Audio Engineering Society

Convention Paper 7056

Presented at the 122nd Convention
2007 May 5–8 Vienna, Austria

The papers at this Convention have been selected on the basis of a submitted abstract and extended precis that have been peer reviewed by at least two qualified anonymous reviewers. This convention paper has been reproduced from the author's advance manuscript, without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. Additional papers may be obtained by sending request and remittance to Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA; also see www.aes.org. All rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Spatial audio rendering using sparse and distributed arrays

Aki Härmä, Steven van de Par, and Werner de Bruijn [1]

[1] *Philips Research Laboratories, Eindhoven, The Netherlands*

Correspondence should be addressed to Aki Härmä (aki.harma@philips.com)

ABSTRACT

A widely distributed but multi-channel audio reproduction system can be used to create dynamic spatial effects for various entertainment and communication applications. In this paper we focus on the follow-me audio effect where the sound source appears moving with the observer who is walking through a hallway or going from one room to another in the home environment. We give an overview of the array theory for the sparse distributed loudspeaker systems, study the binaural properties of the sound field rendered with a sparse line array, and compare two different dynamic rendering techniques in a new type of a listening test.

1. INTRODUCTION

An array of loudspeakers can be used to create a virtual sound source in the environment of the user. There are many different techniques. One can use amplitude panning [1], various types of holophonic reproduction methods such as ambisonics [2] or wave field synthesis (WFS) [3], or adaptive methods such as transaural reproduction [4], or adaptive wave field synthesis [5]. In all cases the listener is assumed to stay more or less in a sweet spot, or at least within a restricted listening area inside a volume enclosed by the loudspeakers. Furthermore, the sources are typically restricted to the space outside this volume. Sometimes this is called the one-room sit-down entertainment scenario.

One can also consider another type of a scenario where the user is free to move in the environment, for example, to walk from one room to another or to go outside of the listening area. In principle, a wave field synthesis system can be designed so large that the listener is always within the assumed listening space. However, this is not always feasible or desired, e.g, in the home environment, because the required number of loudspeakers increases rapidly as a function of the area or the volume of the listening space. In this paper, we study the use of spatial ren-

dering methods for a sparse and widely distributed array of loudspeakers. The most familiar example of such an *ambient* audio reproduction system is the public announcement system of a railway station or an airport. However, while the public announcement system is designed to deliver the same message to everyone in the area the goal of the system considered in this paper is to produce localized sound to a tracked individual in the home environment.

Clearly a sparse array of speakers does not allow for an exact reproduction of the desired wave field for most listening positions. Although the spatial reproduction may not be very accurate, a distributed array of individually driven loudspeakers can be used to create very interesting spatial effects for different types of entertainment applications such as gaming, or in communication applications such as a hands-free telephony.

In this paper the focus is in a spatial audio effect which could be called the *follow-me* audio scenario. Here the goal is to create an illusion that the sound source is moving with the listener when the listener walks, for example, through a hallway or from one room to another. In this paper we limit the study to the case of audio rendering using a line array where the distances between individual loudspeakers are large compared to the wavelength. In Section 2, we introduce some of the basic properties of such arrays and study their properties in a numerical simulation. In Section 3 we introduce two dynamic rendering techniques which are then compared in a listening experiment in Section 4.

2. SPARSE LINE ARRAY

It is easy to create a strong follow-me audio effect using a uniform line array of loudspeakers driven by an identical audio signal in each channel. No processing is needed and the effect works simultaneously for an unlimited number of simultaneous listeners. In principle, playing identical signals from a line array is a way to synthesize an acoustic plane wave field propagating perpendicular to the array. The follow-me effect in the plane wave field is familiar to everyone, for example, in a visual analogue the sun always appears to move along with the observer. Due to the spatial aliasing the wave field created using a sparse line array where the spacing of the loudspeakers is two meters is actually a plane wave only below $f_{\max} = c/(\delta x) = 344/2 = 172$ Hz.

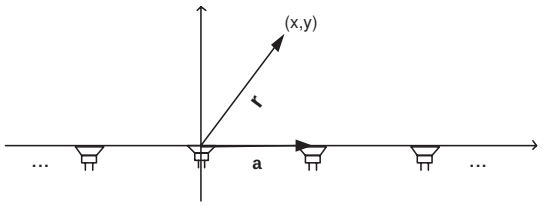


Fig. 1: Infinitely long line array in the free field.

Gardner [6] referred to the perceived position of a sound source in a line array as an illustrative case for summing localization. In his observation the sound is always heard localized to the nearest loudspeaker. However, when the current authors tested this in an anechoic chamber, the result was actually a strikingly smooth follow-me effect for a listener walking by such a line array. In this paper, we first study the *plane-wave* effect of a uniform line array, to understand why it works so well, and later explore alternative methods where the reproduction is truly following the listener.

2.1. Steady-state response of a line array

Let us first study the case of a uniform line array in the free field to characterize the sound pressure level along the path of the listener. The transfer function from an infinitely long line array with spacing of a meters to the position $\mathbf{r} = [x, y]$ in space is given by

$$P(\mathbf{r}, \omega) = \sum_{n=-\infty}^{\infty} A(\omega, \angle(\mathbf{r} - n\mathbf{a})) \frac{\mathbf{e}^{-i\omega(|\mathbf{r}-n\mathbf{a}|)/c}}{|\mathbf{r} - n\mathbf{a}|}, \quad (1)$$

where $A(\omega, \theta)$ is the directionality function of a loudspeaker, and position vectors $\mathbf{r} = [x, y]^T$ and $\mathbf{a} = [a, 0]^T$ are defined as in Fig. 1. In plane-wave rendering, all the loudspeakers are driven by the same input signal $X(\omega)$. Therefore, the pressure signal in a spatial position $\mathbf{r}$ may be written as follows:

$$Y(\mathbf{r}, \omega) = X(\omega) P(\mathbf{r}, \omega). \quad (2)$$

For a white input signal $|X(\omega)|^2 = 1$ the Root-Mean-Square (RMS) value of the pressure signal produced by an infinitely long line array of omnidirectional

speakers $A(\omega, \angle(\mathbf{r} - n\mathbf{a})) = 1$ at position $[x, y]$ may be written in the following form:

$$R(x,y) = \sqrt{\int_{-\infty}^{\infty} \left| \sum_{n=-\infty}^{\infty} \frac{\mathbf{e}^{-i\omega(|\mathbf{r}-n\mathbf{a}|)/c}}{|\mathbf{r}-n\mathbf{a}|} \right|^2 d\omega} \quad (3)$$

Evaluating the square of the series we end up with

$$R(x,y) = \sqrt{\int_{-\infty}^{\infty} (R_W(x,y) + R_P(x,y)) d\omega} \quad (4)$$

$$R_W(x,y) = \sum_{n=-\infty}^{\infty} \left(\frac{1}{|\mathbf{r}-n\mathbf{a}|} \right)^2 \quad (5)$$

$$R_P(x,y) = \sum_{m \neq n} \frac{\mathbf{e}^{-iwA/c}}{A} \frac{\mathbf{e}^{iwB/c}}{B}, \quad (6)$$

where $A = |\mathbf{r} - n\mathbf{a}|$, and $B = |\mathbf{r} - m\mathbf{a}|$. The two terms in (5) represent the (incoherent) summation of signal powers and the summation of pressures denoted by $R_W(x,y)$ and $R_P(x,y)$, respectively. Note that the first term is independent of the frequency while the second term depends on the frequency variable ω. The first term can be written in the following form:

$$R_w(x,y) = \sum_{n=-\infty}^{\infty} \frac{1}{(x-na)^2 + y^2} \quad (7)$$

$$= \frac{\pi \sinh(2\pi y/a)}{ay(\cosh(2\pi y/a) - \cos(2\pi x/a))}. \quad (8)$$

where the last form is based on the series expansions of trigonometric and hyperbolic functions. This is a smooth function along the path of the listener. In fact, we may easily derive that the fluctuations caused by this component are less than one decibel when the distance of the listener from the array is $y > a/1.7$, where a is the spacing of loudspeakers in the array. The magnitude of the term $R_W(x,y)$ is plotted in the top-left panel of Fig. 2.

The frequency-dependent term $R_P(x,y)$ produces large amplitude fluctuations to the listening area. The values of $R_W(x,y)$ at 100, 500, and 5kHz frequencies are illustrated in the other panels of Fig. 2. Clearly, the amplitude pattern is different at different frequencies. Consequently, one can expect fluctuations in the timbre and spatial attributes of the perceived sound field depending on the spectrum content of sound and the position of the observer.

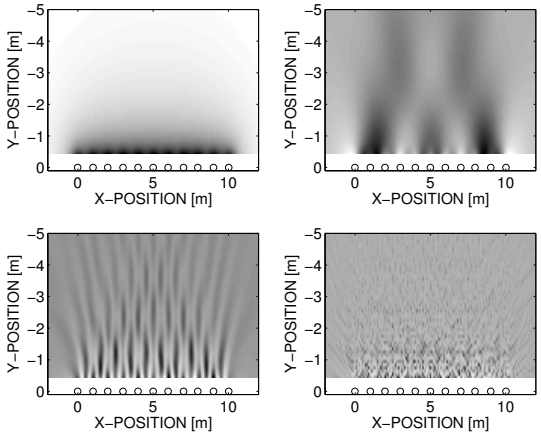


Fig. 2: Values of $R_W(x,y)$ in front of the array (top-left) and $R_p(x,y)$ for 100Hz (top-right), 500Hz (bottom-left), and 5kHz (bottom-right), respectively.

In computing the RMS value (5) we integrate over the frequency variable ω. Most of the complex exponentials in (5) vanish in integration (due to the orthonormality). What are left are non-zero components at discrete locations $x = ka/2$. In those points, the amplitude of $P(\mathbf{r}, \omega)$ is approximately 3dB above the value of $R_W(x,y)$.

Using real loudspeakers where the directivity index $Q > 1$ the picture becomes even more complicated, although, the fluctuations caused by $R_W(x,y)$ terms are then somewhat reduced in amplitude. At high frequencies, the amplitude response depends also on the accuracy in the exact positioning of the speakers. To summarize, we may argue that the distribution of the RMS values of the sound pressure does not suggest a smooth and continuous plane-wave effect.

2.2. Summing localization

The temporal characteristics of the wave field along a line parallel to the array changes from one position to another. This is due to the different arrival times of wave fronts from individual loudspeakers. In spatial hearing, the temporal characteristics of the wave field are of particular interest. Figure 3 illustrates some the time constants for certain perceived spatial effects in a sound field.

The follow-me effect in plane-wave rendering may be associated with the localization dominance and sum-

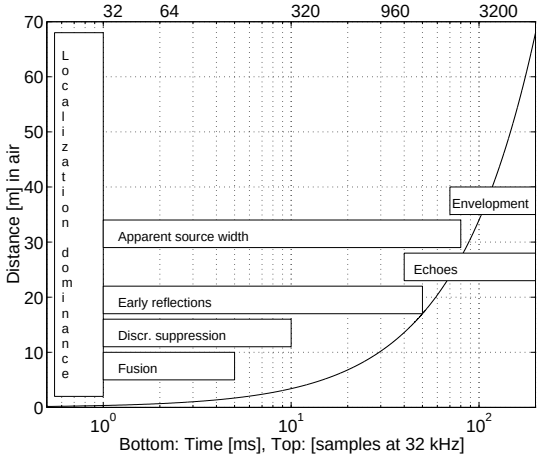


Fig. 3: Some time constants relevant to the spatial perception.

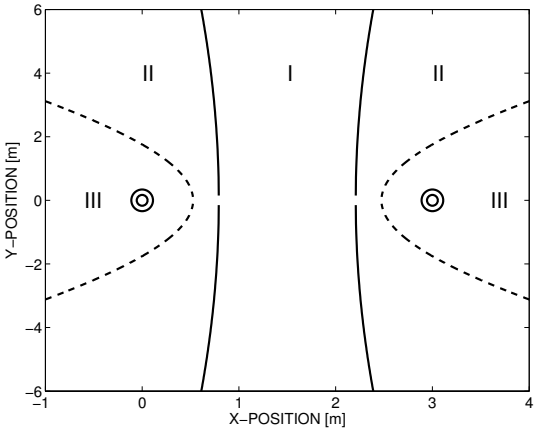


Fig. 4: Regions of summing localization (I) and fusion (I+II) around a pair of stereo loudspeakers. The positions where the lag is 1 ms and 5 ms are represented by solid and dotted curves, respectively.

ming localization mechanisms of spatial hearing [7]. When the lag between the arrival times for a sound from the nearest speaker and the other speakers is less than five milliseconds the sound source is fused in perception into one sound source. When the lag is below about one millisecond the perceived location of the sound source is positioned somewhere between the leading and the lagging sources. But, if the lag is more than one millisecond the position of the leading sound dominates the localization.

The space around a pair of loudspeakers in Figure 4 is divided into three regions. In regions I, II, and III, the time difference between the arrival times of the direct sounds from the two loudspeakers is less than one millisecond, 1-5 milliseconds, and more than 5 milliseconds, respectively. The spacing of the loudspeakers is two meters. In region I, the perceived position depends on the distance between the the two speakers due to the summing localization effect. In region II, the two speaker produce a fused sound image which is positioned close to the leading speaker. In the region III, where the lag is more than 5 milliseconds, a listener tends to hear the two speakers as two separate sources.

In an array of more than two speakers the situation is more complicated. Fig. 5 illustrates the positions in front of the array where the sounds from one, two, or three nearest loudspeakers are heard within one millisecond. If a listener, for example, moves along the line $y = 2$m, the sound source appears positioned to the nearest loudspeaker in all white areas. In moving over a gray area the sound source moves rapidly from one loudspeaker to another. Consequently, one may expect that the movement of the source is perceived somewhat irregular. At the distance of 4.5 meters in Fig. 5 a smoother movement of the source is expected since there are always at least two speakers within one millisecond. Based on trigonometry, the distance from the array where this occurs for an arbitrary uniform line array is given by

$$h = (a^2 - \tau^2)/2\tau, \tag{9}$$

where $\tau = ct = 344 \times 10^{-3}$m is the distance corresponding to the propagation of sound in the air in one millisecond. For example, with a spacing $a = 1$m this distance from the array is slightly more than one meter, but for $a = 3$m the continuous summing localization effect is only obtained at distances over 12 meters from the array.

2.3. Numerical simulations

The analysis based on amplitude fluctuations and the arrival times from different speakers are only a rough characterizations of the listening experience. To get a more detailed picture of the factors affecting the perceived position and movement pat-

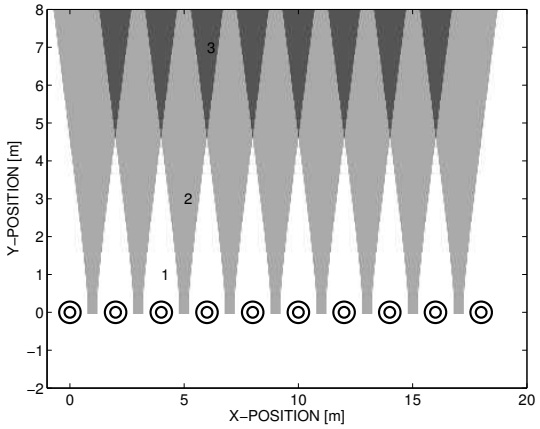


Fig. 5: The gray levels represent the number of loudspeakers audible within the first one millisecond in the area in front of the loudspeaker array. In white, gray, and black areas only the nearest speaker, two nearest loudspeakers, and three nearest loudspeakers are heard in one millisecond, respectively.

tern of the source, we used a computational model of loudspeaker listening. The simulation is based on a model for a loudspeaker response in free field, i.e., the Green's function model of a sound field produced by a piston in an infinite baffle by Stepanishen [8]. The spatial impulse responses are convolved with a set of head-related transfer functions (HRTFs) from the CIPIC database [9]. The pulse responses formed by summing all responses to each ear were presented to a binaural lateralization model that uses time-intensity trading data [10] to combine Interaural Time Delays (ITDs) and Interaural Intensity Differences (IIDs). The model first computes the predicted perceived lateralization for each band, combining ITDs and IIDs. Then the model averages lateralizations across all auditory bands to determine the average lateralization at which the model predicts the sound source to be heard.

The binaural model is now described in more detail. First the pulse responses associated with the left and right ear are transformed to the frequency domain, resulting in two spectra, $l(k)$ and $r(k)$. In order to determine the lateralization in one auditory band, the spectra are weighted with a 4-th order gammatone filter $g_i(k)$ as described in Van de Par et al. [11]. Here i is the index for the auditory filter. First the level differences between the two weighted spectra are determined resulting in an IID expressed in dB:

$$\mathrm{IID}_i = 20\log_{10}\frac{\sum_{k=0}^{N-1}|l(k)g_i(k)|^2}{\sum_{k=0}^{N-1}|r(k)g_i(k)|^2}, \qquad (10)$$

where N is the length of the spectrum.

Secondly the ITD in μs. is determined using:

$$\mathrm{ITD}_i = \frac{10^{-6}}{2\pi f_i}\arg\sum_{k=0}^{N/2-1} l(k)r^*(k)g_i^2(k), \qquad (11)$$

where f_i is the centre frequency (in Hz) of the i-th auditory filter band.

By fitting the time-intensity trading data of Young and Levine (1977) we assume that the same lateralization that is created by a certain IID_c can also be created by an ITD_c assuming that:

$$\mathrm{IID}_c = (3.56\cdot 10^{-8}f_i^2 + 0.0178)\mathrm{ITD}_c. \qquad (12)$$

We assume that for frequencies below 1.5 kHz, both IIDs and ITDs contribute to lateralization, however, above 1.5 kHz we assume that lateralization is dominated by IIDs. In this way an effective interaural intensity difference $\mathrm{IID}_{\mathrm{eff}}$ is determined for frequencies below 1.5 kHz, by assuming that the contributions add linearly:

$$\mathrm{IID}_{\mathrm{eff},i} = \mathrm{IID}_i + (3.56\cdot 10^{-8}f_i^2 + 0.0178)\mathrm{ITD}_i, \qquad (13)$$

while for frequencies above 1.5 kHz we simply assume that $\mathrm{IID}_{\mathrm{eff},i} = \mathrm{IID}_i$.

The second stage of the model assumes that the effective IIDs for each auditory frequency band combine into a single lateralization via a simple averaging operation, thus we use:

$$\mathrm{IID}_{\mathrm{mean}} = \frac{1}{M}\sum_{i=0}^{M-1}\mathrm{IID}_{\mathrm{eff},i}. \qquad (14)$$

Using the data of Yost [12] that gives the dependence of perceived lateralization on IID, we define the perceived sound source angle α in degrees as:

$$\alpha = \frac{90\cdot\mathrm{IID}_{\mathrm{mean}}}{15}, \qquad (15)$$

with the restriction that α is within the interval from -90 to 90 degrees. It should be noted that the model only predicts the lateralization angle at which the sound source is heard; no distinction is made about whether the source is heard in the front or the back direction.

In Fig. 6 results are shown of predicted perceived angles for a listener facing the array, and facing parallel to the array (like in walking by the array). In facing the array, the changes in the angle of the source are large and corresponds to Gardner's observation that the source is mostly localized to the nearest speaker [6]. However, at a larger distance from the array the movement of the source from one speaker to another becomes smooth. This can also be predicted from the chart of Fig. 5, where the number of loudspeakers heard in the first one millisecond increases as a function of the distance from the array.

The bottom panel of Fig.6 shows the predicted directions of sources for a listener facing right parallel to the array. Remember that the model makes no prediction about back-front orientation of the sound source; we assume that sources are heard towards the front of the listener (arrows pointing to the right). The arrows point now much more often perpendicular to the array even at a smaller distance from the array. This result seems to support the observation of the current authors that the plane-wave effect in a sparse line array produces a strong follow-me effect for a listener walking by the array.

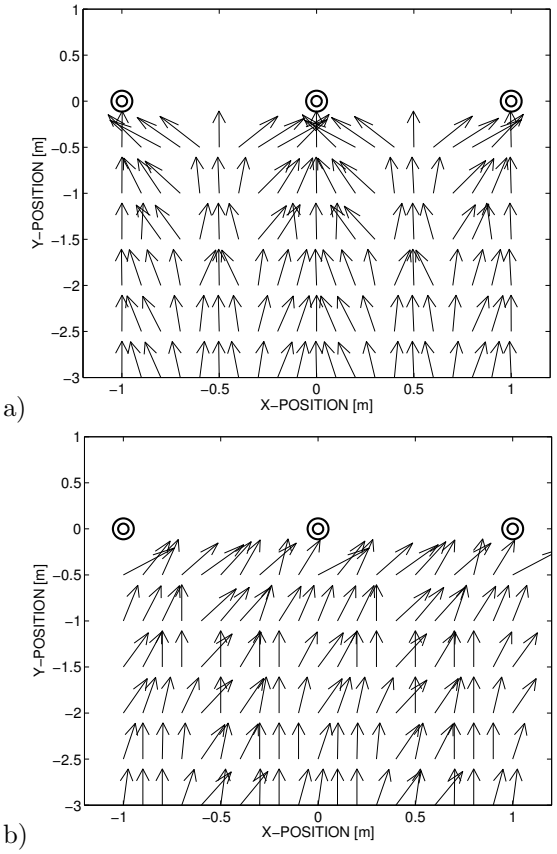


Fig. 6: A prediction of the binaural localization model of the perceived direction of the source in positions in front of a ULA with listeners a) facing the array and b) facing parallel to the array.

3. DYNAMIC SPATIAL RENDERING

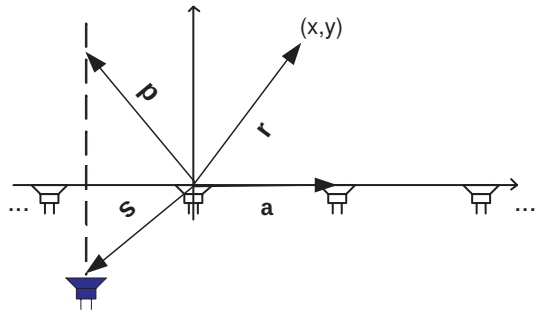
The simple method of playing identical signals from a line array of loudspeakers to create the follow-me effect for a moving observer has many shortcoming. First, for a listener who is not in front of the array but outside of the range spanned by the two ends of the array, the sound appears severely colored due to the comb filtering effect. In many cases, for example, hands-free telephony, it is desired that the sound is following the user. The sound source should appear localized close to the user also for a second person with a different state of motion. For a second person who is not using the application, the follow-me effect may appear disturbing.

A simplest method is to play sounds always only from the nearest loudspeaker (NL). Typically that transition form one speaker to another should be performed smoothly to avoid artifacts related to the onset and offset of audio.

Generally, we may add a weighting function $W(n,\omega)$, where n is the index of a loudspeaker in the array, see Fig. 7. In this way, the spatial response of the array is given by

$$P(\mathbf{r},\omega) = \sum_{n=-\infty}^{\infty} A(\omega, \angle(\mathbf{r}-n\mathbf{a}))W(n,\omega)T(\mathbf{r},\omega), \quad (16)$$

$$T(\mathbf{r},\omega) = \frac{\mathbf{e}^{-i\omega(|\mathbf{r}-n\mathbf{a}|)/c}}{|\mathbf{r}-n\mathbf{a}|} \quad (17)$$



a)

Fig. 7: Definition of vectors in the secondary source model.

3.1. Amplitude panning method

The simplest way to make the sound source move is to use a scalar localized weighting function for the loudspeakers. For example,

$$W(n,\omega) = \begin{cases} 1+\cos((x_p - x)\pi/\beta), & |x - x_p| \leq \beta \\ 0, & \text{otherwise} \end{cases} \quad (18)$$

where x_p is the x-position of the listener, uses a risen cosine window (CW) function to weight loudspeakers that are closest to the listener. Typically the parameter $\beta > 3a$ to make the movement smooth.

3.2. Secondary source model

There are naturally infinitely many weighting functions $W(n,\omega)$ that could be used in the follow-me rendering. One particularly interesting alternative is to use essentially a similar method that is used in wave field synthesis [3] with a much more dense array to synthesize the wave field representing a point source. In WFS, the loudspeaker signals are generated such that they represent the secondary sources (SS) on an infinite surface between a source in position $\mathbf{s}$ and the listener. In the case of a line array we may produce the loudspeaker signals similarly. Using the notation of Fig. 7 the weight function corresponding to the secondary source model is given by

$$W(n,\omega) = \frac{\mathbf{e}^{-i\omega(|\mathbf{s}-n\mathbf{a}|)/c}}{|\mathbf{s}-n\mathbf{a}|}. \quad (19)$$

In the follow-me scenario the x-position of $\mathbf{s}$ is selected such that it is the same as the x-position of the listener $\mathbf{p}$.

The amplitude along the listener's trajectory fluctuates depending on the position of the source $\mathbf{s}$.

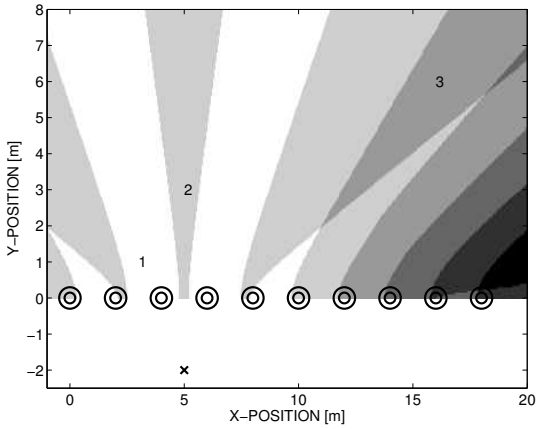


Fig. 8: The different gray levels represent the number of loudspeakers audible within the first one millisecond in the area in front of the loudspeaker in the secondary source model, similarly to Fig. 5.

However, the fluctuations are again relatively small. The plot of the areas where the sounds from one, two, or more loudspeakers can be heard during the first millisecond from the direct sound has an interesting pattern in the secondary source method, see Fig. 8. First, it appears that exactly in the desired listening position the pattern is almost similar to Fig. 5. However, the region where there are several sources within the first millisecond increases significantly for observers not in the listening position.

4. LISTENING EXPERIMENT

To compare the performance of the presented algorithms a listening experiment was developed. The follow-me scenario assumes that the listener is moving. Listening experiments involving moving listeners are rare. Several authors have conducted listening tests to evaluate the performance of head-tracked auralization. For example, Wightman and Kistler conducted a test where the listeners' head movements were either restricted or encouraged in a study on the role of head movements in resolving front-back ambiguities in localization of sound sources [13]. However, the current authors could not find a good example of a listening experiment involving freely-moving tracked listeners.

The listening test setup reported in the current article is quite unconventional. The configuration is illustrated in Fig. 9. The listening test was carried

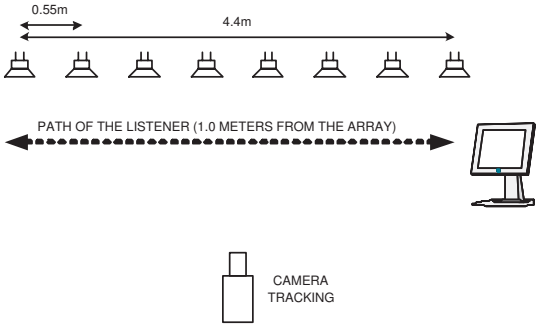


Fig. 9: The listening test setup.

out in a quiet and relatively damped listening room at Philips Research Laboratories, Eindhoven, The Netherlands. Eight small active loudspeakers (M-Audio StudioPro 3) were arranged close to one wall in a line with a 0.55 m spacing between the speakers. In all experiments the user was asked to walk back and forth the path marked in Fig. 9 and give grades for the *smoothness of movement*, *coloration*, *distance of the source*, and *size of the source* using sliders in the graphical user interface of the test program.

The position of the listener moving in front of the array was tracked using a visual user tracking system based on a video camera and it was used to control the dynamic rendering of audio in the follow-me scenario. The playback was active only when the user was in front of the array and the playback was muted when the subject moved to the PC monitor at the end of the line array to give the grades. Four different rendering techniques were tested. In plane-wave rendering (PW) identical audio signal is played from all the speakers. A simple method where the sound was always played from the nearest speaker (NS) was included in the experiment as a reference method. The amplitude panning method based on the risen cosine window (CW) function of Eq. 18 was configured such that the size of the window spanned the range of three loudspeakers. In the secondary source method (SS) the nominal source location $\mathbf{s}$ was one meter behind the array at the same x-position as the listener.

In all methods two different orientations of the listener were studied. In the first set of experiments, the subjects were asked to walk by the array facing parallel to the array (90 degrees). In the second set, the listeners were instructed to walk facing the array (0 degrees). In Figs. 10-11 the the orientation are denoted by numbers 0 and 90.

Three test sequences were used: pink noise sequence, slow harmonic music sample (duet of French Horn and Clarinet), and male speech sequence. The signals were played at the sample rate of 44.1 kHz. The total number of test cases was 12 (four methods and three test sequences). The listeners were allowed to switch between the four methods freely in each sample.

Eight experienced listeners with normal hearing participated in the test. Most listeners were not told in advance that one of the methods was the static plane-wave rendering. The interviews of the subjects, and the authors' own experience suggest that it is difficult to discriminate the static plane-wave rendering from the dynamic rendering methods in this setup unless one manages to somehow confuse the video tracking, e.g., by very fast movements. The follow-me effect is in all cases strikingly clear to the subject. However, for a second non-moving observer in the room, the movement of the sound source with the subject in the dynamic case is clear while in the plane-wave rendering, of course, nothing changes as the subject moves.

The median, and the lower and upper quartile values over all test sequences and all listeners for the *smoothness* of movement are illustrated the top panel of Fig. 10. The highest smoothness grades are obtained in the plan-wave rendering when the listener is facing the array (PW0 condition). The differences between PW0 and methods CW and SS are statistically significant (p-values < 0.01) but the differences between CW and SS are not statistically significant (p-values are 0.1). However, the difference between PW0 and PW90 is statistically significant.

A difference between the two orientations can be expected from the results of the numerical simulation in Fig. 6. However, the results of Fig. 6 actually suggest more smooth movement for a listener parallel to the array. In the 90 degree condition, where the listener is facing parallel to the array, the difference between methods PW, CW, and SS are not statistically significant. That is, the differences between rendering techniques are smaller in the 90 degree condition than in the case where the subject

is facing the array. This supports the observation that spatial cues are more smooth for a listener facing parallel to the array, although, more experiments may be needed.

The nearest speaker rendering (NS) was typically experienced jumping from one speaker to another and it also received the lowest smoothness scores. The difference between 0 and 90 degree orientation of the listener was found statistically significant only in plane-wave (PW) rendering and the secondary source (SS) methods.

The subjects found it difficult to quantify the level of *coloration* in the follow-me effect. There are only few statistically significant differences in results in the lower panel of Fig. 10. The secondary source method with the listeners' orientation parallel to the array appears giving the lowest score (most colored).

The results of the evaluation of the perceived size and the distance of the source are shown in top and bottom panels of Fig. 11, respectively. In NS method the source is almost always heard at the distance of the array (50) and the size is close to 0 (the size of a speaker), that is, the sound is clearly localized to the nearest speaker. In both orientations CW is also perceived close to the position of the array but the size of the source is graded larger. Both PW and SS get larger grades for both size and distance, that is, the perceived source is larger than a loudspeaker and it is localized behind the array. There seems to be no systematic statistically significant trend in the perception of the size or distance of the source for different orientations of the listener.

5. CONCLUSIONS

In this paper we have studied the use of a sparse uniform line array to produce the follow-me effect, an illusion of sound source moving with the moving observer. The *plane-wave* rendering, where the line array of speakers is driven by the identical audio signal produces a strong follow-me effect. Therefore, we first studied this method in detail. It turned out that while the wave field produced using such an array is very nonhomogeneous due to the large spacing of the speakers, the perceptual model of spatial perception predicts a relatively smooth movement of a source with a moving listener. In particular, a difference was found between the cases where the listener is facing the array, and where the listener is facing parallel to the array like in walking by the array. This finding was also supported by the listening test results presented in the paper.

The plane-wave rendering is not a desired technique for several reasons. First, the sound appears very colored for an observer outside of the area spanned by the array. Secondly, the sound appears as a follow-me source to everyone in the environment. Therefore, we introduced three alternative techniques to generate loudspeaker signals so that the virtual source is truly moving with the tracked listener. In the simplest method the sound is always played from the nearest loudspeaker such that there is a smooth time-domain cross-fading from one speaker to another during the transition. The second method is similar to amplitude panning where the a spatial scalar weight function is applied to the array of speakers. Finally, the secondary source method is essentially similar to the wave field synthesis of a point source positioned behind the array [3].

In a listening experiment we compared the static plane-wave rendering and the three dynamic solutions controlled by camera-based real-time tracking of the listener. First, it was found that it is difficult for a listener moving in front of the array to determine if the current setup is static or dynamic. Secondly, a simple nearest speaker method was always heard localized at the nearest speaker. In the listeners' evaluation, the plane-wave rendering got the highest grades for the smoothness of the movement. However, the two more complicated dynamic rendering methods were also given high grades. In the case where the listener is facing parallel to the array the differences between the methods in the *smoothness of movement* are not significant.

In the reported experiment the coloration of sound was not perceived significant in any of the methods. The secondary source got somewhat lower grades than the amplitude weighting method, however the difference is small. Nevertheless, based on the current experiment we can conclude that the computationally light method based on spatial amplitude weighting appears working well in the follow-me effect.

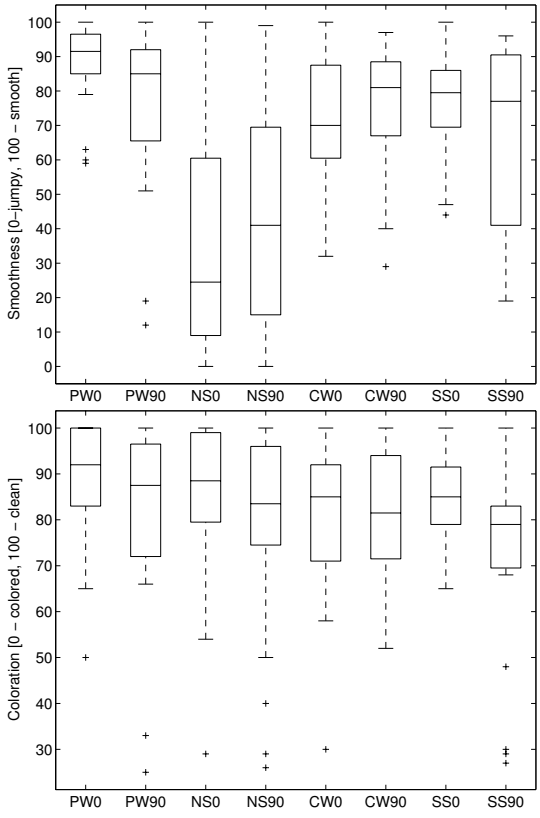


Fig. 10: Results of the listening experiment. Smoothness (top) and the coloration of the source (bottom).

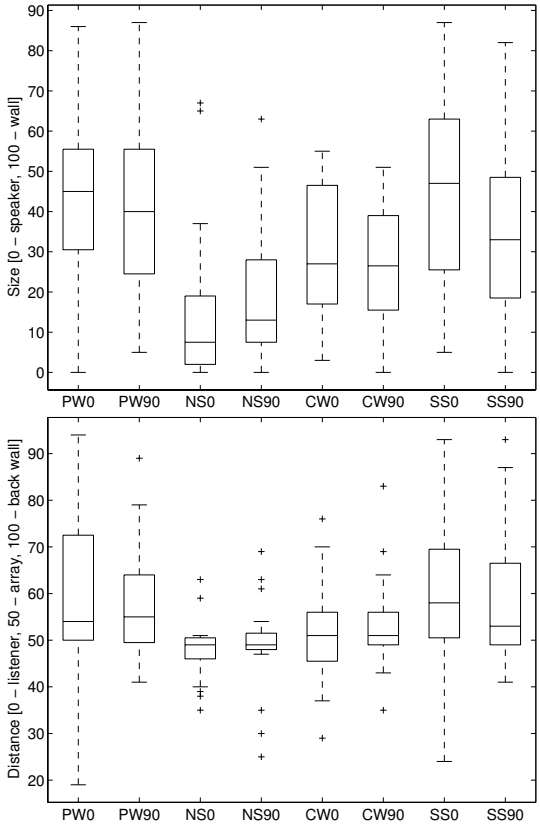


Fig. 11: Results of the listening experiment. Size (top) and the distance of the source (bottom).

6. REFERENCES

[1] V. Pulkki, "Virtual source positioning using vector base amplitude panning," *J. Audio Eng. Soc.*, vol. 45, no. 6, pp. 456–466, 1997.

[2] M. A. Gerzon, "Periphony: width-height sound reproduction," *J. Audio Eng. Soc.*, vol. 21, no. 1, pp. 2–10, 1972.

[3] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," *J. Acoust. Soc. Am.*, vol. 93, no. 5, pp. 2764–2778, May 1993.

[4] O. Kirkeby, P. A. Nelson, F. Orduna-Bustamante, and H. Hamada, "Local sound field reproduction using digital signal processing," *J. Acoust. Soc. Am.*, vol. 100, pp. 15841593, 1996.

[5] P.-A. Gauthier and A. Berry, "Adaptive wave field synthesis with independent radiation mode control for active sound field reproduction: Theory," *J. Acoust. Soc. Am.*, vol. 119, no. 5, pp. 2721–2737, May 2006.

[6] M. B. Gardner, "Some single- and multiple-source localization effects," *J. Audio Eng. Soc.*, vol. 21, no. 6, pp. 430–437, July/August 1973.

[7] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, "The precedence effect," *J. Acoust. Soc. Am.*, vol. 106, no. 4, pp. 1633–1654, October 1999.

[8] P. R. Stepanishen, "Transient radiation from pistons in an infinite planar baffle," *J. Acoust. Soc. Am.*, vol. 49, pp. 1629–1638, 1971.

[9] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The cipic hrtf database," in *Proc. 2001 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2001)*, New Paltz, NY, USA, October 2001.

[10] L.L. Young and J. Levine, "Time-intensity trades revisited," *J. Acoust. Soc. Am.*, vol. 61, pp. 607–609, 1977.

[11] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. H. Jensen, "A perceptual model for sinusoidal audio coding based on spectral integration," *EURASIP Journal on Applied Signal Processing*, vol. 9, pp. 1292–1304, 2005.

[12] W.A. Yost, "Lateralization position of sinusoids presented with interaural intensive and temporal differences," *J. Acoust. Soc. Am.*, vol. 70, pp. 397–409, 1981.

[13] F. L. Wightman and D. J. Kistler, "Resolution of frontback ambiguity in spatial hearing by listener and source movement," *J. Acoust. Soc. Am.*, vol. 105, no. 5, pp. 2841–2853, May 1999.