HUMAN INSPIRED AUDITORY SOURCE LOCALIZATION

Sylvia Kümmel, Eric Haschke and Thorsten Herfet

Telecommunications Lab,
Saarland University,
Germany
{kuemmel,haschke,herfet}@nt.uni-saarland.de

ABSTRACT

This paper describes an approach for the localization of a sound source in the complete azimuth plane of an auditory scene using a movable human dummy head. A new localization approach which assumes that the sources are positioned on a circle around the listener is introduced and performs better than standard approaches for humanoid source localization like the Woodworth formula and the Freefield formula. Furthermore a localization approach based on approximated HRTFs is introduced and evaluated. Iterative variants of the algorithms enhance the localization accuracy and resolve specific localization ambiguities. In this way a localization blur of approximately three degrees is achieved which is comparable to the human localization blur. A front-back confusion allows a reliable localization of the sources in the whole azimuth plane in up to 98.43 % of the cases.

1. INTRODUCTION

Humans can estimate the position of a sound source in the auditory scene quite accurately with a localization blur of approximately two to three degrees in the fore side azimuth plane dependent on the psychological study [1]. This paper describes an approach that tries to mimic this ability of the human auditory system and localizes sound sources in the complete azimuth plane of the auditory scene with a comparable localization blur. In further steps of the project the localization method described in this paper is extended to the multiple source case and is used as input to a following source separation architecture [2], [3] that takes advantage of the positions of the sources in the auditory scene.

To imitate the conditions humans experience in an auditory scene, a movable human dummy head with three degrees of freedom – called Bob – is used. Bob resides in a normal office room and is able to move in any human-like position to investigate the auditory scene around, which is constructed using a standard 7.1 loudspeaker system.

2. HUMANOID SOURCE LOCALIZATION

For source localization the human brain mainly uses Interaural Time Differences (ITD) and Interaural Level Differences (ILD) [1] of the signals received at the left and the right ear, which arise due to the distance between the ears. This spatial separation enables a sampling of the received signals in the auditory space. The solid head between the two ears introduces diffraction and scattering of the sound waves and accounts for significant head shadows at the ear that is turned away from the sound source.

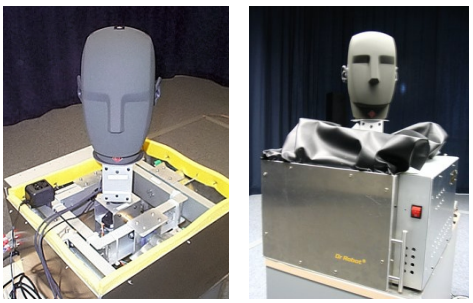

Figure 1: *Bob – the movable human dummy head.*

The spatial separation of the ears and the head shadow not only affect the arrival times of the signals, but also account for interaural level differences of the signals. The signal at the turned away ear has travelled further and so has lost more energy on its way, which leads to slight level differences dependent on the incidence direction. The head shadow contributes additional level differences, which can be up to 25 dB at high frequencies [4]. Analogously to the ITD, the direction dependent ILD can be used to estimate the location of a sound source. But opposed to the ITD values, the ILD values are not well predictable by diffraction theory and depend heavily on the arrival angle, the frequency and the distance of the source [5].

3. ESTIMATION OF ITD AND ILD

The ITD between the two ears can be estimated by correlating the two ear signals and finding the peak in the resulting function. Assume x_L and x_R denote the time domain signal of the left and the right ear. The correlation function is defined as

$$R(l) = \sum_t x_L(t+l) \cdot x_R(t). \tag{1}$$

The interaural time difference is then estimated as the time value corresponding to the highest peak in the correlation function.

The interaural level difference between the left and the right ear signal is computed by subtracting the power of the left signal and the power of the right signal.

$$\Delta L = L_{left} - L_{right}. \tag{2}$$

$$\Delta L = 10 \cdot \log_{10} \frac{\sum_t x_L^2(t)}{\sum_t x_R^2(t)} \quad dB. \tag{3}$$

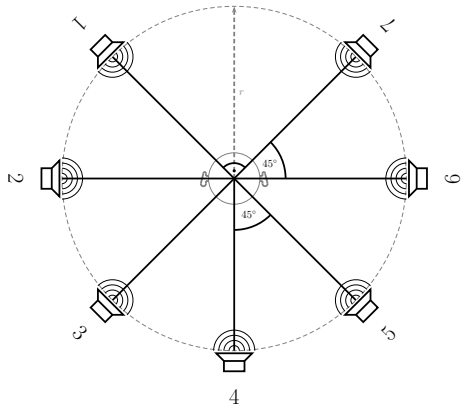

Figure 2: *Schematic view of loudspeaker constellation.*

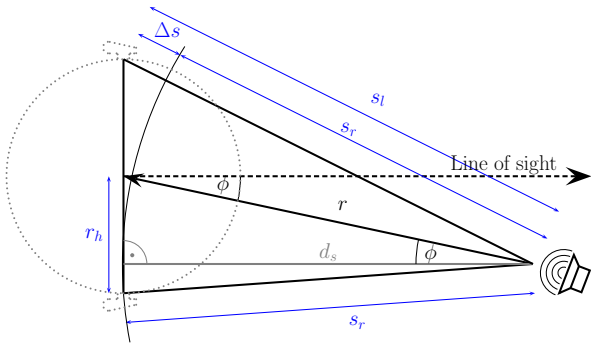

Figure 3: *Free-field propagation of sound waves through a transparent head.*

Current neurophysiological research [6] found evidence that the human brain is also able to perform correlation based methods to estimate the time differences. There are special cells in the inferior colliculus in the brain stem which are maximally sensitive to a specific ITD, independent of the frequency of the incoming signal. There also exist special cells that are tuned to respond to specific ILDs [5].

4. INCIDENCE ANGLE ESTIMATION

This section first describes a new circle-formula for estimating the direction of incidence of the auditory source based on the available ITD which assumes that the sources are located on a circle around the listener. This formula is compared to two standard formulas to transform the ITD values to the estimated incidence direction: the freefield formula and the Woodworth formula. Additionally an approximation of the HRTF for the described scenario is evaluated for its source localization capabilities.

4.1. Formula for circular loudspeaker constellation

In the current scenario, the dummy head Bob is located in the center of a normal office room, while the loudspeakers are arranged on a circle around as depicted in figure 2. For such a scenario, where the sources are assumed to emanate from a circle with a specified radius r, the following approach is used to localize the sources.

Consider the scenario shown in figure 3, where the head is assumed to have radius r_h. The travelled distance of the left and right source to the two ears s_l and s_r differs by

$$\Delta s = s_l - s_r \tag{4}$$

The resulting interaural time difference is estimated by

$$\Delta t = \frac{\Delta s}{c} \tag{5}$$

where $c = 343\frac{m}{s}$ is the speed of sound in air. Assuming that the sound source is always located on the circle, the distance between the sound source and the center of the head is always r. Substituting equation (5) into equation (4) yields

$$s_l - s_r = \Delta t \cdot c \tag{6}$$

The distances of the sound source to the left and the right ear unfold after simple trigonometric transformations to

$$s_l = \sqrt{r^2 + 2 \cdot r_h \cdot r \cdot \sin\phi + {r_h}^2} \tag{7}$$

$$s_r = \sqrt{r^2 - 2 \cdot r_h \cdot r \cdot \sin\phi + {r_h}^2} \tag{8}$$

Inserting equations (7) and (8) into equation (6) yields after rearranging terms a function of ϕ in dependency of the interaural time difference Δt

$$\phi(\Delta t) = \frac{180^\circ}{\pi} \cdot \arcsin\left[\frac{\Delta t c}{2 r r_h}\sqrt{\frac{1}{2}\left(r^2 + {r_h}^2 - \frac{1}{2}\Delta t^2 c^2\right)}\right] \tag{9}$$

4.2. Freefield Formula

Figure 3 shows the scenario when the head is assumed to be a perfect sphere with radius r_h and the sound waves are supposed to be able to travel through the head without diffraction and reflection. The distance of the sound emanating source is assumed to be large compared to the head radius. The travelled distances of the left and right signal differ by an amount

$$\Delta s = s_l - s_r, \tag{10}$$

which is dependent on the incidence direction of the wave front. Assuming a propagation speed of sound in air of $c = 343\ m/s$, the arrival time difference Δt is proportional to Δs:

$$\Delta t = \frac{\Delta s}{c} \tag{11}$$

By applying simple geometric transformations, the interaural time difference between the two ears can be approximated dependent on the incidence angle ϕ and the head radius r_h by the sine law [4]:

$$\Delta t \approx \frac{2 r_h \sin\phi}{c} \tag{12}$$

4.3. Woodworth Formula

In the case of a head modelled as a perfect solid sphere, the sound waves diffract and reflect at the turned-away side. Accounting for the diffraction characteristics, the length of the travelled path of the incident sound wave is longer than in the free-field case. Motivated by this, Woodworth and Schlosberg [7] applied diffraction theory to a completely spherical head, yielding the following formula to approximate the ITD:

$$\Delta t = \frac{r_h(\phi + \sin\phi)}{c} \tag{13}$$

4.4. Head Related Transfer Function

The complete refraction and resonance characteristics of the human dummy head Bob and the media lab can be specified by measuring the Head Related Transfer Function (HRTF) for ITD and ILD. The localization of a source in the auditory scene is then approached by using the HRTF as a table lookup: The ITD and ILD of the left and right signal are computed and compared to the HRTF to find the dedicated position.

HRTFs are very sensitive to changes in the setup and the performance severely degrades, if the recording setup and the application scenario differ [8]. Even for different head positions in the same environment differences in the HRTFs occur (see figure 4).

The HRTF of the human dummy head Bob is measured by playing back white noise from each of the seven loudspeakers (positions depicted in figure 2). For each loudspeaker position, Bob turns from 0° (looking straight to the loudspeaker) to 360° (again looking straight to the speaker) in 1°-steps and records one second of white noise in each case. For each loudspeaker the incoming signal is filtered with a gammatone filterbank of 512 channels. For each of the 512 frequency channels of the resulting cochleagram and each of the 360 source positions, the ITDs and ILDs are computed. The complete HRTFs for ITDs for each loudspeaker position are plotted in figure 4.

To smooth out local variations and to make the HRTF more robust against changes in the environment, a mean HRTF is computed by taking the mean of all seven measured HRTFs. Inspecting the mean ITD-HRTF yields that each channel can roughly be approximated by a sinusoid of a specific frequency and amplitude – as already mentioned by several other researchers (i.e. [9], [4], [8]). An approximate HRTF is then found by fitting each channel to a sinusoidal model:

$$\Delta t(\phi, f) = \alpha_f \cdot \sin(\omega_f \phi),$$

where α_f denotes the frequency dependent scaling factor, and ω_f specifies the frequency dependent sinusoid frequency. The mean HRTF, the approximated HRTF and the resulting error function is plotted in figure 5. Especially in the low frequencies up to 1 kHz, the sinusoidal model fits very well and the error function shows only little deviations. This is consistent with the so called Duplex Theory which describes the human source localization based on the combined evaluation of the physical cues ITD and ILD and was first identified by Lord Rayleigh [10]. In frequencies larger than 1 kHz, the ITD starts to become ambiguous as the wavelength of the signal becomes comparable to the size of the head and the correlation function of the left and right ear signal starts to exhibit several peaks. For frequencies greater than 1 kHz, the ILD cues become a reliable measure of the incidence direction as the signals of short wavelength are not refracted by the human head – they either are reflected completely or pass with little refraction.

In analogy to the approximation of the ITD-HRTF, the ILD-HRTF can be approximated by a sinusoidal model. Viste [4] for example also approximates the ILD with sinusoids of different frequency and amplitude. Inspecting the mean ILD-HRTF however shows additional peaks and valleys in the curves, especially in frequencies higher than 1 kHz. In the case of the dummy head Bob, a model consisting of only one sinusoid does not adequately model the ILD-HRTF. Fitting harmonic Fourier series of length two however already gives a good and simple approximation of the function.

$$\Delta l(\phi, f) = \alpha_f \cdot \sin(\omega\phi) + \beta_f \cdot \sin(3\omega\phi) \tag{14}$$

This model is analog to a model used by Duda et. al [11]. Duda et. al noted that the ILD is periodic in ϕ and approximate the ILD by complete Fourier series expansions.

The result of the approximation is plotted in figure 6. In the low frequencies up 800 Hz, there are only little ILDs as the waves pass through the head without reflection. Above 800 Hz the used model approximates the ILD well up to approximately 1.5 kHz. Above 1.5 kHz additional sinusoidal vibrations occur. Adding further harmonics to the model can eliminate these peaks and valleys and leads in the end case to the model used by Duda et. al [11].

At the current time there are no simple models known that describe the characteristics of the ILD in general [4] as in the case of the ITD, where the Freefield-Formula, the Woodworth-Formula and the Formula for a circular arrangement of the sound sources already give easy and good approximations for the estimation of the incidence direction.

4.5. Results

To estimate the incidence direction of a sound source in the auditory scene recorded by the human dummy head Bob, the five described formulas are compared to each other to find the most suitable and computational manageable algorithm for source localization. Figure 7 shows the results of the five described estimation formulas for sound source locations of -90° to 90°. The values are obtained by taking the average values for 240 speech sources of one second length taken from the speech database CMU Arctic [12] and played back from the specified directions. Each sound source is recorded with a sampling rate of 44.1 kHz. The estimation of the location by the HRTF formulas is performed by computing the average estimated direction of several frequency channels. For the ITD-HRTF the ITD value is used as table lookup to find the most probable incidence direction for all channels between 200 Hz and 1000 Hz, the range where the error function of the approximated ITD-HRTF is almost zero. The ILD-HRTF computes the location of the source analogously but uses only channels from 800 Hz to 1400 Hz as the error function of the ILD-HRTF is small in this range.

The black line shows the optimal results, where each sound source is assigned to its correct incidence direction. The freefield formula (the red line) performs well for incidence directions between -20° and $+20^\circ$ and shows a localization blur of approximately three degree, which is comparable to the human localization blur [1]. For values greater than $\pm 20^\circ$, the algorithm overestimates the directions by up to 40°. The Woodworth formula (the green line) performs similarly and comparable to the Freefield formula, but the overestimation of incidence directions greater than

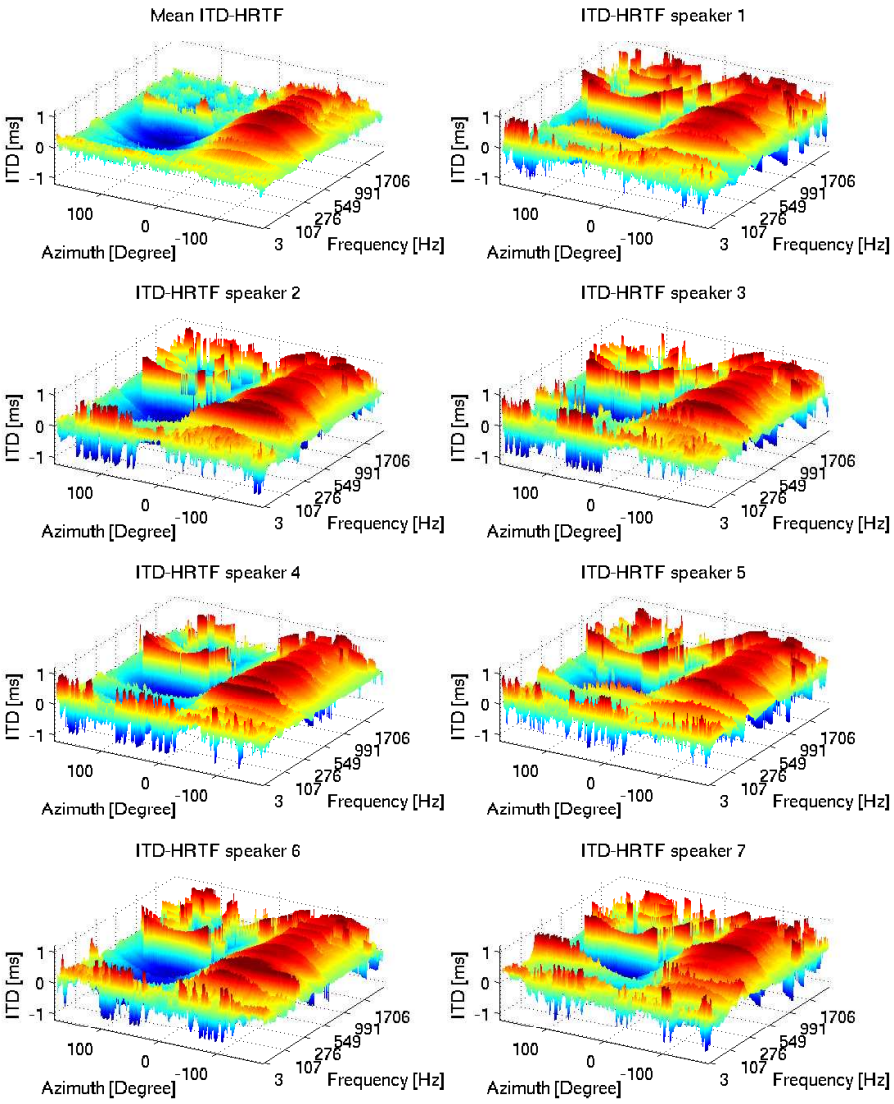

Figure 4: *HRTF for ITD for dummy head Bob residing in a normal office room.*

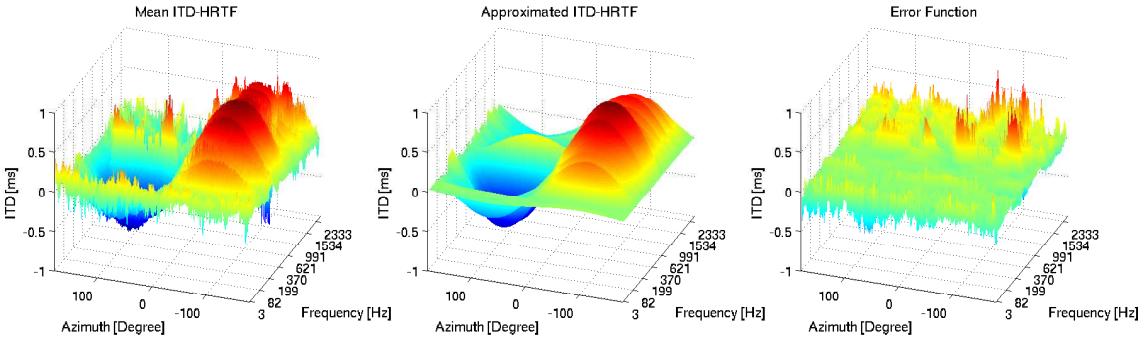

Figure 5: *Mean measured HRTF for ITD and approximated HRTF.*

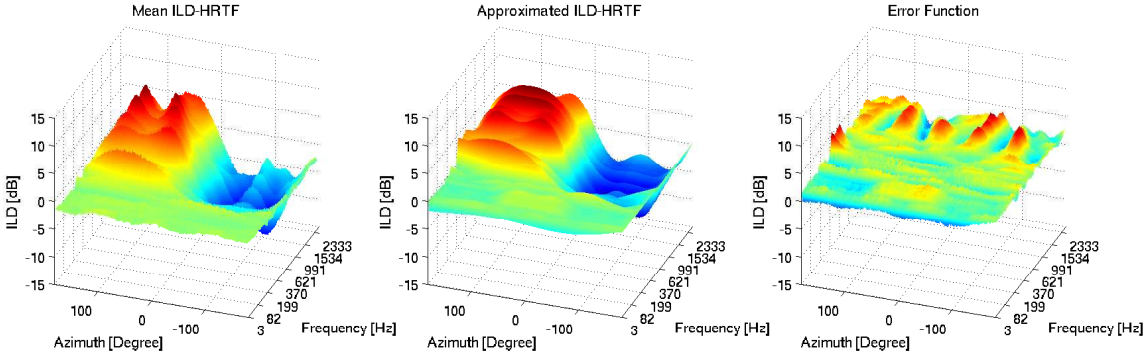

Figure 6: *Mean measured HRTF for ILD and approximated HRTF.*

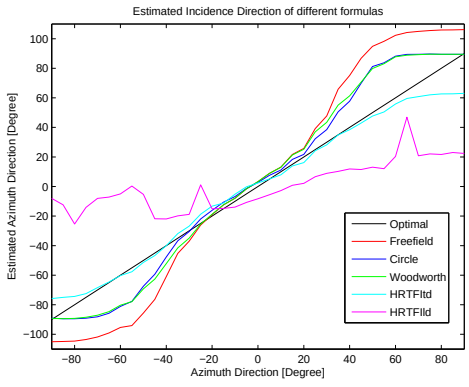

Figure 7: *Comparison of the five described formulas for incidence direction estimation based on the interaural time and level difference.*

$\pm 20^\circ$ is only up to 20°. The formula for the circular arrangement (the blue line) of the loudspeakers performs about 4 to 5° better than the Woodworth and the freefield formula especially in the range between $\pm 20 - -40^\circ$ and therfore should be preferred to cover a higher reliability range. The HRTF of the interaural time differences outperforms the other formulas and estimates the location of the sound source reliably up to $\pm 50^\circ$ with less than three degree localization blur. The HRTF for the ILD performs only poorly and can only be used to derive a coarse direction like "on the left side" or "on the right side".

The presented algorithms for localization deliver reliable incidence direction estimations for source locations between $\pm 20^\circ$ for the formulas and $\pm 50^\circ$ for the HRTFs. The human dummy head Bob should be able to localize sources in the complete horizontal plane. To estimate also incidence directions greater than $\pm 50^\circ$ reliably, the location estimation is used iteratively. In each iteration Bob turns to the estimated direction and refines his computation until a stable result occurs.

Figure 8 shows the results for the different formulas for two and three iterations. The ITD-HRTF formula estimates the directions almost optimal with less than three degree localization blur even in the two iteration case (see left plot), but the accuracy slightly increases in the three iteration case (see right plot).

For two iterations, the ITD formulas refine their accuracy in the range up to $\pm 30^\circ$ and for incidence directions greater than 60° and achieve a localization blur of less than three degree for these directions. In the range between 30° and 60° on both sides of the head, the estimation suffers. While the Circle and the Woodworth formula misjudge the incidence directions by up to 10° in

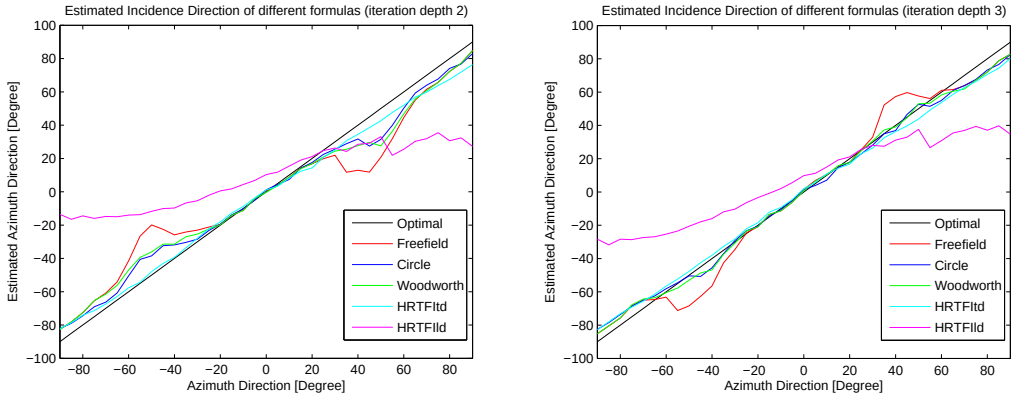

Figure 8: *Results of the iterative localization with two (left plot) and three (right plot) iterations.*

this range, the Freefield formula evaluates falsely up to $30°$. The fall-off of the performance for the Circle, Woodworth and Freefield formula in this range can be traced back to the moderate performance of these formulas in the one-iteration case. Assuming a source position of $45°$, the Freefield formula estimates the position to approximately $90°$ (see graph in figure 7). The head is then moving to position $90°$ and the new relative position of the source is at $-45°$, which in turn leads to a false estimation of $-90°$ and so on and so forth. The performance of the Woodworth and the Circle formula in this range can be explained analogously, but the first estimation is approximately $70°$ and the next iterations can resolve this error as can been seen in the three iteration case. The third iteration increases the accuracy of the estimation and the Woodworth, the Circle and the HRTF-ITD formula perform almost equally.

The results of the ILD-HRTF have been improved compared to the non-iterative case, but cannot keep up with the incidence estimation based on ITD. The performance of the ILD estimation has increased compared to the two iterations case, but is still not comparable to the incidence estimation based on ITD. By using more iteration steps, the ILD estimation can be further refined and an acceptable accuracy is achieved by approximately eight iterations.

To localize also sources in the back of Bob, a front-back discrimination is implemented as described in the next section.

5. FRONT-BACK CONFUSION

All algorithms used for location estimation (except the HRTF) assume a spherical head and a symmetrical setup regarding the front and back direction. These formulas always assume that the sound source is located in front and return an estimated incidence between $\pm 90°$ as they are only invertable in this range. They are not able to distinguish, if a source is coming from the front or the back direction. Signals coming from the back are localized erroneously at the mirrored frontal position.

Assuming a constant distance of the sound sources, there always exist two incidence directions in the horizontal plane that exhibit the same time and level differences as depicted in figure 9. Denoting these directions with ϕ_1 and ϕ_2, the two possible source locations are related to each other by

$$\phi_2 = \phi_1 + 2 \cdot \alpha \tag{15}$$

All formulas – including the HRTF – approximate the ITD and

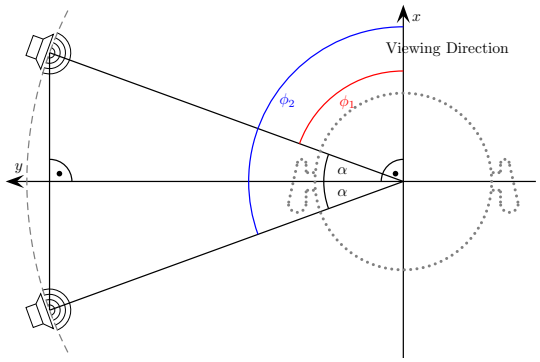

Figure 9: *Front-Back Confusion of single sound source.*

ILD by a kind of sinusoidal function. The inverse table lookup of the time and level differences return degree values between $\pm 90°$, which always corresponds to position ϕ_1. It is therefore not possible to infer with these formulas, if the source is coming from position ϕ_1 or position ϕ_2.

If the distance of the sound source is not constant, the two possible locations of the sound source expand to half-lines emanating in directions ϕ_1 and ϕ_2. If the elevation dimension is additionally regarded, the set of possible locations further expand to the rotational solid of the two half lines, which results in the so called cone of confusion [1]. Experiments with human subjects revealed that for a real – not completely spherical – human head the half lines look more like hyberbolas, which corresponds to an hyperboloid in the three dimensional case [13].

Humans use the frequency and direction dependent filtering of the outer ear to determine if the sound source is coming from the front or the back direction [1]. Additionally humans use head motions to resolve front-back ambiguities [1]. A slight movement of the head yields a specific change in the ITDs and ILDs and is used to estimate, if the source is located in the front or the back direction (compare to figure 10).

The iterative source localization algorithm described in the last section resolves most of the front-back confusion errors by the separate incidence direction estimation in each direction. Figure 11 shows the results of the iterative localization for a variable number of iterations. The algorithm stops, when a stable result is

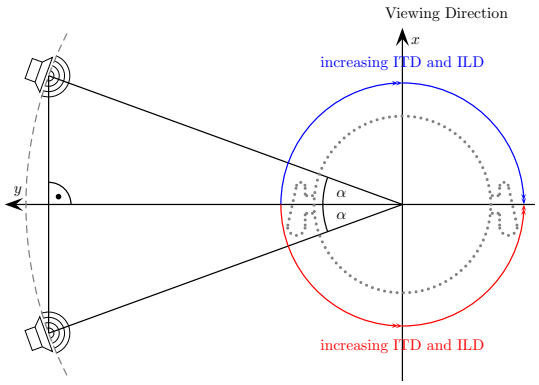

Figure 10: *Change of ITD according to the incidence direction of the sound source.*

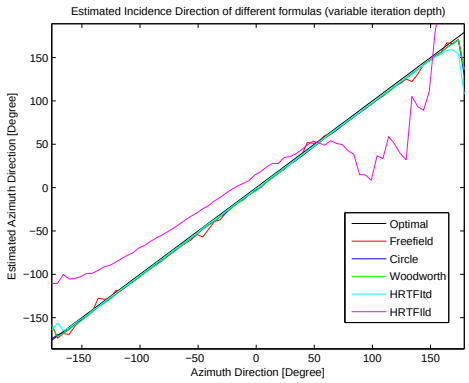

Figure 11: *Results of the iterative localization with front-back confusion for a variable number of iteration (maximal seven iterations).*

Algorithm	Correct Estimated [%]
Freefield	98.24
Circle	97.22
Woodworth	98.43
HRTF-ITD	94.35
HRTF-ILD	96.39

Table 1: *Results of the front-back confusion algorithm for the five described algorithms.*

reached or a maximum of seven iterations has occurred. All formulas perform comparable to the only front-estimation case. The Freefield formula further exhibits the peaks at $\pm 45^\circ$ for the same reason described above. For incidence directions with absolute values greater than 170°, the algorithms sometimes errorsly localize the mirrored location in front, which leads to the decreased mean value in the figure.

For incidence directions with absolute values greater than 170° an additional front-back estimation therefore evaluates if the source is coming from the front or the back. The front-back confusion is resolved by using the moving ability of Bob. During the iterative detection of the location of the source, Bob notices the changes in the ITD and the ILD in every iteration according to the corresponding positions of the head in each iteration.

It has turned out that the ILD changes are more reliable than the ITD changes to estimate the front-back direction. The algorithm first examines the positions of the localization track – the absolute head positions during each iteration. Then for each two iterations, the change of the ILD is assigned to the front or the back direction according to the sign and a final direction is judged based on a major vote.

Problems arise, when there is only one iteration. If the source is i.e. located at 180° and the localization algorithm estimates the incidence direction as 0° in the first iteration, no change in the ILD respectively ITD can be measured. In these cases, where only one valid position is available, the head is moved by a specific amount – i.e. 10° – to one side, the ILD change is measured and the front-back direction is judged.

Table 1 shows the results of the front-back-confusion algorithm for the five localization formulas, when using only the positions and the ILD changes of the first two iterations of each localization. The values are obtained by evaluating 80 speech signals of one-second length, played back from all positions between -20° to 20° and 160° to 200°, which leads to 6400 evaluations of the algorithm for each formula. For the Woodworth formula, the algorithm estimates the front-back direction correctly in 98.43 percent of the cases.

6. CONCLUSIONS & FUTURE WORK

This paper described an approach to localize a single sound source in the complete azimuth plane of an auditory scene using a human dummy head. The presented iterative variants of the algorithms with the front-back confusion are able to localize a sound source in a real reverberant auditory scene with a localization blur of less than three degrees.

Future work especially includes the extension of the localization algorithms to multiple sources. This can for example be accomplished by regarding the complete correlation function of the

two ear signals, not only the highest point. In each iteration, assumptions about the number and the positions of the sources can be computed and refined in each iteration to get a final result.

7. REFERENCES

[1] Jens Blauert, *Spatial Hearing (Revised Edition)*, MIT Press, 1997.

[2] Sylvia Schulz and Thorsten Herfet, "Binaural Source Separation in Non-Ideal Reverberant Environments," in *Proceedings of 10th International Conference on Digital Audio Effects (DAFx-07), Bordeaux, France*, September 2007.

[3] Sylvia Schulz and Thorsten Herfet, "Humanoid Separation of Speech Sources In Reverberant Environments," in *Proceedings of 3rd International Symposium on Communications, Control and Signal Processing (ISCCSP 2008), St. Julians, Malta*, March 2008.

[4] Harald Viste, *Binaural Localization and Separation Techniques*, Ph.D. thesis, Swiss Federal Institute of Technology, Lausanne, June 2004.

[5] D. L. Wang and Guy J. Brown, *Computational Auditory Scene Analysis - Principles, Algorithms, Applications*, IEEE Press, Wiley Interscience, 2006.

[6] R. B. Masterton and T. J. Imig, "Neural mechanisms for sound localization," *Annual Review of Physiology*, vol. 46, pp. 275–287, 1984.

[7] R. S. Woodworth and H. Schlosberg, *Experimental Psychology*, Holt, New York, 1954.

[8] Harald Viste and Gianpaolo Evangelista, "On the Use of Spatial Cues to Improve Binaural Source Separation," in *Proceedings of 6th International Conference on Digital Audio Effects (DAFx-03), London, UK*, September 2003.

[9] E. M. Hornbostel and M. Wertheimer, "Über die Wahrnehmung der Schallrichtung," 1920.

[10] J. W. Strutt (Lord Rayleigh), "On our perception of sound direction," *Philosophical Magazine*, vol. 13, pp. 214 – 232, 1907.

[11] R. O. Duda and W. L. Martens, "Range dependence of the response of a spherical head model," *Journal of the Acoustical Society of America*, vol. 104, no. 5, pp. 3048 – 2058, 1998.

[12] John Kominek and Alan W Black, "CMU ARCTIC databases for speech synthesis," 2003.

[13] Jens Blauert, *Untersuchungen zum Richtungshören in der Medianebene bei fixiertem Kopf*, Ph.D. thesis, Rheinisch-Westfälische Technische Hochschule Aachen, 1969.