Audio Engineering Society

# Conference Paper

Presented at the Conference on
Audio for Virtual and Augmented Reality
2016 Sept 30 – Oct 1, Los Angeles, CA, USA

# Positioning of Musical Foreground Parts in Surrounding Sound Stages

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

[1]*Assessment of IP-based Applications, Technische Universität Berlin, Ernst-Reuter-Platz 7, 10587 Berlin, Germany*
[2]*Audiovisual Technology Group, Technische Universität Ilmenau, Ehrenbergstraße 29, 98693 Ilmenau, Germany*

Correspondence should be addressed to Christoph Hold (`Christoph.Hold@alumni.tu-berlin.de`)

## ABSTRACT

Object-based audio offers several new possibilities during the sound mixing process. While stereophonic mixing techniques are highly developed, not all of them generate promising results in an object-based audio environment. An outstanding feature is the new approach of positioning sound objects in the musical sound scene, providing the opportunity of stable localization throughout the whole listening area. Previous studies have shown that even if object-based audio reproduction systems can enhance the playback situation, the critical and guiding attributes of the mix are still uncertain. This study investigates the impact on listening preference evoked by different spatial arrangements of sound objects, with a strong emphasis on the high-attention foreground objects of the presented music track.

## 1 Introduction

Virtually any new or modern reproduction system provides a high number of surrounding loudspeakers, or other methods for creating a sound scene which surrounds the listener. These high-channel systems can enrich the listening experience, since an increasing channel number with suitable sound mixes can lead to increased preference ratings. This was shown in a previous study comparing two-channel stereophony, 5.1 surround and wave field synthesis (WFS) [1]. Furthermore, previous results connect more loudspeakers to higher emotional reaction, which encourages deep immersion of listeners [2].

Creating spaciousness is one of the goals in the mixing process of popular music. All modern popular music recordings are distinctly modified by post processing, and the majority of spatial components inherent to pop music is artificial. The generated spatial arrangement of instruments is one of the most influential tools in creating spaciousness, envelopment and spatial richness. In this context, several sound-mixing rules evolved during the past decades. While some famous stereo recordings from the 1960s are known for their individual and pronounced left-right instrument panning, today's recordings share strong commonalities.

The new approach of object-based audio evokes, besides a novel positioning method, differently sounding results. Compared with traditional stereophony, these differences manifest themselves in smaller sound centers of instruments, the concept of virtual panning spots [3], and idiosyncratic artifacts for individual rendering methods [4]. During the sound mixing process, the engineer incorporates an adapted workflow, leading to further different results. In a mix, not all elements of the music scene behave equally, and especially leading instruments—in particular lead vocals—play a special role. These leading parts represent the musical foreground, and capture the most listening attention. Notably, they are typically placed around the center position in contemporary stereophonic music. Virtual source positioning is, however, not the only attribute determining whether an object is perceived as musical background or leading foreground. Loudness and dynamic-range compression are able to pull sounds into the foreground, but this effect, utilized in stereophonic mixing, may not directly translate to object-based audio mixing contexts.

Latest research supports tailored audio-object treatment based on underlying categories, one of which is found to be related to background sounds [5]. This forms the counterpart to the attention-grabbing foreground. The presented categories affirm that listeners distinguish between background and other parts. Furthermore, it is proposed to not only categorize audio objects, but also customize them during the production stage and final rendering process. For unfamiliar music, a related study showed a significant influence of the sound mix on both preference and quality ratings, as well as a positive correlation between both ratings [6]. This underlines the importance of mixing, and the way content is processed. It is not clear yet, how traditional habits and sound mixing techniques translate to object-based reproduction systems and applications, especially since latest object-based virtual reality content noticeably exploits its spatial spreading capabilities, often evoking highly scattered results.

This leads to the question of whether the spaciousness induced by the reproduction system, or the influence of the mixing process, is more important for preferences of listeners. During the informal interviews after the listening experiment by Hold et al. [1], some listeners reported a feeling of unpleasantness and annoyance for off-center positioned lead instruments. The observations indicated a specific point, beyond which more spatial spread no longer generates a positive correlation. Similar results were detected for the related attribute "width" [7] and for room acoustics, where only a certain and context dependent interval of spaciousness was perceived as pleasant [8].

The present study investigates the influence on preference for various spatial arrangements of the foreground elements in current popular music. Therefore, different object-based sound mixes of the same pop music piece are assessed within a WFS system. These contain scaled alterations of the positioning of the foreground elements, including both current common practice and more spatially spread and narrowed versions. In order to relate the current findings of variation in mixing to the findings of variation of the reproduction method, the WFS mix, as well as the two-channel stereo mix from the experiment of Hold et al. [1], are included for comparison.

## 2 Methods

### 2.1 Apparatus

The listening test took place in a 83 $\mathrm{m}^3$ acoustically damped listening room (room Calypso in the Telefunken building of TU Berlin). The listeners sat on a heavy chair wearing open headphones (AKG K601) with an attached head tracker (Polhemus Fastrak). They sat in front of a flat screen placed on a small table and were able to choose between a mouse or keyboard for entering their responses.

In a separate room, a computer equipped with a multichannel sound card including D/A converters (RME Hammerfall DSP MADI) played back all sounds. The signals traveled through a head phone amplifier (Behringer Powerplay Pro-XL HA 4700) and analogue cable to the head phones in the listening room, a distance of approximately 5 m.

### 2.2 Stimuli

#### 2.2.1 Audio material and mixing

The audio material consisted of a multitrack recording session with double trackings, mainly for guitars and vocals. It was a moderate tempo pop music piece including deep male vocals, acoustic and electric guitars, bass, drums, shaker and also reverb and delay effects. The author also recorded the track, ensuring no heavy processing was applied during recording.

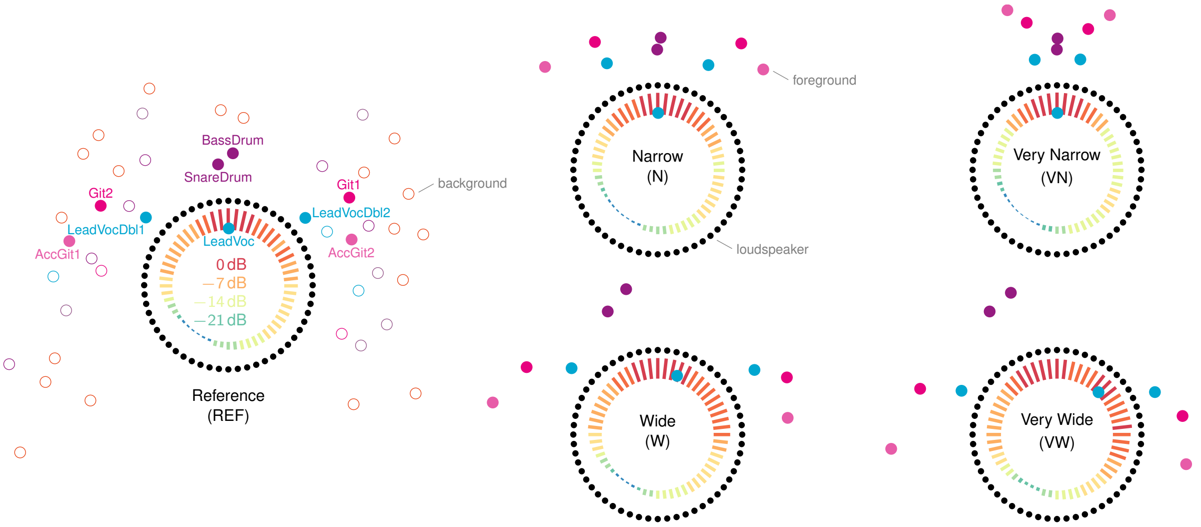

**Fig. 1:** Arrangement of the five different object-based mixes for the WFS system. The positions of the different foreground elements (vocals, drums, guitars) are shown by the filled points relative to the circular loudspeaker array for the five different cases. For the reference mix the sound objects belonging to the background are indicated by the open circles, and omitted for the other conditions as they remained the same. For every loudspeaker the colored bar towards the center of the loudspeaker array displays the root mean square magnitude of its activity averaged over the time.

Starting from a reference WFS mix (REF), created for an earlier experiment [1], two different mixes with narrower foreground elements and two mixes with wider foreground elements were produced. The choice of which audio objects belong to the musical foreground is highly content dependent, but was quite obvious in the present music piece. The decision taken for “vocal, drums, guitar” also matches the three most mentioned instruments regarding “like” and “quality” from a previous study by Wilson and Fazenda [6]. Here, the foreground instrument “drum” was represented by the sound objects “bassdrum” and “snaredrum” and its remaining parts were considered to belong to the background. Besides the lead tracks, common pop music practice includes vocal and guitar (harmony-) double tracks, which were also displaced accordingly.

REF represents a very common and modern variant with lead tracks in the center, guitar tracks positioned to the side and double tracks spread symmetrically, compare Fig. 1. This arrangement is similar to the stereo mix also created for an earlier experiment [1], however moderately wider. The narrow version (N) moves all tracks towards the center, chiefly the double-tracked guitars. The very narrow (VN) mix consists of a center foreground base a little narrower than stereo. In the wide (W) mix, the foreground objects are gently pulled apart from the center of the scene, retaining an appropriate and symmetrical impression. Hence, the lead vocal and drum tracks are shifted inversely, with the drums on the left and lead vocals on the right side. In the very wide (VW) mix, some guitar parts finally appear from behind the listeners. The background part of the piece, including reverbs and delays, remain at their reference position in every WFS mix.

The mixes are available in an object-based format as metadata [9] and signal feeds [10]. The finished mix is available under Hold et al. [11]. Note, that the finished mix starts at 84 s of the original recording. In the listening test the first 30 s of those mixes were used. The extract included pre-chorus and chorus, starting at a point consistent with musical phrasing with one bar leading into the pre-chorus. It was chosen as the chorus contains all the foreground instruments and as it contains a transition into the chorus, allowing participants to hear the processing characteristics in two slightly different settings. Figure 1 shows the activity of each loudspeaker averaged over these 30 s.

### 2.2.2 Rendering and binaural synthesis

The WFS mixing was monitored on a circular loudspeaker array consisting of 56 loudspeakers (Elac 301), bolstered by a subwoofer (Genelec 7060A). The loudspeaker array had a diameter of 3 m and was located in a 54 $\mathrm{m}^3$ acoustically damped listening room (room Pinta in the Telefunken building of TU Berlin). An open source WFS renderer (SoundScape Renderer [12], with compensated distance dependent amplitude decay [1]) computed the loudspeaker driving signals that were then stored as sound files.

The actual experiment was not conducted with the real loudspeaker array, but with a dynamic binaural simulation [13] of an anechoic version of the same loudspeaker array. This facilitates auditory modeling of the data as the model then has access to the same ear signals as the participants during the test. For the dynamic binaural synthesis one binaural room scanning (BRS) file was created for every loudspeaker [14] with a resolution of 1° utilizing high resolution head-related impulse responses [15]. During playback the binaural synthesis software (SoundScape Renderer [12]) convolved every BRS file with the corresponding loudspeaker driving signals, which were summed and returned as headphone signals. The binaural renderer updated the ear signals depending on the head orientation of the listeners.

The same loudness of the binaural signals for the different conditions was ensured by correcting the signals for a head orientation of 0° applying a loudness model (non-stationary Zwicker function of the Genesis loudness toolbox 1.2).

## 2.3 Participants

21 Participants (9 females; age range: 23-53; mean age: 29) were recruited. They self-reported no hearing loss or hearing disturbances. Informed written consent was obtained from each participant, and they received a financial compensation. The study received ethical approval from the Technische Universität Berlin Ethics Committee (RA_01_20140422).

## 2.4 Procedure

This study was part of a larger listening test in which other parameters of the mix including compression and reverberation were varied. Pairs belonging to one mix parameter were always grouped together, but the appearance of those groups was randomized. The group containing the mix changes in spatial arrangement consisted of 15 trials, with the whole experiment of 107 trials lasting 45 minutes. In each trial, participants were presented with a pairwise comparison of two temporally aligned clips of music, between which they could switch back and forth. Playback stopped after the end of the 30 s long extract and an answer had to be given to advance to the next trial following an inter-trial interval of one second. Participants could submit their answer before the end of the trial, given that they had heard a minimum of five seconds and had heard each of the two stimuli at least two times. Before the start of the experiment, participants practiced the paradigm twice with the experimenter. Here, another extract from the song was played and one of the tracks was presented at $-6$ dB.

At the end of the experiment the participants completed a verbal survey asking for average daily hours spent listening to music and favorite music genres. Furthermore, participants were asked:

1) *When comparing a pair of stimuli, what did you pay attention to or which attributes of the mix triggered your decision?*

2) *Try to explain reverberation, compression and equalization with respect to music production. Do you have expertise in sound mixing?*

These survey responses were recorded by the experimenter.

## 2.5 Statistical analysis

Suppose there is a number of musical pieces A, B, and C which should be assessed by listeners regarding their preferences. The advantage of the paired comparisons method to achieve this lies in its very few assumptions about the underlying process leading to the choices of the listeners. It is able to measure choices by the listeners, like circular triads where A is preferred over B, B over C and C over A. This can be a completely reasonable choice for stimuli that vary in different aspects. If instead a ranking of the stimuli or a preference rating on a scale is applied, it is already assumed that the rankings lie on a one dimensional perceptual scale [16]. The pair-wise comparison circumvents this restriction and allows an analysis of the underlying dimensions afterwards.

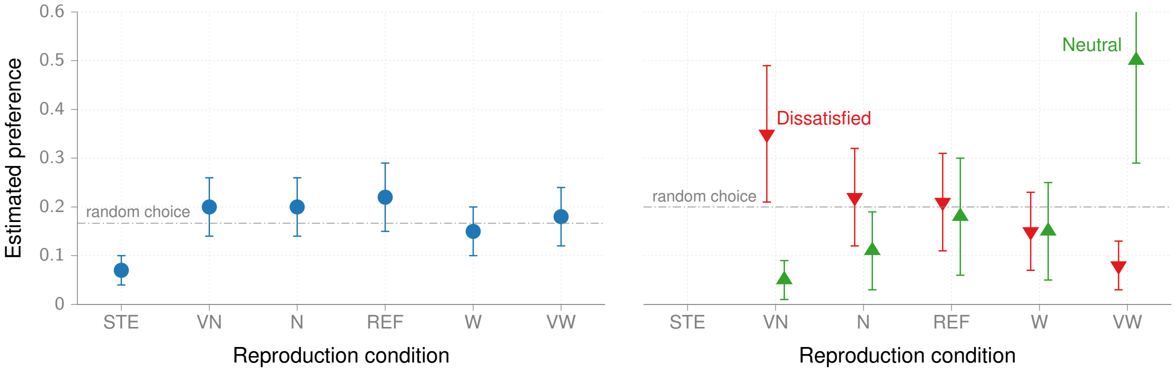

**Fig. 2:** Ratio scale of preference for stereo and five different WFS mixes with changes in the positioning of the foreground elements. The preference is provided as the probability of each condition together with its 95% confidence interval. The conditions were stereo (STE), and five WFS mixes very narrow (VN), narrow (N), reference (REF), wide (W), very wide (VW). In the left graph the results for all listeners are presented, while in the right graph the results are split into the two groups of listeners: dissatisfied or neutral regarding laterally displaced lead vocals.

An indication of a higher dimensional perceptual space is the systematic appearance of a high number of triads. Triads can also stem from inconsistent individual choice behavior and can occur in a non-systematic way when there is no agreement among the listeners. Counting the triads only provides a descriptive measure of the underlying choice process. In order to classify whether the appearance of triads is systematic, a statistical test is required. This can be achieved by fitting a Bradley-Terry-Luce (BTL) model [17] to the data. A $\chi^2$ goodness of fit test compares the estimated BTL model against an ideal saturated model for the paired comparisons. The BTL model holds and is not rejected, as long as the corresponding $p$-value does not drop below 0.1 [19]. If the BTL model holds it indicates that no systematic deviations from a one-dimensional perceptual space occur and estimates the choices of the listener on a ratio scale for that dimension [18].

The current study used the BTL implementation in R (eba-package) after Wickelmaier and Schmid [19]. The BTL values were normalized to sum up to unity and present the probability that a given condition was preferred.

## 3 Results

For the evaluation, the ratings were aggregated over all participants. The number of inconsistent triads was 56. In order to test if a ratio scale could be obtained from the preferences, a BTL model was applied to the paired comparison data and then verified. In the following, every discussed BTL model fits the data, this means its $p$-value of the corresponding $\chi^2$-test never drops below 0.1 and the model is therefore not rejected. The estimated preference was 0.07 for STE, 0.22 for REF, 0.2 for VN, 0.2 for N, 0.15 for W, and 0.18 for VW. Figure 2 displays the results together with their 95% confidence intervals. It is obvious, that all WFS conditions are preferred over stereo, even the VN mix which has a narrower foreground spread than STE. The different WFS mixes are rated equally with only a slight preference towards REF and against W.

To get further insights into the different WFS mixes, STE is excluded from the analysis below. Even if the BTL model fits the data including STE it might be the case that other perceptual dimensions are included in its rating. This seems very likely, as STE is a completely different reproduction method, with general multi-dimensional perceptual differences to WFS.

In order to judge whether there was systematic disagreement between the participants in their ratings, Kendall's coefficient of concordance $w$ was calculated [20]. It ranges from 1 (maximum agreement) to a minimum very close to zero, which is $w_{\min} = -0.05$ in our scenario. The corresponding $p$-values indicate how likely the agreement between judges is by chance, derived from a $\chi^2$-test. For all observations regarding positioning, a relatively low agreement of $w = 0.05$ ($p = 0.01$)

is found, which indicates that there might by a systematic disagreement between the participants.

To analyze this disagreement, the participants are split into different groups based on the questionnaires. For Question 1, twelve participants reported they were dissatisfied when lead tracks—especially the lead vocals—were shifted outside the center. Nine subjects did not report any attributes directly related to positioning. This leads to grouping *dissatisfied* and *neutral*. Additionally, a second grouping was analyzed. If participants were able to answer at least three of the four points in Question 2, they were categorized as *experts*. This resulted in five *expert* and 16 *naive* listeners.

Regarding group effects, the likelihood ratio of individually calculated BTL models reveal whether two groups of subjects rated differently. This compares whether the combination of both group model likelihoods is significantly higher than the likelihood of the model calculated from the entire population, as this distance is approximately $\chi^2$-distributed. For expert versus naive listeners, there is no significant difference between both groups, according to $p = 0.14$ from the corresponding $\chi^2$-distribution.

For the two groups of participants that reported to be either neutral or dissatisfied with laterally shifted lead vocals, a significant difference was found ($p < 0.01$). This difference was quantified via Kendall's rank correlation coefficient $\tau$ which describes the ordinal association between paired group outcomes. Correlation between participants neutral or dissatisfied with laterally shifted lead vocals is $\tau = -0.77$ with $p < 0.01$ ($H_0$ : *The actual* $\tau$ *is* 0). This negative correlation is also visible in Fig 2; the preferences rise inversely to one another.

## 4 Discussion

In the authors' previous work [1, 21], reproduction-specific mixes of different multi-track musical pieces were created, specifically tailored to stereo, 5.1-surround and WFS reproduction. Stereo was detected to be inferior to the higher channel systems, especially to WFS. It is possible that the object-based mix for WFS was simply mixed better, biasing responses towards WFS. The present study tries to quantify this for the mix attribute spatial foreground arrangement for the music track created for the previous study.

The results reject this argumentation and underline the previous results. The participants favored even the least preferred WFS version over stereo and hence positioning, if any, is not advantageously applied in WFS. Presumably an attribute independent from foreground positioning, that remains constant to some extent, ensures the high WFS ranking.

However, there was low agreement of the listeners regarding the optimal spatial arrangement. From the results of the survey, participants were divided into two groups resembling a neutral and dissatisfied attitude towards laterally displaced lead vocals. Those two groups rated almost inversely on the WFS conditions, while agreeing on the reference (REF), there was strong disagreement on the very wide (VW) and very narrow (VN) versions. The very wide spread foreground elements, which is an untraditional variant nowadays, may have aroused interest and thus get preferred by some listeners. Besides novelty, wide source spreading likely produces higher separation. Although this is avoided in many cases, it could have been pleasant in this particular scenario. Separated content appears more prominent, perhaps this interacts positively with other characteristics of the mix.

The upcoming content production for spatial audio systems faces the challenge to serve both listener groups. One approach might be to apply a wide spatial arrangement, but leave the lead vocals in the center. An option might be allowing the user to switch between different versions of the mix, as user interaction can be another application of object-based audio.

## 5 Summary

The results of this study support that it is worth considering object-based audio and multi-channel audio systems, since stereo cannot keep up with the tested WFS system. The investigation further constitutes one step towards connecting object-based audio mixing techniques and their impact on listener preference. Here, the results highlight that the balance between spatial richness and unpleasant performance of a mix seems to depend strongly on the listener. Some listeners clearly favor wide spatial arrangements where others seem to stick to the classical foreground arrangements in pop, even if they prefer more spacious audio systems.

## 6 Acknowledgements

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

## References

[1] Hold, C., Wierstorf, H., and Raake, A., "The Difference Between Stereophony and Wave Field Synthesis in the Context of Popular Music," in *140th Conv. Audio Eng. Soc.*, paper 9533, 2016.

[2] Västfjäll, D., "The Subjective Sense and Experienced Emotions in Auditory Virtual Environments," *Cyberpsychol. Behav.*, 6(2), pp. 181–188, 2003.

[3] Theile, G., Wittek, H., and Reisinger, M., "Potential wavefield synthesis applications in the multichannel stereophonic world," in *24th Int. Conf. Audio Eng. Soc.*, paper 35, 2003.

[4] Wierstorf, H., *Perceptual Assessment of sound field synthesis*, Ph.D. thesis, Technische Universität Berlin, 2014.

[5] Woodcock, J., Davies, W. J., Cox, T. J., and Melchior, F., "Categorization of broadcast audio objects in complex auditory scenes," *J. Audio Eng. Soc.*, 64(6), pp. 380–394, 2016.

[6] Wilson, A. and Fazenda, B., "Relationship Between Hedonic Preference and Audio Quality in Tests of Music Production Quality," in *QoMEX*, pp. 1–6, 2016.

[7] Wilson, A. and Fazenda, B., "Perception & evaluation of audio quality in music production," in *DAFx 2013*, 2013.

[8] Västfjäll, D., Larsson, P., and Kleiner, M., "Emotion and Auditory Virtual Environments: Affect-Based Judgments of Music Reproduced with Virtual Reverberation Times," *Cyberpsychol. Behav.*, 5(1), pp. 19–32, 2002.

[9] Hold, C. and Wierstorf, H., "Object-based audio scene files for variations of the spatial arrangement in pop mixes for Wave Field Synthesis," 2016, doi:10.5281/zenodo.61110.

[10] Hold, C. and Wierstorf, H., "Signal feeds for creating the music mixes for comparison of wave field synthesis, surround, and stereo," 2016, doi: 10.5281/zenodo.55718.

[11] Hold, C., , Nagel, L., Raake, A., and Wierstorf, H., "Variations of pop mixes for Wave Field Synthesis," 2016, doi:10.5281/zenodo.61000.

[12] Geier, M., Ahrens, J., and Spors, S., "The SoundScape Renderer : A Unified Spatial Rendering Methods," *124th Conv. Audio Eng. Soc.*, paper 7330, 2008.

[13] Horbach, U., Karamustafaoglu, A., Pellegrini, R., Mackensen, P., and Theile, G., "Design and Applications of a Data-based Auralization System for Surround Sound," in *106th Conv. Audio Eng. Soc.*, paper 4976, 1999.

[14] Wierstorf, H., "Binaural room scanning files for a 56-channel circular loudspeaker array," 2016, doi:10.5281/zenodo.55572.

[15] Wierstorf, H., Geier, M., and Spors, S., "A Free Database of Head Related Impulse Response Measurements in the Horizontal Plane with Multiple Distances," in *130th Conv. Audio Eng. Soc.*, eBrief 6, 2011.

[16] Kendall, M. G. and Smith, B. B., "On the method of paired comparisons." *Biometrika*, 34(Pt 3-4), pp. 324–345, 1947.

[17] 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.

[18] Choisel, S. and Wickelmaier, F., "Evaluation of multichannel reproduced sound: Scaling auditory attributes underlying listener preference," *J. Acoust. Soc. Am.*, 121(1), pp. 388–400, 2007.

[19] Wickelmaier, F. and Schmid, C., "A Matlab function to estimate choice model parameters from paired-comparison data." *Behav. Res. Meth. Ins. C.*, 36(1), pp. 29–40, 2004.

[20] Legendre, P., "Species Associations: The Kendall Coefficient of Concordance Revisited," *J. Agr. Biol. Envir. St.*, 10(2), pp. 226–245, 2005.

[21] Hold, C. and Wierstorf, H., "Music mixes for comparison of wave field synthesis, surround, and stereo," 2016, doi:10.14279/depositonce-5173.