The ratio between the systems is better provided by the Bradley-Terry-Luce model.

The applied Bradley-Terry-Luce model passed the test, whether it fits the data and satisfies basic requirements. This means the p-value of the corresponding χ^2 distributions never drops below 10%, therefore the model is not rejected [16]. The stimuli are furthermore acceptably distinguishable, which corresponds to the low number of 21 measured circular triads (Stereo>Surround>WFS>Stereo) over all 144 measurements.

Figure 4 presents the calculated results, for the overall system performances in 4a and separated for each track presented on each system in 4b. The dots mark the calculated abilities with their 95% confidence interval bars. Figure 4a highlights that Stereo is inferior to the higher channel systems. The distance between Surround and WFS is remarkably smaller. Figure 4b indicates that for 'Track A', the systems perform more similar than for the other tracks, with the largest deviations for 'Track C'. During the bridge part of 'Track B', WFS is no longer preferred and rated similar to Surround. Nevertheless, only in one out of six cases, the WFS ranking is lower than the Surround ranking.

4 Discussion

This section first discusses the observed system differences that became obvious during the mixing process. It also shows the contrasting shades of stereophony and WFS and the high dependency on the presented content. The listening test underlines the superior WFS performance.

4.1 Observed System Differences – Engineers View

The spaciousness increases with the number of speakers involved. The step from Stereo up to Surround is much stronger than from Surround up to WFS. Still, the performance of Stereo is remarkable when listening in the sweet-spot. With a decent mix, it creates a fairly

good sense of immersion, though the actual sound only comes from the frontal plane. Regarding spatial performance, the possibilities differ due to the underlying reproduction concepts. Phantom sources outside the frontal sound stage are unstable and should be avoided for sharp localization [17, 4.2], but WFS allows stable positioning of virtual sound sources in the whole listening area. Localization in general is much more stable outside the sweet-spot in WFS, which corresponds to the results of previous listening tests, e.g. [5].

An aspect that emerges more during the mixing process is the way the individual sound sources interact with each other. Current sound engineering often wants to achieve a dense sound, meaning all the individual instruments should glue together and form a uniform appearance. This shall indeed not result in unsatisfactory separation, in particular the lead vocals must always be apparent and clear. The object oriented approach of WFS, with ideally one virtual source per instrument, produces a high separation between those. This pure point source representation is not always desired [11] and complicates the interlocking. Both stereophonic systems behave differently on this point, in a direct comparison the individual tracks seem per se less separated and the sound stage less transparent. Additionally, many mixing techniques and processing tools that produce these more glued sound, such as *bus-compression*, are developed for Stereo. Hence, it is harder to achieve separation in stereophony and it is harder to achieve interaction in WFS contrastingly, even though virtual panning spots provide a suited compound of both.

Spatial aliasing of WFS introduces spectral and temporal artifacts which have to be minded during the production process. It differs from the expected sonic quality and tone of stereophonic systems. As shown in Wierstorf et al. [6], it is likely to experience high frequency alteration and also increasing energy at higher frequencies. Since tone shaping is part of the mixing process (compare figure 3, EQ), the result played out in WFS may differ from stereophony. Micro dynamics, the dynamics within one single stroke of one instrument, and especially transients behave more critical in WFS, due to temporal smearing. Since this temporal smearing is particularly pronounced for focused sources [7], positioning of percussive material as focused sources should be avoided.

4.2 Listening Test

The previously described reproduction properties and benefits are experienced at the production side by the sound engineer. The listening test investigated the experience of naive listeners. The outcome of the listening test shows that Stereo is inferior to the higher channel systems, which corresponds to previous results [2]. Regarding the system ranking and overall performance, Surround and WFS most likely benefit from their better spatial performance. WFS affirms the assumed superiority. It is conspicuous that the preferred systems change clearly with the presented music content. Most likely different content yields different shades of the systems. The high density of the created complex sound scenes creates many opportunities to pay attention to a variety of details. After the mixing stage remaining, often subtle, WFS Artifacts are not always realized by naive listeners. Certainly it can be assumed that not every mix pleased every lister and sometimes too much spaciousness seems counterproductive. It is not clear yet, which exact attributes trigger the decision.

5 Summary

During the comparison preparation of the contrasting reproduction systems Stereo, Surround and WFS, the demand for proper sound mixes emerged in order to compare these systems meaningful. Creating those complex popular music sound scenes for a scientific comparison means to achieve a balance between demonstrating the maximally available performance of each system and still maintain comparability. A workflow draft for multi system mixing and its basic components are exposed, just as some experienced limitations. Nevertheless, it has been shown that the adherent artifacts of WFS can be handled and turned out to be less disturbing than expected. The major influence of observed sound quality is still based on the sound mixing skills and WFS is a suitable reproduction system for pop music content.

The results highlight that WFS with proper adjusted mixes can enrich the listener experience in popular music compared to stereophony, although the full potential of music production and mixing for WFS, for example in terms of dynamic spatial effects, were not even included in the test. Previously described artifacts of WFS were identified during the mixing process, but affected the overall results less than expected.

6 Acknowledgements

This research has been supported by EU FET grant TWO!EARS , ICT-618075.

References

[1] Shim, H., Oh, E., Ko, S., and Park, S. H., "Perceptual Evaluation of Spatial Audio Quality," in *Proceedings of the 129th Convention of the Audio Engineering Society*, p. Paper 8300, 2010.

[2] Västfjäll, D., "The Subjective Sense and Experienced Emotions in Auditory Virtual Environments," *CyberPsychology & Behavior*, 6(2), 2003.

[3] Rumsey, F., *Spatial Audio*, Focal Press, Oxford, 2001.

[4] Geier, M., Ahrens, J., and Spors, S., "Object-based Audio Reproduction and the Audio Scene Description Format," *Organised Sound*, 15(03), pp. 219–227, 2010.

[5] Wierstorf, H., Raake, A., and Spors, S., "Localization of a virtual point source within the listening area for Wave Field Synthesis," in *Proceedings of the 133rd Convention of the Audio Engineering Society*, p. Paper 8743, 2012.

[6] Wierstorf, H., Hohnerlein, C., Spors, S., and Raake, A., "Coloration in Wave Field Synthesis," in *Proceedings of the 55th International Conference of the Audio Engineering Society*, pp. Paper 5–3, 2014.

[7] Wierstorf, H., Raake, A., Geier, M., and Spors, S., "Perception of focused sources in Wave Field Synthesis," *Journal of the Audio Engineering Society*, 61(1-2), pp. 5–16, 2013.

[8] ITU-R, "Multichannel stereophonic sound system with and without accompanying picture BS Series," *International Telcommunication Union Reccomendations BS. 775-3*, 3, p. 23, 2012.

[9] Geier, M., Ahrens, J., and Spors, S., "The SoundScape Renderer: A Unified Spatial Audio Reproduction Framework for Arbitrary Rendering Methods," in *Proceedings of the 124th Convention of the Audio Engineering Society*, p. Paper 7330, 2008.

[10] Source Material, "'Track A': Telefunken Elektroakustik- TheBrew_WhatIWant, 'Track B': own production (unpublished), 'Track C': MedleyDB- HopAlong_SisterCities, 'Track D': MedleyDB- Lushlife_ToynbeeSuite," 2015.

[11] Theile, G., Wittek, H., and Reisinger, M., "Potential wavefield synthesis applications in the multichannel stereophonic world," in *Proceedings of the 24th Conference of the Audio Engineering Society*, p. Paper 35, 2003.

[12] Vickers, E., "The Loudness War," *Journal of the Audio Engineering Society*, 59(5), pp. 346–351, 2011.

[13] Zwicker, E. and Fastl, H., *Psycho-Acoustics - Facts and Models*, Springer, Berlin, 1999.

[14] Genesis, "Loudness Toolbox for Matlab 1.0," 2009.

[15] Bradley, R. and Terry, M., "Rank analysis of incomplete block designs: I. The method of paired comparisons," *Biometrika*, 39(3/4), pp. 324–345, 1952.

[16] Wickelmaier, F. and Schmid, C., "A Matlab function to estimate choice model parameters from paired-comparison data." *Behavior research methods, instruments, & computers*, 36(1), pp. 29–40, 2004.

[17] Theile, G., *On the Localisation in the Superimposed Soundfield*, Ph.d. thesis, Technische Universität Berlin, 1980.