

COVER PAGE

Document downloaded by @DAEL

Thu Nov 26 16:54:41 2026
For personal use

When automatic English translation is provided, only the original document is authentic.
The EAA cannot be held responsible of any translation error

Bibliographical reference

Elevation in Wave-Field Synthesis Using HRTF Cues, Jose J. Lopez, Maximo Cobos and Basilio Pueo, Acta Acustica **vol. 96** (Number 2), 2010, pp. 340-350

DOI
https://doi.org/10.3813/AAA.918283

ACTA ACUSTICA UNITED WITH ACUSTICA
Vol. 96 (2010) 340 – 350
DOI 10.3813/AAA.918283

Elevation in Wave-Field Synthesis Using HRTF Cues

Jose J. Lopez[1)], Maximo Cobos[1)], Basilio Pueo[2)]
[1)] Institute of Telecommunications and Multimedia Applications (iTEAM), Universidad Politécnica de Valencia, Spain. jjlopez@dcom.upv.es
[2)] Signals, Systems and Telecommunications Group, Polytechnic School, University of Alicante, Spain.

Summary
In this paper, spectral elevation cues of Head-Related Transfer Functions (HRTF) are used in conjunction with Wave-Field Synthesis (WFS) to produce the sensation of elevated virtual sources in a realistic way. This approach allows WFS to simulate elevated virtual sources above and below the horizontal plane, counteracting the clear disadvantage of WFS compared to other spatial sound systems that provide elevation with additional loudspeaker arrays, e.g. Ambisonics or 10.2 Surround. Therefore, different elevation filterbanks have been computed from several HRTF databases to obtain pure elevation cues. According to the results of our experiments, the proposed approach is a powerful technique for including elevation effects in common WFS systems without the need for using elevated loudspeakers or changing the reproduction set-up.

PACS no. 43.60.-c, 43.66.Pn

1. Introduction

Human hearing plays a major role in the way our environment is perceived. One of the most important cues in spatial perception of sound is localization. Generally, sound is perceived in three dimensions, width, height and depth all of which are necessary to achieve a natural perception of sound. [1]. In fact, natural sound scenes are made from different sounds, which may be perceived individually or as different entities where complex grouping mechanisms are involved.

For that reason, the accurate recreation of sound scenes aiming at stimulating source localization in a natural way is one of the main objectives of audio reproduction systems. Among them, WFS is a spatial reproduction technique that, by analogy to holography and in base to the Huygens principle, can synthesize an acoustic field inside a volume from the pressures obtained in the surrounding surface. However, in practice, most WFS scenes are synthesized by reproducing isolated tracks corresponding to different audio events as point sources or plane waves radiated from a set of virtual sources placed at user-defined locations.

WFS was first proposed with application to 3D sound by Berkhout [2]. The synthetic wave front is created by loudspeaker arrays that substitute the individual secondary sources. The main advantage of these systems is that the acoustic scene has not sweet spot. When listeners move inside the listening area, the perspective of the auditory scene changes in a convincing way according to its relative position to the virtual source. However, the recreation of a true natural wave field can be fulfilled with certain restrictions. Huygens principle needs to be discretized in practice, which means that an infinite plane of continuous secondary source distribution is replaced by a number of finite line arrays of equidistant discrete loudspeakers. These practical constraints yield to amplitude and frequency artifacts, and restrict localization to the horizontal plane only.

Several methods for dealing with the problems mentioned above can be found in the literature [3]. However, the restriction of reproduction in the horizontal plane only is the most important and difficult problem to solve in terms of source localization. This is a clear disadvantage of WFS in comparison to other spatial sound systems that provide elevation, as Ambisonics or the 10.2 cinema surround sound system. For these reasons, different solutions have already been proposed to overcome this problem. For example, putting a linear array on the ceiling or using two parallel linear arrays located at different elevation angles. Unfortunately, the phantom effect does not work in elevation as good as in azimuth so these systems do not always provide the desired quality [4]. It was also reported that subjects did not perceive a well-defined phantom image between the two loudspeakers but either heard the sound coming from one of the two loudspeakers or perceived an unclear image. More sophisticated and accurate localization cues are those provided by Head-Related Transfer Functions (HRTF) [5], which describe how a given sound is filtered by diffraction and reflection properties of the head, pinna, and torso, before the sound reaches the

Received 7 May 2009,
accepted 12 November 2009.

© S. Hirzel Verlag · EAA

eardrum and inner ear. Pre-filtering effects are very dependent on the direction of arrival of the incoming sound. These effects play a very important role in source localization, particularly in the determination of source elevation. [6]. The idea of using HRTF cues for providing an enhanced reproduction of sound including elevated sources is nothing new. In the mid 70s, Jens Blauert patented a binaural synthesis method that made use of HRTFs [7]. Moreover, he continued working on the idea, studying the influence that some specific frequency ranges had on the perception of elevation [8].

In this paper, the use of HRTF spectral elevation cues in conjunction with WFS is proposed with the aim of producing the sensation of elevated virtual sources. Azimuth localization is achieved with the usual WFS system, but elevation is simulated by means of a filtering stage prior to WFS rendering. Different elevation filterbanks have been computed from several HRTF databases. Also, a normalization step for removing cues which are not directly related to elevation, has been performed in order to obtain pure elevation cues.

The structure of this paper is a follows. In section 2, a brief introduction to state of the art in 3D sound is presented. In section 3 the proposed approach is introduced, describing how to achieve elevated sound sources in WFS with HRTF elevation cues. Experiments and results are discussed in section 4 and finally the conclusions of this work are summarized in section 5.

2. Elevation in 3D sound systems

Elevation of sound sources in spatial audio systems has not been discussed in depth in engineering and research. In this section, a brief review on the procedures by which current multichannel spatial sound systems manage virtual elevated sources is provided, in order to improve the findings of our investigation.

2.1. 3D audio system overview

The goal of audio reproduction systems has always been to accurately recreate a given sound scene in terms of source localization. Over the last few decades, surround sound has played an increasingly important role in the entertainment industry, as well as in the field of multimedia, reproducing sound in a way which is more "natural" with the aim of enhancing the listening experience. While high-fidelity audio devices currently demonstrate flat frequency responses and very low levels of noise and distortion, reproduction issues involving spatial quality or spatial fidelity are still an active research field.

2.1.1. Ambisonics

Ambisonics is a technique developed in the early 70s [9], that provides a way to encode three dimensional soundfields, usually by recording. These encoded soundfields can then be reproduced over various different speaker arrangements, which is known as Ambisonic decoding. An advantage of Ambisonics reproduction is that it is based on solid mathematics. The accuracy in the reproduction of the soundfield is given by the Ambisonic order, related to the order of a spherical harmonic decomposition of the soundfield. Whereas zeroth order corresponds to Mono reproduction, the most common first order form is known as B-Format. This format uses four channel encoding, corresponding to the instantaneous sound pressure (w), and the three components of its gradient (x, y, and z) which are related to the particle velocity at a point in space. The loudspeaker signals are derived by using a linear combination of these four channels, where each signal is dependent on the actual position of the speaker in relation to the center of an imaginary sphere, the surface of which passes through all available speakers. In more advanced decoding schemes, spatial equalization is applied to signals to account for the differences in the high and low-frequency sound localization mechanisms in human hearing. Current research in Ambisonics reproduction is related to higher-order Ambisonics (HOA), where more channels than in the first-order B-Format are used [10]. Although most HOA studies have been focussed on the directional encoding of plane waves, near field effects should not be ignored [11].

2.1.2. Vector base amplitude panning

Vector Based Amplitude Panning (VBAP) is a method for positioning virtual sources to arbitrary directions using a setup of multiple loudspeakers [12]. VBAP is based on amplitude panning, so the same sound signal is applied to a number of loudspeakers with appropriate non-zero amplitudes. The number of loudspeakers used in VBAP is arbitrary, however, an increasing number of loudspeakers yields to a better localization of virtual sources [13]. With 2-D setups, VBAP is a reformulation of the existing pairwise panning method. However, it can be generalized for 3-D loudspeaker setups as a triplet-wise panning method. A sound signal is then applied to one, two, or three loudspeakers simultaneously.

2.1.3. Surround systems

Multichannel surround sound systems have the capability of improving the spatial attributes of the reproduced sound scenes, facilitating the perception of sound sources coming from different directions.

They have always been closely related to the cinema industry and, in the last decades, several formats which have intended to emphasize the audiovisual experience have appeared. Despite the number of surround formats available, elevation has never received special attention. The 10.2 surround system is a recent format that proposes to place two height loudspeakers, left and right, above the audience [14]. These two loudspeakers are then used to create the reflection of a virtual source behind the screen, whilst the rest of the loudspeakers recreate other major reverberations from the sides and back of the room. The system is made up of 14 discrete channels, including two LFE channels.

The 22.2 surround system proposed by Hamasaki [15] uses 24 speakers (including two sub-woofers) arranged in

three layers. This system constitutes the sound proposal for the new experimental video format called *Super Hi-Vision*. The loudspeakers are distributed in the three layers as follows: nine loudspeakers above the ear level, ten loudspeakers at ear level, and five loudspeakers below the ear level. Although a very good spatial impression can be obtained with such a number of channels, there is also a big difficulty regarding post-production since the system still relies on a channel-based conception of sound scenes.

2.1.4. Wave field synthesis

Although, today, five channel systems are a consolidated standard in multichannel audio, there is an increasing interest in emerging reproduction systems based on sound field rendering. The most popular of these systems is Wave-Field Synthesis (WFS), a spatial reproduction system capable of synthesizing an acoustic field in an extended area by means of loudspeaker arrays. This property makes it possible to synthesize a sound scene with the correct spatial characteristics for each listener [16],[17],[18] regardless of their position in the listening area. Therefore, every listener perceives his/her own perspective of the reproduced scene and the experience is closer to that of natural hearing.

WFS is based on the wave theory concept of Huygens: all points on a wave front serve as individual point sources of spherical secondary wave fronts. A large number of small and closely spaced loudspeakers are used to apply this principle to practical acoustic systems. Each loudspeaker in the array is fed with a corresponding driving signal calculated by means of algorithms based on the Kirchhoff-Helmholtz integrals and Rayleigh representation theorems [2]. The theory of WFS was recently revisisted in [19]. The desired wave-field is achieved by the superposition of the sound fields generated by each loudspeaker. This technique enables an accurate representation of the original wave-field with its natural temporal and spatial properties in the entire listening space. WFS is also capable of synthesizing virtual sources both in front of and behind the array, and with a certain directivity characteristic [20]. These properties make WFS a powerful spatial sound reproduction system. However, creating a copy of a sound field is not completely possible due to some practical constraints:

- *Discretization* of an ideal continuous secondary source distribution to a loudspeaker array, which leads to spatial aliasing and results in both spatial and spectral errors in the synthesized sound field at high frequencies.
- *Finiteness* of the array, which leads to truncation effects and results in diffraction waves that cause after-echoes and pre-echoes.
- *Restriction* to a line loudspeaker array in the horizontal plane instead of a planar array, which leads to amplitude errors and restricts the localization to the horizontal plane.

2.2. Localization mechanisms and HRTFs

Binaural localization is based on the comparison of the auditory inputs from both our two ears. Azimuth localization is performed via the “Interaural Time Difference” (ITD) and the “Interaural Level Difference" (ILD) cues [8]. The primary biological binaural cue is ITD and it is related to the time delay that appears when sound reaches the near ear and the far ear. The ILD cue, less significant, is given by the reduction in loudness when the sound reaches the far ear. Due to the in-plane position of the human ears, these cues will only aid in localizing the sound source in the azimuth plane, and not in elevation. To get instantaneous localization in more than two dimensions from ITD and ILD, two or more detectors are required. However, there are complex variations in the degree of attenuation of a sound when travelling from the source to the eardrum. These frequency-dependent attenuations are related to both azimuthal angle and elevation and can be summarized in the HRTF. The HRTF is a transfer function that, for a certain angle of incidence, describes the sound transmission from a free field to a point in the ear canal of a human subject [21]. As a result, an estimation of the source location in azimuth and elevation can be performed with wideband sounds. Additional information can also be found by moving the head, so that the HRTF for both ears changes in a way known by the subject. In [22], a spatial audio system was proposed based on the assumption that the spectral cues are common in any sagittal planes, as shown in Figure 1. In that system, sound localization in any direction was achieved by combining HRTF filtering in the median plane and interaural differences. However, subjects’ own HRTFs were used in the listening tests and therefore, not a single HRTF model was considered for their experiments.

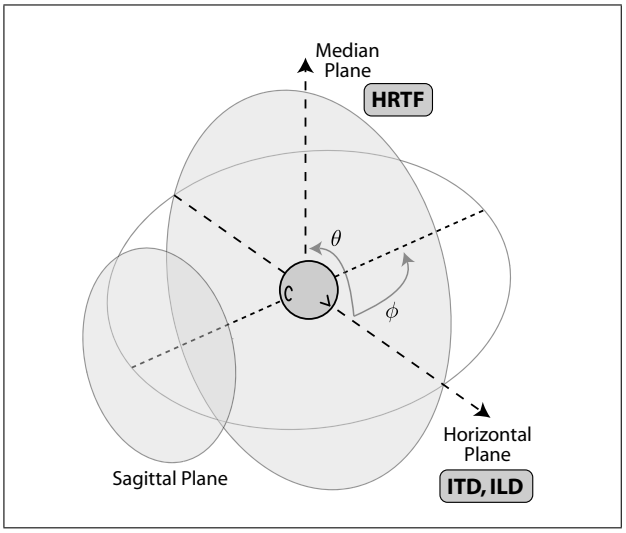


Figure 1. System geometry.

Several attempts have been made to model HRTFs, both to understand their behavior and to simplify the binaural synthesis process. The task of modeling HRTFs has always encountered four major problems [23]:

1. Difficulty of approximating the effects of wave propagation and diffraction by simple, low-order parameterized filters.

2. Complicated joint dependency of the HRTF's on azimuth, elevation and range.
3. Lack of a quantitative criterion for measuring the accuracy of an approximation
4. Increased variability from person to person in HRTF.

As shown in [24], the pinnae play a critical role in determining elevation, so that two major ridges in the outer ear act like reflecting surfaces, producing multipath echoes whose timing gave the cues for elevation. Various models have been proposed in the literature [25][26]. However, whereas azimuth effects are readily modeled, accurate modeling of elevation cues is still a challenge.

2.3. Binaural reproduction systems issues

It is known that 3D audio systems based on binaural reproduction by using HRTF show different problems that still need to be solved. Many listeners experience "front-to-rear confusion", that is when untrained listeners hear auditory events behind them when they were intended to be exactly in front of them. Although sophisticated ear models for dummy heads have been constructed to be optimal in the acoustic sense, there remain unavoidable errors due to the uniqueness of each listener's head. Other models considering ITD and ILD cues produce the so called "lateralization" and "inside the head effect", especially when using headphones [27]. In addition, the incompatibility in the reproduction of dummy head signals over loudspeakers is another classical problem: loudspeaker reproduction of binaural signals introduces crosstalk, where the left channel signal intended for the left ear will also be heard by the right ear and vice versa. This non desired effect may be eliminated by prefiltering the binaural signal with an inverse system called a crosstalk cancelling filter [28]. The resulting playback signals will partly cancel each other, leaving only the intended binaural signals at the ears of the listener. Crosstalk cancelled loudspeaker reproduction of binaural signals is typically designed for anechoic conditions. However, this processing works correctly only if the room in which sound is played back has highly sound absorbing walls and the listener is in a predetermined position.

Surprisingly, most of the described problems of binaural systems are related to azimuth localization of the sources. Although a lot of effort in binaural hearing has been focussed on azimuth localization, the elevation problem has not yet been extensively studied. Although the research community has tried to model individualized HRTFs, the exact nature of these cues is not yet well understood. Refer to [29] and [30] for some proposed HRTF models.

3. Proposed hybrid approach

WFS has been shown to provide excellent localization accuracy in the horizontal plane. WFS is capable of creating accurate wave fronts and, as a consequence, accurate virtual source directions below the aliasing frequency limit. Although incorrect directional cues can occur from contributions above the aliasing frequency, they may produce only a negligible decrease of the locatedness of the virtual source [3]. Nevertheless, a more important problem to solve is the restriction to azimuth only localization.

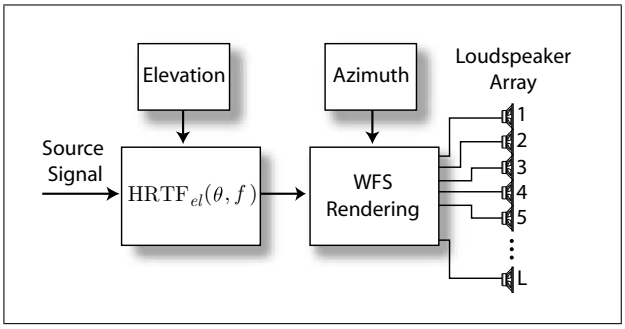


Figure 2. Proposed HRTF-WFS Hybrid System.

In this section, the proposed approach, where azimuth localization is performed by using the basic WFS algorithm and elevation is simulated by source filtering, is described in detail.

3.1. System overview

The hybrid approach to incorporate elevation in WFS is based on two stages. The first stage manages the elevation of the source, the angle θ shown in Figure 1. Thereafter, the proposed system filters the virtual source signal according to this angle, as depicted in the first block of Figure 2. Then, this filtered signal is channeled to a standard WFS stage, where the distance and azimuth angle ϕ of Figure 1 is processed accordingly to the WFS algorithm.

The following requirements were considered in the design of the system:

- Elevation should be perceived in any location inside the listening area in order to be consistent with the main WFS advantage.
- Computational cost must be acceptable.
- Elevation effects should be perceived by all individuals.
- WFS should remain unchanged if no elevation is considered in the positioning of the virtual sources.

In order to fulfill the above requirements, a number of aspects were also considered and will be discussed in the following. Firstly, if the sources have an azimuth angle different from $\phi = 0$, then some elevation will still be perceived. This is due to the fact that elevation spectral cues are common in any sagittal plane [22]. Secondly, as these spectral cues are the same in any sagittal plane, only a reduced set of filters are required. In addition, resolution in the median plane does not need to be very accurate, so the computational cost is not a limiting factor. Thirdly, several HRTF databases were analyzed in order to look for common spectral features in the median plane. Although the aim is not to provide an accurate universal HRTF that works perfectly for any individual, some common features may cause a perceivable elevation effect which can be sufficient for many applications [31]. Finally, filtering does not affect the sources when they are located in the horizontal plane, and so no modifications to the original WFS are introduced in this case.

3.2. Selection of HRTF elevation cues

As discussed in section 2.2 and verified in numerous studies, there is a large variability in the HRTF from one person to another. Besides, this variation is particularly marked in elevation cues, which are highly dependent on the ear morphology [24] and also the distance and shape of shoulders [29].

For that reason, when implementing the system, it is required to find a set of elevation cues that are as universal as possible. There were some attempts to achieve this goal with encouraging results [32], but the results were not as satisfactory as with the use of individualized HRTF.

Since one of the most important advantages of WFS is the possibility of providing an immersive sound scene for several listeners located in the listening area simultaneously, a universal HRTF is needed in our proposed system. This is necessary in order not to lose the basic WFS multi-listener feature.

Therefore, a method has been developed, that allows us to extract these cues from HRTF databases from multiple subjects. There are different research groups that have been devoted to collect HRTF measurements from a large number of subjects and have kindly been made available to the scientific community. For this study, two of these public HRTF database have been selected, one provided by IRCAM [33] and another by CIPIC [34]. In section 3.2.2, the procedure to obtain universal HRTF in elevation from these measures will be discussed.

On the other hand, there are software plug-ins and sound equipment that comprise a supposedly universal HRTF. Among them, the *Roland Sound Space Processor RSS-10* is a singular device that provides a specialization based on HRTF which is quite convincing [35]. In order to achieve a proper comparison, this reference processor has been employed to obtain the impulse responses at different positions of elevation in the median plane.

Finally, in section 3.2.4 a number of synthetic HRTF are presented. These are based on the use of 5 second-order IIR cascade filters, named second order sections (SOS), that approximate the response of the RSS-10, albeit with a very low computational cost and in a fully parametric mode.

3.2.1. Isolation of HRTF elevation cues

Assuming far field conditions, the HRTF is a function of the arrival direction of the source (ϕ, θ) and frequency f, expressed as HRTF(ϕ, θ, f). Moreover, there is also a different HRTF for the right and left ears, but they can be considered to be symmetric in ϕ. This is not the case of angle θ, where the asymmetry of the human body in this direction is clear. Since only HRTF elevation cues are of interest for our system, just the values in the median plane ($\phi = 0$) will be sufficient. It is important to note that, even in the case of a source located in $\phi = \theta = 0$, the HRTF will not be a neutral function in frequency, but it will have a filtering effect for that direction. Although this effect would be important for headphone binaural reproduction systems, it will not be of interest for systems with frontal loudspeaker reproduction, as this effect will be naturally produced in the sound path from the loudspeakers to the listener. Therefore, it is necessary to eliminate the filtering effect produced by a head exposed to a front coming sound and retain only the filtering effects due to elevation cues. For this purpose, the HRTF must be normalized as follows:

$$\mathrm{HRTF}_{el}(\theta, f) = \frac{\mathrm{HRTF}(0, \theta, f)}{\mathrm{HRTF}(0, 0, f)}, \tag{1}$$

where $\mathrm{HRTF}_{el}(\theta, f)$ is a neutral function in frequency for sources with $\theta = 0$, i.e. $\mathrm{HRTF}_{el}(0, f) = 1$. Taking into account the diagram shown in Figure 2, this function results totally coherent with the hybrid system involving HRTF and WFS, as it will be the same as the conventional WFS system when the source is located in the horizontal plane.

3.2.2. Universal HRTF cues for elevation

In the last subsection, the procedure to extract pure HRTF cues to be used in a hybrid system has been discussed. However, one important problem here is to find which is the best HRTF to choose. The ear measured from a single subject is not sufficient for obtaining a robust and universal system due to the variability existent among subjects [36]. In addition, the aim is not to look for a system that chooses the HRTF that best fits one listener, as this approach would deteriorate the WFS capability of creating an extended immersive listening area for a large audience. Therefore, it is necessary to obtain a set of universal elevation cues. This is a different problem to that of finding a complete universal HRTF, as only some common features in the median plane responses are looked for. Whereas several attempts to find a universal HRTF can be found in the literature [37], there are few works which describe common elevation cues over a large number of subjects.

In this paper, the procedure used to obtain a universal elevation HRTF is based on averaging the databases of different audiences, from both IRCAM and CIPIC. For the IRCAM database, 49 listeners were used, after two discordant listeners were removed. For the CIPIC database all 45 available listeners were used. Each of the databases contained the HRIR impulse responses obtained from the HRTF for all space positions, sampled at 44.1 kHz and with a length of 128 samples. For this work, only the HRTF of the median plane, null azimuth, are to be considered. The elevation step used for the measures were not constant for all databases: for the IRCAM database, a 15 degrees step was employed, whereas for the CIPIC, this value decreased to 11.25 degrees. However, as it will be discussed, this is not an issue when later comparing the final results graphically.

For each of the listener's HRTF and for each step in the median plane, the final HRTF was obtained upon calculating the FFT of the provided HRIR. As explained in section 3.2.1, the effect of head filtering for front sources has been eliminated for each listener by dividing each frequency bin by the frontal HRTF, according to (1). With

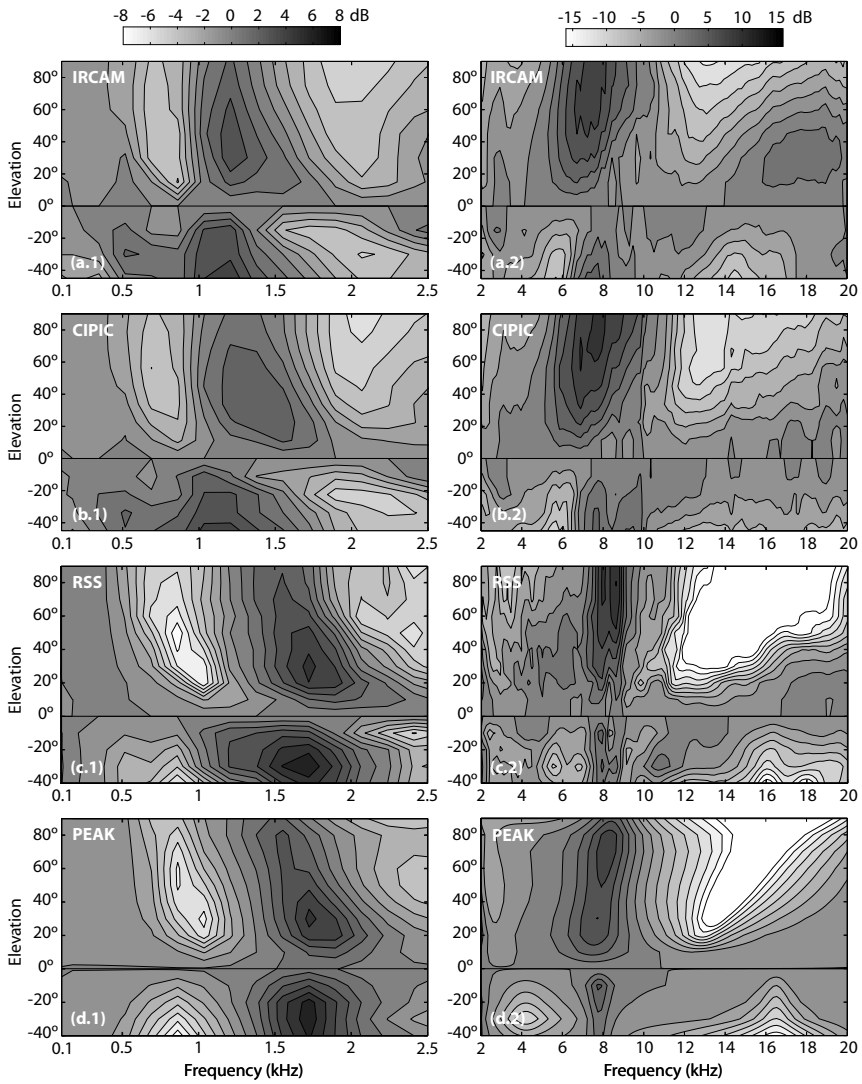


Figure 3. HRTF magnitude in the median plane for two frequency ranges (Low-frequency in column 1. High-frequency in column 2). (a) Computed average values for the IRCAM HRTF database. (b) Computed average values for the CIPIC HRTF database. (c) Measured values for the RSS HRTF. (d) Synthesized HRTF using peak filters.

the resulting values for each listener, an averaged spectral absolute value was made, as expressed in (2):

$$\mathrm{HRTF}_{med}(\theta, f) = \frac{1}{N} \sum_{i=1}^{N} \left| \mathrm{HRTF}_i^{el}(\theta, f) \right| . \qquad (2)$$

Once the HRTF_{med} is computed for each elevation, their minimum-phase HRIR are obtained using the method described in [38]. Since HRIR are often minimum- or quasi-minimum-phase functions, a more accurate impulse response is expected to be obtained, in comparison with a linear phase response. In the resulting averaged HRIR, a length of 128 samples at 44.1 kHz was considered, which is the same as they were originally.

3.2.3. Comments on the obtained HRTF

On the one hand, the computed average values for IRCAM and CIPIC HRTF databases are shown in Figures 3a and b. These were calculated as the FFT of the averaged provided HRIR, as described in the preceding section. For clarity, they were represented in two frequency ranges. On the left side, the low frequency range between 100 Hz and 2.5 kHz is plotted, where the effects of shoulder reflections are dominant, as described in [29]. In this case, the HRTF peaks and dips are restricted in the range of ±8 dB. On the right side, the high frequency range between 2 and 20 kHz is represented, where the elevation effects are mainly due to the pinnae. In this case, peaks and dips are of greater level, about ±15 dB. As discussed in section 3.2.1, these peaks and dips are not of absolute value in the HRTF, but relative to the HRTF front direction, according to (1).

A high degree of similarity can be observed when comparing the average HRTF responses from both databases, as shown in Figures 3a and b. Some peaks are noticeable within the two frequency ranges, which depend on the central frequency, Q factor and gain. For example, at low frequencies, a large peak between 1 and 1.5 kHz is observed, whereas at high frequencies, another peak is present between 7 to 8 kHz. Also, there are a number of similar dips for both frequency ranges. Both databases were formed by considering a set of approximately 40 subjects. Although some differences are encountered, one could expect that

both would lead to the same result if the number of subjects was increased and the sample population was similar. Factors such as diversity of ages, races or genders, or the limited scanned population may be the reasoning for the variation between the two results.

On the other hand, the measured values for *RSS* HRTF are shown in Figure 3c. To measure such elevation HRTF provided by the RSS device, a zero degree azimuth was set and the HRIR values were measured in 10 degrees elevation steps. To that end, a MLS sequence was injected to the device input and then the corresponding output was captured and deconvolved to obtain an accurate response with a high signal to noise ratio. In concordance with the previously described case, the response was truncated to 128 samples. As shown in Figure 3c, the low-frequency peak discussed before is shifted upwards, between 1.5 and 2 kHz. The high-frequency peak observed in the previous HRTF is also shifted to 8 kHz and it is much narrower, and very concentrated around this frequency. Moreover, for the three HRTF, it is also observed that increasing elevation always causes an attenuation of high frequencies between 12 and 20 kHz.

3.2.4. Synthetic HRTF

The three elevation HRTF obtained above are defined by their impulse response HRIR for each of the elevation measurement. For their use in real-time systems, a FIR filtering of length equal to the impulse response must be carried out. Moreover, if the aim is to simulate a source at an elevation not covered in the measurements, an interpolation between two closest available elevations, above and below the desired one, must be done. Although thanks to the current processing capacity, a FIR filter of 128 samples is not an insurmountable problem, it remains interesting to have a computationally more efficient filtering.

Through the results previously obtained and the observation of the HRTF plots in Figure 3, two peaks and three dips can be clearly identified in the HRTF magnitude plot for both positive and negative elevations. This applies to both the average HRTF responses of the IRCAM and CIPIC databases and to the measured RSS response. Therefore, it seems reasonable to establish the hypothesis that we can model the elevation response on the median plane through five peak filters, two with positive gain and three with negative gain for both positive and negative elevations.

A peak filter is a second-order parametric filter, which enhances or attenuates with a gain G (dB) at a bandwidth centered at a frequency f_c, and with a quality factor Q. Equation (3) shows the normalized analog prototype peak filter where A is the square root of the filter gain in volts.

$$H_{\text{peak}}(s) = \frac{s^2 + s\frac{A}{Q} + 1}{s^2 + s\frac{1}{AQ} + 1}. \tag{3}$$

This minimum phase filter will be transformed to the discrete domain $H(z)$ by bilinear transformation, and the computation of the digital filter coefficients a_1, a_2, b_0, b_1 and b_2 from the filter parameters f_c, Q, G can be found in [39]. Therefore, a high efficient structure using five peak filters in cascade can be used to model the HRTF response in elevation.

At this point, the next problem to solve is the estimation of the 3 parameters f_c, Q and G for each of the five filters and for any elevation. This is a minimization process, and is not part of this work. In [40], the authors have described a method that fits 3 polynomial functions for each of the 3 parameters that describes a peak filter. The aim is to adjust the filter response to a target response depending on the elevation. Using this algorithm, a filter with five peak sections that corresponds to the frequency response measured in the RSS has been computed.

With the 15 resulting polynomial functions, computed for both positive and negative elevations, a similar response to the three previous HRTF is plotted in Figure 3d. The resulting plot is very similar to that of the RSS, but with very low computational cost and without the need of discretizing elevations. This means that the IIR filter coefficients can be obtained continuously for any elevation and always having smooth transitions. Moreover, computing the three filter parameters through the peak polynomial model as a first step, and later obtaining the five parameters for each second order section from them, ensures good smooth transitions.

4. Experiments

In the previous sections, we have discussed how to combine horizontal and vertical source spatialization by using WFS and HRTF elevation cues, respectively. Afterwards, the elevation cues have been computed, justifying their preprocessing and obtaining their values in 4 different ways. In the first two approaches, an average HRTF calculated from measured databases was considered. In the third approach, a commercial device was measured and analyzed. Finally, a parametric IIR filter model has been proposed, whose coefficients depend on elevation.

At this point, it is necessary to validate the assumptions made and validate the different obtained elevation cues to test if the system works correctly, and to study their behavior individually. To do this, the best way is to conduct a subjective test using a jury to evaluate the system in-situ.

4.1. Aim of the experiments

Before starting with the description of the experiments, it is important to make clear that the system proposed in this paper is not intended to provide accurate localization of virtual sources in elevation, but to give the listeners the possibility of experiencing elevation effects. The purpose is to provide an improvement to WFS regarding its limitation to reproduce elevated sources, since outperforming elevation-oriented systems such as VBAP, Ambisonics or advanced Surround (10.2 or 22.2) is completely out of our scope. That is the reason why we preferred to evaluate relative and dynamic elevation in the experiments below,

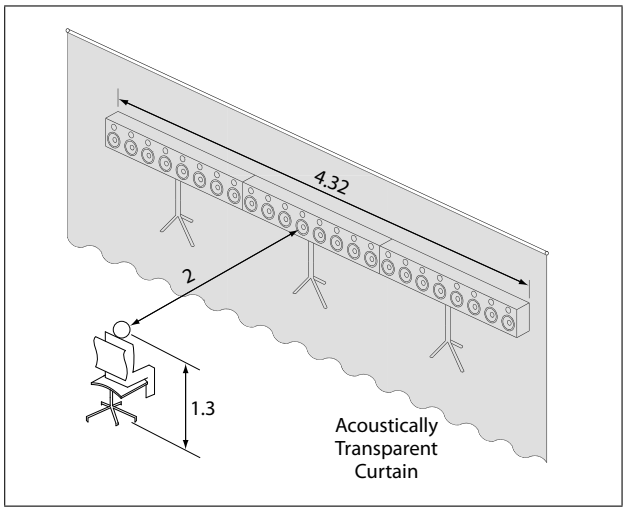


Figure 4. Experiments Set-up. Distances in meters.

rather than measuring absolute localization. Thus, the experiments here described were designed with the aim of testing the capability of the system to provide the sensation of having sources below or above the horizontal plane, to place virtual sources one below/over the other, and to simulate dynamic movements in the median plane.

4.2. Listening tests

Several listening tests were carried out using a panel of 12 subjects. Four of them were researchers familiarized with spatial sound systems and the rest were recruited students with no experience in subjective assessment methodologies. The experiments were carried out using a 24 loudspeaker WFS array. This array is placed inside a recording studio, which is acoustically treated to get a T_{60}@1000Hz < 0.25 s. The volume of this room is 96 m^3 an its floor size is 4 m by 9.1 m. The background noise inside the studio was measured, obtaining SPL values below 25 dB(A). The experimental set up is depicted in Figure 4. Each element of the array is a two-way system, using a 5"1/2 woofer and a 1" tweeter, with a crossover frequency of 2.5 kHz. The loudspeaker separation of the array is 180 mm. The spatial aliasing frequency for this arrangement in the worst case is about 1 kHz. No special aliasing improvement techniques were used in the rendering process. As shown in Figure 4, the array used in the experiments was a simple line of loudspeakers. Therefore, the basic WFS driving function described in the original WFS papers from Berkhout can be simply applied without modifications [17]. Moreover, this WFS prototype is already configured to equalize in real time the loudspeakers employing digital filters, taking into account not only the transducer response, but also the -3 dB filtering effect needed for WFS rendering. As a result, our system has a flat frequency response considering a 1/3 octave frequency resolution in the listener position. Room reflections and spatial aliasing introduce sound coloration, but the 1/3 octave smoothed response is almost flat around the listener location.

Table I. Listening tests details.

WFS Set-Up	
Geometry	Linear
Number of loudspeakers	24
Separation distance	18 cm
Spatial aliasing frequency	∼1 kHz
Loudspeaker type	two-way: 5"1/2 and 1"
Crossover frequency	2500 Hz
Equalization	Yes
Virtual Source Type	Point sources
Room Details	
Reverberation Time	T_{60}@1000Hz < 0.25 s
Volume	96 m^3
Floor size	4 × 9.1 m
Background noise	< 25 dB(A)
Test and Subjects	
Subjects	Ages 22-40; hearing thresholds ≤ 20 dB
Test signals	Pink noise bursts of 1 s duration and 0.5 s gaps
Loudness	65 dB(A) when elevation is 0°

An acoustically transparent curtain was used to hide the loudspeakers and avoid the ventriloquism effect. The virtual sound sources used in the experiments were rendered in WFS as point sources. The test signals consisted of a set of three broad-band pink noise bursts of 1 second duration with silence intervals of 0.5 seconds in between. The loudness of the test signal was adjusted to be 65 dB(A) at the listening position when elevation was 0°.

The experiments below were repeated for each of the considered responses (IRCAM, CIPIC, RSS and PEAK). The desired HRTF filtering is performed directly in the sources before applying the WFS algorithm. This saves computational time in comparison to filtering each loudspeaker signal. For the three first HRTF banks (IRCAM, CIPIC and RSS), FIR filtering was performed. For PEAK, IIR filtering was applied. Table I shows the details of the listening tests.

4.3. Experiment 1: Upwards/downwards discrimination

In the first experiment, the discrimination capability between upwards and downwards moving sources was studied. For each of the databases considered, four stimuli were generated. They consisted of the reference test signal (pink noise) moving in the following directions:

1. Upwards from the horizontal plane to 40°.
2. Downwards from 40° to the horizontal plane.
3. Upwards from the horizontal plane to -40°.
4. Downwards from -40° to the horizontal plane.

Before starting the test with a panel of 12 subjects, some initial informal experiments were carried out with a reduced group of four listeners. Since no official recommendations were found matching our requirements, we

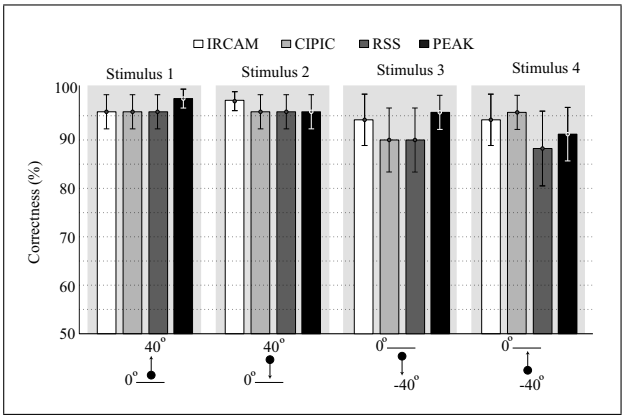


Figure 5. Results for the Upwards/Downwards Discrimination Experiment. Hit rates for the different elevation systems and source movements. Limits indicate 95% confidence intervals.

thought that a preliminary test was necessary for making sure that the designed tests really addressed our research goals. These pre-tests consisted in the same experiments described in the paper, but with longer duration due to the comparison of added elevation pairs. These preliminary results showed that elevation over ± 45 degrees was no longer perceived, and just an excessive source colouration was reported by the listeners. Besides, there is no negative elevations below 40 degrees in any of the considered databases. For that reason, it was decided to limit the test to ±45 or ±40 degrees, according to the database, given the difficulty of identifying higher elevations in this initial test.

Each stimulus was randomly presented to the listener 5 times, making a total of 20 assessments per listener. For each sound presented, the subjects had to decide if there was either an upwards movement or a downwards movement, with the possibility of listening to them several times if needed. It is important to note that this experiment not only provides an absolute cue for elevation, but the observed changes in the filtered noises can help to simulate movements in sagittal planes. Therefore, having a sensation of elevation is easier.

Figure 5 shows the results obtained for this experiment, where the bars indicate the grade of success obtained by the listeners in identifying correctly the direction of movement of the stimulus 1 to 4 of the above list. It can be observed that all the systems performed significantly well.

For movements above the horizontal plane, the results are always above 95% for all HRTF banks. In the case of movements below the horizontal plane, the results are slightly worse, but also provided about 90%. These results are quite encouraging since at least, it can be distinguished whether a source is moving up or down in the proposed system, which would allow such systems to use special effects in movies or video games.

4.4. Experiment 2: Higher/lower discrimination

In this experiment, the capability of identifying which sound is at a higher position from two successive sounds filtered with different θ responses was studied. To accomplish this, the jury must have the same sound excitation, but filtered for two different elevations. Then, the subject has to answer which of the two seems to be higher, the first or the second one. The sounds were again two pink noise bursts corresponding to different values of angle θ. In the experiments all the pair combinations of available elevations in the range of ±45 or ±40 degrees were used, but differentiating two tests, one for positive and another for negative elevations. Combinations were chosen randomly for each subject using a software test that automates the process.

Figure 6 shows different matrices representing the hit rate obtained for each pair of stimuli in the different elevation systems. Rows indicate the θ position of the first sound and columns indicate the θ position of the second sound. White cells indicate that all the subjects successfully perceived which sound was at upper and which was at lower positions. Note that smaller rates are achieved when the relative distance of the two sounds is small, as the filters used for simulating elevation are more similar in these cases.

Limiting the test at elevations no higher than 45 degrees, the results show quite encouraging results in the graphs, although there is confusion in some stimuli for static sources. It is also worth to mention that although not reflected in the results, the overall reduction in the rate of success is caused primarily by the repeated failure of some members of the jury, rather than dispersed failures. Perhaps this is because in these members, the head, torso and ear anthropology is quite different than that calculated in other cases.

Although not studied in this paper, it seems reasonable to assume that, following the experiments of sources in motion, a subject could readily identify the elevation of a static source if it has previously had a movement of ascent or descent. Therefore, the presented results would improve if the two sources moved in elevation before becoming static.

Finally, it is fair to mention that for a fixed source to a certain position x, y, z, when the listener moves away or approaches the virtual source, this virtual source should change its elevation relative to the subject, since the angle at which the listener perceived it also varies, unless the source is very distant, where little variation occurs. This misalignment is an inherent small drawback of this system. However, it is as minor problem since the localization accuracy in elevation is not very large in humans [8] and because the benefits found in this system are greater than that of its disadvantages.

5. Conclusion

In this work, a novel system that combines the use of HRTF spectral elevation cues jointly with WFS has been proposed in order to conform a full 3D sound reproduction system, which is a powerful and practical solution to the well-known limitations regarding elevation in

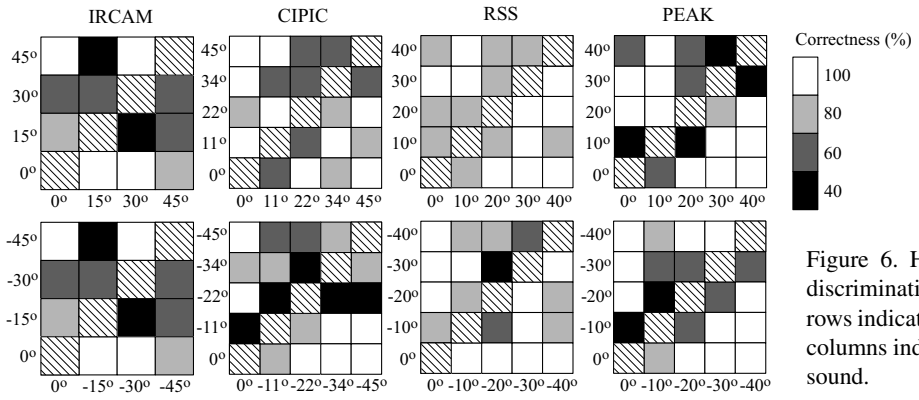


Figure 6. Hit rate for the Higher / Lower discrimination experiment. For each matrix, rows indicate θ position of the first sound and columns indicate the θ position of the second sound.

WFS systems. The distribution of the sources in the horizontal plane is carried out by means of WFS providing high-quality localization accuracy and a large sweet spot, whereas the sensation of elevation is provided by a filter that simulates quasi-universal HRTF elevation cues.

Different contributions have been described throughout this paper. Firstly, not only the idea of a combined HRTF-WFS system has been proposed, but also elevation cues have been analyzed with the aim of providing a way to integrate them into a conventional WFS reproduction system. On the other hand, our own set of elevation cues has been obtained by merging different types of existent approaches: an exhaustive search over two large public databases obtained by different laboratories and the measures performed over an electronic device that provides full 3D reproduction. Finally, a low-complexity model for the synthesis of elevation cues on the basis of IIR filtering has been achieved.

The proposed approach has been tested over a dedicated WFS system and two sets of subjective assessments have been conducted. The results obtained from these listening tests showed the viability of the system, which provided very acceptable performance in most cases.

Despite that elevation is not always perfectly perceived by all the listeners, the performance improves when motion sources are considered and more than 90% of the subjects are able to describe their movement. Regarding static sources, the results are also more than acceptable. Although no absolute elevation has been measured, the proposed approach has shown to be very useful in providing elevation effects in WFS without the need for using additional loudspeakers or changing the reproduction set-up. A quantitative evaluation of absolute localization accuracy would be interesting in the line of this research and that is something we plan to address in our future work.

References

[1] F. Rumsey: Spatial audio. Focal Press, 2001.

[2] A. J. Berkhout, D. de Vries, P. Vogel: Acoustic control by Wave Field Synthesis **93** (1993) 2764–2778.

[3] H. Wittek: Perceptual differences between wavefield synthesis and stereophony. PhD thesis, School of Arts, Communication and Humanities, University of Surrey, October 2007.

[4] W. P. J. deBrujin, M. Boone: Application of wave field synthesis in life-size videoconferencing. Proceedings of the 114th AES Convention.

[5] C. I. Cheng, G. H. Wakefield: Introduction to head-related transfer functions (HRTF's): Representations of HRTF's in time, frequency, and space (invited paper). Proceedings of the 107th Convention of the Audio Eng. Soc., New York, USA, 1999.

[6] P. M. Hofman, J. G. A. Van-Riswick, J. Van-Opstal: Relearning sound localization with new ears. Nature Neuroscience **1** (1998).

[7] J. Blauert, G. Boerger, L. Peter: Method and arrangement for controlling acoustical output of earphones in response to rotation of listener's head. United States Patent 3962543, 1976.

[8] J. Blauert: Spatial hearing. MIT Press, Cambridge, MA, 1997.

[9] M. A. Gerzon: With-height sound reproduction. Journal of the Audio Engineering Society **21** (1973) 2–10.

[10] D. Malham: Higher order ambisonic systems for the spatialisation of sound. Proceedings of the Int. Computer Music Conf. (ICMC99), Beijing, China, 1999.

[11] J. Daniel: Spatial sound encoding including near field effect: Introducing distance coding filters and a viable, new ambisonic format. Proceedings of the 23rd International Conference of the Audio Eng. Soc., Copenhagen, Denmark, 2003.

[12] V. Pulkki: Virtual sound source positioning using vector base amplitude panning. Journal of the Audio Engineering Society **45** (1997) 456–446.

[13] V. Pulkki: Localization of amplitude-panned virtual sources. II: Two- and three-dimensional panning. Journal of the Audio Engineering Society **49** (2001) 753–767.

[14] T. Holman: 5.1 Surround Sound: Up and running (2nd edition). Focal Press, 2007.

[15] K. Hamasaki, K. Hiyama, R. Okumura: The 22.2 multichannel sound system and its application. 118th Conv. Audio Eng. Soc., Barcelona, Spain, 2005, number 6406.

[16] A. J. Berkhout: A holographic approach to acoustic control. Journal of the Audio Engineering Society **36** (1988) 977–995.

[17] A. J. Berkhout, D. de Vries, P. Vogel: Acoustic control by wave field synthesis. Journal of the Acoustical Society of America **93** (1992) 2764–2778.

[18] M. M. Boone, E. N. G. Verheijen, P. F. van Tol: Spatial sound field reproduction by wave field synthesis. Journal of the Audio Engineering Society **43** (1995) 1003–1012.

[19] S. Spors, R. Rabenstein, J. Ahrens: The theory of Wave Field Synthesis revisited. 124th Conv. Audio Eng. Soc., Amsterdam, The Netherlands, 2008, number 7358.

[20] E. Corteel: Synthesis of directional sources using wave field synthesis, possibilities, and limitations. EURASIP J. Appl. Signal Process. **18** (2007).

[21] H. Moller, M. F. .Sorensen, D. Hammershoi, C. B. Jensen: Head-related transfer functions of human subjects. Journal of the Audio Engineering Society **43** (1995) 300–321.

[22] M. Morimoto, M. Itoh, K. Iida: 3-D sound image localization by interaural differences and the median plane HRTF. Proceedings of the 2002 International Conference on Auditory Display, Kyoto, Japan, 2002.

[23] R. O. Duda: Modeling head related transfer functions. Proceedings of the 27th Asilomar Conference on Signals, Systems and Computers, Asilomar, CA, USA, 1993.

[24] D. W. Batteau: The role of the pinna in human localization. Proceedings of the Royal Society of London **168** (1967) 158–180.

[25] C. P. Brown: Modeling the elevation charactersitics of the head-related impulse response. Master's thesis, San Jose State University, 1996.

[26] C. P. Brown, R. O. Duda: An efficient HRTF model for 3-D sound. Proceedings of the IEEE ASSP, 1997.

[27] W. M. Hartmann, A. Wittenberg: On the externalization of sound images. J. Acoust. Soc. Am. **99** (1996) 3678–3688.

[28] B. S. Atal, M. R. Schroeder: Simulation of room acoustics using electronic computers. Gravesaner Blätter **27/28** (1996).

[29] C. Avendano, V. R. Algazi, O. R. Duda: A head-and-torso model for low-frequency binaural elevation effects. Proceedings of IEEE WASPAA'99, New Paltz, NY, USA, 1999.

[30] R. Algazi, R. O. Duda, P. Satarzadeh: Physical and filter pinna models based on anthropometry. Proceedings of the AES 122nd Convention, Vienna, Austria, 2007.

[31] A. Kulkarni, H. S. Colburn: Role of spectral detail in sound-source localization. Nature **396** (1998) 747–749.

[32] H. L. Han: Measuring a dummy head in search of pinna cues. Proc. of the 90th Conv. of the Audio Eng. Soc., 1991, number 3066.

[33] IRCAM: LISTEN HRTF database. available online at: http://recherche.ircam.fr/equipes/salles/listen/, 2003 (last update).

[34] V. R. Algazi, O. R. Duda, D. M. Thompson, C. Avendano: The CIPIC HRTF database. Proceedings of IEEE ICASSP'01, New Paltz, NY, USA, 2001.

[35] C. Chan: Sound localization and spatial enhancement realization of the Roland sound space processor. Proc. of the 91st Conv. of the Audio Eng. Soc., New York, October 4-8, 1991.

[36] E. M. Wenzel, M. Arruda, D. J. Kistler, F. L. Wightman: Localization using nonindividualized head-related transfer functions. Journal of the Acoustical Society of America **94** (1993) 111–123.

[37] D. J. Kistler, F. L. Wightman: A model of head-related transfer functions based on principal component analysis and minimum-phase reconstruction. Journal of the Acoustical Society of America **91** (1992) 1637–1647.

[38] A. V. Oppenheim, R. W. Schafer, J. R. Buck: Discrete-time signal processing. 2 edition. Prentice Hall, 1999.

[39] U. Zölzer: Digital audio signal processing. John Wiley & Sons, New York, NY, 1997.

[40] J. J. Lopez, M. Cobos, B. Pueo: Efficient implementation of the HRTF elevation cues by means of parametric IIR filters **2010** (submitted).