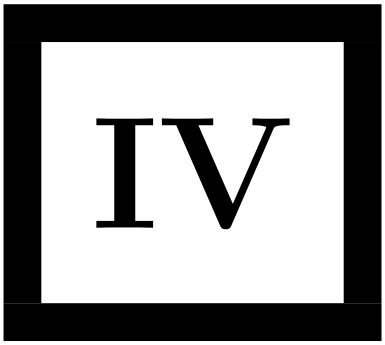

Publication IV

Hirvonen, T. and Pulkki V., ”Perception and Analysis of Selected Auditory Events with Frequency-Dependent Directions”, Journal of the Audio Engineering Society , Vol 54, No. 9, Sep 2006.

Perception and Analysis of Selected Auditory Events with Frequency-Dependent Directions*

TONI HIRVONEN AND VILLE PULKKI
(toni.hirvonen@tkk.fi) (ville.pulkki@tkk.fi)

Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology, FI-02015 TKK, Finland

Relatively little is known about complex auditory events caused by multiple simultaneous sources. In order to gain insight into this topic, the perception of wide-band (200–1179-Hz) noise and click train stimuli was examined with subjective tests focusing on perceived spatial distribution. By reproducing different frequency bands of the stimuli from loudspeakers at different azimuth directions, the spatial content of the overall stimulus was varied in 15 test cases. The subjects were required to indicate those loudspeakers that they perceived as radiating sound in each case. The results suggest that the highest and lowest frequencies of the stimuli were more perceptually significant than the middle frequency region. The test cases were never perceived as being more than half the actual width of the source ensemble. The order of the critical-band signals in the loudspeaker setup had a minor effect on the overall width. When a click train stimulus was used instead of continuous noise, the perceived width was reduced significantly. Cross-correlation-based auditory modeling techniques were also examined for their ability to predict the subjective results and were found to be not entirely suitable for the purpose.

0 INTRODUCTION

As we rarely experience complete silence, everyday life presents us with many different listening situations. The sound waves arriving at the eardrums are usually combinations of direct sounds and reflections emerging from several sources in different directions, and in many cases the sound sources themselves are large compared to the wavelength of the sound. The resulting auditory events can be called complex as opposed to listening to a single, pointlike sound source under free-field conditions [1]. Despite the fact that most natural sounds are complex, much of the traditional auditory research has been focused on simpler cases, relying on the somewhat questionable principles of scientific reductionism.

This paper investigates the perceptual issues related to complex auditory events caused by multiple simultaneous sources. Such events often occur in everyday spatial hearing situations such as multichannel reproduction. Even single loudspeakers commonly house multiple elements, and thus different frequencies of the sound arrive from different directions. In the experiments presented here the stimuli were created by presenting different sounds from multiple azimuth directions in an anechoic environment using many loudspeakers in a wide horizontal sector. This research is focused on the perceived auditory width of sound, on the perceived spatial distribution of sound, as well as on auditory modeling.

This introduction outlines the previous research and open questions related to the topic. Section 1 details the methods used, as well as giving the hypothesis for the present research. Listening test results are presented and analyzed in Section 2. Section 3 investigates the use of computational auditory models in analyzing the subjective test cases used here. Finally a summary is given in Section 4.

0.1 Perception of Multiple Simultaneous Sound Sources

This section considers the perceptual attributes that are reasonable when describing the spatial perception of wide, multiple-source sounds. First some basic concepts should be defined. In acoustics, sound source traditionally refers to the physical world, where a physical entity emits acoustic waves in the audible range, whereas the internal percept produced by the senses is called image, event, or object [2]. These concepts have, however, raised many philosophical questions [3] and are not straightforward in all situations, such as when classifying simultaneous sounds that are not similar.

A loudspeaker is commonly considered a single sound source. Attributes that effectively define the spatial perception of a single sound source are its apparent location and size, usually just its width. However, if several nearby

*Manuscript received 2005 December 14; revised 2006 May 10 and June 16.

loudspeakers emit sound, one may find it hard to distinguish the individual loudspeakers clearly. Rather, the entire loudspeaker setup might be better characterized as an ensemble consisting of several source elements [4]. This paper does not consider the perceptual integration and segregation of sources into different auditory objects, as no reliable methods exist for this purpose. Instead, a source ensemble is considered to produce a single, horizontally wide sound event. In this case the concept of pointlike localization breaks down and is no longer a very interesting attribute.

Perceived width is a more important attribute in describing complex multiple-source auditory events. Traditional research on width or spaciousness has been focused on concert hall and listening room acoustics, where early reflections play an important role in this respect [6]. This paper investigates auditory perception solely in the absence of reflections, where width has been shown to be a function of loudness, duration of sound, frequency, and interaural characteristics [7], [8]. Only the latter attribute varies in the test cases utilized here, and its effects are the least understood.

While perceived width describes the overall spatial extent of the sound, it does not describe the distribution of the directions from where the sound is perceived to arrive. In this paper the additional concept of spatial distribution is adopted to characterize this aspect.

Thus the two main perceptual attributes of interest in this paper are perceived auditory width and perceived spatial distribution. Details on how these attributes are defined are presented in Sections 1.5 and 2. Not much research has been done on these topics in situations where the elements of an ensemble produce different sounds. Although some experiments, for example, with multichannel audio systems [5] exist, the perceived spatial distribution of the sound of the corresponding events is not well known.

0.2 Modeling Auditory Perception

The second part of this paper investigates the use of auditory models in describing the perception of multiple-source auditory events. There exist numerous computational models that characterize the periphery, and partly also the middle stages following the cochlea, of the auditory nervous system. Most of them stem from the cross-correlation-based model suggested by Jeffress [9], and produce a three-dimensional correlogram that is a function of frequency and internal delay between the two input signals. By investigating the correlogram, these models are able to accurately predict simple phenomena, such as lateralization of pure tones and bandpass noise, and specific binaural masking level difference cases ([10] and references therein). However, with more complicated input signals, and attributes other than lateralization, the interpretation of the correlogram output becomes more difficult [11].

When the ear input signals ascend toward the auditory cortex, the neural processing involved is less well known. The fitting term "higher level processing" is often used in connection with the cortex and even processes beyond the auditory periphery. When auditory models fail, multidimensional perceptual attributes traditionally have been examined with indirect psychological methods. Numerous papers on auditory scene analysis have combined brain measurements with psychological tests ([12], [13] and references therein). However, these research branches focus prominently on the segregation of temporal streams using time-variant stimuli and not so much on perceived spatial distribution of complex stimuli.

0.3 Previous Research by the Authors

The present authors have investigated the perception of complex multiple-source stimuli in the past [14], [15], and this paper partly represents an attempt to further investigate the questions raised by the results. The previous experiments [15] presented the subjects with multiple simultaneous sounds using a nine-loudspeaker grid in the azimuth sector between −22.5° and 22.5° in an anechoic environment. Each loudspeaker radiated a different critical band of the overall broad-band noise stimuli, and the spatial order of the bands was varied from case to case. If the possible segregation of different loudspeaker signals is disregarded, this kind of stimulus can be interpreted as a broad-band noise, whose directional cues suggest different azimuth directions at different critical bands.

Fig. 1 illustrates a typical stimulus used in the previous experiments by indicating how the implied direction of the localization cues changed as a function of frequency. In this particular example the highest and lowest frequency bands were presented from adjacent loudspeakers, and there was a distinctive jump or discontinuity within a small azimuth angle around 420 Hz. In the experiments the subjects were required to indicate the perceived center of the sound event, as well as which loudspeakers were perceived as radiating sound.

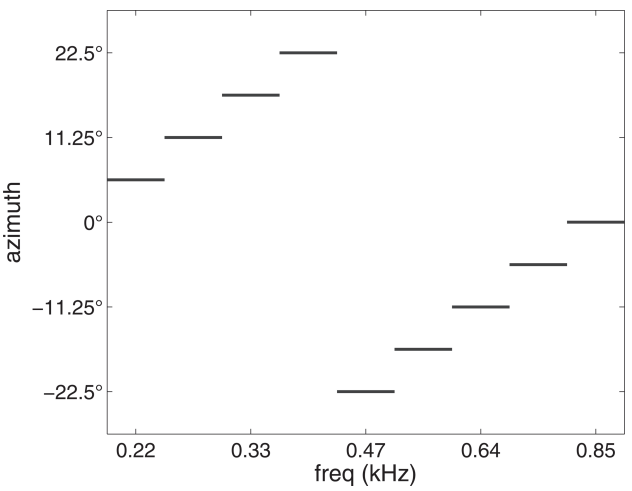

Fig. 1. Example spatial configuration of frequency bands for stimuli used in previous experiments by authors. A Gaussian broad-band noise signal was divided into narrower bands, which were routed to loudspeakers at different azimuth angles. The total broad-band stimulus was formed by playing all narrow-band signals simultaneously. — implied azimuth direction of localization cues of overall stimulus as a function of frequency.

The previous results presented several new questions. The lowest and highest frequency bands of the stimulus frequency range were perceptually important, that is, the subjects often perceived the sound as emanating from their direction. This was especially true when the two bands were presented from adjacent loudspeakers, as in Fig. 1. In this paper we wish to further evaluate the relative importance of the lowest and highest frequencies compared to the middle region. In addition it was unclear whether a "frequency discontinuity," where a large shift in frequency occurs within a small azimuth angle, was perceptually significant because of some perceptual interaction between spatially nearby critical bands. Could, for example, perceived width be affected by the spatial direction of the different bands or are the bands processed separately regardless of their spatial attributes, as in traditional auditory models? Furthermore, as the previous experiments utilized only noise stimuli, the effect of different stimulus types was left unclear.

1 METHODS

In this section we present the research hypotheses and questions as well as the test method utilized.

1.1 Test Setup

As mentioned, this research is limited to the horizontal plane under anechoic conditions. The test setup, illustrated in Fig. 2, consisted of a loudspeaker grid with 11 loudspeakers. All loudspeakers were suspended in front of the listener along a line approximately 2 m long, and the differences in distance were compensated for with delays so that the sound from all loudspeakers arrived at the listening position at the same time. The loudspeakers covered the azimuth sector from −45° to 45° symmetrically at eye level. Thus the spacing between adjacent loudspeaker centers was 9°, as seen from the optimal listening position, that is, the sweet spot. The loudspeakers were visibly numbered from 1 to 11, with 1 being the leftmost speaker at −45°, 6 the middle speaker at 0°, and 11 the rightmost speaker at 45°. The loudspeakers were also angled so that all of them faced the sweet spot. Genelec model 8030 A active monitors were used in all positions of the setup.

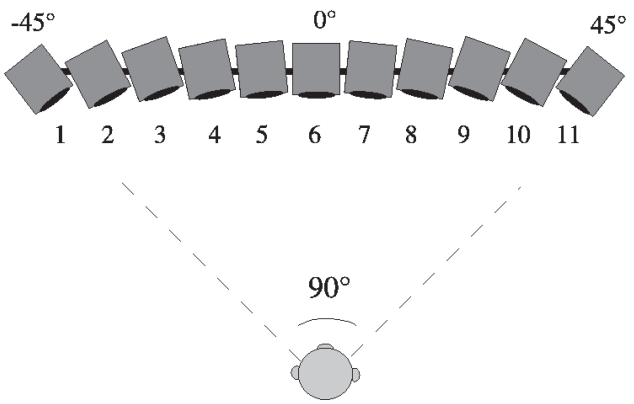

Fig. 2. Schematic of listening test setup. Subjects listened to a grid of 11 numbered loudspeakers in horizontal plane spanning 90° azimuth.

The listening setup was constructed inside an anechoic chamber with a lower frequency limitation of 90 Hz. The test was controlled via a desktop computer (Apple Mac G4) equipped with a multichannel audio system. All control equipment was outside the anechoic chamber in order to minimize background noise. Only a wireless keyboard used to register the subjects' responses was placed inside the anechoic chamber. The frequency response of each loudspeaker in the setup was measured at the sweet spot and found to be flat within ±0.75 dB in the test frequency range.

The present setup is similar to the one used in the previous experiments, which had nine loadspeakers in the −22.5° to 22.5° azimuth range [15]. However, the subjects had commented that discriminating between closely placed loudspeakers in a narrow sector was sometimes difficult. Here the azimuth range is extended to 90° and the azimuth gap between adjacent loudspeakers is increased. These changes should ease the discrimination task.

1.2 Subjects

Fifteen subjects participated in the experiments. They were acoustics students aged 20 to 30 years. None reported any hearing defects.

1.3 Stimuli

The test utilized both continuous Gaussian noise and click train stimuli. In order to prevent waveform learning effects, the noise stimuli were 60-s-long segments sampled randomly from a frozen 2-min noise sample each time the stimulus was repeated. The noise stimulus also had a 100-ms raised cosine fade-in and fade-out. The click stimuli were trains of frozen filtered clicks with an interclick interval of 330 ms.

The basic idea of the test was to present consecutive nonoverlapping frequency bands of the overall stimulus from different azimuth directions using the loudspeaker setup shown in Fig. 2. The bandwidth of the total stimulus, that is, the combined signals of all loudspeakers, was 200–1179 Hz in all test cases. This frequency range consists of 11 critical bands, as measured by the equivalent rectangular bandwidth (ERB) scale [16], the lowest band being 200–249 Hz.

In the noise-stimulus cases the different ERB-band signals were filtered from a white Gaussian noise signal with a real-time window-based fast Fourier transform (FFT) filtering system. FFT filtering allowed for precise band-limiting so that the magnitude responses of the different passbands were very close to rectangular and 1 ERB wide in the noise test cases.

The ERB-band click signals were created prior to the test by filtering a unit impulse with 200-tap FIR filters whose passbands were centered at ERB-band center frequencies. However, the 3-dB passbands were effectively wider than in the noise cases, approximately 3 ERB, which results in frequency overlap between different ERB-band click signals. This was deemed acceptable for the click cases, as longer filters would have resulted in longer click samples that would no longer be perceptually impulse-like, which was a priority here. Now the click length was 200 samples, that is, 9 ms.

The levels of the ERB-band noise signals were aligned for loudness prior to the experiments, with informal subjective comparison of all samples. Three persons agreed independently with the final alignment. The rationale behind this procedure was to give all ERB bands in the experiment as equally perceptual weights as possible. The filtered clicks were perceived equally loud without additional alignment, probably because of their short lengths.

A stimulus created using the method described can be thought of as a broad-band sound source whose interaural cues imply a different direction as a function of frequency. Interaural time difference (ITD) and interaural level difference (ILD) imply the same direction within the frequency band of each loudspeaker. The implied direction changes abruptly in 9° steps when moving from one loudspeaker to another. Because the arrival times of the sound from all loudspeakers to the sweet spot were very close (less than 0.1 ms), the precedence effect can be excluded.

1.4 Test Cases and Hypotheses

In this section the hypotheses and research questions for the present research are stated, and the test cases selected to investigate them are presented. A total of nine different frequency-band configurations were selected. Fig. 3 illustrates the spatial configurations of the cases so that the implied direction of the interaural cues is represented by a solid line as a function of frequency.

In case 1, for example, the lowest band is routed to the loudspeaker in the direction of −45°, and the azimuth angle increases with frequency so that the highest band is presented from 45° azimuth. In case 2 all ERB-band signals are presented from the middle loudspeaker at 0° to evaluate the minimum perceived width and the accuracy of the test setup. In contrast case 3 uses only the outer two loudspeakers. In general cases 1–3 can be considered references for the test system.

As the first research question we wish to examine the relative importance of the highest and lowest frequencies in a complex broad-band stimulus. Previous research implied that the low and high frequencies of a noise stimulus whose localization cues change as a function of frequency are perceptually significant [15]—the subjects perceived the sound emanating prominently from the direction of these frequencies. It is hypothesized that these bands have a notably higher effect on the perceived directional distribution of sound than the middle frequencies between them. Test cases 4–6 in the middle row of Fig. 3 examine this hypothesis—the frequency content of the middle loudspeaker is increased in varying steps, while the spatial extent of the broad-band sound source remains at 90°. Thus cases 1, 4, 5, and 6 form a series where the purpose is to examine at which point the 0° direction becomes more perceptually dominant than the outer loudspeakers. It should be noted that as more ERB-band signals are routed to the middle loudspeaker, the sound pressure level (SPL) of this loudspeaker signal is also increased.

Second, the effect of the spatial context of the frequencies in a complex stimulus is examined. Specifically we wanted to know whether varying the order of the ERB-band signals abruptly in the loudspeakers of the test setup would alter the perception of width. Based on the previous experiments, it is hypothesized that the significance of this effect can be shown. Cases 7–9 investigate how abrupt changes in localization cues are perceived, and whether

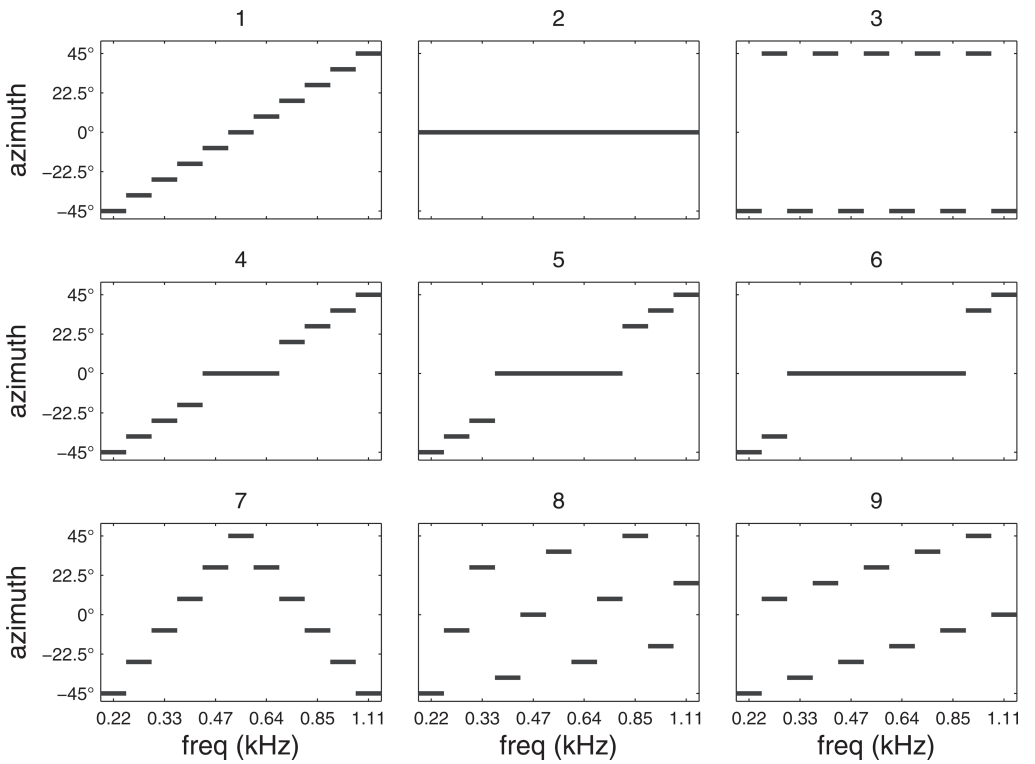

Fig. 3. Spatial configurations of critical bands of broad-band stimuli used in different test cases. Titles indicate case numbers. — active loudspeakers as a function of frequency.

they increase the perceived width of the overall stimulus. If this were the case, there would have to be some contextual interaction between the critical bands of the auditory system. Compared to case 1, cases 7–9 have an increasing number of large, abrupt changes in localization cues as a function of frequency. Also noteworthy is that cases 1, 8, and 9 all have one ERB-band signal routed to each loudspeaker, only in different orders. This way the effect of the spatial order of the frequencies can be examined when all loudspeakers radiate an equal number of ERB bands.

Third, the effect of the stimulus type is examined. Three test cases were composed using both continuous Gaussian noise and click train stimuli. There exist no experimental data that directly suggest that the different stimuli would not be perceived equally wide. However, the perception of width is possibly related to other topics where differences between stimulus types have been found, such as spectral and temporal integration. In addition to the nine noise test cases, cases 1, 2, and 9 were also repeated using an ERB-band-filtered click train in place of the ERB-band noise signals. The total number of test cases was thus increased to 12.

In addition to examining subjective results, we are also interested in how Jeffress-based auditory models can be used to simulate the perception of complex stimuli. Although Section 0.2 detailed some of the shortcomings of the model, it is assumed that the higher level processing in these cases is strongly based on the output of the lower levels of hearing. Thus if the model is accurate, its output could be used to extract a spatial distribution similar to the listening test results.

1.5 Procedure

As mentioned, the two test attributes of interest are the perceived width and the spatial distribution of the test cases. In practice the spatial distribution was evaluated by indicating the elements, that is, the loudspeakers, of the source ensemble that were perceived as radiating sound. In any one case the subjects could choose any number of loudspeakers between 1 and 11 with no restrictions. They could, for example, select only the outer loudspeakers. Width, on the other hand, is here defined as the number of the 11 loudspeakers that were perceived to emit sound. As the source elements need not to be adjacent, this definition is not so much related to the horizontal extent but rather is used to indicate the sum of the widths of all auditory objects produced by the ensemble.

In the test the subjects were seated in a chair with a headrest facing the middle loudspeaker (number 6) at 0°. The subjects were told to utilize the headrest to keep their heads toward the middle loudspeaker during the test. The subjects' responses were recorded via a wireless keyboard that they held during the test. The subjects could see the loudspeakers, and number keys 1–11 were used to indicate the corresponding loudspeakers. Using these keys, the subjects could mark the loudspeakers they perceived as emitting sound. Unintentional keystrokes could also be corrected.

In the beginning of the listening test the subjects went through a familiarization phase that included an explanation of the test task, listening to five random samples in the presence of the conductor. Care was taken not to bias the listeners while discussing the task, and it was emphasized that the test has no "correct" answers as such. Rather, it was important that the listener would carefully analyze whether each loudspeaker emitted sound in each test case. The sound level was kept constant during the test (approximately 70 dB, A-weighted).

The stimulus in each trial was looped for the time it took the subjects to give their responses. All subjects evaluated the 12 test cases six times, resulting in 72 evaluations per person. The test was performed in two similar sessions of $3 \times 12 = 36$ cases each, with a break between them. The order of the samples in each session was randomized. One case was evaluated in less than a minute on average. The whole test took usually 1.5 hours.

2 RESULTS

Fig. 4 presents the results of the listening test for each test case. The spatial order of the frequency bands in each case is shown similarly as in Fig. 3. In addition the number of responses given for each loudspeaker is represented by 11 vertical bars on the left of each case panel. The bars represent the raw subjective data. All subjects' results have been included, and the results are not averaged or scaled between subjects. The total cumulative loudspeaker responses can be interpreted as spatial distributions whose peaks indicate the direction where the sound was heard most prominently. The title above each panel indicates the stimulus type (noise or click train), as well as an additional measure in parentheses: the average number of loudspeakers (ls) marked to emit sound in each test case. Although this number can be seen as a qualitative indicator of the perceived width of the sound, the measure does not take the perceived direction of the loudspeakers into account. Crosses on the left of each panel mark the loudspeaker positions where the the distribution deviates from the mean of each case, that is, from a uniform distribution as detailed in the following section.

2.1 Statistical Analysis

In this section the data are analyzed with some commonly used statistical hypothesis tests. Two questions arise when examining the empirical result distributions in Fig. 4: do the distributions differ 1) between cases, and 2) from a uniform distribution, where the probability p that a given loudspeaker is perceived to emit sound is 1/11 for all loudspeakers? To examine these questions, Kolmogorov–Smirnov (K–S) goodness-of-fit tests between two distributions were utilized. The K–S test is assumed to be a useful nonparametric method in this analysis, as it is sensitive to differences in both location and shape of the empirical cumulative distribution functions of the two samples (test case results). The K–S test gives a p value for the null hypothesis that the two distributions are equal, which is usually rejected if $p \leq 0.05$. This limit is also utilized here.

The first question is addressed by comparing the empirical distributions of all test cases with one another, which results in 66 comparisons. The results of this comparison are presented in the lower left triangle matrix of Table 1. It can be seen that most of the p values indicate significant dif-

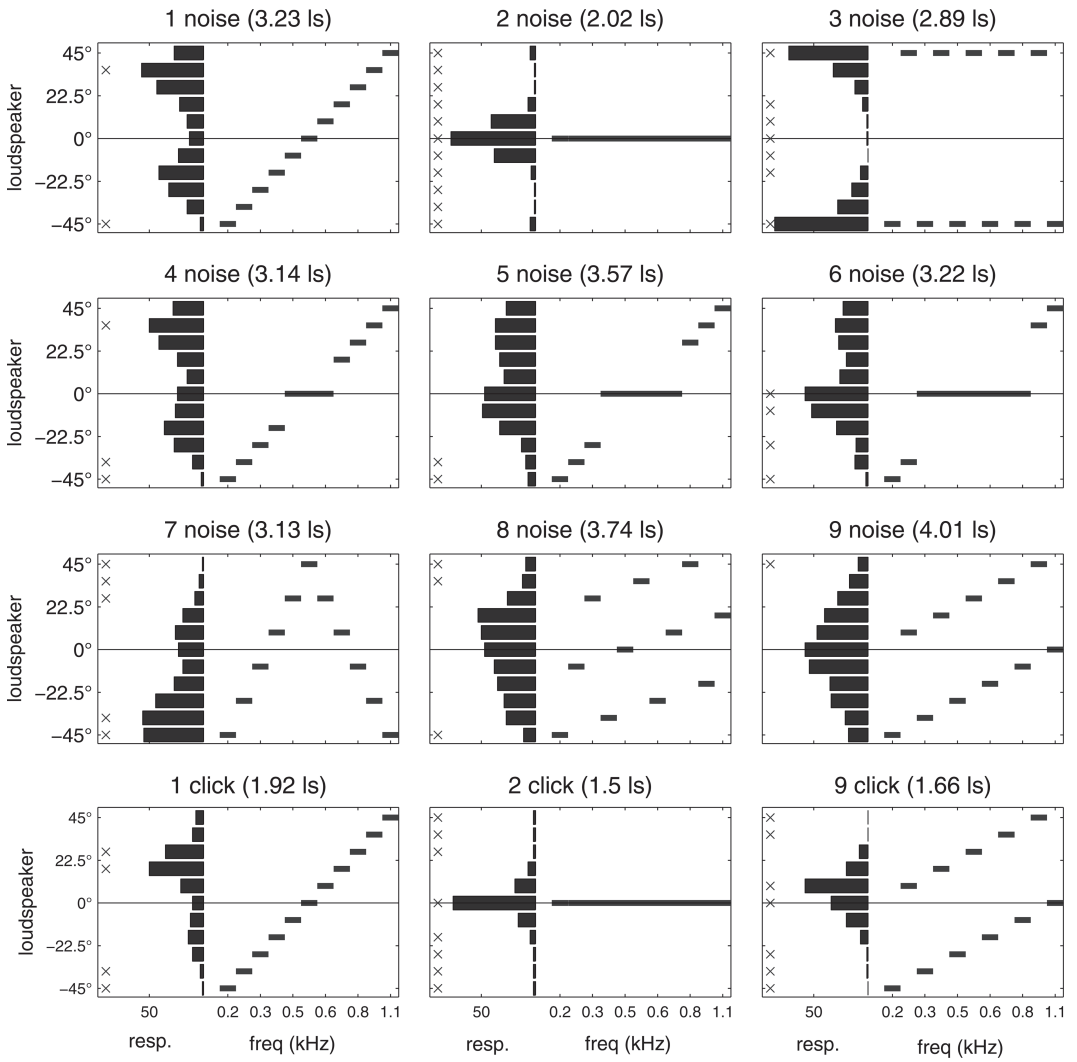

Fig. 4. Listening test results for individual test cases. Actual directions of different ERB-band signals are indicated by lines as in Fig. 3. Left—11 bars indicate number of responses given to corresponding loudspeakers; ×—loudspeaker locations with significant peaks or valleys. Titles indicate case number, stimulus type, and average number of loudspeakers marked.

Table 1. Results of students *t* tests and Kolmogorov–Smirnov tests for significance of differences*

	Student's *t* Tests for Perceived Numbers of Loudspeakers											
Case	1n†	2n	3n	4n	5n	6n	7n	8n	9n	1c‖	2c	9c
1n†		*<0.01*	*0.03*	0.44	*0.03*	0.94	0.46	*<0.01*	*<0.01*	*<0.01*	*<0.01*	*<0.01*
2n	*<0.01*		*<0.01*	*<0.01*	*<0.01*	*<0.01*	*<0.01*	*<0.01*	*<0.01*	0.55	*<0.01*	*0.02*
3n	*<0.01*	*<0.01*		0.15	*<0.01*	0.09	0.20	*<0.01*	*<0.01*	*<0.01*	*<0.01*	*<0.01*
4n	0.89	*<0.01*	*<0.01*		*<0.01*	0.66	0.94	*<0.01*	*<0.01*	*<0.01*	*<0.01*	*<0.01*
5n	*0.02*	*<0.01*	*<0.01*	0.06		*0.04*	*0.02*	0.36	0.06	*<0.01*	*<0.01*	*<0.01*
6n	*<0.01*	*<0.01*	*<0.01*	*<0.01*	0.37		0.58	*0.01*	*<0.01*	*<0.01*	*<0.01*	*<0.01*
7n	*<0.01*	*<0.01*	*<0.01*	*<0.01*	*<0.01*	*<0.01*		*<0.01*	*<0.01*	*<0.01*	*<0.01*	*<0.01*
8n	*<0.01*	*<0.01*	*<0.01*	*<0.01*	*<0.01*	*<0.01*	*<0.01*		0.16	*<0.01*	*<0.01*	*<0.01*
9n	*<0.01*	*<0.01*	*<0.01*	*<0.01*	*<0.01*	*<0.01*	*<0.01*	0.67		*<0.01*	*<0.01*	*<0.01*
1c‖	*<0.01*	*<0.01*	*<0.01*	*<0.01*	*<0.01*	*<0.01*	*<0.01*	*<0.01*	*<0.01*		*<0.01*	*0.04*
2c	*<0.01*	0.83	*<0.01*	*<0.01*	*<0.01*	*<0.01*	*<0.01*	*<0.01*	*<0.01*	*<0.01*		0.15
9c	*<0.01*	*<0.01*	*<0.01*	*<0.01*	*<0.01*	*<0.01*	*<0.01*	*<0.01*	*<0.01*	*<0.01*	*<0.01*	
	Kov–S Tests for Distributions											
Uniform Distributions	*<0.01*	*<0.01*	*<0.01*	*<0.01*	*<0.01*	*<0.01*	*<0.01*	*<0.01*	*<0.01*	*<0.01*	*<0.01*	*<0.01*

* Italic—significant *p* values.
† Noise cases.
‖ Click cases.

ferences between cases, except for five comparisons. The lowest row of Table 1 presents the K–S test results between each case and a uniform distribution. It can be seen that the results suggest that the similarity hypothesis with a uniform distribution should be rejected in all test cases.

As the K–S test with a uniform distribution also does not tell which individual loudspeaker locations are the most significant, two-tailed confidence intervals for a binomial distribution with equal probabilities ($p = 1/11$) for all loudspeaker locations were calculated in each case. This distribution represents the situation where all loudspeaker locations were marked by chance, and thus any deviation beyond the confidence intervals can be interpreted as either prominent sound or lack of sound perceived from that direction. The significant loudspeakers are marked with crosses in the left of each panel in Fig. 4.

In addition to the analysis of spatial distributions, paired Student's *t* tests were performed between cases in order to establish whether they have different widths or numbers of perceived loudspeakers. All subjects' data were again used so that the number of marked loudspeakers in each trial was used as the dependent variable. These data are presented in the upper right triangle matrix of Table 1. Again, most cases (all but 13 comparisons) have significant differences in the number of perceived loudspeakers compared to most other cases. The data were tested in all comparisons with the Jarque-Bera test for goodness-of-fit to a normal distribution and found mostly suitable. Exceptions were the data of noise case 6 and click case 1, whose abnormality was caused by the fact that on a few occasions some subjects had marked most of the loudspeakers as emitting sound, whereas in most trials, only a few loudspeakers were so marked.

2.2 Discussion

This section discusses the relevant empirical and statistical analysis results based on Fig. 4 and Table 1. Not every individual statistically significant result will be covered here. A noteworthy, general aspect of the results is that the sound was on average perceived to radiate from fewer than half the loudspeakers, even in cases where all loudspeakers were in fact radiating sound. This result is similar to those of our previous studies and, without considering actual auditory object analysis, implies that different ERB-band signals were integrated together spatially.

Let us first examine the three "reference" noise cases 1, 2, and 3 in more detail. In the first case the subjects on average marked 3.23 loudspeakers as emitting sound. The perceived spatial distribution shows two peaks despite the fact that the ERB bands are routed evenly to all loudspeakers in the order of increasing frequency. The K–S test indicates a deviation from a uniform distribution, and the confidence intervals of the binomial distribution show one distinct maximum. The peak is located at 36° azimuth and ERB-band center frequency 970 Hz. According to our first hypothesis, the middle frequency range is less perceptually salient. The result seems to confirm this somewhat, as the distribution is low around 0° in this case.

Case 2 presented all frequency bands of the stimulus from the loudspeaker 6 (0°). Interestingly the subjects on average marked two loudspeakers as radiating sound. Some even confused the sound as emanating from the direction of the two outer loudspeakers of the setup. However, the accuracy of the test system can be said to be adequate, since the distribution is symmetric, and it distinctly peaks at 0°. Compared to case 1, presenting all ERB-band signals from a single loudspeaker instead of 11 reduced the perceived number of loudspeakers only by one, but a *t* test indicates that the difference is significant. The results for case 3 are similar, only the outer loudspeakers dominate perception.

In cases 4–6 the number of ERB-band signals in the middle loudspeaker is increased progressively, effectively increasing the SPL of this loudspeaker as well. The extended frequency content and loudness should at some point cause the middle loudspeaker to stand out perceptually. It can be seen that in order to get a visible distribution peak for this direction, approximately a little less than one-half of the ERB-band signals must be routed from this loudspeaker (case 5). Only with more than one-half the frequency bands is the frontal direction notably prominent (case 6). It is also interesting that the number of prominent peaks beyond the confidence intervals does not exceed one, although there seems to be a notable peak in the lower frequencies, which seems to shift from −22.5° to 0°. K–S test results also support the original hypothesis: the overall distributions of cases 1 and 4 are similar, whereas cases 5 and 6 differ from case 1 but are similar to each other. Thus our first hypothesis is valid in this respect. Of these four cases, case 5 has notably more perceived loudspeakers than the others. This is logical in the sense that the middle frequencies do not completely overpower the lower and higher frequencies, or vice versa, and thus both are perceived.

The effect of abrupt changes in the localization cues can be seen from cases 7–9. The perceived number of loudspeakers compared to case 1 increases significantly in cases 8 and 9. As both cases 1 and 9 used one ERB-band signal per loudspeaker in the setup, the second hypothesis in Section 1.4 can be accepted, although the effect is not large. Interestingly case 7 is an exception to the pattern of increasing number of perceived loudspeakers, even though here most loudspeakers radiate two ERB-band signals. This is suspected to be caused by the fact that the important lower and high frequencies are presented from the same loudspeaker, and by the lack of abrupt changes in the cues in this case. As opposed to cases 1 and 7, cases 8 and 9 show no prominent peaks in the perceived spatial distribution, indicating that the segregation into clearly separable spatial objects was harder. The K–S test results, which indicate that cases 8 and 9 have similar, more uniform distributions and that case 1 differs from them, support the hypothesis.

Finally the lowest row in Fig. 4 shows the results for cases 1, 2, and 9 when using click train stimuli. In all three cases the perceived number of loudspeakers is significantly less than with the corresponding noise stimuli, de-

spite the fact that the click signals of different frequency bands were approximately 3 ERB wide and overlapping, as explained in Section 1.3. The increased frequency content would be expected to increase the perceived width compared to the noise cases. Furthermore, all distributions show only a single sharp peak. Especially case 9 is noteworthy: it was perceived notably wide (four loudspeakers) with the noise stimulus, but with the click stimulus there is no significant difference in width between cases 1, 2, and 9. It must be concluded that the click train stimulus is perceived significantly narrower than continuous noise. It seems that the subjects perceived more of a single point-like source in all click cases, regardless of the spatial configuration. This coincides with findings where the efficient spectral integration was confined to a narrow time window on the order of 30 ms [17]. Thus the frequency bands of the short click stimuli were integrated more than the bands of the noise stimuli.

3 BINAURAL MODELING

This section examines how well the current cross-correlation-based auditory modeling techniques can account for the results obtained from the listening tests. The test cases used in the listening tests (see Fig. 3) were simulated and processed computationally. The model output was then analyzed for its ability to predict the subjective spatial loudspeaker distribution in each of the 12 cases.

3.1 Auditory Model

The cross-correlation model used in this study is that implemented in Akeroyd [18]. The general structure of the auditory model is similar to the rest of the Jeffress-based models reviewed, for example, in [10]. It can roughly be divided into preprocessing, correlogram calculation, and possible postprocessing.

The preprocessing stage includes filtering the ear input signals to different critical bands using a gammatone filter bank [19]. In this analysis we used the density of one filter per ERB band covering the entire stimulus range. After the critical-band filtering the transduction of sound waves to neural impulses in the cochlea is simulated. The authors tested all transduction methods available in the toolbox, including relatively simple techniques as well as a more complex hair-cell model suggested by Meddis et al. [20]. It was found that simple half-wave rectification of the input signal produced the most adequate results in these cases in the sense that the peaks of the model output were the sharpest. This issue is discussed further in the following section. Thus this method is considered in the remaining analysis.

The neural output is used to calculate the three-dimensional cross correlogram, which gives the cross correlation as a function of frequency and internal delay. The delay represents the position of the output in the Jeffress delay line. Sounds arriving from outside the median plane (0° azimuth) have interaural differences in arrival time at the eardrums. Thus the internal delay value coincides with the azimuth direction of the sound. A typical use of the Jeffress auditory model is to determine the perceived direction by locating the maximum or average of the three-dimensional correlogram mesh.

3.2 Simulation and Discussion

As its input the auditory model used takes a two-channel signal that represents the ear inputs. The test cases used in the listening test were simulated by using head-related transfer functions (HRTFs) publicly available in the CIPIC HRTF database [21]. Solely measurements from subject 165 were used in this research. Different ERB-band signals in each test case were filtered with the HRTFs corresponding to the azimuth directions of the loudspeakers in the real test setup and summed. The levels of the simulated signals were adjusted to correspond to the sound level of the real test cases (70 dB, A-weighted).

Fig. 5 is an example correlogram of the simulated case 1. It can be seen that the peaks of the correlogram occur at larger internal delay values as the frequency increases. This coincides with the directions of the actual ERB-band signals in the real test case. However, since the subjects perceived the sound as emitting from fewer than four loudspeakers in the actual test case, the model seems to give predictions that are too accurate. If the maxima of the correlogram are interpreted as perceptually salient, the peaks in most frequency bands indicate that the corresponding direction should be prominent. Also, the model produces two peaks at the highest ERB bands due to aliasing—above 1 kHz the correlogram has maxima at both ±450 μs. This does not correspond well with the subjective results (for example, case 7). The reason for the aliasing phenomenon is due to the relationship between the sound wave length and the distance between the ears, which at high frequencies results in an ambiguous ITD cue. It should be noted that the actual auditory system includes many unknown higher level processes that might process the correlogram further and overcome these problems.

In the following we analyze the average cross correlogram across the entire frequency range utilized. In order to perform the comparison between subjective results and simulations, it was necessary to establish the internal delay values that correspond to each loudspeaker direction. To accomplish this, a wide-band (200–1179-Hz) Gaussian noise signal was simulated to emanate from the directions corresponding to each loudspeaker, and the maximum of the resulting average correlogram peak was interpreted as the internal delay value corresponding to the loudspeaker. The internal delay span between the centers of the outer loudspeakers of the setup was found to be −508 to 451 μs, that is, there is a slight asymmetry toward the negative delay values due to HRTF measurement inaccuracy.

The thin lines in Fig. 6 show the cross-correlogram curves averaged directly over the frequency channel outputs of the auditory model. The heavy lines illustrate an alternative method to be discussed in the following paragraphs. The subjective spatial distributions for each test case are given by bars corresponding to the 11 loudspeakers in the listening setup. The maximum values of the subjective and modeled distributions are normalized to be equal.

When investigating Fig. 6 it can be seen that the cross correlograms averaged over frequency (thin lines) are relatively flat and in most cases indicate only a single peak. Furthermore, the modeled distributions are rather similar in many cases, for example, 4, 5, 6, and 7. It is difficult to determine the perceptually salient loudspeaker directions with this method.

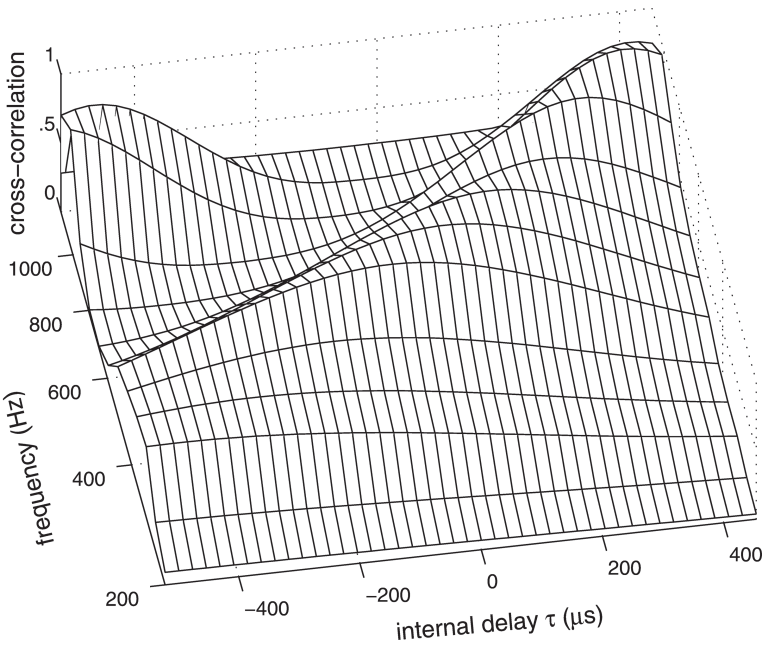

Fig. 5. Simulated auditory model output for case 1, showing cross correlation between two ear signals as a function of critical frequency band and internal delay.

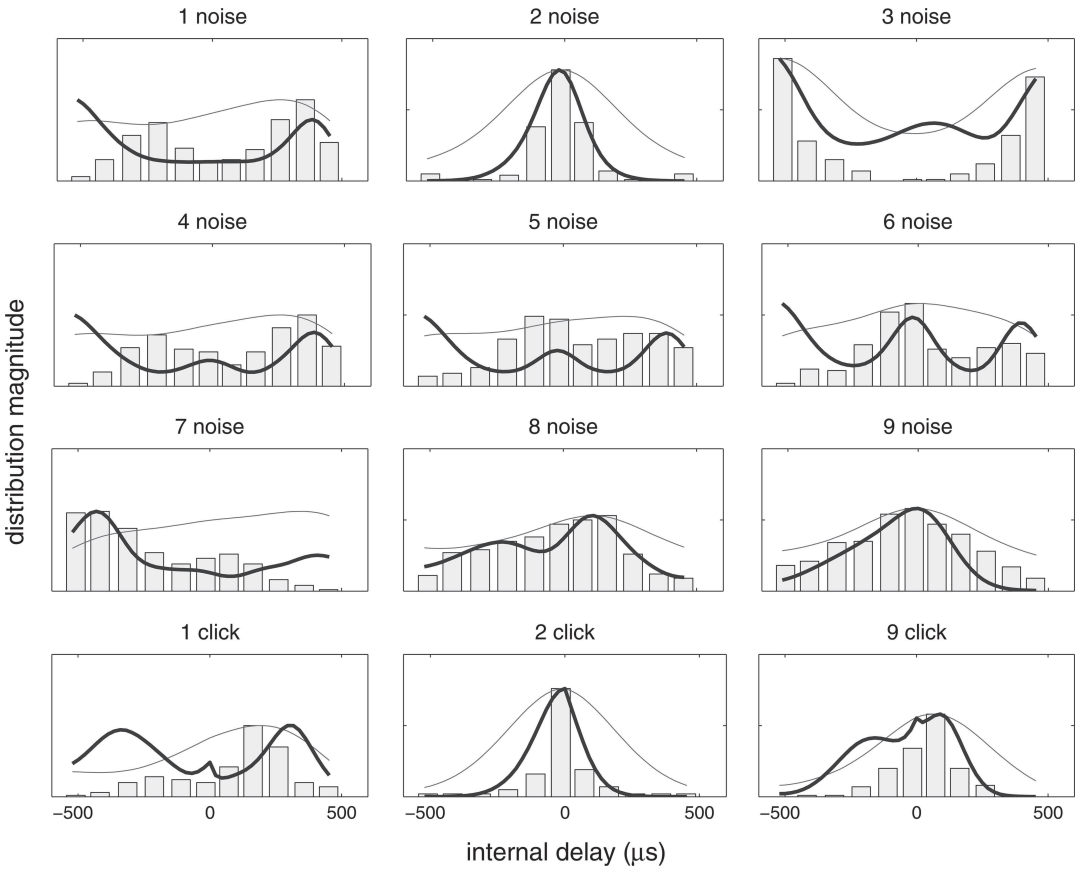

Fig. 6. Simulated auditory model outputs compared with listening test results. Cross-correlation-based model was used to calculate average correlogram over frequency. — basic output of model, ▬ alternative method where correlation curves of each frequency channel are sharpened prior to averaging. Subjective loudspeaker distributions are given by bars located in internal delay axis corresponding to loudspeaker azimuth angles.

We have implemented an alternative method with the goal of extracting the distribution features more precisely. The underlying problem with the previous method is that the correlogram peaks at low frequencies are very broad. This results in poor spatial resolution. Here the correlogram at each frequency band is sharpened by raising it to powers in the range of 1 to 25, depending on the frequency. A similar method has been used by Shear, only with a constant power factor at all frequencies [22]. This method produces equally sharp cross-correlogram peaks at all frequencies. The different frequencies are further weighted according to our previous research, where the highest and lowest frequencies were found more salient [15]—before calculating the across-frequency averaged correlogram in each case, the correlograms at the highest and lowest frequency bands were multiplied by 10. All parameters were adjusted iteratively and their exact values are omitted here, since the purpose is only to illustrate an example procedure of how the Jeffress model can be extended to describe complex listening situations similar to the current cases.

As can be seen from Fig. 6, the average correlogram curves produced by the modified method correspond to the subjective distributions in the sense that they generally have similar numbers of distinctive peaks. Especially the rise of the distribution around the 0° direction is identical in cases 4–6. The features of the subjective distributions are in some cases modeled well.

However, a problematic aspect of the results is that modeled cases 1, 4, 5, and 6 have a notable peak in the leftmost direction, which does not coincide with the subjective results. This phenomenon is the result of a general feature in cross-correlation-based models: at higher frequencies the correlogram produces peaks that are close to one another in the internal delay axis due to aliasing. This can be seen from Fig. 5 for case 1. At the two highest ERB bands the correlogram seems to suggest both directions −45° and 45°, although the sound is simulated to emanate from −45°. Similar to case 1, cases 4, 5, and 6 also present the highest band from the leftmost loudspeaker. With the alternative modeling method the high frequencies are given much weight. This phenomenon thus explains the prominent peaks in the predicted distributions in the −45° direction that do not coincide with the subjective results in these four cases.

Some model implementations have tried to rectify the aliasing problem by weighting the peaks more near the median plane [23]. However, here case 3 in particular would be problematic. In light of these results the suitability of cross-correlation-based auditory models for simulating the perception of complex multiple-source ensembles is questionable if further "higher level" processing is not applied to the model output. Recent results have even suggested that the cross-correlation mechanism is not physiologically plausible [24].

When using the click stimuli, simulated cases 1 and 9 deviate from the subjective results as being notably broader in azimuth. As mentioned, spectral integration is more efficient for short transient stimuli than for continuous sound. The fact that the modeling method used does not take this into account is the probable reason for the results. Additional inconsistencies in the click cases are also caused by the aliasing factors discussed in the previous paragraph.

4 SUMMARY

This paper studied the perception and modeling of complex sound events caused by multiple simultaneous ERB-band sources in different directions. A loudspeaker setup spanning the azimuth sector between −45° and 45° with 11 loudspeakers was constructed in an anechoic chamber. Eleven ERB-band signals in the range of 200–1179 Hz were routed to varying loudspeakers, and the subjects were required to indicate the loudspeakers they perceived as radiating sound. All subjects' responses form a spatial distribution in each test case. In addition, the average number of marked loudspeakers was examined as an indicator of perceived width.

The directional configuration of the critical bands was varied in 12 separate test cases designed to test specific hypotheses. When comparing the perceptual saliency of the middle versus the high/low frequency range it was found that the middle range had a notably smaller effect on the perceived directional distribution of sound.

Also, the order of the ERB-band signals in the loudspeakers was found to have a small but significant effect on the perceived width of the sound event. Cases with larger abrupt changes in the localization cues were perceived slightly wider than those in which the cues changed more moderately as a function of frequency. This confirms the existence of some form of interchannel processing between auditory critical bands in the auditory system.

Three cases were tested with both continuous noise and click train stimuli. It was found that the click cases were perceived notably narrower, mostly as radiating from one or two loudspeakers. However, the perceived number of loudspeakers radiating sound was not significantly greater than one-half the actual number, even in the noise cases. This indicates that some frequency bands were integrated together spatially.

Simulations of the test cases implemented with a cross-correlation-based auditory model were also examined and compared to the subjective results. The simulation results were not found suitably similar in all cases, mainly because the spatial distributions obtained from the simulations were too similar in different test cases. It is hypothesized that the cross-correlation model, as applied here, does not account for all mechanisms of auditory perception that are used to process stimuli such as presented in this study.

5 ACKNOWLEDGMENT

The authors wish to thank Juha Merimaa for his insightful comments. This work was supported by the Academy of Finland under project 105780 and by the Emil Aaltonen Foundation.

6 REFERENCES

[1] M. B. Gardner, "Image Fusion, Broadening, and Displacement in Sound Localization," *J. Acoust. Soc. Am.,* vol. 46, pp. 339–349 (1969).

[2] J. Blauert, *Spatial Hearing,* rev. ed. (MIT Press, Cambridge, MA, 1997).

[3] T. D. Griffths and J. W. Warren, "What Is an Auditory Object?," *Nature Neurosci.,* vol. 5, pp. 887–892 (2004).

[4] F. Rumsey, "Spatial Quality Evaluation for Reproduced Sound: Terminology, Meaning, and a Scene-Based Paradigm," *J. Audio Eng. Soc.,* vol. 50, pp. 651–666 (2002 Sept.).

[5] S. Zielinski, F. Rumsey, and S. Bech, "Subjective Audio Quality Trade-Offs in Consumer Multichannel Audio-Visual Delivery Systems. Part I: Effects of High Frequency Limitation," presented at the 112th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 50, p. 513 (2002 June), convention paper 5562.

[6] M. Barron and A. H. Marshall, "Spatial Impression Due to Early Lateral Reflections in Concert Halls: The Derivation of a Physical Measure," *J. Sound Vib.,* vol. 77, pp. 211–232 (1981).

[7] E. G. Boring, "Auditory Theory with Special Reference to Intensity, Volume, and Localization," *Am. J. Psychol.,* vol. 37, pp. 157–188 (1926).

[8] D. Perrot and T. Buell, "Judgments of Sound Volume: Effects of Signal Duration, Level, and Interaural Characteristics on the Perceived Extensity of Broadband Noise," *J. Acoust. Soc. Am.,* vol. 72, pp. 1413–1417 (1981).

[9] L. A. Jeffress, "A Place Theory of Sound Localization," *J. Comp. Physiol. Psychol.,* vol. 41, pp. 35–39 (1948).

[10] R. M. Stern and C. Trahiotis, "Models of Binaural Perception," in *Binaural and Spatial Hearing in Real and Virtual Environments,* R. H. Gilkey and T. R. Anderson, Eds. (Lawrence Erlbaum, Mahwah, NJ, 1997), pp. 499–531.

[11] C. Faller and J. Merimaa, "Source Localization in Complex Listening Situations: Selection of Interaural Cues Based on Interaural Coherence," *J. Acoust. Soc. Am.,* vol. 116, pp. 3075–3089 (2004).

[12] E. S. Sussman, "Integration and Segregation in Auditory Scene Analysis," *J. Acoust. Soc. Am.,* vol. 117, pp. 1285–1292 (2005).

[13] S. L. MacCabe and M. J. Denham, "A Model for Auditory Streaming," *J. Acoust. Soc. Am.,* vol. 101, pp. 1611–1621 (1997).

[14] V. Pulkki and T. Hirvonen, "Localization of Virtual Sources in Multichannel Audio Reproduction," *IEEE Trans. Speech, Audio Process.,* vol. 13, pp. 105–119 (2005).

[15] T. Hirvonen and V. Pulkki, "Localization and Perceived Width of Sound Sources with Frequency-Dependent Direction," *Acta Acustica—Acustica,* to be published.

[16] B. R. Glasberg and B. C. J. Moore, "Derivation of Auditory Filter Shapes from Notched-Noise Data," *Hear. Res.,* vol. 47, pp. 103–138 (1990).

[17] W. A. C. van der Brink and T. Houtgast, "Spectro-Temporal Integration in Signal Detection," *J. Acoust. Soc. Am.,* vol. 88, pp. 1703–1711 (1990).

[18] M. A. Akeroyd, "Binaural Cross-Correlogram Toolbox for MATLAB," http://www.lifesci.sussex.ac.uk/home/Michael_Akeroyd/download2.html (2001).

[19] M. Slaney, "An Efficient Implementation of the Patterson–Holdsworth Filter Bank," Tech. Rep. 35, Apple Computer (1993).

[20] R. Meddis, M. Hewitt, and T. M. Shackleton, "Implementation Details of a Computational Model of the Inner-Haircell/Auditory-Nerve Synapse," *J. Acoust. Soc. Am.,* vol. 87, pp. 1813–1816 (1990).

[21] 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 Electroacoustics* (New Paltz, NY, 2001 Oct.), pp. 99–102.

[22] G. D. Shear, "Modeling the Dependence of Auditory Lateralization on Frequency and Bandwidth," Master's thesis, Dept. of Electrical and Computer Engineering, Carnegie-Mellon University, Pittsburgh, PA (1987).

[23] R. M. Stern and G. D. Shear, "Lateralization and Detection of Low-Frequency Binaural Stimuli: Effects of Distribution of Internal Delay," *J. Acoust. Soc. Am.,* vol. 100, pp. 2278–2288 (1996).

[24] B. Grothe, "Sensory Systems: New Roles for Synaptic Inhibition in Sound Localization," *Nature Neurosci.,* vol. 4, pp. 540–550 (2003).

THE AUTHORS

T. Hirvonen

Toni Hirvonen was born in Vaasa, Finland, in 1976. He received his M.Sc. (E.E.) degree from Helsinki University of Technology (TKK), Espoo, Finland, in 2002.

Since 2003 he has been working in the TKK Laboratory of Acoustics and Audio Signal Processing, conducting postgraduate research and studies. His main research topics are spatial hearing, auditory modeling, and audio reproduction.

The biography of Ville Pulkki was published in the 2006 January/February issue of the *Journal.*