7 - 12 SEPTEMBER 2014, KRAKÓW
FORUM ACUSTICUM

Localization Properties of Data-based Binaural Synthesis including Translatory Head-Movements

Fiete Winter, Frank Schultz, and Sascha Spors
Institute of Communications Engineering, University of Rostock, R.-Wagner-Str. 31 (H8), D-18119 Rostock, Germany

Summary
Binaural synthesis of plane wave decomposed spherical microphone data using head-related transfer functions (HRTFs) is a well-known approach for auralization. Rotational head movements can be considered by dynamic rotation of the HRTF dataset. Translatory head movements are coped by spatio-temporal shifts of the individual planes waves.
This paper analyses this auralization method with respect to the localization of sound sources. The algorithm's performance is evaluated with respect to the accuracy achieved by a binaural model utilized for localization. The influence and effects of spatial sampling (number of plane waves) and the translatory head-movement are investigated. In addition, the modal resolution is taken into account.

PACS no. 43.66.Ba, 43.60.Fg

1. Introduction

Binaural synthesis utilizing head-related transfer functions (HRTFs) is a common approach to auralize acoustic sources. HRTFs represent the acoustic free field transmission path from the source to the outer ears. Due to a varying anatomy, HRTFs differ amongst invididuals. In addition, HRTFs are dependent on the head/body-orientation and position with respect to the source. While transfer functions in free-field conditions are usually termed HRTFs, binaural room transfer functions (BRTFs) include additional room reflections. For a virtual acoustic scene, left and right ear drum signals are rendered by filtering an anechoic signal of a virtual source with the left and right ear HRTF/BRTF.
In order to enable head-rotations and translatory movement of the listener, dynamic binaural synthesis selects the actual HRTF/BRTF accordingly. This requires a dataset containing a (densely) sampled grid of HRTFs/BRTFs for all possible head-rotations and translatory shifts. Due to the room reflections, BRTFs are not invariant with the respect to rotations/translatory shifts. Therefore, the measurement effort for the required dataset would be considerably, especially for BRTFs. Binaural synthesis using HRTFs/BRTFs suffers from two limitations: The rendering of many sound sources or diffuse sound fields cannot be realized efficiently due to the large number of required BRTF datasets. Measuring individual BRTFs (for each listener and each individual room) is technically demanding and very time consuming.
A combination of sound field analysis (SFA) techniques and HRTF-based binaural synthesis can be used to overcome these limitations. In SFA diverse approaches are known that decompose a captured sound field into plane waves. Spherical microphone arrays exhibit properties, which are independent of the direction of the impinging waves and are therefore preferred for the analysis of multi-path sound fields. Due to the geometry, the captured sound field may be decomposed into surface spherical harmonics which is known as modal beamforming. An alternative decomposition can be realized by so called delay-and-sum beamforming. Practical implementations are limited with respect to the number of sampling positions of the spherical microphone array and equipment self-noise. As a result, data-based binaural synthesis may suffer from inaccuracies resulting from limited spatial bandwidth and noise amplification. Combining SFA and HRTFs, the sound pressure at the left respectively right ear for an actual head-orientation is given as the superposition of the associated (far-field) HRTFs filtered by the plane wave expansion coefficients of the captured sound field. For translatory motion the plane wave expansion is first computed with respect to the

center position of the microphone array. This reference point can then be shifted by applying a phase shift to the plane wave expansion coefficients in the spatial frequency domain. If this phase shift is applied before the filtering process with the far-field HRTFs and the subsequent superposition, the ear-signals for a translatory shifted head position are yielded [1]. Hence, translatory head-movements of the listener can be considered explicitly with no additional measurement effort.

The paper analyses the localization properties of the introduced approach utilizing a binaural model. First the mathematical background of this approach is recollected. Then the influence of the translatory head-movement and resolution of the plane wave decomposition (spatial resolution) on the localization performance is investigated. The effects of a limited modal resolution are analysed in addition for the modal beamforming technique. Although the modal resolution is closely related to the number of microphones, discrete sampling positions are out of the scope of this paper and array apertures are assumed to be continuous.

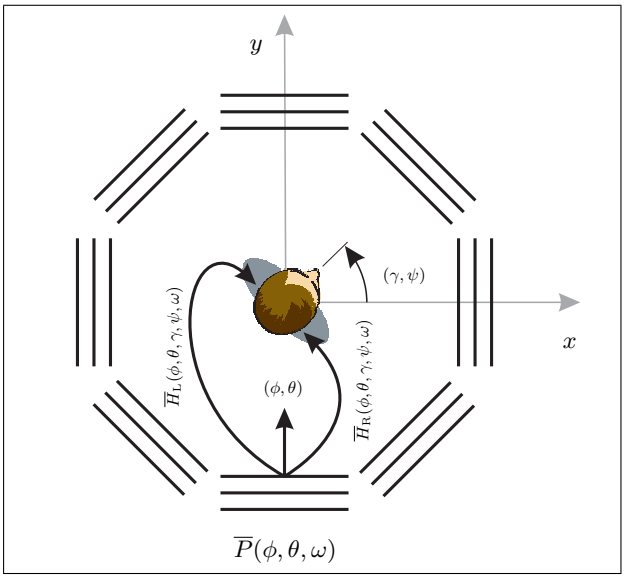

Figure 1. Data-based binaural synthesis using a plane wave expansion of the virtual sound field. For illustration, the filtering of the left/right HRTFs by the plane wave expansion coefficients $\overline{P}(\phi, \theta, \omega)$ is shown only for one particular direction. The z-axis points upwards.

2. Nomenclature

$\mathbf{x} = (x, y, z)^T = \|\mathbf{x}\| \cdot (\cos\alpha \sin\beta, \sin\alpha \sin\beta, \cos\beta)^T$ with $\|\mathbf{x}\| = r = \sqrt{x^2 + y^2 + z^2}$ describes a spatial position vector. The wavespace vector is defined as $\mathbf{k} = (k_x, k_y, k_z)^T = \|\mathbf{k}\| \cdot (\cos\phi \sin\theta, \sin\phi \sin\theta, \cos\theta)^T$ with $\|\mathbf{k}\| = \omega/c = \sqrt{k_x^2 + k_y^2 + k_z^2}$ and the speed of sound $c = 343$ m/s. The angular temporal frequency $\omega = 2\pi f$ is related to the temporal frequency f. The head looking direction is expressed by the unit length vector $\mathbf{n}_{\mathrm{H}} = (\cos\gamma \sin\psi, \sin\gamma \sin\psi, \cos\psi)^T$. The standard convention for spherical coordinates with the angles azimuth $\alpha, \phi, \gamma \in [0, 2\pi)$ and colatitude $\beta, \theta, \psi \in [0, \pi]$ is used. Within the paper, the wave vector $\mathbf{k}$ shall point into direction of where the wave is coming from. To relate the sound pressure $p(\mathbf{x}, t)$ with its temporal spectrum $P(\mathbf{x}, \omega)$ the Fourier transform convention $p(\mathbf{x}, t) \propto \int P(\mathbf{x}, \omega)\, \mathrm{e}^{+\mathrm{i}\,\omega\, t} \mathrm{d}\omega$ with the imaginary unit $\mathrm{i}^2 = -1$ is utilized. Thus, a unit amplitude plane wave is characterized by $\mathrm{e}^{+\mathrm{i}\mathbf{k}_{\mathrm{PW}}^T \mathbf{x}} \cdot \mathrm{e}^{+\mathrm{i}\,\omega\, t}$ with the radiation direction $-\mathbf{k}_{\mathrm{PW}}$. The spherical Bessel function of first kind and the spherical Hankel function of second kind with order n are denoted by $j_n(\cdot)$ and $h_n^{(2)}(\cdot)$, respectively. The spherical surface harmonics are defined as

$$Y_n^m(\alpha, \beta) = (-1)^m \sqrt{\frac{2n+1}{4\,\pi} \cdot \frac{(n - |m|)!}{(n + |m|)!}} \times \mathrm{P}_n^{|m|}(\cos\beta)\, \mathrm{e}^{+\mathrm{i}\, m\, \alpha} \quad (1)$$

[2, eq. (2.1.59)] with $n \in \mathbb{N}_0$, $-n \leq m \in \mathbb{Z} \leq +n$ and the associated Legendre polynomials $\mathrm{P}_n^{|m|}(\cdot)$.

3. Data-based Binaural Synthesis

3.1. Plane Wave Decomposition

For the proposed approach we utilize the plane-wave decomposition (PWD) on spherical apertures. According to [2, 4], any sound field $P(\mathbf{x}, \omega)$ can be represented as a superposition of plane waves, if the spherical region of interest of radius R is solenoidal. This superposition reads

$$P(\mathbf{x}, \omega) \propto \int_0^{2\pi} \int_0^{\pi} \bar{P}(\phi, \theta, \omega)\, \mathrm{e}^{+\mathrm{i}\mathbf{k}^T \mathbf{x}} \sin\theta\, \mathrm{d}\theta \mathrm{d}\phi \quad (2)$$

denoting the spectrum of a plane wave emerging from $\mathbf{k}(\phi, \theta)$. In the context of this paper, two beamforming methods to acquire the plane wave decomposition of a sound field are briefly reviewed. The modal beamformer (MB) [5] utilizes the spherical harmonics transform (SHT)

$$\mathring{P}_n^m(\omega) \propto \int_0^{2\pi} \int_0^{\pi} P(\mathbf{x}, \omega)\, Y_n^m(\alpha, \beta)^*\, \sin\beta\, \mathrm{d}\beta \mathrm{d}\alpha \quad (3)$$

of a soundfield evaluated on a spherical surface ($||\mathbf{x}|| = R = \mathrm{const}$). The coefficients of the modal plane wave decomposition are related to the SHT coefficients by

$$\bar{P}_{\mathrm{MB}}(\phi, \theta, \omega) \propto \sum_{n=0}^{N_{\mathrm{SHT}}} \sum_{m=-n}^{n} \frac{\mathring{P}_n^m(\omega)}{d_n(\omega)} Y_n^m(\phi, \theta) \quad (4)$$

with an appropriate so called radial filter $d_n(\omega)$. The filter depends on the characteristics of the spherical

Table I. c_n coefficients for modal and delay-and-sum beamformer utilizing an open sphere of radius R with a continuous distribution of pressure microphones. r_s denotes the distance of the point source from the array center.

Beamformer	Wave	c_n	N	Reference
modal	plane	1	N_{SHT}	[3]
modal	spherical	$-\mathrm{i}\frac{\omega}{c}h_n^{(2)}\left(\frac{\omega}{c}r_s\right)\frac{1}{4\pi\,\mathrm{i}^n}$	N_{SHT}	[2, 3]
delay&sum	plane	$\left\lvert 4\pi\mathrm{i}^n j_n\left(\frac{\omega}{c}R\right)\right\rvert^2$	∞	[3]
delay&sum	spherical	$-\mathrm{i}\frac{\omega}{c}h_n^{(2)}\left(\frac{\omega}{c}r_s\right)4\pi\mathrm{i}^{-n}\left\lvert j_n\left(\frac{\omega}{c}R\right)\right\rvert^2$	∞	[2, 3]

microphone array. In this paper an open sphere with pressure sensors is considered. Its radial filter is defined as $d_n(\omega) = 4\pi\mathrm{i}^n j\left(\frac{\omega}{c}R\right)$ [4]. A finite SHT-order N_{SHT} limits the modal bandwidth and is connected to a limited number of microphones. The delay-and-sum beamformer (DSB) [3] realizes a spatially full-band plane-wave decomposition

$$\bar{P}_{\mathrm{DSB}}(\phi,\theta,\omega) \propto \int_0^{2\pi}\int_0^{\pi} P(\mathbf{x},\omega)\mathrm{e}^{-\mathrm{i}\mathbf{k}^T\mathbf{x}}\sin\beta\,\mathrm{d}\beta\mathrm{d}\alpha. \quad (5)$$

It can be considered as a high frequency and/or farfield approximation of the modal beamformer for $N_{\mathrm{SHT}} \to \infty$ [1]. For continuous spherical microphone arrays, an analytical solution of the PWD coefficients for plane and spherical sound field is given by

$$\bar{P}(\phi,\theta,\omega) = \sum_{n=0}^{N} c_n \frac{2n+1}{4\pi} P_n(\cos\Theta) \quad (6)$$

with the coefficients c_n and the summation order N stated in tab. I. Θ denotes the angle between the look direction of the plane wave decomposition (ϕ,θ) and the direction of where the plane/spherical wave is coming from.

Due to the characteristics of the spherical hankel function $h_n^{(2)}(\cdot)$ numerical instabilities occur if summing up results for high orders n and/or small arguments. Bernschütz et al. [6] have presented a soft-knee amplitude limitation technique for bandlimited modal beamforming. The amplitude limited hankel function reads

$$\tilde{h}_n^{(2)}(x) = \frac{2\eta}{\pi}\frac{h_n^{(2)}(x)}{\lvert h_n^{(2)}(x)\rvert}\arctan\left(\frac{\pi}{2}\frac{\lvert h_n^{(2)}(x)\rvert}{\eta}\right). \quad (7)$$

The amplification threshold $\eta = 10^{(a/20)}$ is defined by its equivalent a in dB. For the modal beamformer with an order up to $N_{\mathrm{SHT}} = 23$ a suitable value of $a = -20$dB has been found. Since the DSB has no bandlimitation a numerically stable realization for spherical waves can not be established with this approach. It is furthermore questionable, if the delay-and-sum beamformer itself is suitable for non-plane wave scenarios, since it represents high frequency/farfield approximation. However, a detailed analysis on the effects of amplitude limitation is out the focus of this paper.

3.2. Auralization

In order to auralize the plane wave coefficients $\bar{P}(\phi,\theta,\omega)$, a virtual head is set up at $\mathbf{0} = (0,0,0)^T$. The unit amplitude plane waves $\mathrm{e}^{+\mathrm{i}\mathbf{k}^T\mathbf{x}}$ in (2) are replaced by the respective far-field HRTFs $\bar{H}_{\mathrm{L,R}}(\phi,\theta,\gamma,\psi,\omega)$ of the virtual head. For a certain head orientation (γ,ψ) the sound pressure

$$P_{\mathrm{L,R}}(\gamma,\psi,\omega) \propto \int_0^{2\pi}\int_0^{\pi} \bar{P}(\phi,\theta,\omega) \times \bar{H}_{\mathrm{L,R}}(\phi,\theta,\gamma,\psi,\omega)\sin\theta\,\mathrm{d}\theta\mathrm{d}\phi \quad (8)$$

at the left/right ear is given by the superposition of the HRTFs filtered by the plane wave expansion coefficients $\bar{P}(\phi,\theta,\omega)$. An illustration is given in fig. 1. Due to the finite resolution of HRTF datasets, (8) has to be discretized resulting in a summation over discrete azimuth-/elevation-angles. The total number of angles is denoted as N_{PW}.

3.3. Translatory Head-Movements

To apply a translatory head-movement the position of the listener for which the PWD (2) was generated has to be shifted from the origin to a desired coordinate $\mathbf{x}_{\mathrm{T}} = (x_{\mathrm{T}}, y_{\mathrm{T}}, z_{\mathrm{T}})^T$. The method to auralize the $\bar{P}(\phi,\theta,\omega)$ for $\mathbf{x}_{\mathrm{T}}$ is briefly outlined in this section. For a detailed description of this technique and its properties the reader is referred to [1].

The PWD coefficients $\bar{P}(\phi,\theta,\omega)$ are related to the temporal impulse response of the beamformer via the temporal inverse Fourier transform

$$\bar{P}(\phi,\theta,\omega) \;\bullet\!\!-\!\!\circ\; \bar{p}(\phi,\theta,t). \quad (9)$$

Spatial shifting to $\mathbf{x}_{\mathrm{T}}$ results in a temporal shift of the impulse response $\bar{p}(\phi,\theta,t)$ taking the shifting theorem of the Fourier transform

$$\bar{P}(\phi,\theta,\omega)\,\mathrm{e}^{+\mathrm{i}\mathbf{k}^T\mathbf{x}_{\mathrm{T}}} \;\bullet\!\!-\!\!\circ\; \bar{p}\left(\phi,\theta,t+\frac{\mathbf{k}^T\mathbf{x}_{\mathrm{T}}}{\omega}\right) \quad (10)$$

into account. Thus, translatory movements can be realized by manipulating the phase of the PWD coefficients. With the reference point $\mathbf{0} = (0,0,0)^T$ of the

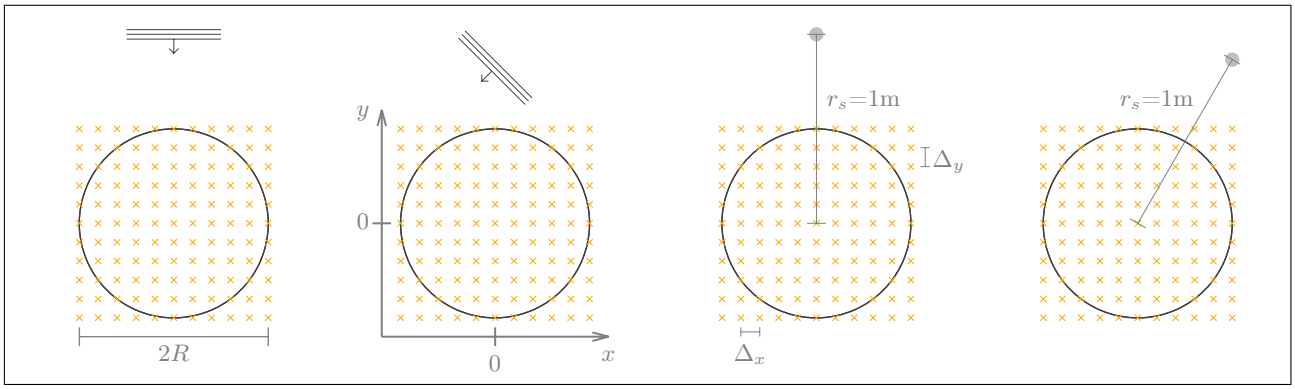

Figure 2. Localisation experiments are performed for a regular xy-grid of translatory shifts $\mathbf{x}_\mathrm{T}$ (orange crosses) with a resolution of $\Delta_x = \Delta_y = 0.1$m. The virtual head look direction remains $\mathbf{n}_H = (0, 1, 0)^T$. The four evaluation examples include a plane wave with $\phi_\mathrm{PW} = 90°, 45°$ and a spherical wave steming from $\mathbf{x}_s = (0, 1, 0)^T, (0.5, 0.8660, 0)^T$. The solid black circle illustrates the surface of the spherical microphone array with a radius of $R = 0.5$m. r_s denotes the distance between array center and the point source.

virtual head the sound pressure at the left/right ear reads

$$P_{\mathrm{L,R,T}}(\gamma, \psi, \omega, \mathbf{x}_\mathrm{T}) \propto \int_0^{2\pi} \int_0^{\pi} \bar{P}\,(\phi, \theta, \omega)\, \mathrm{e}^{+\mathrm{i}\mathbf{k}^T \mathbf{x}_\mathrm{T}} \times \bar{H}_{\mathrm{L,R}}(\phi, \theta, \gamma, \psi, \omega) \sin\theta \,\mathrm{d}\theta \mathrm{d}\phi \tag{11}$$

for translation of the head to $\mathbf{x}_\mathrm{T}$. Considering time-discrete data with sample index $k \in \mathbb{Z}$ the sound pressure impulse response is derived by the inverse discrete time fourier transformation (DTFT)

$$p_{\mathrm{L,R,T}}(\gamma, \psi, k, \mathbf{x}_\mathrm{T}) \propto \int_{-\pi}^{\pi} P_{\mathrm{L,R,T}}(\gamma, \psi, \omega, \mathbf{x}_\mathrm{T})\, \mathrm{e}^{+\mathrm{i}\Omega k}\, \mathrm{d}\Omega \tag{12}$$

with the normalized angular frequency $\Omega = 2\pi \frac{f}{f_s}$.

4. Evaluation

The proposed approach is evaluated by numerical simulations. The four setups being part of the investigations are illustrated in fig. 2. The evaluation is based on the plane wave expansion $\bar{P}\,(\phi, \theta, \omega)$ computed for the reference position $\mathbf{x} = (0, 0, 0)^T$. The coefficients are derived using the delay-sum-beamformer (5) and the modal beamformer (4) with different modal resolutions $N_\mathrm{SHT} = 3, 5, 10, 23$. Analytic solutions solution are given by (6) and tab. I The coefficients are computed for $\theta = 90°$ and $\phi = -180° + n\,\Delta_\phi$ with $n = 0, ..., N_\mathrm{PWD} - 1$. The resolution $\Delta_\phi = 360°/N_\mathrm{PWD}$ of the azimuth angle is determined by the number of plane waves N_PWD. The head orientation is fixed to $\psi = 90°$ and $\gamma = 90°$. In addition, a grid of translatory shifts $\mathbf{x}_\mathrm{T}$ around the center of the microphone array is applied for each of the expansions. For each shift the resulting sound pressure impulse responses $p_{\mathrm{L,R,T}}(\gamma, \psi, k, \mathbf{x}_\mathrm{T})$ at the left/right ear are computed according to (11) and (12). KEMAR HRTFs [7] measured at 3m distance were utilized for this. The length of the impulse responses is limited to 2^{12} samples. All experiments are executed for a temporal sampling frequency of $f_s = 44.1$ kHz.

In order to evaluate the localisation properties the binaural model of Dietz et al. [8] extended by Wierstorf et al. [9] is used. The ear signals needed as an input for the binaural model are generated via a filtering of an anechoic test signal with the sound pressure impulse responses $p_{\mathrm{L,R,T}}(\gamma, \psi, k, \mathbf{x}_\mathrm{T})$. A bandpass filtering, gamma-tone filtering, half-wave rectification and a lowpass filtering is applied to both channels separately during monaural preprocessing. Afterwards, the interaural time difference is computed for each frequency band and converted to localization azimuths via a lookup-table. This table is generated a-priori with the same HRTF dataset used for the experiments. The median azimuth of all (equally weighted) frequency bands is taken as the localization result. An outlier-rejection is applied during this process in order to disregard directions which differ 30° from the median. Both, the original and the extended model, are freely available as part of the *Auditory Modelling Toolbox* [10].

5. Results & Discussion

At first, the effects of modal bandlimitation on the localization performance in combination with the translatory shift are illustrated in fig. 3. As already stated in [11], the localization is more accurate for sources which are placed in the look direction of the virtual head. A limited resolution of the modal beamformer leads to a loss of directivity. Localization errors increase when shifting the head position orthogonal to the incidence direction of the wave field. This effect

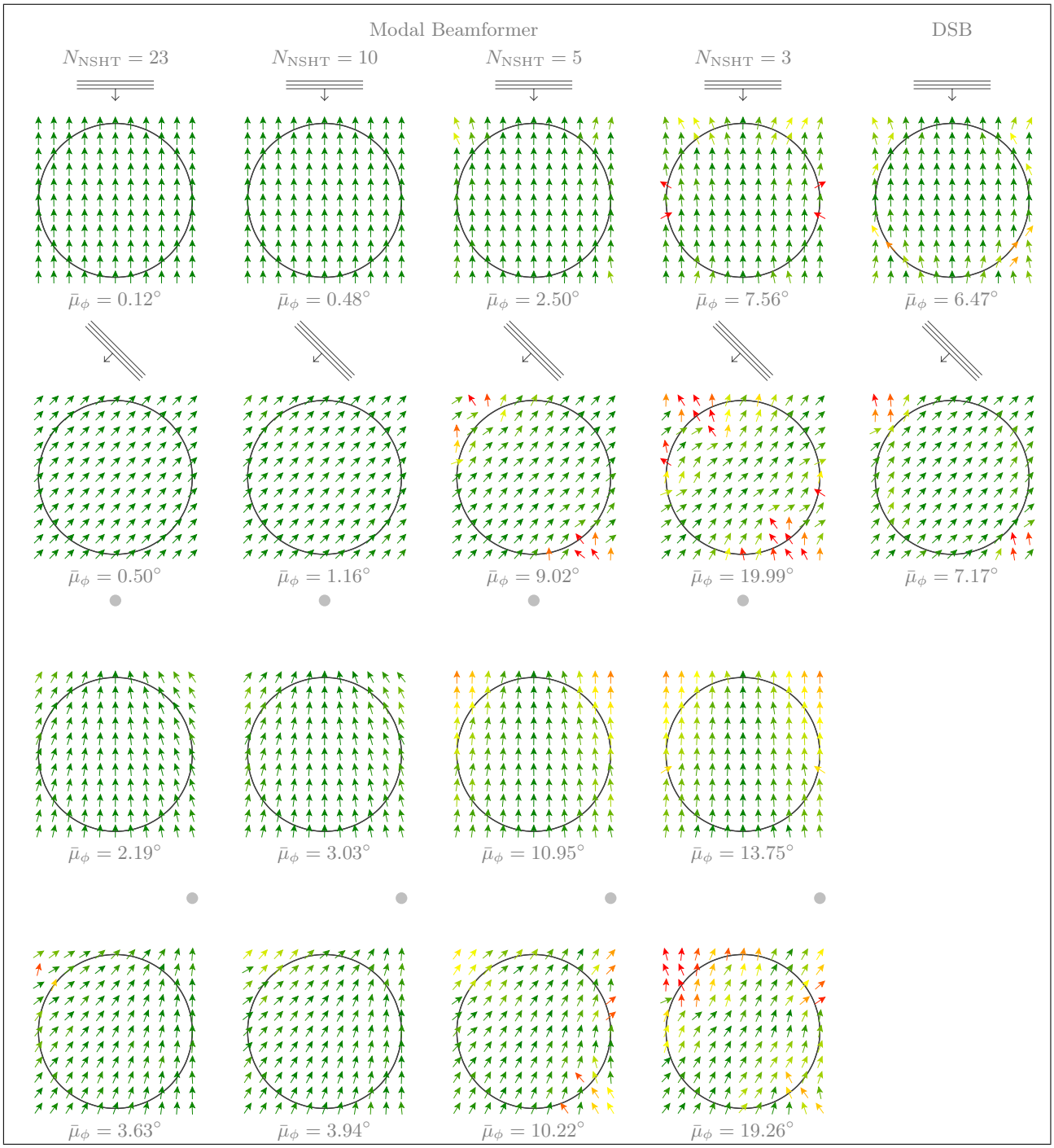

Figure 3. Each row show the localization results for one of four test scenarios introduced in fig. 2. The first four columns represent the modal beamformer with different modal resolutions. The results for the delay&sum beamformer are presented in the last column. The direction of the arrows indicate the localization result of the binaural model while their colors encode the absolute localization error. The mean absolute localization error $\bar{\mu}_\phi$ is given under each plot. $N_{\mathrm{PW}} = 360$.

becomes more significant with a decreasing modal resolution. Secondly the interaction between the resolution of the plane wave decomposition N_{PW} and the modal resolution N_{SHT} is analysed. Due to the modal bandlimitation the degrees of freedom of the plane wave expansion is fixed to $N_{\mathrm{DOF}} = 2N_{\mathrm{SHT}} + 1$ when considering only the horizontal plane ($\theta = 90°$) for the PWD [11]. Although the PWD coefficients can be computed for arbitrary directions, utilizing more then N_{DOF} plane waves does not lead to a significant benefit in terms of localization (see fig. 4). It is also shown that the localization performance decreases noticeable for $N_{\mathrm{PW}} < N_{\mathrm{DOF}}$.

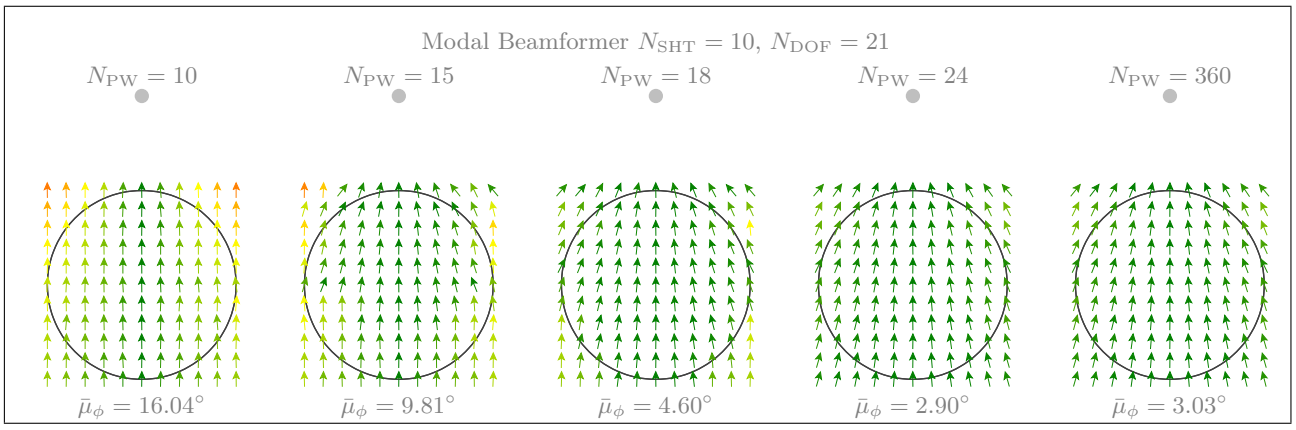

Figure 4. Localization results are illustrated for the third test scenarios introduced in fig. 2. The modal resolution is fixed, while the number of plane waves is varied.

6. Summary and Conclusion

This paper evaluates the localization properties of data-based binaural synthesis using a plane wave decomposition of an incident soundfield. Utilizing spherical microphone arrays, this method can be applied to separate the soundfield acquisition from the HRTF capture. Therefore, individualized binaural synthesis is done more easily by exchanging listener dependent HRTF datasets. Translatory movement of the listeners head is included due to the spatial shifting property of plane waves. A limited number of microphones and a finite resolution of HRTF datasets imply pratical limitations to this technique. The effects of these limitations to the perceptual localization of sound sources were evaluated by a binaural model. The modal resolution, which is closely connected to the number of microphones, significantly influences the valid area of translatory head movement, where an accurate localization is possible. The resolution of the HRTF dataset necessary for localizing sources accurately depends on the degrees of freedom of the recorded soundfield, which are directly connected to the modal resolution. The localization accuracy does not benefit from a supersampling in terms of using more plane wave coefficients. Although the results of the binaural model have been validated by hearing experiments in multi-channel audio reproduction, additional validation for binaural synthesis is desired. Beyond that, other perceptional aspects like coloration or distance perception remain for future research.

Acknowledgement

This research has been supported by EU FET grant Two!Ears, ICT-618075.

References

[1] Schultz, F.; Spors, S. (2013): "Data-based binaural synthesis including rotational and translatory head-movements." In: *Proc. of 52nd Intl. Aud. Eng. Soc. Conf. on Sound Field Control - Engineering and Perception, Guildford, UK.*

[2] Gumerov, N.A.; Duraiswami, R. (2004): *Fast multipole methods for the Helmholtz equation in three dimensions.* Oxford: Elsevier Science.

[3] Rafaely, B. (2005): "Phase-mode versus delay-and-sum spherical microphone array processing." In: *IEEE Signal Process. Letters*, **12**(10):713–716.

[4] Williams, E.G. (1999): *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography.* London: Academic Press.

[5] Park, M.; Rafaely, B. (2005): "Sound-field analysis by plane-wave decomposition using spherical microphone array." In: *The Journal of the Acoustical Society of America*, **118**(5):3094–3103.

[6] Bernschütz, B.; Pörschmann, C.; Spors, S.; Weinzierl, S. (2011): "Softlimiting der modalen Amplitudenverstärkung bei sphärischen Mikrofonarrays im plane wave decomposition Verfahren." In: *Proc. of the 37th Deutsche Jahrestagung für Akustik (DAGA), Düsseldorf, Germany*, 661–662.

[7] Wierstorf, H.; Geier, M.; Spors, S. (2011): "A free database of head related impulse response measurements in the horizontal plane with multiple distances." In: *Proc. of 130th Aud. Eng. Soc. Conv., London, UK.*

[8] Dietz, M.; Ewert, S.D.; Hohmann, V. (2011): "Auditory model based direction estimation of concurrent speakers from binaural signals." In: *Speech Communication*, **53**(5):592 – 605.

[9] Wierstorf, H.; Raake, A.; Spors, S. (2013): "Binaural assessment of multichannel reproduction." In: J. Blauert, ed., *The Technology of Binaural Listening*, Modern Acoustics and Signal Processing, 255–278, Springer Berlin Heidelberg.

[10] Søndergaard, P.; Majdak, P. (2013): "The auditory modeling toolbox." In: J. Blauert, ed., *The Technology of Binaural Listening*, Modern Acoustics and Signal Processing, 33–56, Springer Berlin Heidelberg.

[11] Spors, S.; Wierstorf, H.; Geier, M. (2012): "Comparison of modal versus delay-and-sum beamforming in the context of data-based binaural synthesis." In: *Proc. of 132nd Aud. Eng. Soc. Conv., Budapest, Hungary.*