Evaluation of sound field synthesis techniques with a binaural auditory model

Marko Takanen[1], Hagen Wierstorf[2], Ville Pulkki[1], and Alexander Raake[2]

[1]*Aalto University School of Electrical Engineering, Department of Signal Processing and Acoustics, Espoo, Finland*

[2]*Technische Universität Berlin, Telekom Innovation Laboratories, Berlin, Germany*

Correspondence should be addressed to Marko Takanen (`marko.takanen@aalto.fi`)

ABSTRACT

Wave field synthesis and Ambisonics strive to reconstruct a sound field within a listening area using the interference of loudspeaker signals. Due to the spatial sampling, an error-free reconstruction is not achieved within the entire listening area and consequently, the perceived quality of the reproduction may be impaired. Specifically, sound events may be localized incorrectly and the individual loudspeaker signals may result in perceived coloration. Here, a binaural auditory model was employed to predict the localization error in several off-center-listening positions and to visualize coloration artifacts. The model outputs provided good match for perceptual data from previously conducted listening tests, verifying the applicability of the model to evaluate the reconstructed sound fields.

1. INTRODUCTION

Spatial sound reproduction techniques employing several loudspeakers have become feasible over the past years. Currently, there are various techniques that strive to produce a plausible spatial impression using a multichannel loudspeaker setup. One approach is to record each sound event separately and to position the events around the listener using panning algorithms. In addition, the old idea of Steinberg & Snow [1] to synthesize a whole sound field is exploited in sound field synthesis techniques that employ a large number of loudspeakers [2].

Optimization of the above-mentioned techniques requires tuning of several parameters affecting the overall performance. Computational auditory models could provide an appealing alternative to the direct use of human listeners by predicting the effects of the modifications on the perception. The authors of the current paper have independently of each other applied binaural models to predict the spatial perception of different reproduction techniques. A modified version of the binaural model by Dietz *et al.* [3] predicted the perceived direction of synthesized sources in sound field synthesis [4]. Spatial artifacts introduced by nonlinear time-frequency-domain techniques were evaluated in [5] with a model [6] that emulates the functionality of the nuclei in the auditory pathway based on neurophysiological data.

The present study combines challenging results from listening experiments in sound field synthesis with predictions of the binaural model [6]. The goal is to see if this model is able to give meaningful insights also in the perception of sound field synthesis methods.

2. METHODS I: SYNTHESIS TECHNIQUES

The basic principle of sound field synthesis is that the sound pressure in an extended audience area is given by the sound pressure at its boundary. In practice, the pressure on the boundary is controlled via loudspeakers. The usage of loudspeakers opposes a spatial sampling process on the boundary which results in errors in the synthesized sound field if the distance between the loudspeakers is too large. For example, loudspeaker distances below 1 cm are needed to achieve an error free sound field up to a frequency of 20 kHz. This is not achievable at the moment and distances around 15 cm between the loudspeakers are employed. With such a setup, the sound field is correct only up to frequencies of around 1.5 kHz. Due to the dominance of the low frequencies for localization of a sound source the directional perception of such sound fields is in most cases not impaired [4]. What is impaired by the errors in the higher frequencies is the timbre of the synthesized sound field [7].

This study focusses on loudspeaker setups that employ even larger distances between the loudspeakers. In all

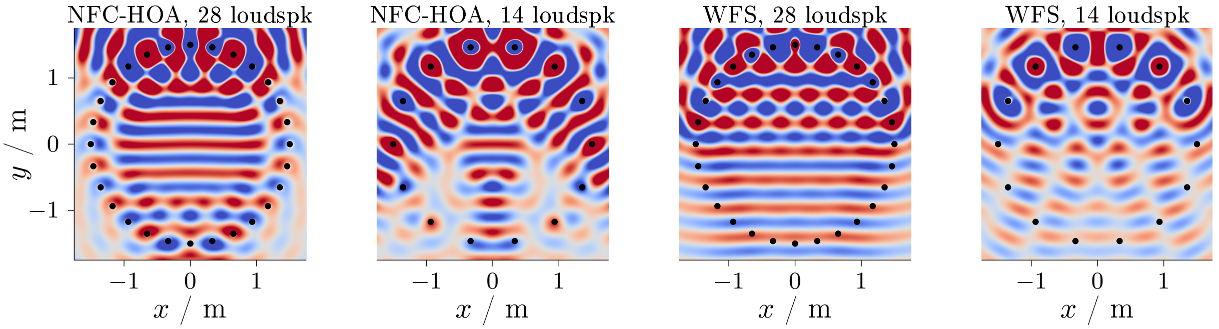


Fig. 1: Sound pressure of a synthesized monofrequent plane wave going into the direction $(0, -1)$. The plane wave is synthesized by band-limited NFC-HOA and WFS and has a frequency of 800 Hz. The black dots indicate the loudspeaker positons.

cases, a circular loudspeaker array with a diameter of 3 m is used, but the number of loudspeakers is either 14 or 28, corresponding to either 67 cm or 34 cm spacing between the adjacent loudspeakers. For these cases, errors in the synthesized sound field are expected even below 1 kHz. Suppressing such errors can be achieved by distributing them differently in the synthesized sound field. One option is derived with near-field compensated higher order Ambisonics (NFC-HOA). In NFC-HOA, it is possible to limit the order of the spherical harmonics to half the number of applied loudspeakers for a circular loudspeaker array. This effectively minimizes the errors in the synthesized sound field at the center of the listening area. Figure 1 shows the synthesized sound field of a mono-frequency plane wave with a frequency of 800 Hz traveling into the direction $(0, -1)$. The plane wave was synthesized by band-limited NFC-HOA and in comparison with Wave Field Synthesis (WFS) without band limitation. In the last case, the errors in the sound field are more equally distributed in x-direction.

The error free region in the sound field is relatively small already at 800 Hz for NFC-HOA employing 14 loudspeakers. This raises the question how the errors affect the localization of the plane wave outside of the central listening position. The next sections will present listening results that investigated this question and the results of the binaural model for the same question.

3. METHODS II: AUDITORY MODEL

The computational evaluations were performed using a count-comparison principle [8] based binaural auditory model [6]. As depicted in Fig. 2, the model, implemented in Matlab, takes digitized binaural signals as input and processes them first through the two periphery models, one in each hemisphere. The periphery model emulates the functionality of the cochlea with a nonlinear time-domain model [9] and the conversion of the basilar membrane movement into neural signals with the inner hair cell model by Meddis [10]. Following the neuroanatomy of the auditory pathway, the outputs of the two periphery models are then projected into the functional models of the medial and lateral superior olives (MSO and LSO, respectively), that decode the binaural cues in the input signal according to the count-comparison principle. In addition to the MSO and LSO models, the binaural model contains a psychoacoustically-motivated wideband MSO model to account for the human selectivity to envelope-ITDs in broadband sounds (see, e.g. [11]).

The subsequent steps in the binaural auditory model were designed based on psychoacoustical knowledge about spatial sound perception with the aim to form a topographically organized map. This is accomplished using the directional cues to map the periphery model outputs onto an one-dimensional binaural activity map. The directional cues are obtained by combining the MSO and LSO model outputs, separately for each hemisphere, while emulating the tendency of the auditory system to favor off-median plane cues in localization [12]. In addition, the dominant role of onsets on localization [13] is emulated by enhancing the visibility of the onsets on the binaural activity map. As an output, the model produces a binaural activity map, where the left/right acti-

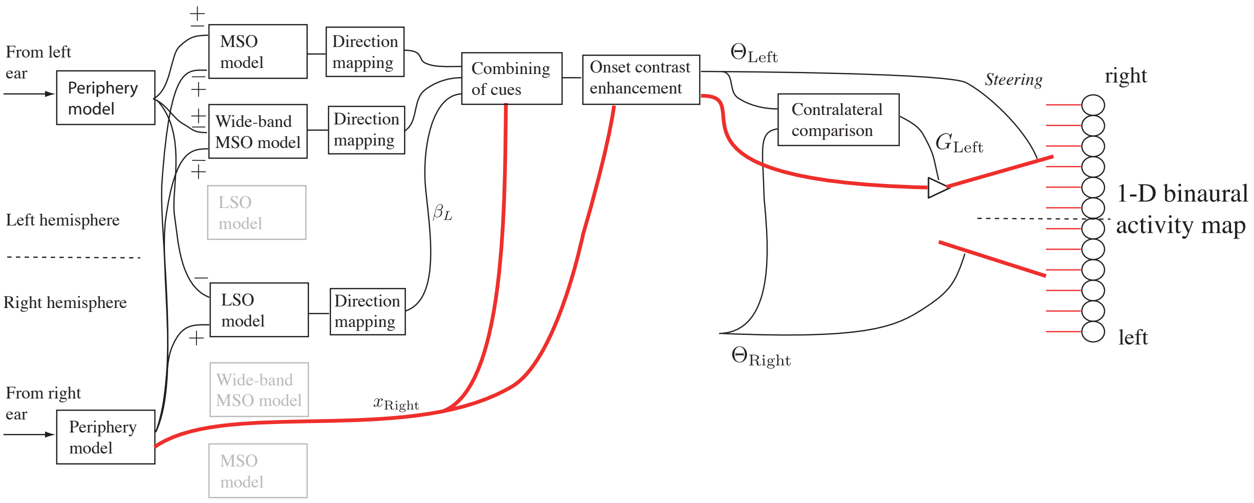


Fig. 2: Illustration on how the binaural activity map is formed from the left and right ear canal inputs. Only the pathways to the activation projected to the left hemisphere are shown for simplicity.

vation location on the map indicates the spatial arrangement of the sound scenario, and different colors are used to represent different frequency regions for visualization purposes (see Fig. 4(i)). The functionality of the model is described in detail in [6].

4. METHODS III: EVALUATIONS

Several binaural listening scenarios were simulated to evaluate the sound field synthesis capabilities of WFS and NFC-HOA. The scenarios of this study were selected among the ones employed in the listening tests [14] in order to compare the model outputs to the perceptual data. That is, to test whether the artifacts seen in the binaural activity maps are in accordance with the results of the listening tests. Both the listening tests and the evaluations of this study used binaural stimuli where the different scenarios were simulated using head-related transfer functions (HRTFs). The similarity of the stimuli enables direct comparison between the model outputs and the perceptual data. The binaural stimuli were obtained with the Sound Field Synthesis Toolbox [15].

In each scenario, WFS or NFC-HOA was used to synthesize the sound field using a circular loudspeaker array consisting of either 14 or 28 loudspeakers (see Fig. 1). Two different kind of sound fields were to be synthesized – a point source at the direction of 0° from a listener at a central position, and an impulse-like sound at the direction of $-36°$.

The first-named sound field condition were included to evaluate the effective listening area of different sound field synthesis techniques, and the resulting activity maps were compared to results of a listening test [14]. In this test, the subjects were requested to indicate the perceived direction of the sound source in corresponding conditions. The listening test contained also a follow-up experiment, where the participants were asked to indicate whether they heard one or two sound sources. Similarly to the listening test, the size of the effective listening area of each synthesized sound field was evaluated here by simulating scenarios where the listener would be sitting at different positions along the horizontal axis, the center being at $(0,0)$, and the furthest off-center condition being at $(-1.25,0)$, i.e. at 0.25 m distance from the nearest loudspeaker. A 200-ms-long pink noise signal, with a 10-ms-long linear rise and decay was used both as the sound propagating as a plane wave and as the sound emitted by the point source.

The last-named sound field condition is related to another listening test [7], where the participants rated the impairments of different samples, i.e. synthesized sound fields, in comparison to the known reference that was a point source. The impairments were to be rated in terms of coloration aspects. Several different stimuli (music, pink noise, and speech) were used in the listening test. However, a different kind of stimulus needed to be used in this study in order to evaluate such impairments based on the binaural activity map provided by the model. Consequently, the simulated scenario consisted of a point-source, at the direction of $-36°$ from the listener's point of view, emitting a 4-ms-long white noise burst with a 2-ms-long linear rise and decay.

5. RESULTS AND ANALYSIS

Point source

A sound emitted by a point source is perceived from the same *position* in the given environment, but the perceived *direction* of the auditory event changes as the listener moves within the listening area. In the evaluated scenario, the point source is located at $(0, 2.5)$, i.e. directly in front and at 2.5 m distance from the central listening position. Hence, the synthesized sound field should evoke the perception of an auditory event directly in front of the listener when he or she is located at the center. A listener on the left from the central position should perceive the event on the right. The results of the listening test [14] show that the sound event is indeed localized accurately in all reconstructed sound fields when the listener is located at the central position (Fig. 3).

The techniques differ in the size of the listening area within which the sound field is reconstructed accurately. The listening area appears to be the smallest with the 7th-order NFC-HOA as, e.g. at $(-0.5, 0)$, the sound event is localized to the front of the listener, and not towards the correct location of the sound event like in the case of the 14th-order NFC-HOA. Interestingly, listeners positioned at 1.0 m distance from the center perceived two auditory events when listening to the 7th-order NFC-HOA reproduction (see Fig. 3). As expected, the 14th-order NFC-HOA reproduction achieves a larger effective listening area than the 7th-order does – the localization error for the 14th-order Ambisonics is still small at $(-0.5, 0)$ and increases substantially only at the two furthest off-center positions, namely at $(-1.0, 0)$ and $(-1.25, 0)$ (Fig. 3). The effective listening area seems to be wider in WFS as the sound event is localized closer to the desired direction with both WFS reproductions also at the furthest off-center positions [14].

The corresponding activity maps in Fig. 4 also visualize several related spatial artifacts in the reconstructed sound fields. For instance, the activation evoked by the 7th-order NFC-HOA spreads over a large area in Fig. 4(b) but does not shift away from the center like it does in Fig. 4(c). The activity map in Fig. 4(d) contains two activation areas, one in each hemisphere, reflecting the above-mentioned perception of two auditory images. Similarly, Figs. 4(e–f) show increased activation on the left side that is caused by the signals emitted by loudspeakers closest to the listener in the 14th-order NFC-HOA reproduction at the furthest off-center positions.

The activity maps seem to be in good agreement with the perceptual data. However, the agreement needs to be validated via numerical analysis. Here, the approach utilized by Stern & Colburn [16] is exploited and the mass centroid of the model output is used to predict the perceived direction of the sound event. In [16], such an approach yielded accurate predictions for lateralization-matching experiments. In this case, the mass centroid is acquired from the histogram

$$H(l) = \frac{\sum_{t=0}^{T}\left(\sum_{z=1}^{6} M(l,t,z)\right)}{\max\left[\sum_{t=0}^{T}\left(\sum_{z=1}^{6} M(l,t,z)\right)\right]} \qquad (1)$$

that is computed to describe the average distribution of the activation on the map M. Here, l denotes the left-right location on the map with the discrete values $-1, -0.9, \ldots, 1$, z describes the frequency region (see Fig. 4(i)), and t denotes the time. As the strength of the image on the map depends on the level of the periphery model output at the given frequency band [6], the histogram may be considered as the energy-weighted average of the activations at different frequency regions at different time instants. The resulting histograms in Fig. 5 illustrate how densely the activation is concentrated at a specific area on the corresponding activity map shown in Fig. 4. Subsequently, the mass centroid

$$C = \frac{\sum_{l=-1}^{1} H(l)l}{\sum_{l=-1}^{1} H(l)} 90° \qquad (2)$$

is computed based on the histogram. Here, the mass centroid is transformed into azimuthal direction with the multiplication by 90°. Upon acquisition of the mass centroid values for each reconstructed sound fields, similar values were computed for the corresponding reference conditions having the point source at $(0, 2.5)$ and the listener at $(0, 0)$, $(-0.5, 0)$, $(-1, 0)$, and at $(-1.25, 0)$. Similarly as before, the conditions were simulated using the Sound Field Synthesis Toolbox [15] to obtain the binaural stimuli that were processed with the binaural auditory model and the mass centroids were computed based on the activity maps provided by the model. The difference between the values acquired for the reconstructed sound fields and for the corresponding reference conditions is used as a measure to predict the localization error.

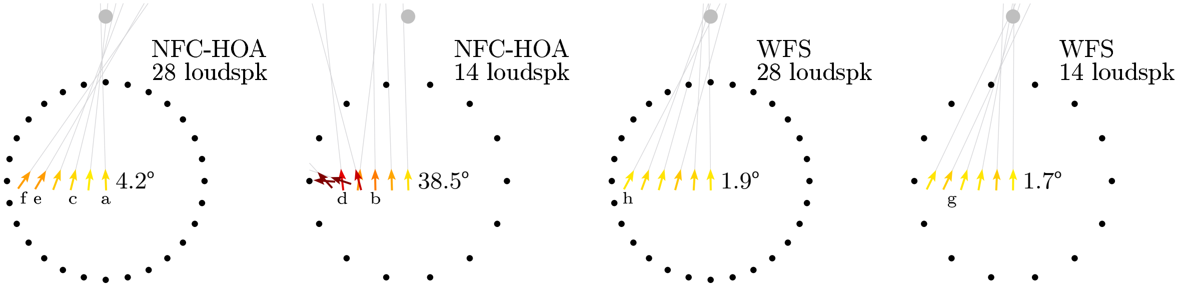


Fig. 3: Localization results for a point source synthesized with different sound field synthesis methods. The arrows are pointing towards the direction a listener, placed at the center of the arrow, perceived the corresponding auditory event. The number to the right of the arrows is the mean absolute deviation of the perceived direction from the desired one. The corresponding modeling graph is indicated by the small letter under the arrow.

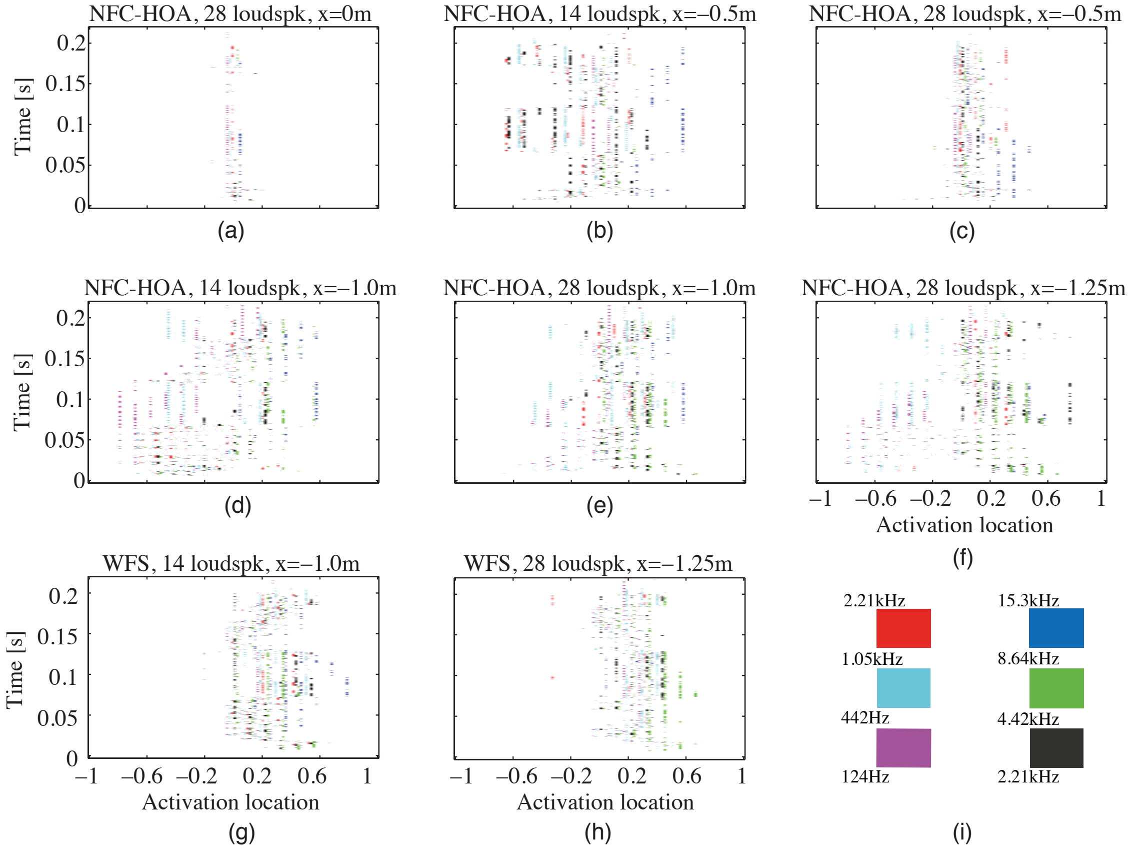
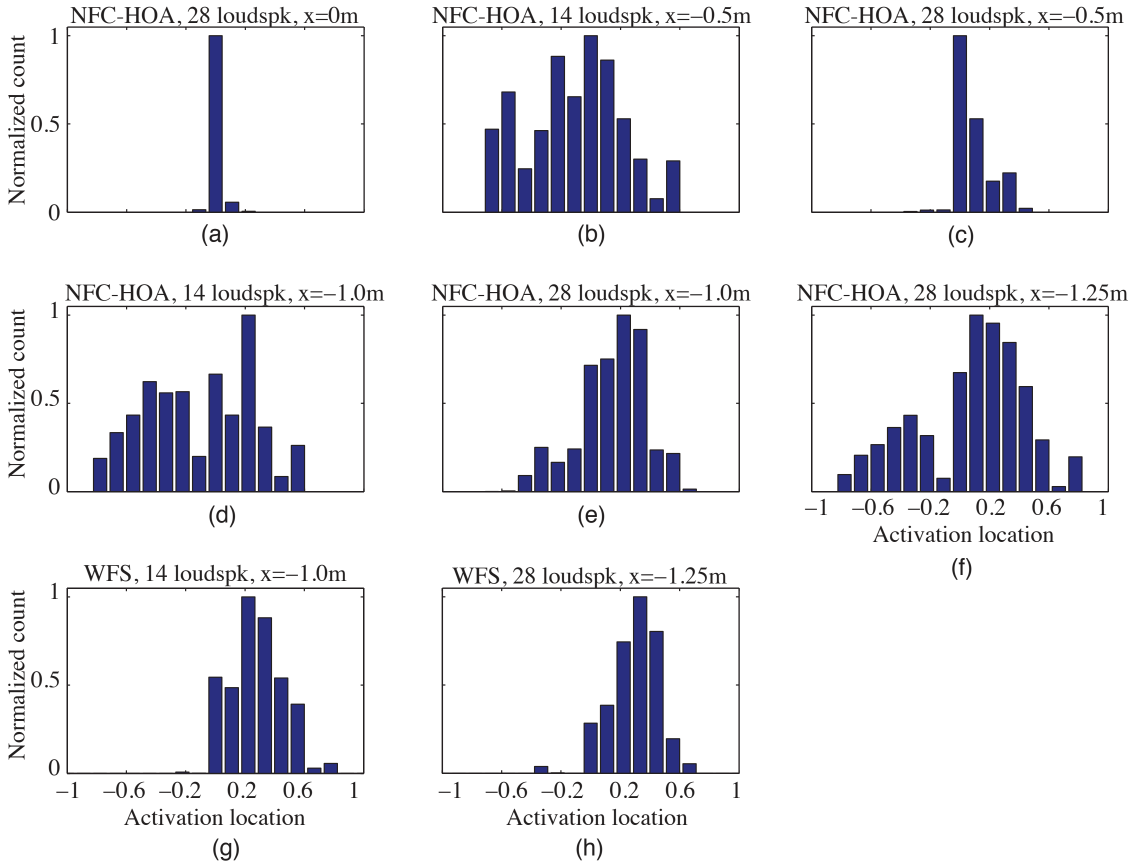


Fig. 4: Binaural activity map for a point source at $(0, 2.5)$ when the listener is simulated to be at $(0, 0)$ or at various off-center positions, and the sound field is synthesized either with WFS or NFC-HOA.

Table 1 summarizes the predicted values and the corresponding localization errors [14].

The model predicts the perceived direction quite well for all conditions besides (e) and (f) which are the NFC-HOA conditions with 28 loudspeakers and a listener



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 5: Histograms of the activation on the activity maps shown in Fig. 4. The highest peak in each histogram has been normalized to have the value of 1 to ease visual comparison.

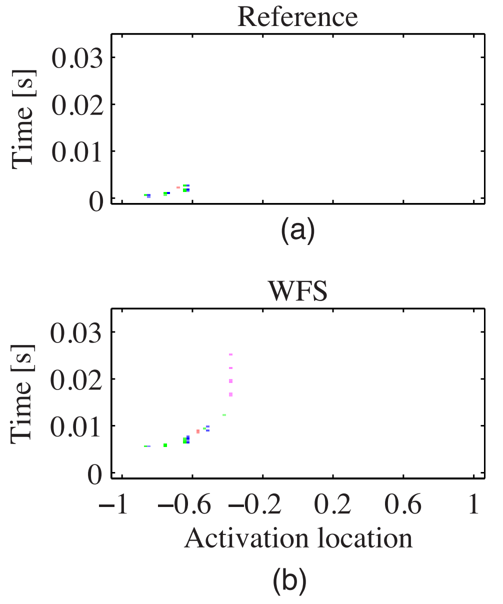
Table 1: Localization errors of the listening experiment [14] and the corresponding predictions for the eight evaluated scenarios illustrated in Fig. 3.

#	Loc. error - model	Loc. error - experiment
(a)	-0.48	1.75
(b)	15.79	12.05
(c)	-2.20	-2.31
(d)	32.10	33.79
(e)	9.05	-8.61
(f)	16.03	-8.47
(g)	-3.01	-3.28
(h)	-3.86	-0.76

placed to the side. The mean absolute deviation of the predicted direction from the perceived one is 7° considering all conditions. Conditions including NFC-HOA with low orders are still a challenging task for binaural models. The same conditions were fit into another binaural model that predicted results for WFS with high accuracy [14]. It achieved a mean absolute deviation of 8° showing that the appllied conditions are challenging.

Impulse-like sound

The above-mentioned small differences between the expected and the perceived direction of a sound event in WFS are achieved partly due to the precedence effect (for a review, see [17]). That is, WFS is able to reconstruct the first wavefront correctly and therefore, the sound event is localized to the expected direction despite the later arrival of interfering sounds from individual loudspeakers [18]. Although the localization is not impaired by these excess sounds, the sound event is perceived as colored [18, 7].



(a)

(b)

Fig. 6: Binaural activity map for (a) an impulse in the direction of $-36°$, and (b) the corresponding synthesis with WFS employing 28 loudspeakers.

This study investigates whether the binaural auditory model can detect the excess sounds arriving after the first wavefront and consequently, to evaluate coloration artifacts in WFS. Figure 6 shows the activity maps obtained for the reference, i.e. an impulse-like sound at $-36°$, and for the sound field reconstructed with WFS. The activity map also shows that the first wavefront of the reconstructed sound field does indeed evoke activation in the same location as does the reference stimulus. However, also the later arriving interfering loudspeaker signals can be seen to evoke activation on the map (Fig. 6(b)). Hence, the model output is in accordance with the perception.

6. DISCUSSION

Binaural auditory models could provide an interesting tool for developers of spatial sound reproduction techniques, by being able to predict the effects of different modifications on the quality perception. The model used in this study aims to ensure its general applicability in various tasks via accurate emulation of the functionality of the nuclei in the auditory pathway. The abovementioned evaluations verified that also sound fields reconstructed with WFS and NFC-HOA can be evaluated with such a model in a manner that the model outputs are in line with the human perception. The model was able to visualize both spatial and coloration artifacts in the reconstructed sound fields. Moreover, numerical values computed based on the activity maps predicted accurately the perceptual data about the about the localization error in the corresponding scenarios. Thus, the study presents one of the first successful attempts to evaluate sound field synthesis techniques based on both spatial and timbral aspects using a binaural auditory model.

However, a visual inspection of the resulting binaural activity map is needed to evaluate whether the output of the model matches with the human perception. Visual comparisons are also required for detecting artifacts in the reconstructed sound fields. However, as demonstrated above, the effect of these artifacts on some perceptual aspects can already be analyzed numerically based on the activity map. Still, more general instrumental evaluation of the reconstructed sound fields requires development of high-level algorithms that would do predictions based on the activity map. In addition, the spatial resolution of the model needs improvement.

7. CONCLUDING REMARKS

Wave Field Synthesis and Ambisonics can reconstruct a sound field within a listening area by exploiting the interference of loudspeaker signals. Both techniques can achieve the larger effective listening area the denser loudspeaker arrays are used since the spatial aliasing is reduced. The traditional approach to evaluate the performance of these techniques consist of either visual or instrumental inspection of the reconstructed sound fields. Here, an alternative approach was used. Sound fields reconstructed with WFS and NFC-HOA were evaluated with a binaural auditory model [6] and the model outputs were compared to listening test results. Specifically, several binaural listening scenarios were simulated using HRTFs and the resulting binaural signals were processed with the model to obtain binaural activity maps for the different scenarios. The activity maps showed artifacts in the reconstructed sound fields at off-center-listening conditions, being in good agreement with the perceptual data about the evoked spatial impression [14]. A numerical analysis verified the accuracy of the model prediction about the localization error caused by the artifacts. In addition, the activity maps visualized how the individual loudspeaker signals result in audible coloration artifacts in WFS [7], although the first wavefront is reconstructed correctly. Consequently, it was demonstrated that the binaural auditory model can be used to evaluate the performance of sound field synthesis techniques, and to aid in the further development of such techniques.

ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement No. 240453, from the Academy of Finland, and DFG RA 2044/1-1.

8. REFERENCES

[1] J. C. Steinberg and W. B. Snow, "Symposium on wire transmission of symphonic music and its reproduction in auditory perspective: Physical Factors," *Bell System Technical Journal*, vol. 13, no. 2, pp. 245–58, 1934.

[2] S. Spors, H. Wierstorf, A. Raake, F. Melchior, M. Frank, and F. Zotter, "Spatial Sound With Loudspeakers and Its Perception: A Review of the Current State," *Proceedings of the IEEE*, vol. 101, no. 9, pp. 1920–38, 2013.

[3] M. Dietz, S. D. Ewert, and V. Hohmann, "Auditory model based direction estimation of concurrent speakers from binaural signals," *Speech Comm.*, vol. 53, pp. 592–605, May 2011.

[4] H. Wierstorf, S. Spors, and A. Raake, "Binaural Assessment of Multichannel Reproduction," in *The Technology of Binaural Listening* (J. Blauert, ed.), pp. 255–278, Springer-Verlag, Berlin, Germany, 2013.

[5] M. Takanen, O. Santala, and V. Pulkki, "Binaural Assessment of Parametrically Coded Spatial Audio Signals," in *The Technology of Binaural Listening* (J. Blauert, ed.), pp. 333–358, Springer-Verlag, Berlin, Germany, 2013.

[6] M. Takanen, O. Santala, and V. Pulkki, "Visualization of functional count-comparison-based binaural auditory model output," *Hear. Res.*, vol. 134, pp. 147–163, Mar. 2014.

[7] H. Wierstorf, C. Hohnerlein, S. Spors, and A. Raake, "Coloration in wave field synthesis," in *Proc. AES 55th Conference*, (Helsinki, Finland), Aug. 27-29 2014.

[8] G. von Békésy, "Zur Theorie des Hörens. Über das Richtungshören bei einer Zeitdifferenz oder Lautstärkeungleighheit der beiderseitigen Schalleinwirkungen," *Physik. Zeitschr.*, pp. 824–835, 857–868, 1930.

[9] S. Verhulst, T. Dau, and C. A. Shera, "Nonlinear time-domain cochlear model for transient stimulation and human otoacoustic emission," *J. Acoust. Soc. Am.*, vol. 132, pp. 3842–3848, Dec. 2012.

[10] R. Meddis, "Simulation of mechanical to neural transduction in the auditory receptor," *J. Acoust. Soc. Am.*, vol. 79, pp. 702–711, Mar. 1986.

[11] C. Trahiotis and R. M. Stern, "Lateralization of bands of noise: Effects of bandwidth and differences of interaural time and phase," *J. Acoust. Soc. Am.*, vol. 86, pp. 1285–1293, Oct. 1989.

[12] W. A. Yost, "Lateral position of sinusoids presented with interaural intensive and temporal differences," *J. Acoust. Soc. Am.*, vol. 70, pp. 397–409, Aug. 1981.

[13] M. Dietz, T. Marquardt, N. Salminen, and D. McAlpine, "Emphasis of spatial cues in the temporal fine structure during the rising segments of amplitude-modulated sounds," *PNAS.*, pp. 1–6, Aug. 2013.

[14] H. Wierstorf, A. Raake, and S. Spors, "Localization in Wave Field Synthesis and higher order Ambisonics at different positions within the listening area," in *DAGA*, 2013.

[15] H. Wierstorf and S. Spors, "Sound field synthesis toolbox," in *Proc. AES 132nd Convention*, (Budapest, Hungary), Apr. 26-29 2012. eBrief No. 50.

[16] R. M. Stern and H. S. Colburn, "Theory of binaural interaction based on auditory-nerve data. IV. A model for subjective lateral position," *J. Acoust. Soc. Am.*, vol. 64, pp. 127–140, Jul. 1978.

[17] R. Litovsky, S. Colburn, W. A. Yost, and S. Guzman, "The precedence effect," *J. Acoust. Soc. Am.*, vol. 106, pp. 1633–1654, Oct. 1999.

[18] E. Corteel, "Equalization in an Extended Area Using Multichannel Inversion and Wave Field Synthesis," *J. Audio Eng. Soc.*, vol. 54, pp. 1140–1161, Dec. 2006.