TRANSAURAL 3-D AUDIO WITH USER-CONTROLLED CALIBRATION

Adrian Jost, Jean-Marc Jot

Creative Advanced Technology Center. 1600 Green Hills Rd. Scotts Valley, CA 95067, USA
ajost@atc.creative.com, jmj@atc.creative.com

ABSTRACT

A calibration method allowing users to customize the loudspeaker layout for 2-, 4-, and 5.1-channel playback, and to steer the "sweet spot" to the position of the listener's head is presented. The method, which is applied to a computationally efficient transaural 3D audio system for dynamic spatialization of multiple sound sources, is based on user interaction and auditory feedback. The robustness of the auditory sensation is analyzed for small displacements of the listener near the sweet spot. A modification of the system permits continuous adjustment of the sweet spot size by the listener. The modification limits the artifacts due to the transaural processing for positions away from the sweet spot. For wide settings, the system gradually reduces to a discrete amplitude panning system.

1. INTRODUCTION

The earliest known transaural reproduction system was described by Atal and Schroeder in 1963 [1]. The system, which equalized for the loudspeaker-to-ear transfer functions and cancelled crosstalk signals, provided a foundation for transaural synthesis. Computationally efficient crosstalk canceller topologies were proposed later by Iwahara and Mori [2] or Cooper and Bauck [3].

A common feature of all transaural playback systems is that they are optimized for a given reproduction layout, and their performance is sensitive to displacements of the listener's head or of the loudspeakers. Desktop computer or home theater users may find the prescribed speaker and "sweet spot" placement impractical due to space limitation and will often disregard the recommended layout if no flexibility is offered. Even small displacements can cause phase reversals of the cancellation signals at certain frequencies, thus rendering the crosstalk cancellation ineffective or even creating audible artifacts. Recently, Gardner completed an in-depth theoretical and experimental study of transaural reproduction and designed an adaptive transaural system capable of steering the sweet spot to the position of the listener's head under control of a position sensor [4].

In this paper, we describe a 3D positional audio system and a calibration utility that allow users to customize the loudspeaker location for 2-, 4-, and 5.1-channel reproduction. The directional and transaural processing filter parameters are adjusted by means of listening tests performed by the user. The calibration procedure also lets users continuously adjust the "size" of the sweet spot in order to provide extended freedom of movement, or accommodate a larger audience. This adjustment essentially controls a trade-off between reduced artifacts over a wider listening area or more accurate positional reproduction at the "center" of the sweet spot. For wider settings, the system gradually reduces to a discrete amplitude panning system as described e.g. in [5]. The investigations and developments described in this paper have application in the personal computer audio industry, as well as any system rendering 3D positional audio over standalone loudspeakers, wearable loudspeakers, or loudspeaker chairs.

2. AN EFFICIENT SYSTEM FOR 3D AUDIO OVER LOUDSPEAKERS

Figure 1 describes a transaural spatialization system for desktop 3D audio and home theater, designed for computationally efficient dynamic spatialization of multiple sound sources. It combines three stages:

a) The encoding stage is specific to each sound source and uses a conventional discrete amplitude panning method designed for 6 loudspeakers surrounding the listener in the horizontal plane (front left and right, side left and right, rear left and right – or 6 directions respectively numbered 1, 2, 6, 3, 5, 4).
b) The decoder (or binaural synthesis stage) receives the summed outputs of all encoders and produces a two-channel submix via 6 pairs of HRTF filters (L_i, R_i), i = 1..6. Each pair of filters reproduces a "virtual speaker" (VSP) [6].
c) The decoded signal passes through a transaural cross-talk canceller (TACC) using the asymmetric variant of the Iwahara-Mori topology, as proposed by Gardner [4].

The TACC topology shown in Figure 1 cancels the crosstalk from each loudspeaker to the contralateral ear, but does not correct for the ipsilateral transfer functions. Consequently, the VSPs in the binaural synthesis stage use free-field equalized HRTFs:

$$L_i(z) = \underline{L}_{i/1}(z)\, z^{-m_{L,i/1}} \text{ and } R_i(z) = \underline{R}_{i/2}(z)\, z^{-m_{R,i/2}},$$

where $\underline{L}_{i/j}(z)$ denotes the minimum-phase transfer function derived from the ratio of the left-ear HRTF magnitude frequency spectra for direction i and j, while $m_{L,i/j}$ denotes a delay derived from the excess-phase difference of these same HRTFs [7,4].

The two feedback branches of the TACC contain the interaural transfer functions (ITFs) for the front left and right directions, $L_2(z)$ and $R_1(z)$. As a result, it can be verified that the complete playback system, combining the three stages described above, verifies the "discreteness" condition: a sound panned to the direction of a loudspeaker in the encoder will feed only that loudspeaker at the output of the TACC.

This system can be extended to 4-speaker playback by duplicating the cross-talk cancellation stage (front TACC and rear TACC), connecting the front VSP outputs exclusively to the front TACC, the rear VSP outputs exclusively to the rear TACC, and the side VSP outputs, scaled down by 3 dB, to both the front and rear TACCs. This forms a 4-speaker transural playback system, which also verifies the conditions for discreteness.

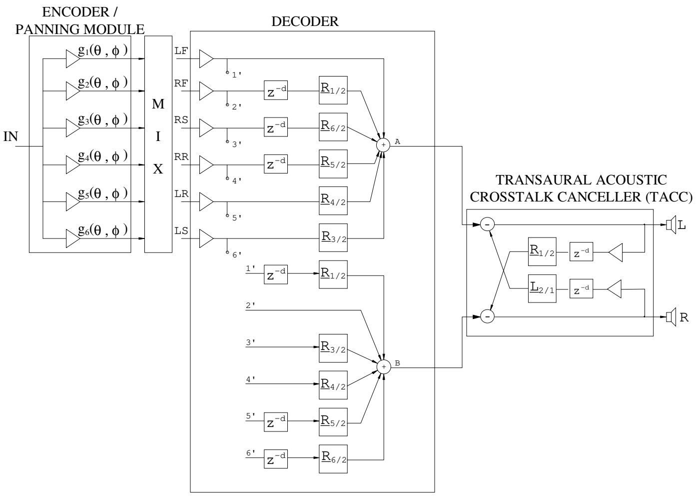

Figure 1. Spatializer signal flow graph

3. AUTOMATIC LOUDSPEAKER PLAYBACK SYSTEM CALIBRATION

As a general principle, the calibration of a loudspeaker playback system involves the equalization of the different loudspeakers in level and spectrum and their time alignment with respect to a reference listening position. This can be addressed automatically by measuring an acoustic transfer function from each loudspeaker to a microphone placed at the reference listening position (notional position of the center of the head). The acoustic level, spectrum, time delays, and phase inversion (e.g. wiring inverted) can be determined and equalized using a single microphone.

However, in order to optimize playback with a transaural system, it is also necessary to calibrate for the angular position of the loudspeakers and for head-related parameters (primarily the listener's head size). For determining the angular location of each loudspeaker relative to the listener, two pressure capture points, in or near the ear canals of the listener, are needed. A head model can then be used to map interaural delays to the direction of incidence. Simultaneously, an inversion of the channels with respect to the median plane can be corrected. To correct a channel inversion relative to the frontal plane, three microphones are needed, unless the listener rotates by 90 degrees to repeat the calibration. If for some reason the listener is not satisfied with the automatically calibrated settings, additional control on the system must be provided, in which case a manual calibration is still needed. In this paper, we present an alternative approach to automatic calibration which we will call "user-controlled calibration".

4. USER-CONTROLLED CALIBRATION FOR TRANSAURAL AUDIO PLAYBACK

The purpose of user-controlled calibration is twofold. First, it aims at optimizing the binaural and transaural synthesis for a given loudspeaker layout. Second, it attempts to determine the physical parameters (distances and angles) forming that layout. These physical parameters which describe the signal processing parameters in the transaural crosstalk canceller (TACC) and in the binaural synthesis, are adjusted indirectly by the user through simple listening tests and continuous adjustment with real-time auditory feedback. Additional hardware (e.g. microphones) is not necessary.

4.1. Calibration procedure

In this section, we review the successive steps of the calibration procedure, and detail how the physical parameters and the algorithm parameters are corrected.

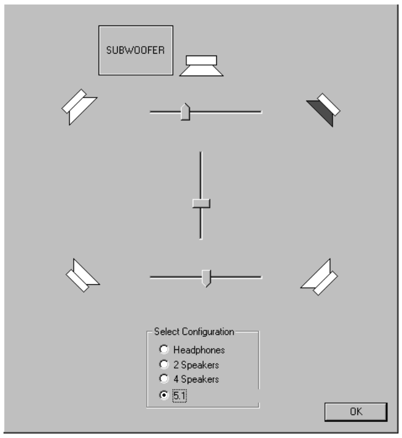

Figure 2. User Interface snapshot

1. <u>Speaker assignment</u>: We first verify that each loudspeaker is connected to the appropriate channel. Misconnections can be corrected in software so that the user does not need to change the wiring. Figure 2 shows a possible user interface. A sound is sent to each channel at a time, and the listener clicks on the icon representing the active loudspeaker.

2. <u>Volume balance</u>: Test signals A and B are two similar but uncorrelated sounds band-limited to medium frequencies with identical spectrum and level. The test sounds are panned into a pair of speakers (i.e. A is fed to front left, B to front right). They start playing when the user grabs a balance slider located between two speaker icons on the user interface. The balance controls the relative gain (frequency independent) between the two speakers.

3. <u>Polarity correction</u>: This step ensures that the loudspeaker wiring polarity is correct. This is critical for transaural crosstalk cancellation to operate. The test signal A = B contains mostly low and medium frequencies. When the user clicks on a loudspeaker icon, the phase of that channel is inverted and the polarity displayed. The correct setting yields a louder sound image (especially at low frequencies) that is more sharply focused.

4. Delay alignment: Sound A = B is a sharp transient sound (e.g. finger snap). The balance control adjusts the relative delay between the channels. The listener should hear a single, sharp sound when the correct setting is reached.

5. Fine volume balance: The test signal A = B contains mostly medium frequencies. The balance controls the same gain as in step 2, but a finer tuning can be achieved by adjusting for the apparent direction of the monaural sound to appear in front, rather than comparing the loudness of two uncorrelated sounds.

6. Low-frequency volume balance: The test signal A = B contains mostly low frequencies. The balance controls the low-frequency gains of the respective channels and keeps the sum of the *amplitude* gains constant at low frequencies ($A_{LF}+B_{LF}$ = constant).

7. High-frequency volume balance: The test signal A = B contains mostly high frequencies. The balance controls the high-frequency gains of the respective channels and keeps the sum of the *power* gains (squared amplitude gains) constant at high frequencies. ($A_{HF}^2+B_{HF}^2$ = constant).

8. Speaker angular position and head size: This step, described in section 4.2, aims at optimizing the sharpness of lateral sound images. It involves the adjustment of the binaural and transaural synthesis delays, which are then mapped to the loudspeaker angular position and head size.

In a 4-speaker configuration, the rear speakers are also shown on the user interface with two additional sliders. One slider is used to adjust time alignment, level and spectrum between the rear pair of speakers whilst the front speakers are muted. Next, the second slider allows adjustment of the front-back balance and alignment with all four speakers active. In a 5.1 configuration, step 3 is repeated with the subwoofer connected. At step 5, the test signal C = 1/sqrt(2)*(A+B) is fed to the center speaker. The center channel level is adjusted until it has the same perceived loudness as A and B together.

4.2. Crosstalk canceller calibration

In this section, we detail the calibration of step 8 (section 4). We adjust the delays in the binaural synthesis stage, i.e. the ITD for each VSP, and the delays in the TACC branches. We map these delays to the loudspeaker angular position and head size using Woodworth's formula (Equation 1), where a is the head radius [m], c the sound velocity [m/s], θ the incidence angle [rad].

$$ITD = \frac{a}{c}\left[\theta + \sin(\theta)\right] \qquad (1)$$

Is it accurate to use Woodworth's ITD model, which is frequency independent, as the mapping function? The answer depends on how ITD is extracted from HRTF data. In the binaural synthesis stage, we separate HRTFs into minimum-phase and non-minimum-phase components using the Hilbert transform [7]. The ITD extracted by factoring out the excess phase corresponds to the high-frequency asymptotic value of ITD, which is well approximated by the Woodworth formula [8]. The expected increase of group delay at low frequencies remains in the minimum-phase component of the HRTF.

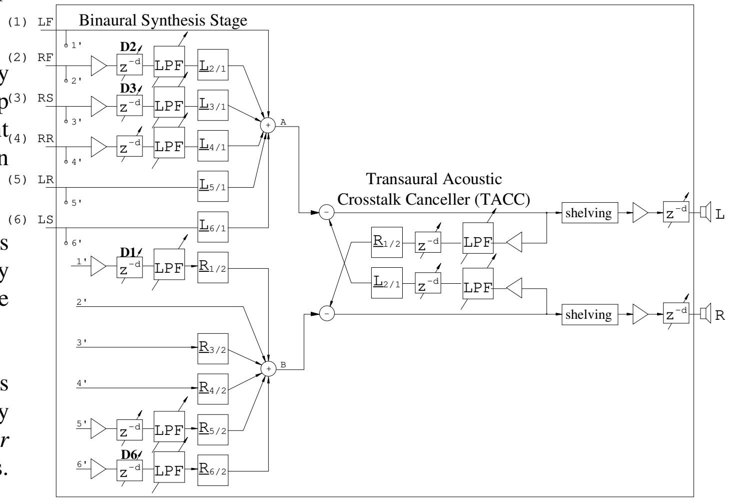

Figure 3. Two-speaker spatializer signal flow graph with calibration elements. The front speaker delay lines feeding the contra-lateral HRTFs are D1 and D2. They correspond to *ITDfront* and are equal to the delay lines in the TACC branches. The side VSP delay lines corresponding to *ITDside* are D3 and D6.

$$ITDfront_{L,R} = \frac{a}{c}\left[\theta front_{L,R} + \sin(\theta front_{L,R})\right] \qquad (2)$$

$$ITDside = \frac{a}{c}\left(1 + \pi/2\right) \qquad (3)$$

An asymmetric loudspeaker layout requires three listening tests to determine head size and angular position of two front loudspeakers relative to the listener. First, a monaural test signal is panned to a side VSP (line 3 or 6 on Figure 3) while the crosstalk canceller delays and the front speaker delays are adjusted. The optimal value of ITD for the front speaker location is obtained when crosstalk is minimized. This translates into a maximal lateral impression. Another technique to identify the optimum ITD based on real-time acoustic feedback is to impose zero on one binaural channel, and to adjust the delays for best cancellation at one ear. This test signal produces a somewhat unpleasant, unnatural effect. Second, the test signal is panned to the opposite VSP (line 6 or 3 on Figure 3) and the same experiment is repeated to determine the delay lines for that side. Finally, with the test signal panned to the side (3) or (6), *ITDside* is adjusted by changing the binaural synthesis delays only. The ideal head size value is the smallest value that positions the sound image to the side. Although larger values of head size preserve the lateral image, exaggerating the head size parameter distorts the 3D space, causing circular panning around the head to be irregular (slower on sides, faster in front and rear). We noticed that the sequence of calibration is also relevant. Adapting the TACC delays first allows a more accurate head size adjustment. Once the head size is adjusted, the ITD for other VSP (e.g. rear VSP in a 2- speaker configuration) can be computed. If a rear crosstalk canceller is used in 4- and 5.1-speaker configurations, the first two adjustments are repeated, with the test signals feeding the rear TACC only.

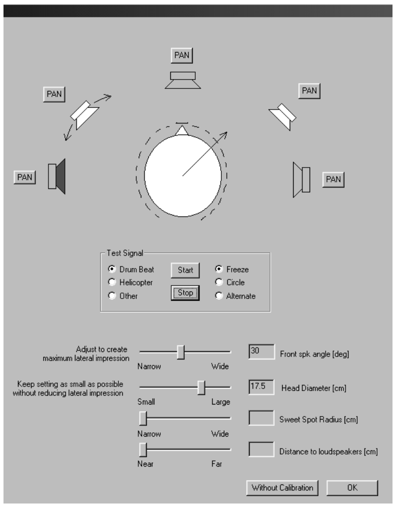

Figure 4. User Interface snapshot (step 8)

ITDside and *ITDfront* are not coupled and can be separately adjusted. However, a user-interface may be more attractive if it displays physical parameters instead of signal processing parameters. Head size and angular position are coupled since a is also used to compute the speaker angle θ. While adjusting a, the new value of θ satisfying equation 2 can be calculated using the Newton optimization method, applied to a single-variable non-linear equation [9]. The Newton method was chosen for its quadratic convergence. Its performance is robust in this case because the starting point (initial guess for θ) is not too far from the solution.

Knowledge of the loudspeaker locations also allows us to re-compute the panning laws, optimizing these for the new speaker locations. For greater accuracy, the inter-aural HRTFs $\underline{R}_{1/2}$ and $\underline{L}_{2/1}$ can be replaced, and all VSP free-field equalized to the direction of the loudspeakers.

5. SWEET SPOT SIZE ADJUSTMENT AND ROBUSTNESS

At this point of the calibration procedure, the sweet spot is steered to the listener's position. In this section, we first analyze the robustness of the perceived sensation as the listener moves out of the sweet spot. We then suggest a technique to vary the size of the sweet spot continuously, from very narrow with optimal transaural audio, to a wide audience area and elimination of all transaural processing. Our results rely on numerical simulation, using a database of HRTFs provided by the University of California, Davis [10].

5.1. Simulation model

Figure 5 shows the simulation model. The main elements, from left to right are (1) the binaural input signal (u_L,u_R), (2) the transaural acoustic crosstalk canceller with matrix transfer function **T**, (3) the acoustic transfer matrix **A**, (4) the free-field equalization **F**. Ideal cancellation of crosstalk is achieved when the product **T**·**A**·**F** is equal to the identity matrix. This is not possible in practice due to inter-individual differences in HRTFs, imperfect listener placement, and to a lesser extent, reverberant listening environment. The simulation takes into account these elements with the exception of the reverberation of the listening space. The TACC filters are designed from the HRTF data of one individual and remain static as the listener moves. The acoustic crosstalk transfer functions, L_1, R_1, L_2, and R_2, which form the matrix **A**, correspond to a different individual and are updated according to listener position. The lateral displacement dx and front-back displacement dz of the listener are mapped to the angular position of each loudspeaker, which is used to retrieve the two nearest HRTFs in the database. We interpolate the two minimum-phase magnitude spectra linearly on the dB scale. The interpolated excess phase, or ITD, is represented by an integer delay and a fractional delay d approximated by a first-order all-pass filter $FD(z) = (1 + a\, z^{-1})/(a + z^{-1})$, $a = (1-d)/(1+d)$.

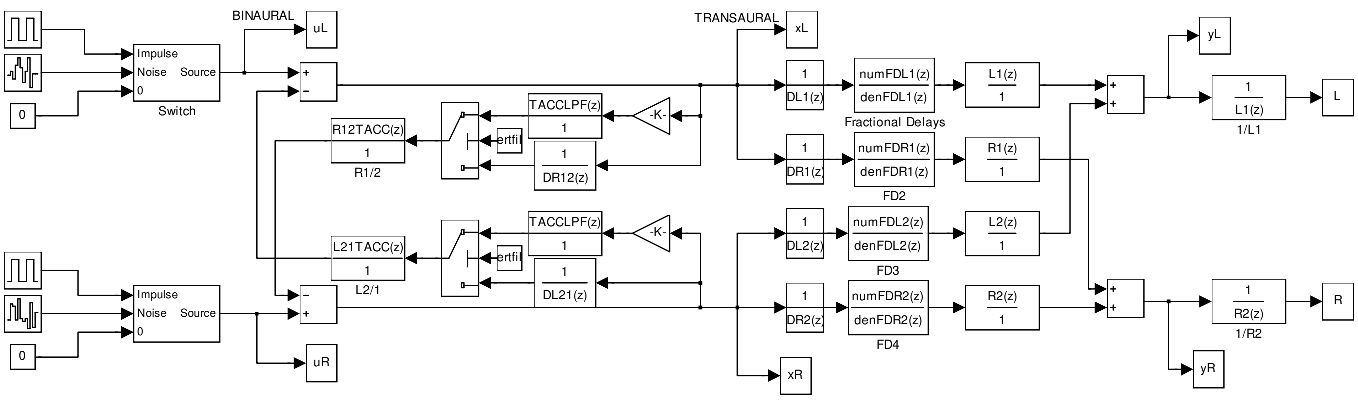

Figure 5. Matlab Simulink numerical model for acoustic crosstalk cancellation.

The following analysis of the sweet spot robustness compares the product **T·A·F** against the identity matrix. For this purpose, we analyze the responses at both ears (output signals *L* and *R*) when imposing an impulse on one binaural input channel (u_L or u_R) and zero on the other. We refer to this experiment as "channel separation". As illustrated in Figure 6, we observe, for various displacements (*dx*, *dz*) of the listener, the ipsilateral response (which should be a perfect pulse to ensure perfect timbre transparency), and the contralateral response (which demonstrates the amount of undesired crosstalk vs. frequency).

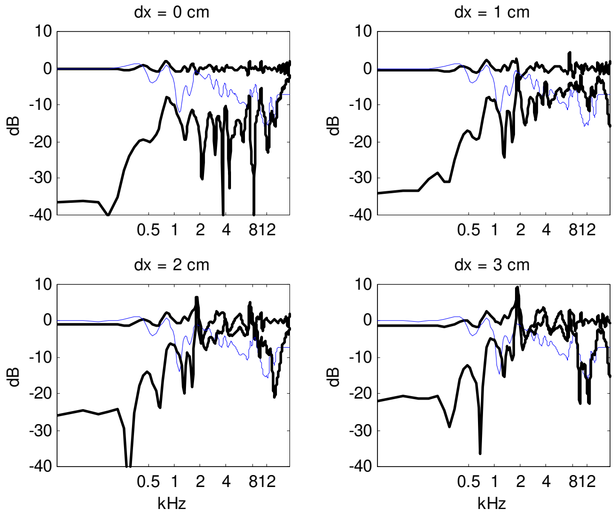

Figure 6. Channel separation for three lateral displacements off the sweet spot center. Thick lines: ipsilateral and contralateral frequency responses. Thin line: natural head shadowing.

At the center of the sweet spot (*dx*=*dz*=0), crosstalk cancellation is effective up to circa 6 kHz where inter-subject differences are minor. As shown on Figure 6, it is best to let the natural baffling effect of the head attenuate crosstalk above 6 kHz rather than trying to cancel it and causing audible artifacts on both the ipsilateral and contralateral ear. In a static design, the ITFs $\underline{R}_{1/2}$ and $\underline{L}_{2/1}$ can simply be low-pass filtered at 6 kHz, as proposed in [4]. Unless a tracking device is utilized to adapt the TACC, movements of the listener cause phase reversals of the cancellation signals at even lower frequencies. The frequency up to which cancellation is efficient rapidly drops to 1500 Hz as the listener moves out of the sweet spot.

Figure 7A shows the energy at the contralateral ear for one octave centered at 4 kHz as a function of listener displacement. For lateral displacements, the energy variation is significant. The intensity oscillates between –25 and –3dB. The acoustic sensation experienced by a listener is a succession of destructive and constructive interference, which is very unpleasant and tiring.

5.2. A spatializer with adjustable sweet-spot size

The simulation results of Figure 6 and Figure 7 show that, for frequencies higher than about 1500 Hz, the cross-talk cancellation network actually increases the amount of undesired crosstalk significantly and causes timbre distortions, with significant fluctuations across positions of the listener. Although the presence of the TACC improves the reproduction for low frequencies, it is detrimental at medium and high frequencies.

In order to address this issue, we propose a modified implementation of the spatializer of Figure 1 as follows:

a) The TACC is band-limited as described in [4], by inserting a linear-phase low-pass filter in cascade with the ITFs $L_2(z)$ and $R_1(z)$. Furthermore, in order to maintain the discreteness property (see section 2), the same low-pass filter is applied to the ITFs $L_2(z)$ and $R_1(z)$ in the binaural synthesis stage.

b) This low-pass filter is continuously variable via a "sweet-spot size" control parameter. This parameter has the effect of reducing the cutoff frequency *fc* of the low-pass filter from 6 kHz to 1500 Hz for increasing size. Due to its limited order, the filter cannot be used to fully remove the cancellation branches (i.e. *fc* cannot be reduced to 0 Hz). Hence, the filter also incorporates a variable frequency-independent attenuation *g* to progressively remove the crosstalk canceller.

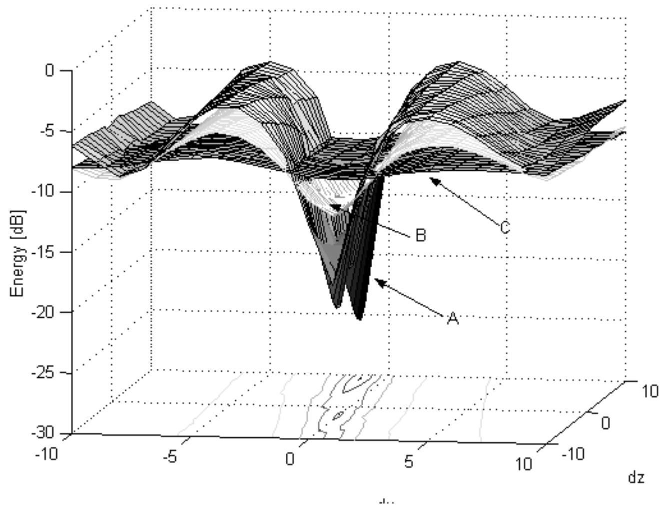

Figure 7. Energy at contralateral ear, 1 octave centered at 4kHz. Displacement expressed in centimeters.

As shown on Figure 8, inserting the filter removes the crosstalk canceller artifacts above *fc* at non-centered positions, while reducing cancellation at the center of the sweet spot. The ipsilateral ear receives a less colored sound, and the contralateral response above *fc* reduces to the natural head shadow. The energy variation (Figure 7, plot B and C) flattens as the cutoff is lowered. In other words, for a given variation of energy, the sweet spot widens. The compromise is a higher energy level at the center of the sweet spot. Channel separation plots on a wide frequency range (Figure 9) present a more complete picture of the sweet spot robustness. The two plots (*fc* = 6 kHz and *fc* = 1500 Hz) intersect at about 1.5 cm lateral displacement. For smaller displacements, the 6 kHz cutoff yields better separation, but offers a very sensitive sweet spot (steep slope). For displacements beyond this intersection, the lower cutoff yields a more robust sweet spot with better channel separation (e.g. 2.7dB or a 40% increase at dx=5cm).

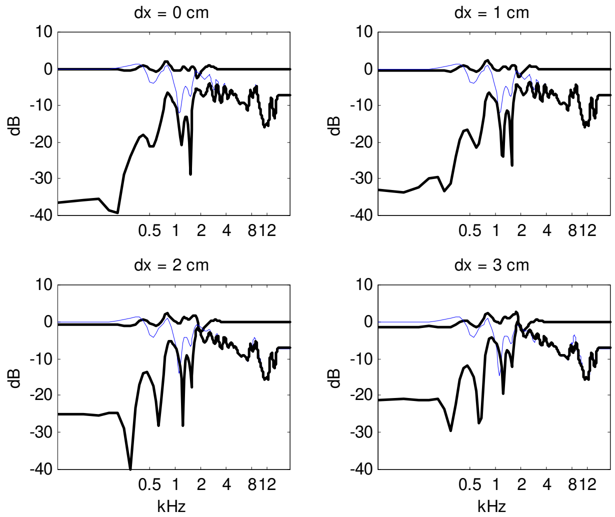

Figure 8. Channel separation. Thick lines: ipsilateral and contralateral frequency responses. Thin line: natural head shadowing. A low-pass filter with cutoff at 1500 Hz is inserted in the TACC branches.

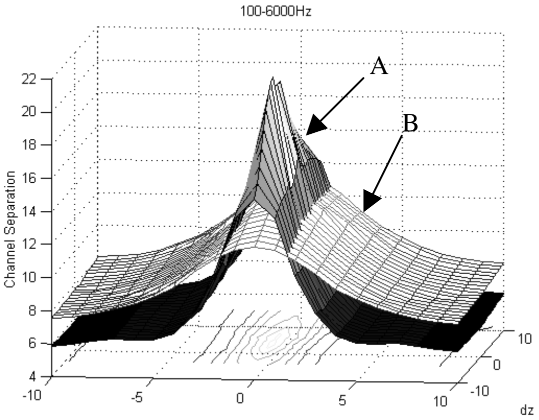

Figure 9. Average of the Left to Right and Right to Left channel separation, weighted by the Bark scale in the range 100-6000 Hz. Plot A: fc = 6 kHz. Plot B: fc = 1500 Hz.

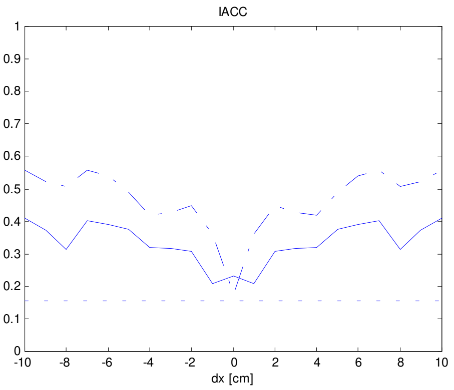

Figure 10. IACC as a function of lateral displacement. Dash-dotted: fc = 6 kHz. Continuous: fc = 1500 Hz. Dashed: IACC of the two binaural signals.

Running the simulation with two uncorrelated noise signals on the binaural channels allows plotting the inter-aural cross-correlation (IACC), a measure of similarity of the signals at the ears defiened as

$$IACC = \frac{\max\limits_{k=-1mss..1mss} |\varphi_{LR}(k)|}{\sqrt{\varphi_{LL}(0)\cdot\varphi_{RR}(0)}} \qquad (4)$$

where φ_{LR} and φ_{xx} are the cross- and auto-correlation functions. Figure 10 indicates that band-limiting the crosstalk cancellation to 1500 Hz allows better reproduction of the spatiousness of a reverberant sound field over a wider listening area.

6. CONCLUSIONS

A calibration procedure for transaural 3-D audio and associated signal processing techniques allowing users to choose a flexible loudspeaker location for 2, 4 and 5.1 channel reproduction is presented. The procedure leads to the knowledge of loudspeaker angular position, allowing optimizing panning laws, binaural, and transaural synthesis. Sweet spot robustness has been analyzed. Inserting a low-pass filter in the cross-talk cancellation loop widens the size of the sweet spot. Increasing its size compromises the sharpness of lateral sound images, but improves their robustness and reduces crosstalk cancellation artifacts. A technique to vary the sweet spot size continuously from very narrow (desktop computer applications) to a wide audience area (home theater situation) is suggested, where the crosstalk canceller and binaural processing are progressively removed.

7. REFERENCES

[1] Schroeder M.R., Atal B.S., 1963 Computer Simulation of Sound Transmission in Rooms. IEEE Int. Conv. Record, 7.

[2] Iwahara, M., Mori T., 1978, Stereophonic sound reproduction system. US Patent 4,118,599.

[3] Cooper D.H.,Bauck J.L., Prospects for Transaural Recording, J. Audio Eng. Soc., Vol. 37, no. 1/2, 1989.

[4] Gardner W.G. 1997, 3-D Audio Using Loudspeakers. M.I.T Media Laboratory, Published by Kluwer Academic.

[5] Pulkki V. 1997, Virtual sound source positioning using vector base amplitude panning., J. Audio Eng. Soc. 25, 4.

[6] Jot J-M., Approaches to binaural synthesis, 105th Conv. of the Audio Eng. Soc. Sept 1998

[7] Jot J.-M., Larcher V., Warusfel O. 1995. Digital signal processing issues in the context of binaural and transaural stereophony. In Proc. 98th Conv. Audio Eng. Soc., (preprint 3980).

[8] Kuhn G. F. 1977. Model for the interaural time differences in the azimuthal plane, JASA Vol. 62 No. 1, July 1977

[9] Nocedal J., Wright S., 1999, Numerical Optimization, Springer Verlag ISBN:0-387-98793-2

[10] Algazi V.R., Avendano C., Thompson D. Subject and Measurement Position Dependence in Binaural Signal Acquisition, J. Audio Eng. Soc., Nov 1999.

[11] Damaske P.,Ando Y., Interaural cross-correlation for multichannel reproduction, Acustica 27: 232-238,1972