SPATIAL TRACK TRANSITION EFFECTS FOR HEADPHONE LISTENING

Aki Härmä and Steven van de Par

Philips Research, Eindhoven, The Netherlands
aki.harma@philips.com

ABSTRACT

In this paper we study the use of different spatial processing techniques to create audio effects for forced transitions between music tracks in headphone listening. The audio effect encompasses a movement of the initially playing track to the side of the listener while the next track to be played moves into a central position simultaneously. We compare seven different methods for creating this effect in a listening test where the task of the user is to characterize the span of the spatial movement of audio play list items around the listener's head. The methods used range from amplitude panning up to full Head Related Transfer Function (HRTF) rendering. It is found that a computationally efficient method using time-varying interaural time differences is equally effective in creating a large spatial span as the full HRTF rendering method.

1. INTRODUCTION

What users commonly do when listening to an audio play list or CD is to jump from one item to another item by pressing the 'Next', or 'Previous' button of the player. This may be performed anywhere between the start and the end of an item and it is implemented in basically all audio players is that the current item is muted and the new track starts playing.

In this paper we study a class of spatial transition effects for headphone listening. The goal is to produce the impression that one track goes physically away and another track comes in. For example, the current music track moves far away to the right and another track slides in from the left hand side. This type of effect has earlier been proposed for surround audio playback with loudspeakers [1] but, to our knowledge, not for headphone listening.

The approach is to position the audio source into a simulated loudspeaker-listener scenario where the virtual loudspeaker, and the listener's ears have well-defined geometric positions. Once this is done, we can move the virtual loudspeaker to arbitrary positions resulting in a perceived movement of the audio sources. In swapping from one audio item to another, the simulation can be performed such that a virtual loudspeaker playing Item 1 is moved far to the left from the user's ears and another loudspeaker playing Item 2 is carried in from the right to the desired playback position. For simplicity, we consider only monophonic audio material in this paper, but the same approach can be used also for stereo or multichannel material by creating multiple virtual loudspeakers.

These effects can be created combining many different methods such as amplitude, or phase panning, HRTF filtering, or room simulation. In the current paper we introduce seven different combinations of algorithms for spatial track transition which differ in computational complexity. Using a new type of listening test, where the subject indicates the movement trajectories of the audio items, we evaluate the effectiveness of each of the methods in creating a large span of perceived auditory movement.

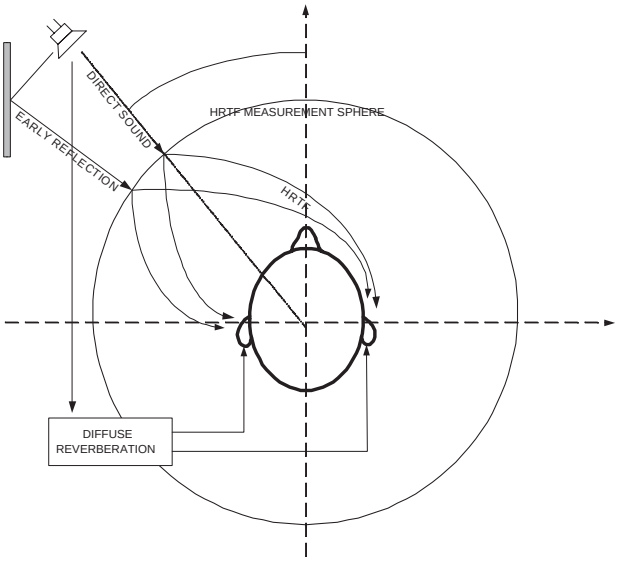

Figure 1: A simulation for a loudspeaker-listener system.

2. THE ACOUSTIC MODEL

The generic model is based on the source-medium-receiver model of binaural simulation [2]. Here, the source is represented by a virtual loudspeaker, the medium is a model of room acoustics, and the receiver is a pair of virtual microphones representing the listener's ears in the room. The model for the medium contains the sound propagation in the room, and the receiver model takes into account the orientation of the listeners head and the Head Related Transfer Functions (HRTFs). This signal processing model is illustrated in Fig. 1 and is similar to those presented by many authors earlier for binaural or transaural listening, see e.g., [3, 2, 4].

The directionality of the source has not been incorporated into the current model, but we assume that the source is an ideal omnidirectional loudspeaker.

In reality, the head-related transfer functions (HRTF) represent impulse responses measured from a limited number of source positions on a sphere with the center position in the center of the listener's head. For example, in the CIPIC data used in the current paper, the radius of the measurement sphere (a hoop where the speakers were fixed) was one meter [5]. The HRTF measurement sphere is shown in Fig. 1. Consequently, the model for a direct sound from a source is a cascade of a direct path filter from the source location to the surface of that sphere, and a HRTF filter from the sphere to the listener's ears. In the room model a limited

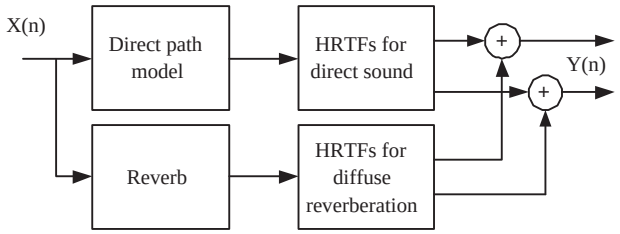

Figure 2: A simulation for a loudspeaker-listener system.

number of early reflections from the room surfaces are often added and they are convolved using HRTFs representing the angles of arrival for each individual reflection, see, e.g, [6]. Finally, the diffuse reverberation is modelled using a filter representing the reverberation and a pair of HRTFs representing the diffuse-field responses to the two ears, that is, means of HRTFs in the horizontal plane.

In this paper, we consider a simplified model consisting only of the direct path and the diffuse reverberation path. The block diagram is illustrated in Fig. 2.

The simulation of a moving sound source can be directly implemented using the system of Fig. 2. For example, the Doppler effect results automatically from changing the delay of the direct path propagation filter as a function of the simulated location of the source. In models for moving sources where the propagation delay has not been implemented, the Doppler effect is sometimes implemented as a separate computational operation, such as frequency modulation in [7] or pitch shifting, to allow better control over the effect.

A signal $x(t)$ played from a virtual loudspeaker is captured using a virtual microphone on the HRTF sphere. The direct sound signal before HRTF filtering is then given by:

$$y_0(t) = x(t,d) * \frac{\delta(t-T)}{d}, \tag{1}$$

where the asterisk denotes convolution, δ is the Dirac's function, $T = d/c$, where c is the speed of sound, d is the distance between the source and its nearest point on the HRTF surface.

For notational convenience we move to the frequency-domain representation of (1):

$$Y_0(\omega) = X(\omega,d)F(\omega,d), \tag{2}$$

where the capital letters denote the Fourier transforms of the parts of (1) and $F(\omega,d) = e^{-i\omega T}d^{-1}$.

Combining all paths from Fig. 2 we may write the synthesis formula (1) in the following form:

$$\begin{bmatrix} Y_l(\omega) \\ Y_r(\omega) \end{bmatrix} = X(\omega,d)\left(F(\omega,d)\begin{vmatrix} H_l(\omega,\alpha) \\ H_r(\omega,\alpha) \end{vmatrix} + \begin{vmatrix} R_l(\omega) \\ R_r(\omega) \end{vmatrix}\right), \tag{3}$$

where $R_l(\omega)$ and $R_r(\omega)$ are the HRTFs to the left and right ear of the listener, respectively, and which depend only on the angle of arrival of the sound α. The model of the reverberation has been integrated with the diffuse field HRTFs into filters $R_l(\omega)$ and $R_r(\omega)$, which are then independent of the source position. In a compact matrix notation we may write this in the following form:

$$\mathbf{Y}(\omega) = X(\omega,d)\mathbf{G}(\omega,\alpha). \tag{4}$$

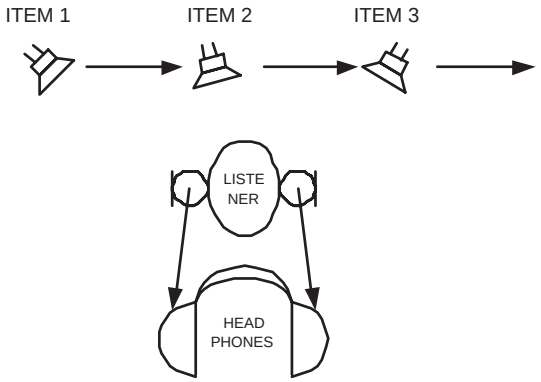

Figure 3: A simulation of a spatial track transition.

3. TRACK TRANSITION EFFECTS

The basic track transition effect studied in the current paper is illustrated in Fig. 3, where virtual loudspeakers representing different audio items that flow past the user. The simulation produces typical spatial audio cues and additionally, due to the direct path delay operator d, a Doppler effect, which is expected to contribute to the perceived illusion of the movement of a source.

4. EXPERIMENTAL SETTINGS

The model of Fig. 2 is computationally expensive mostly due to the HRTF filtering. In a dynamic transition effect, the filter coefficients need to be continuously updated to follow the angle α of the sound source. In practice this requires continuous interpolation of the responses to reduce artifacts related to switching filters. The model of the reverberation is another expensive part because it typically requires implementation of a high-order FIR filter.

In this paper, we study seven different systems that differ in the degree to which simplifications have been made to the signal processing model. They are all studied in the configuration illustrated in Fig. 3, where the monophonic sources move along a line from the left to the right such that, when allowed by the method, sources pass the listener at a constant speed at $\alpha = 0^o$ at the distance of $r = 2$ meters from the listener. The block diagrams of the methods are shown in Fig. 4.

In the pure amplitude panning method (a) the gains for the two ear signals are given by:

$$\mathbf{G}_a(\omega) = \begin{vmatrix} g_n \\ g_0 g_n \end{vmatrix}, \tag{5}$$

where:

$$g_0 = 10^{\frac{14}{40\pi}\arctan(p(n)/r)} \text{ and } g_n = \frac{1}{\sqrt{1+g_0^2}}, \tag{6}$$

where $p(n)$ is position of the sound source as a function of the sampling number n. The equation approximates the listening test data on binaural lateralization of a source in dichotic listening with only level differences between the two ears [4].

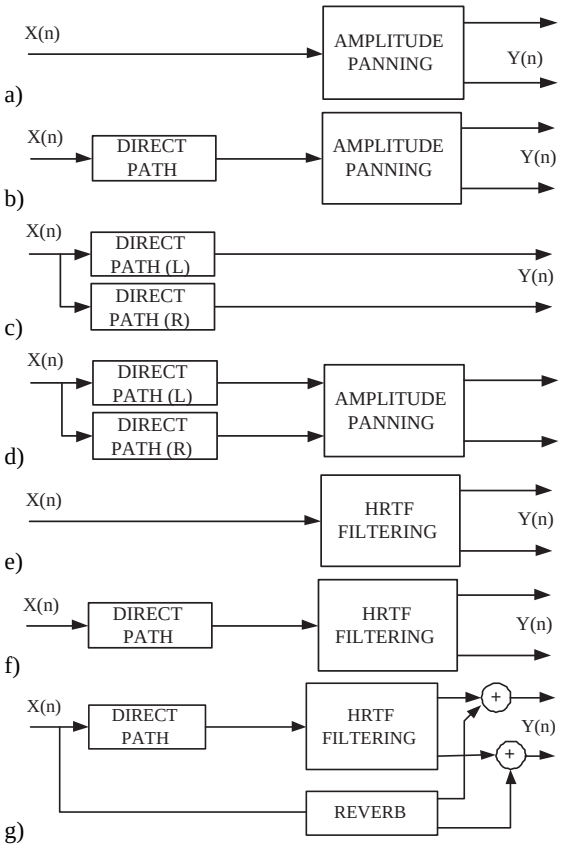

Figure 4: The models for spatial track transition studied in this paper.

The second model (b) combines the amplitude panning with a single direct path model. It can be written in the following form:

$$\mathbf{G}_b(\omega, n) = F(\omega, \sqrt{r^2 + p^2(n)}) \left| \begin{array}{c} g_n \\ g_0 g_n \end{array} \right|, \qquad (7)$$

where the only difference to (5) is the direct path model $F(\cdot)$, which in the case where $x(n)$ changes over time produces a Doppler effect. The fractional delays were implemented in the time domain using the sixth-order Lagrange FIR interpolator. The third model (c) is essentially a free-field model for a listener with an acoustically transparent head. The synthesis formula is given by:

$$\mathbf{G}_c(\omega, n) = \left| \begin{array}{c} F(\omega, \sqrt{r^2 + (p(n) + h/2)^2}) \\ F(\omega, \sqrt{r^2 + (p(n) - h/2)^2}) \end{array} \right|, \qquad (8)$$

where h is the distance between the two ears of the listener. In our simulations we used $h = 0.2$ m, which is somewhat larger than the human average.

The next model (d) is obtained by including a very simple model for the head shadowing to model (c). In fact, the head-shadowing model is exactly the same as the binaural amplitude panning used in (a)-(b), and it yields:

$$\mathbf{G}_d(\omega, n) = \left| \begin{array}{c} g_n F(\omega, \sqrt{r^2 + (p(n) + h/2)^2}) \\ g_0 g_n F(\omega, \sqrt{r^2 + (p(n) - h/2)^2}) \end{array} \right|. \qquad (9)$$

Note, that this model implements an approximative HRTF model, producing the same ITD and ILD cues at all frequencies with one delay and one gain coefficient per channel.

In model (e) we use the HRTF data from the CIPIC database (subject 31) [5]. The azimuthal set of HRTFs at the frontal area was augmented by a number of interpolated HRTF impulse responses. The interpolation was performed linearly in the frequency domain separately for magnitude and unwrapped phase responses. The synthesis equation is given by:

$$\mathbf{G}_e(\omega, n) = \left| \begin{array}{c} H_l(\omega, \mathrm{atan}(p(n)/r)) \\ H_r(\omega, \mathrm{atan}(p(n)/r)) \end{array} \right|, \qquad (10)$$

and the convolutions were implemented efficiently using the FFT overlap-add techniques.

The next model (f) incorporates the direct path model to the HRTF model:

$$\mathbf{G}_f(\omega, n) = F(\omega, \sqrt{p^2(n) + r^2} - r_{\mathrm{hrtf}}) \mathbf{G}_e(\omega, n), \qquad (11)$$

where $r_{\mathrm{hrtf}} = 2$ m is the radius of the HRTF measurement sphere of Fig. 2.

Finally, model (g) includes a model of the diffuse reverberation:

$$\mathbf{G}_g(\omega, n) = \mathbf{G}_f(\omega, n) + \left| \begin{array}{c} R_l(\omega) \\ R_r(\omega) \end{array} \right|, \qquad (12)$$

where the filters $R_l(\omega)$ and $R_r(\omega)$ are synthetic pink noise sequences with the temporal envelope from a real room impulse response with the reverberation time of $T_{60} = 1.0$ s.

5. LISTENING TEST

The purpose of the algorithms introduced above is to provide a spatial experience of a movement of an audio source in headphone listening. Generally, the hearing mechanism does not seem to be sensitive to the movement itself [8]. It is often suggested that the percept of the movement is a consequence of observing the source first at one position, and then at another position. However, there is evidence on brain areas that are actually sensitive to the movement of sound sources [9].

The just noticeable difference for the velocity of a source is typically in the range of 4-9 degrees per second [9] or 1.5 to 4.6 m/s [10] for a source moving a linear trajectory 5 meters in front of the listener. In most studied on just-noticeable differences (jnds) of velocity perception [11, 10] it has been found that the most important cues for the velocity are the Doppler effect and the changes in the overall loudness. The binaural cues including interaural time (ILD) and level differences (ILD) are weaker cues in the velocity discrimination. However, in another experiment where the listeners' task was to indicate the point where a moving source is closest to the listener suggested overall loudness to be the most important cue, followed by dynamic ITD cue, and only then the Doppler effect [12]. In the current article, the primary goal is to create an illusion of a large movement with a low-complexity algorithm, therefore the velocity or the temporal position of a source are not necessarily as important as the perceived distance the source has travelled, that is, the range of the movement.

In this paper, we developed a listening test that aims at depicting the subjective experience of a movement of a source in the case where the user is *scanning* over a sequence of three consecutive samples in a playlist of audio items. A similar movement pattern where the sound source moves along a linear horizontal trajectory

1.5 meters in front of the listener was used in all methods. The movement pattern was such that each new source appeared at the distance of 20 meters to the right of the listener moving to the front of the listener in two seconds, stopping at the front for one second, and then moving in two seconds 20 meters to the left. In methods where the rendering depends only on the angle of the source, such as amplitude panning and pure HRTF rendering, only the angle derived from the position of the source was used. Since the overall loudness has been found to be a dominating cue in listening experiments with moving sources [12, 11, 9, 10] an identical amplitude weighting as a function of the position was used with all the methods.

The test was performed in a sound insulated booth using Beyerdynamic DT990 headphones. Ten subjects participated in the experiment. The test material consisted of five playlists of three audio items each. Three of the playlists represented three-second excerpts from samples of different music genres (rock, pop, rap), one playlist consisted of uncorrelated pink noise sequences, and finally one playlist had rich harmonic tone complexes at three different fundamental frequencies. In the listening test, subjects were asked to listen to a sequence and then draw, using the computer mouse, their subjective impression of the path of the sequence of three audio items on a chart illustrated in Fig. 5X). We projected each drawing to a bitmap of 40×40 pixels for analysis.

6. RESULTS

To compare the span of path assessments in the different methods we computed pixel-wise 2D histograms over all listeners and play lists. The 2D histograms for the seven methods are shown in Figs. 5a-g. The figures show that there are differences between the subjective assessments of the transition paths in the methods discussed in this paper. In Method (a), the path is mainly judged inside the listener's head, while in other methods the span of the effect appears larger. The marginal histogram plotted at the bottom of each panel also suggests that the histograms are tilted towards the right, even if the movement path was symmetrical from the left to the right. This is an interesting finding.

It was found that there are significant differences between individual listeners. The differences are probably largely due to the differences in the ways how individual subjects mapped the auditory experience to a visual geometric form.

In order not to be influenced by these individual differences we decided to convert the results to a relative scale with a pairwise comparison of the individual path drawings for the different rendering methods. Table 1 gives the percentage for the probability that a rightmost point in a path in method X (row) is farther to the right than the corresponding point in the path for method Y (column) in one listener. Comparing methods (a) and (b) in Table 1, we see that the percentages are almost 50%, which means that the methods are essentially similar in the span to the right hand side. The percentages that the path spans farther to the right in methods c-g is 78-94% over the methods a-b. The method (c) gives a higher percentage over method a) than method (b) does. Both methods (b) and (c) contain the distance model. In method (b) the effect causes only the Doppler effect, while method (c) creates an interaural time difference which changes dynamically over the transition path.

It is interesting to note that a computationally light method (d) combining dynamic interaural time difference (with a transparent head model) and amplitude panning gets very similar rankings

A/B|	a	b	c	d	e	f	g
a	0	48	78	82	96	88	82
b	52	0	82	78	84	92	84
c	16	18	0	56	78	72	56
d	14	22	40	0	78	68	48
e	4	14	22	22	0	30	28
f	10	8	28	32	66	0	40
g	18	16	42	50	72	58	0

Table 1: Probability that path produced using transition A has a larger span to the *right* than transition B.

with the most complex method (g). The results suggest that the pure HRTF rendering (method e) gives systematically the largest spatial span to the right. In fact, methods (f) and (g), which add the dynamic distance model, and diffuse reverberation to the pure HRTF model get lower gradings than model (e).

Table 2 gives the percentages for the leftmost point in the path. The results support the observation that the span of the path in the amplitude panning models (a-b) is smaller than in the other methods with interaural time difference cues. However, the percentages are now lower. Comparing the method (a) to method (b), and (e) to (f) suggests that the use of the one-channel distance effect in the form of a Doppler effect in fact decreases the span of the path to the left.

A/B|	a	b	c	d	e	f	g
a	0	40	58	72	74	66	58
b	58	0	72	74	72	72	60
c	40	24	0	68	68	68	60
d	26	26	32	0	62	54	46
e	24	20	30	34	0	44	40
f	32	20	28	44	54	0	42
g	38	30	40	50	52	54	0

Table 2: Probability that path produced using transition A has a larger span to the *left* than transition B.

In the frontal area, see Table 3, it seems that the pure HRTF rendering again gives the largest span to the front. The low score of method (d) in the frontal area is an unexpected result because it gives a large range in right-left direction.

A/B|	a	b	c	d	e	f	g
a	0	60	62	46	66	62	60
b	32	0	56	44	60	44	70
c	34	38	0	38	54	40	58
d	38	52	56	0	60	54	64
e	28	38	36	28	0	42	48
f	34	46	42	34	52	0	60
g	26	26	36	30	40	38	0

Table 3: Probability that path produced using transition A has a larger span to the *front* than transition B.

The intended path passed the users face 1.5 meters at the front of the listener. However, several listeners papered a path behind

the head, too. The comparison in Table 4 suggests that the localization at the back of the head, or behind the head was strongest in amplitude panning methods (a-b), and somewhat increase also in method (d). For example, in 66% of the cases the path drawn for the method (d) span to the back more than the path for the same playlist in method (e).

A/B|	a	b	c	d	e	f	g
a	0	48	30	34	24	24	30
b	46	0	22	28	16	26	22
c	62	72	0	60	32	44	46
d	58	68	40	0	28	40	38
e	64	76	62	66	0	50	56
f	66	66	46	50	38	0	44
g	66	70	46	54	32	48	0

Table 4: Probability that path produced using transition A has a larger span to the *back* than transition B.

7. CONCLUSIONS

In this paper we have studied seven different techniques for the dynamic rendering of sound sources in spatial track transition. In particular, we have focused on a track transition effect where one song comes from the left hand side of the user and disappears to the right. The techniques represent different levels of computational complexity. The simplest techniques are based on dynamic amplitude panning, that is, multiplication of the signal with a scalar coefficient. The most complicated reference method combines HRTF filtering, the Doppler effect, and room reverberation.

The listening tests suggest that the amplitude panning techniques give generally a narrow range of the effect and the image is often inside the head. The plain HRTF processing gives the largest span in the lateral direction. However, a simple method combining delay panning and amplitude panning appears almost equally powerful for the creation of left-right transition effects. However, the HRTF method appears giving a slightly larger span in the front left direction and possibly better externalization.

The addition of room reverberation produced an interesting effect but the benefits are not obvious in the current results. It appears that for most listeners the method combining HRTF filtering and reverberation gave smaller left-right span that the pure HRTF filtering.

In the listening tests the goal was to compare the different methods by the perceived spatial span of the transition effect. This is a different task from the subjective evaluation of the velocity of a source [11, 9, 10] or the closest point in the movement trajectory [12]. In velocity discrimination studies the Doppler effect and the overall loudness have been found to be more important than the binaural cues such as ILD or ITD. The spatial delay panning methods aim at creating a distance cue. In all cases the transition effect was tuned in such a way that it created an audible Doppler effect during the transition of an audio playlist item from left to the right. The Doppler effect was audible in five out of the seven methods. If was found that when the Doppler effect appeared without associated binaural time difference cues, it had almost no influence on the perceived left-right span. In particular, the difference between amplitude panning with and without the Doppler effect was small but the difference between the amplitude panning with the Doppler effect, and the method where amplitude panning was combined with time-varying interaural time-differences was significant. The results seem to suggest that the interaural time-differences actually play a more important role in the perceived span of a transition than the Doppler effect.

From the results of the current listening test we may conclude that the dynamics of interaural time-differences are important in producing a large spatial span for the track transition effects. In addition, a very simple method based on a computationally very efficient simplified sound propagation model to the two ears of a listener gives almost equally good results in the span of the movement effect as a more complicated method based on the measured head-related transfer functions.

8. ACKNOWLEDGEMENTS

The authors are grateful to Armin Kohlrausch for fruitful discussions related to the design of the test setup, Bert den Brinker and Othmar Schimmel for help in improving the manuscript, and finally the listening test subjects for their fine illustrations of auditory motion patterns.

9. REFERENCES

[1] T. Herberger and T. Tost, "System and method for generating sound transitions in a surround environment." US Patent 2005/0047614A1, 2005.

[2] L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen, "Creating interactive virtual acoustic environments," *J. Audio Eng. Soc.*, vol. 47, no. 9, pp. 675–705, 1999.

[3] D. R. Begault, *3-D sound for virtual reality and multimedia*. New York, USA: Academic Press, 1994.

[4] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. Cambridge, MA, USA: The MIT Press, 1999.

[5] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proc. IEEE WASPAA'2001*, (New Paltz, NY, USA), October 2001.

[6] J.-M. Jot, "Scene description model and rendering engine for interactive virtual acoustics," in *AES 120th Conv. preprint 6660*, (Paris, France), May 2006.

[7] J. M. Chowning, "The simulation of moving sound sources," *J. Audio Eng. Soc.*, vol. 19, pp. 2–6, January 1971.

[8] D. W. Grantham, *Hearing (ed. B. J. C. Moore)*, ch. Spatial hearing and related phenomena, pp. 297–339. Academic Press, 1995.

[9] S. Carlile and V. Best, "Discrimination of sound source velocity in human listeners," *J. Acoust. Soc. Am.*, vol. 111, pp. 1026–1035, February 2002.

[10] T. Kaczmarek, "Auditory perception of sound source velocity," *J. Acoust. Soc. Am.*, vol. 117, pp. 3149–3156, May 2005.

[11] R. A. Lutfi and W. Wang, "Correlation analysis of acoustic cues for the discrimination of auditory motion," *J. Acoust. Soc. Am.*, vol. 106, pp. 919–928, August 1999.

[12] L. D. Rosenblum, C. Carello, and R. E. Pastore, "Relative effectiveness of three stimulus variables for locating a moving sound source," *Perception*, vol. 16, pp. 175–186, 1987.

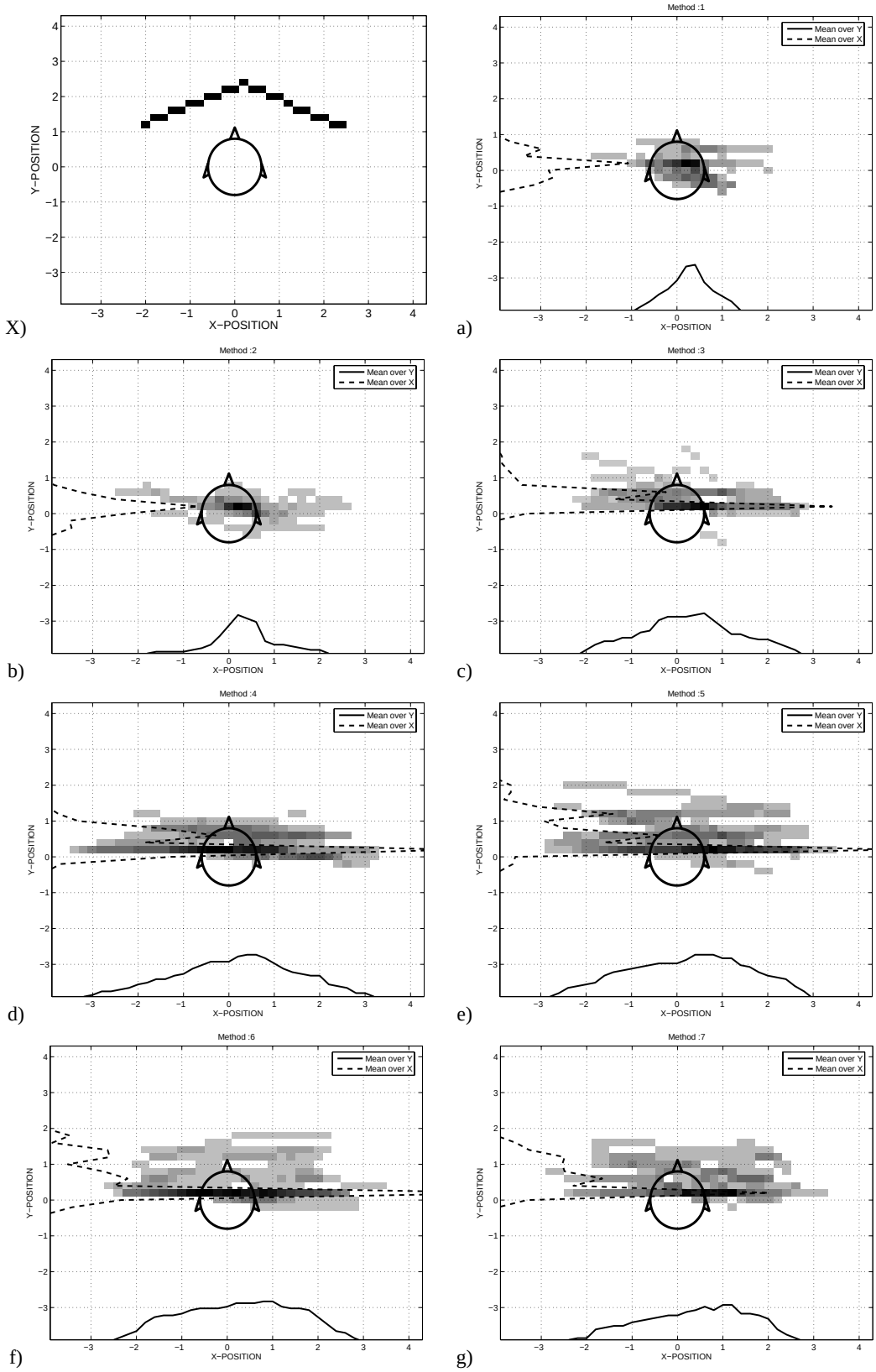

Figure 5: X) An example of a drawing of a user for one playlist and rendering method. a)-g) Histograms of subjective path assessments over all subjects and play lists for all the seven methods. A dark color indicate that the path is often drawn through the pixel.