Localization with binaural recordings from artificial and human heads

Minnaar, Pauli; Olesen, Søren Krarup; Christensen, Flemming; Møller, Henrik

Published in:
Journal of the Audio Engineering Society

Publication date:
2001

Localization with Binaural Recordings from Artificial and Human Heads*

PAULI MINNAAR, *AES Member,* **SØREN KRARUP OLESEN, FLEMMING CHRISTENSEN,** *AES Member,* **AND HENRIK MØLLER,** *AES Member*

Department of Acoustics, Aalborg University, DK-9220 Aalborg, Denmark

Previous experiments have shown that localization with binaural recordings made with artificial heads is inferior to localization in real life and also to localization with recordings made in the ears of selected humans These results suggest that artificial heads may be improved. A new experiment was made, employing recordings from two human heads and seven artificial heads some of which had been developed recently. The listening room setup from previous experiments was used and 20 listeners participated. As in the earlier experiments, more directional errors were seen with binaural recordings than in real life. A clear learning effect was seen over five days, emphasizing the need of a balanced experimental design. The new results show that artificial heads are still not as good for recording as a well-selected human head, although some of the new heads come close The accumulated results from the present and four earlier studies provide sufficient statistics to conclude that there are significant differences between some currently available artificial heads.

0 INTRODUCTION

The aim of binaural sound reproduction is to provide at the eardrums of a listener the sound pressure that would have been there in a listening situation. Assuming that air-conducted sound at the eardrum is the only input to the hearing, the complete auditory experience will be reproduced. Thus all information about an acoustical event, including spatial aspects, can be obtained with two channels only. Binaural recordings can be made with small microphones in the ear canals of a person. Alternatively an artificial head may replace the person.

Binaural *recording* is distinct from the method of generating binaural signals by convolving a monophonic signal with head-related transfer functions (HRTFs), which is referred to as binaural *synthesis*. Binaural recording can only be used to reproduce acoustical events that have taken place in reality somewhere, sometime, whereas binaural synthesis can be used to create imaginary events, for instance, in virtual reality applications. Binaural recordings are also restricted with respect to interactive, dynamic cues. Such cues cannot be used, since the listener has no control over head position and orientation during the recording. As a consequence, the head must be kept stationary during reproduction.

Since it is inconvenient to make binaural recordings in the ears of people, an artificial head is more often used in practice. An artificial head is sometimes referred to as a manikin (mannequin) or a dummy head. It may or may not include a torso. The physical shape of the artificial head should be derived so that the recordings reproduce the acoustical events as accurately as possible for a population of listeners—presumably by modeling the acoustics of an average or typical human. Unfortunately measurements of HRTFs made on people show large differences between individuals; see, for example, Møller et al. [1]. Therefore the design of an artificial head for a population is the challenging problem of finding acoustically relevant features to be maintained.

A well-designed artificial head is needed if binaural recordings are to find widespread application. This has been recognized by several companies and research institutions that have developed artificial heads. The heads must be evaluated in listening experiments in which recordings are made and played back under strictly controlled circumstances. The heads should not only be compared to each other but also to human heads and to real life. Since considerable inter- and intrasubject

* Presented at the 105th Convention of the Audio Engineering Society, San Francisco, CA, 1998 September 26–29
Manuscript received 2000 June 26

variance is generally seen in the results, many listeners and repetitions of stimuli are required, and statistical tests need to be employed to permit comparisons.

0.1 Previous Work

The ability of a binaural sound reproduction system to convey spatial information is most often tested by localization experiments. Traditionally tests have been made under anechoic conditions, but more recently tests have also been made in "ordinary" rooms. The binaural recordings are played back through headphones, or through loudspeakers employing crosstalk cancellation. Listeners are required to keep their heads still during sound presentation. The localization performance of listeners is studied when listening to the binaural recordings, and sometimes also when listening in real life to the sound field that was recorded.

A substantial amount of literature exists on localization with binaural recordings. Møller et al. present thorough reviews of five studies with recordings from human heads [2] and of 18 studies with recordings from artificial heads [3]. Therefore an account of the general literature will not be repeated here. However, since the experiment described in this paper follows directly from the experiments reported by Møller et al. [2]–[4], their main results are summarized here.

All experiments employed a setup of 19 loudspeakers placed around a chair in a standard listening room. A signal was input to each of the loudspeakers in turn, and binaural recordings were made with human and artificial heads in the chair. These recordings were played back over equalized headphones to a listener whose head was in the same location as the recording heads had been. Furthermore, real-life experiments were made where sounds were presented directly to the listener through the loudspeakers. In each case the listener identified the loudspeaker from which he or she perceived the sound.

In the first investigation [2], recordings were made in eight subjects' ears. Everyone's recording was then played back to him- or herself through headphones equalized on their own ears (individual recordings). There was no significant difference between localization in real life and when listening to the recordings. When the subjects listened to the recordings made in the ears of other people (nonindividual recordings), localization was generally poorer. The study therefore suggested that localization is best when using our own head and ears for recording.

In the subsequent investigation [4], 20 subjects listened in real life and to binaural recordings made with 30 human heads. As in the previous experiment, the localization was poorer when listening to nonindividual recordings. The recording heads were ranked according to the overall number of errors for the group of listeners. Recordings from the head that gave the best performance (AVH) were used in a separate experiment, and the good localization with this head was confirmed. This indicated that if nonindividual recordings originate from a carefully selected human head, it is possible to reduce the number of errors substantially.

In the next investigation [3], eight subjects listened to recordings made through the built-in microphones of eight artificial heads. Furthermore, 20 subjects listened to recordings with microphones mounted flush with the entrances to the blocked ear canals of 10 artificial heads (including the previous eight). The localization with the recordings was significantly poorer than in real life. No significant difference was found between results with the two microphone techniques. This demonstrates that the recording point does not influence the localization when headphone equalization is done for the recording point used.

The same 20 listeners participated in the experiment with the 30 human heads [4] and the 10 artificial heads [3]. A comparison of the results showed that 60% of the human heads were better for recording than even the best artificial head. It was concluded that this finding should encourage the design and production of better artificial heads.

0.2 Heads Included in the Present Investigation

Since the study by Møller et al. [3] a number of new artificial heads have become available, and existing ones may have been improved. It is the aim of this study to compare seven of these artificial heads. Furthermore they will be compared to two human heads selected from Møller et al. [4]. Some of the artificial heads are commercially available whereas others only exist as a few samples used in research laboratories. The recording heads are described briefly in the following text.

0.2.1 Knowles Electronics Inc.—KEMAR

The Knowles Electronics manikin for acoustic research (KEMAR) is a well-known artificial head, much used in research. It has ear simulators according to IEC 711 [5] and ANSI S3.25 [6], and it conforms to geometrical and acoustical requirements of IEC 959 [7] as well as geometrical and acoustical (sound pickup) requirements of ITU-T P.58 [8]. In the experiment reported by Møller et al. [3] the manikin was tested with four different pinnae (KEMAR 1–4). Since KEMAR 2 (with DB065/066 pinnae) is the most common, it was employed in the current experiment. Brüel & Kjær 4158 and 4159 ear simulators with preamplifiers (for right and left ears, respectively) and a Brüel & Kjær Nexus 2690 two-channel conditioning amplifier were used.

0.2.2 Georg Neumann GmbH—KU100

The KU100 artificial head from Neumann, like its predecessors KU80 and KU81, finds most of its application in the recording industry. It is the only head in this investigation without a torso. It has a fixed set of pinnae and comes with built-in microphones and preamplifiers. The microphone signals were fed into a two-channel Rostec LMA 4 amplifier to obtain a line-level signal.

0.2.3 Brüel & Kjær A/S—4100

The Brüel & Kjær 4100 artificial head has the same external geometric shape as the Brüel & Kjær 4128 and 5930 tested in previous experiments [3]. However, the

ear canals, microphones, and preamplifiers are different, and the 4100 has a jacket and an adjustable neck ring. During recording the jacket was used and the neck was in an upright position. The built-in microphones and preamplifiers were used with a Brüel & Kjær Nexus 2690 conditioning amplifier. In the 4128 version, the Brüel & Kjær head has ear simulators according to IEC 711 and ANSI S3.25, and it conforms to the acoustical requirements of IEC 959 (but not the geometrical ones), and the geometrical and acoustical requirements of ITU-T P.58.

0.2.4 Head Acoustics GmbH—HMS II

The Head Acoustics artificial head that forms part of the HMS II measurement system is well known in the engineering community. It has a stylized (mathematically describable) head and pinnae. In the version used, the HMS II has ear simulators according to IEC 711 and ANSI S3.25, and it conforms to the acoustical requirements of IEC 959 (but not the geometrical ones), and the geometrical and acoustical (sound pickup) requirements of ITU-T P.58. The signals from the built-in microphones and preamplifiers were fed into a Brüel & Kjær Nexus 2690 conditioning amplifier. The system has an option of equalization during recording which was not used.

0.2.5 Cortex Electronic GmbH—MK1

The MK1 artificial head from Cortex Electronic has an articulated neck and hips. The external shape of the head, torso, and pinnae follows the geometrical descriptions in IEC 959 and ITU-T P.58, but since the head does not have ear simulators, it does not conform to the remaining requirements of the documents. The MK1 has built-in microphones, amplifiers, and analog-to-digital (A/D) converters, and provides a digital signal for recording. Equalization for a specific Sennheiser HE 60 electrostatic headphone, supplied by the manufacturer, was used during recording, implying that recordings were preequalized for this headphone.

0.2.6 Aachen University—ITA

The artificial head developed at the Institute of Technical Acoustics at Aachen University in Germany, ITA, has a hard plastic head and shoulders and human-like pinnae. The built-in microphones and amplifiers were used to obtain a line-level signal. The built-in equalization and A/D converters were not employed during recording.

0.2.7 Aalborg University—VALDEMAR

The artificial head developed at our own laboratory is named VALDEMAR, after the Danish inventor Valdemar Poulsen who invented and patented the magnetic recording principle in 1898. The head and the torso have been designed from acoustical measurements on 40 humans and from anatomical data. The pinnae are casts of a human pinna (subject DOL, included in this study; see Section 0.2.8). Small electret microphones (Sennheiser KE 4-211-2) were inserted into earplugs with the diaphragms facing outward, and mounted flush with the ear canal entrances. A custom-made preamplifier was used. (A report on the construction of the head is under preparation [9]).

0.2.8 Human Heads AVH and DOL

The two people used as recording heads are denoted by AVH and DOL. In the study by Møller et al. [4], where 20 people listened to recordings made with 30 people's heads, recordings of AVH gave the lowest number of median-plane errors, and recordings from DOL ranked fourth. In an inspection of acoustical measurements on 40 human pinnae, DOL's pinna was found to best represent the characteristics of human pinnae. (A report of details of the measurements is under preparation [10].) As in the case of VALDEMAR, Sennheiser KE 4-211-2 microphones were mounted flush with the blocked ear canal entrances. AVH and DOL looked straight ahead and sat perfectly still during the recording.

1 METHODS

The experiments were carried out in a listening room, where 19 loudspeakers were located around the listener. Short segments of speech or noise were presented to the listener either directly from the loudspeakers or indirectly as a binaural recording made in the same setup and reproduced by means of headphones. In both cases the loudspeakers were visible to the listener who had to keep the head still during sound presentation. The task of the listener was to identify the loudspeaker from which he or she perceived the sound. The experiments, therefore, did not aim at measuring absolute localization judgment, but rather the ability to identify a sound source in a 19-alternative forced-choice task. Since the experiments were done in a normal listening room (as opposed to an anechoic environment), the reflections from the room boundaries particular to each loudspeaker were available to aid localization.

1.1 Listening-Room Setup

The setup was made in a listening room complying with IEC 268-13 [11], but without a carpet. A chair was placed 2 m from the back wall facing down the length of the room. The chair was adjusted for each listener to obtain the same position for the middle of the head. This point served as reference for the locations of loudspeakers, which were placed on stands around the room with their main axis pointing to the reference point.

The loudspeaker units were 70-mm Vifa M10MD-39 drivers mounted in 155-mm-diameter hard plastic balls. Thirteen of the loudspeakers were 1 m from the reference point at the following locations: straight in front, in front up 45°, in front down 45°, straight behind, behind up 45°, behind down 45°, left, left up 45°, left down 45°, right, right up 45°, right down 45°, and above. In addition, loudspeakers were placed straight in front at distances of 1.7, 2.9, and 5 m, as well as 45° to the right at distances of 1, 1.7, and 2.9 m. For the two directions where loudspeakers were behind each other (front and 45° to the right), the loudspeakers were slightly dis-

placed vertically by less than the minimum audible angle in order to reduce disturbance of the direct sound (as described by Nielsen [12], [13]). The setup of the loudspeakers around a listener is shown in Fig. 1.

1.2 Procedure

As mentioned, the listeners sat in the setup and listened to the sound either directly from the loudspeakers or as binaural recordings played back through headphones. Except for the headphone used when listening to the binaural recordings, the setup was the same at all times. The stimuli were played automatically by the control room computer, which also registered all the responses during a session. No feedback was given to the listeners.

A small "traffic light" prompted to the listener prior to each stimulus. During a stimulus the listener had to look straight ahead, keep the head still, and listen to identify the loudspeaker from which he or she perceived the sound. A response was submitted by pressing with a pen on a 200- by 210-mm electronic tablet holding a schematic drawing of the loudspeaker setup (see Fig. 1).

The experimenter monitored the position of the listener's head by means of two cameras in the listening room. In general, listeners managed to replace the ears very accurately before each stimulus. The experimenter interrupted the session if head movements were detected during stimuli, or if the listener pressed a stop button, such as to report of an unintended response. Interruptions occurred very rarely, and the experiment was always resumed after a short communication with the subject.

1.3 Stimuli

Two stimuli were used—speech and noise. The speech was a 2.2-s sentence from a female speaker, recorded in an anechoic room at a distance of 1 m. Recording equipment included a Brüel & Kjær 4145 1-inch microphone, a Brüel & Kjær 2660 preamplifier, a Brüel & Kjær 2636 measuring amplifier, and a TC electronic finalizer, used as A/D converter at a sampling rate of 48 kHz and interfacing the AdB Digital Multi!Wav audio I/O board of the control room computer.

The noise signal was white noise (hardware generated and A/D converted as the speech) with a duration of 1 s and faded in and out over 0.05-s intervals.

The speech and white noise signals had been chosen in a pilot experiment, which also included pink noise (faded in and out) and unfaded white noise. Very little difference in localization, if any, was observed between these signals.

1.4 Real-Life Playback

Signals were played back through a Tracer Technologies Big DAADI D/A converter and a Pioneer A-616 power amplifier (modified to maintain a unity gain). The computer controlled a relay unit that channeled the signal to one of the loudspeakers at a time.

The loudspeakers were equalized between 300 Hz and 19 kHz with respect to their mean on-axis free-field response (average of amplitude in Pa/V) by means of a 128th-order minimum-phase FIR filter. For the speech, scaling was made to achieve a unity gain through the system as a whole, thus ensuring that the direct sound from a loudspeaker corresponded to that of the person speaking at natural level from the same location in the listening room. For the white noise, scaling was made to achieve the same digital root-mean-square value after equalization as that of the speech signal. The filters were applied off-line to the recorded sound files to produce new "preequalized" files that were used during playback.

1.5 Making Binaural Recordings

Recordings were made of the stimuli from each of the loudspeakers with each of the artificial and human heads.

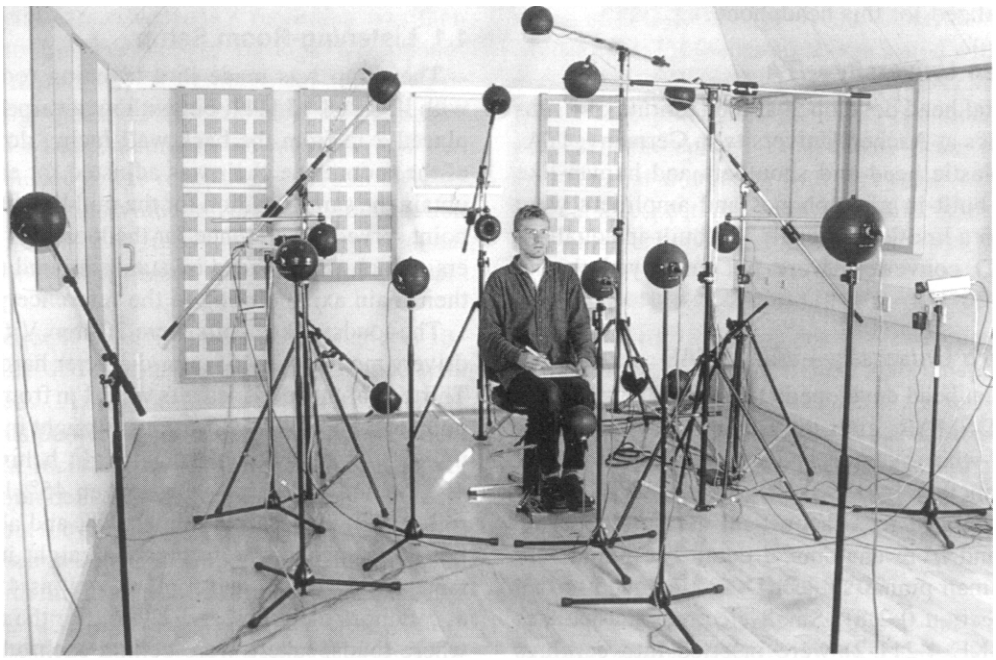


Fig. 1 Experimental setup of loudspeakers in listening room. Ears of recording heads and listeners were in same position.

A recording head was seated in the chair, which was adjusted to have the head at the reference position. In cases where the torso was too short, stacks of paper were placed on the chair. Since the KU100 head had no torso, it was mounted on a microphone stand fixed to the backrest of the chair.

The stimulus was presented to one of the loudspeakers exactly as in real-life playback, and the resulting sound was recorded by the microphones in the ears of the head. The line-level signal from the head went through the A/D converter (TC electronic finalizer) to the control room computer. (For the MK1 head the signal was already digital and went directly to the computer.)

The following steps were taken by a dedicated computer program: select a loudspeaker through the relay unit, start recording the microphone signal, send a stimulus signal to the loudspeaker 0.1 s later, stop the recording 0.4 s after the end of the stimulus (to ensure that the reverberation tail was recorded). This was done for the loudspeakers in turn, giving 19 two-channel recordings per head. Each recording was stored as a separate sound file sampled at 48 kHz.

1.6 Playback and Equalization of Binaural Recordings

A Sennheiser HE 60 electrostatic headphone with a Sennheiser HEV 70 headphone amplifier was used in all experiments. The amplifier, which was modified to have a fixed gain, was fed with the line-level signal from the Tracer Technologies Big DAADI D/A converter.

The playback system should provide at the eardrums the sound pressures that would have been there had the listener been in the original sound field. To meet this requirement it is not sufficient to present simply the recorded signals through the playback chain. It was shown by Møller [14] that the total transmission should include equalization for 1) the transfer function of the recording microphone and 2) the transfer function of the headphone measured at the point in the ear canal where the recording is made. These two transfer functions were measured together by measuring the electrical transfer function from headphone terminals to microphone output terminals while the headphone was positioned at the recording head.

For each head and ear, the logarithmic mean of the magnitudes of five measurements (with headphone repositioning) were used in a Yule–Walker design [15] of a 32nd-order IIR filter. The gains of the filters were adjusted to give a total gain of unity, thus producing sound at a natural level. The filters were applied off-line to the recorded files to produce new files that were used directly during playback.

The procedure differed slightly for the recording heads VALDEMAR, AVH, and DOL. Instead of measuring the headphone on the recording head it was measured on 20 humans with the recording microphones placed as during recording (blocked ear canal entrance). For both the left side and the right side, the logarithmic means of the magnitudes of five measurements on each of 20 humans (100 measurements per side) were used to determine the equalization and the gain. Thus the filters used for these three recording heads were the same.

The procedure was also different for the MK1 head. These recordings were preequalized as determined by the manufacturer for a specific set of headphones of the same make and type, which was used in the experiment. For this head the off-line equalization filters only equalized for the (small) difference between the headphone actually used in the experiment and the one for which the supplier had made the preequalization (both measured five times on the head).

1.7 Listeners

Twenty listeners (10 male and 10 female) participated. They were between 20 and 30 years old and they all had controlled normal hearing. None of the listeners had participated in the localization experiments previously, and none had been used for measuring the headphone transfer functions used for the equalization.

1.8 Experimental Design, Main Experiment

Ten different playback conditions were used: real life and binaural recordings from each of nine recording heads. Each of these conditions was combined with a stimulus type (speech or white noise) to obtain a session. In each session the 19 loudspeakers were repeated five times, and the order of these 95 stimuli was randomized for each listener.

Each listener was exposed to all 20 combinations of playback condition and stimulus type. The order was determined from a Latin-square crossover design which balances not only order effects but also carryover effects (see, for example, [16]).

The duration of a session was approximately 10–12 minutes, and for each listener the 20 sessions were presented over five days, four on each day. Two listeners attended at a time—alternating between sessions in the listening room and breaks of similar duration in a nearby waiting room.

1.9 Experimental Design, Learning Experiment

In order to permit an evaluation of learning effects, extra real-life sessions were inserted for all listeners just prior to every day's program of the main experiment. Only speech was used for these sessions. Including the learning experiment, every visit to the laboratory took approximately two hours.

1.10 Familiarization

Prior to the days of the main and the learning experiments, one day was used for audiometric testing and familiarization with the experiments to follow. A written instruction introduced the listeners to the general procedure. Then the experimenter pointed out the positions of the loudspeakers in the listening room, and listeners practiced submitting responses for real-life and headphone listening. The experimenter was present in the listening room during this. The day ended with two listening sessions similar to those of the real experiments

(each consisting of 95 stimuli): a real-life and a binaural, the order being balanced across listeners. The responses from the familiarization day were not used.

2 RESULTS AND DISCUSSION

Results are shown in stimulus and response plots, and statistical analyses are made for errors, divided into four error categories. Sound sources that produce the same interaural time difference (for the direct sound) are positioned on an approximate cone, called a cone of confusion. If a response is not on the same cone as the stimulus, it is termed *out-of-cone* error. When errors are made by confusing directions on the same cone, it is termed *within-cone* error, except for that special "cone" formed by the median plane, in which case it is called *median-plane* error. A response given in the same direction as the stimulus, but at an incorrect distance, is a *distance* error.

The number of errors in a certain category follows a binomial distribution. Fisher–Irwin tests [17] are used to test the null hypothesis that errors in a certain category observed in two conditions come from the same distribution. Two conditions are said to be significantly different when the null hypothesis is rejected on either a 5%, 1%, or 0.1% level. Also analyses of variance are carried out for those results that fulfill the preconditions for such analyses (error percentages for median-plane and distance errors).

2.1 Effect of Stimulus

Initially the results from the speech and white noise stimuli were compared. Median-plane errors were not significantly different for any condition, that is, real life or any recording head (two-sided Fisher–Irwin tests, 5% significance level). For the other error categories the same was seen for almost all conditions. Therefore the data for the speech and white noise stimuli have been pooled—giving 3800 responses for every condition.

These results correspond well with the results in the pilot experiments used to select stimuli, where little or no differences were observed between four stimuli (see Section 1.3).

2.2 Main Experiment

The "raw" localization data for each of the 10 conditions are shown in Figs. 2–6. Answers are given as circles in a 19 by 19 matrix with the stimulus on the abscissa and the response on the ordinate. The area of each circle is proportional to the number of answers for the particular combination of stimulus and response.

In the case of real life [Fig. 2(a)] most of the answers lie on the diagonal, indicating correct answers. Some median-plane errors are seen, though, such as a stimulus of BACK HIGH that results in an ABOVE response, and FRONT LOW and FRONT (1 M) that are perceived as BACK. Also distance errors occur in real life; for instance, the position −45° (1.7 M) confused with −45° (2.9 M) and FRONT (2.9 M) given as FRONT (1.7 M) or (5.0 M).

A striking increase in the number of errors is evident in the plots for the recording heads. This applies to all heads and to nearly all directions. Only the directions LEFT and RIGHT are nearly always correctly perceived. It is believed that this is due to a dominating role of the interaural time difference, which leaves no ambiguity for these directions.

Errors such as LEFT HIGH perceived as LEFT (out-of-cone error) and LEFT HIGH perceived as LEFT LOW (within-cone error) are typical for the recording heads, but nearly never seen in real life. Also many median-plane errors occur with the recording heads, for instance, FRONT LOW perceived at BACK as well as a variety of confusions between FRONT HIGH, ABOVE, and BACK HIGH. The reader is encouraged to study more directions in detail.

Table 1 lists the errors for every condition, split up into the four error categories. Errors are given as num-

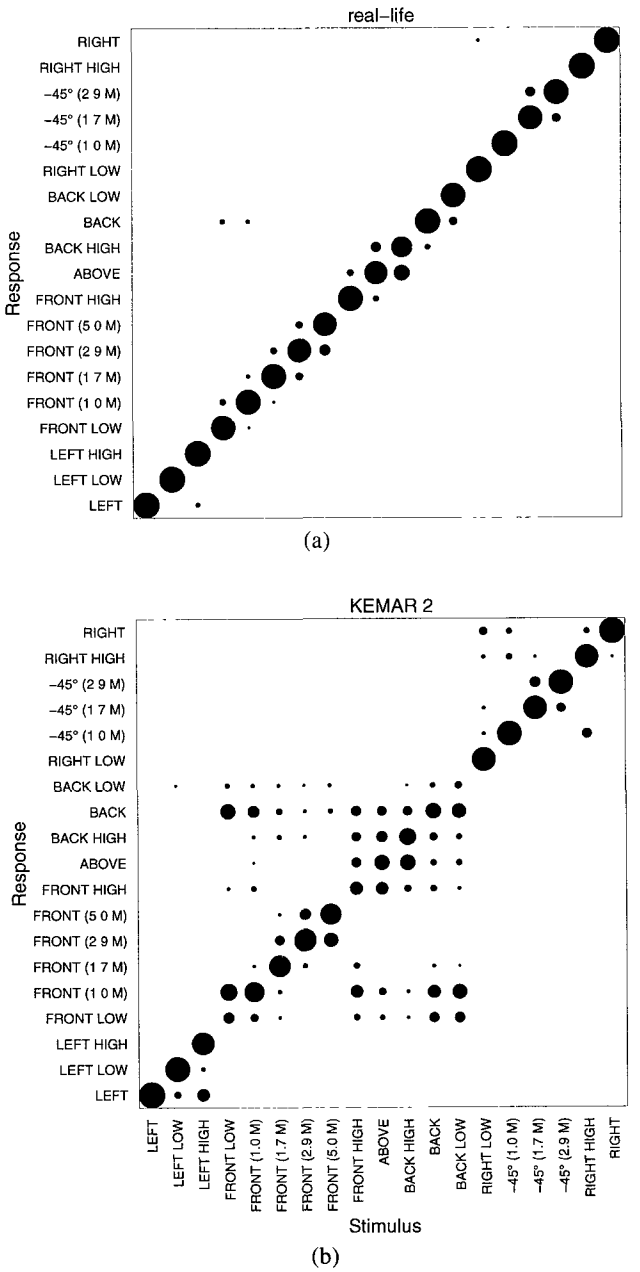


Fig 2 Main experiment: real life and KEMAR 2 (3800 stimuli in each frame). Area of each circle is proportional to number of answers given for particular combinations of stimulus and response.

bers and as a percentage of the potential number of errors in the category. One-sided Fisher–Irwin tests show that results from all recording heads are significantly different from real life on the 0.1% level for all directional errors and at least on the 5% level for distance errors.

The data in Table 1 are displayed graphically in Fig. 7. In the median plane clearly the lowest number of errors occurred for real-life listening followed by the two human heads. Among the artificial heads VALDEMAR and ITA gave the lowest number of median-plane errors, followed by KU100. Notice, though, that KU100 shows the largest number of out-of-cone and within-cone errors. This observation is not surprising, since it is the only head without shoulders and torso.

Distance errors are remarkably similar independent of the condition, real life included. This suggests that the directional filtering of the recording head is less critical for the perception of distance, and that other cues play an important role. Nielsen [12], [13] showed that the room has a significant impact on the perception of distance and proposed that the ratio between direct sound and sound reflected from the room boundaries plays a central role. As the source moves away from the listener, this ratio generally decreases. It is natural to assume that to a large extent the listener is able to distinguish between direct sound and reflections, even if the artificial head does not provide a perfect reproduction of direction. Thus the perception of distance is only marginally affected by imperfections of the artificial head.

2.3 General Validity of Results

In a strict sense the Fisher–Irwin test is only valid for the particular group of listeners who participated in the

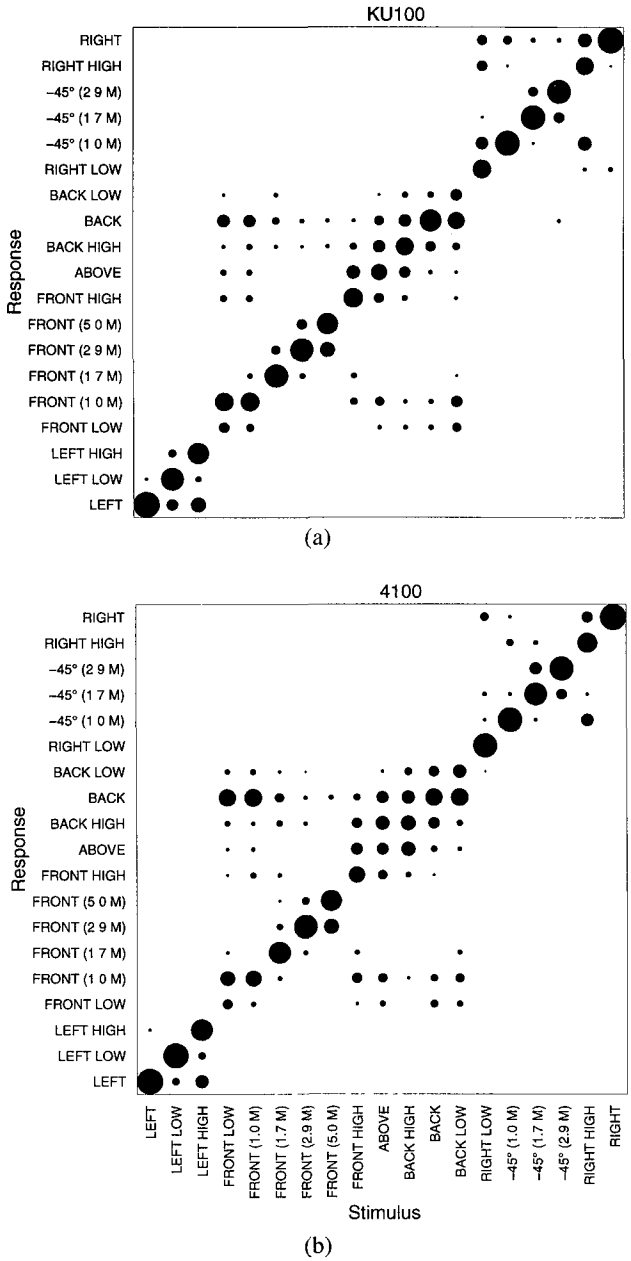


Fig. 3. Main experiment: KU100 and 4100 (3800 stimuli in each frame).

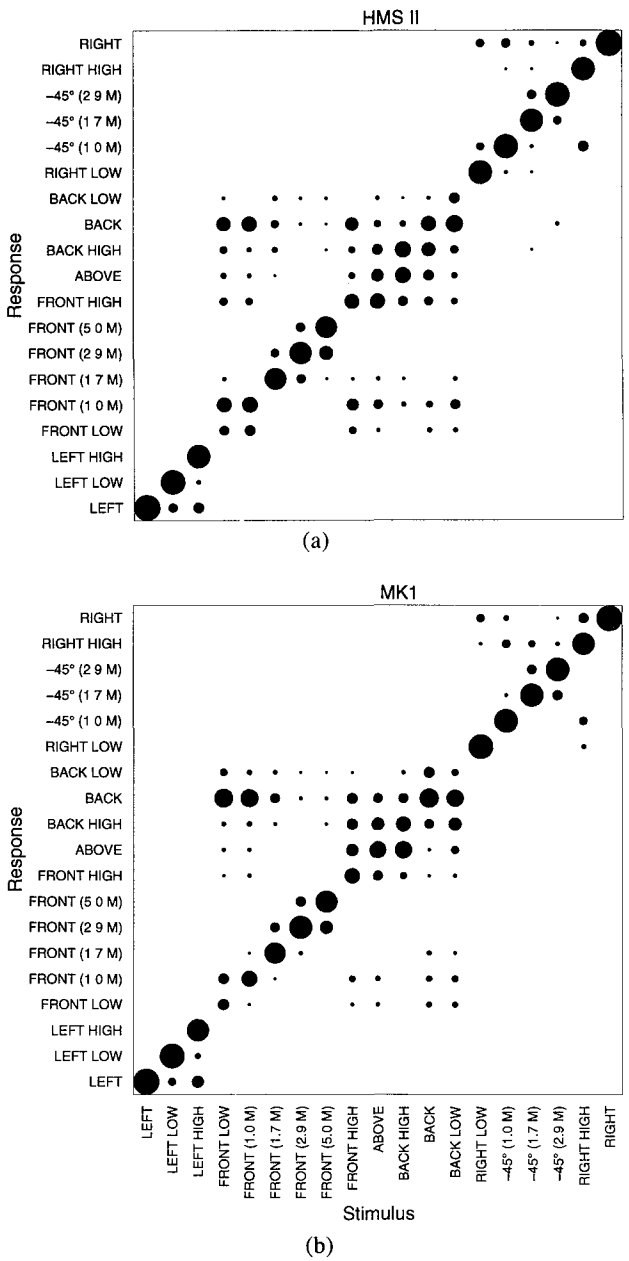


Fig 4 Main experiment: HMS II and MK1 (3800 stimuli in each frame).

experiment. In order to evaluate the results with regard to their general validity for a population, analyses of variance (ANOVA) were also carried out. Since the analyses were made mainly to test a possible discrimination between recording heads, real-life data were not included.

A precondition for an analysis of variance is that all conditions in the experiment produce results that follow normal distributions with the same variance. Bartlett's test of homogeneity of variance [18] was made for error percentages in each category, and the hypothesis of equal variance was accepted at the 5% level for median-plane and distance errors. It was rejected for out-of-cone and within-cone errors—not unexpectedly, since there are only few of these errors for some conditions, and when the mean error percentage is close to zero, the variance will be as well. As a consequence of this lack of homogeneity of variance, analyses of variance were made only for median-plane and distance errors.

Error percentages were analyzed in a two-way analysis of variance with recording head as fixed factor and listener as random factor. This corresponds to a one-way analysis with repeated measures. The outcome of the analyses is given in Table 2. For the median-plane errors the p value for the recording head factor is very small, indicating that the results did not come from the same distribution, that is, recording heads are significantly different. On the other hand, for distance errors the null hypothesis is not rejected, indicating that the recording heads offer comparable distance localization. The p value for the listener factor is very low for both types of errors, clearly indicating that differences between listeners are substantial.

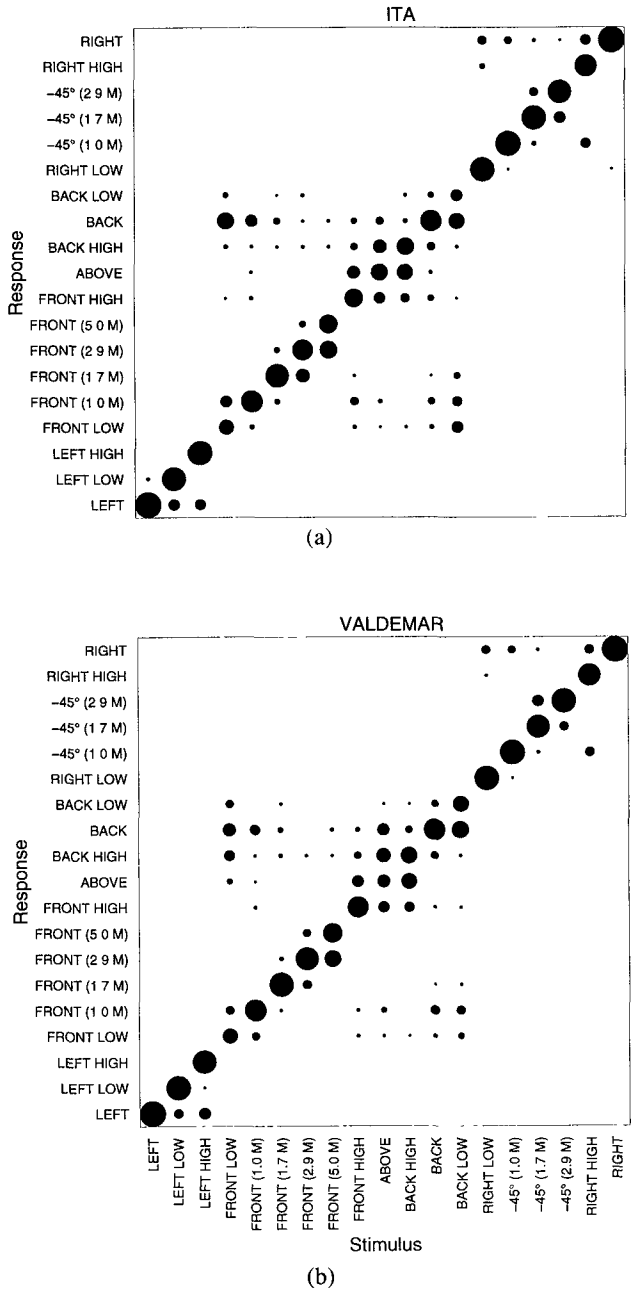


Fig. 5. Main experiment: ITA and VALDEMAR (3800 stimuli in each frame).

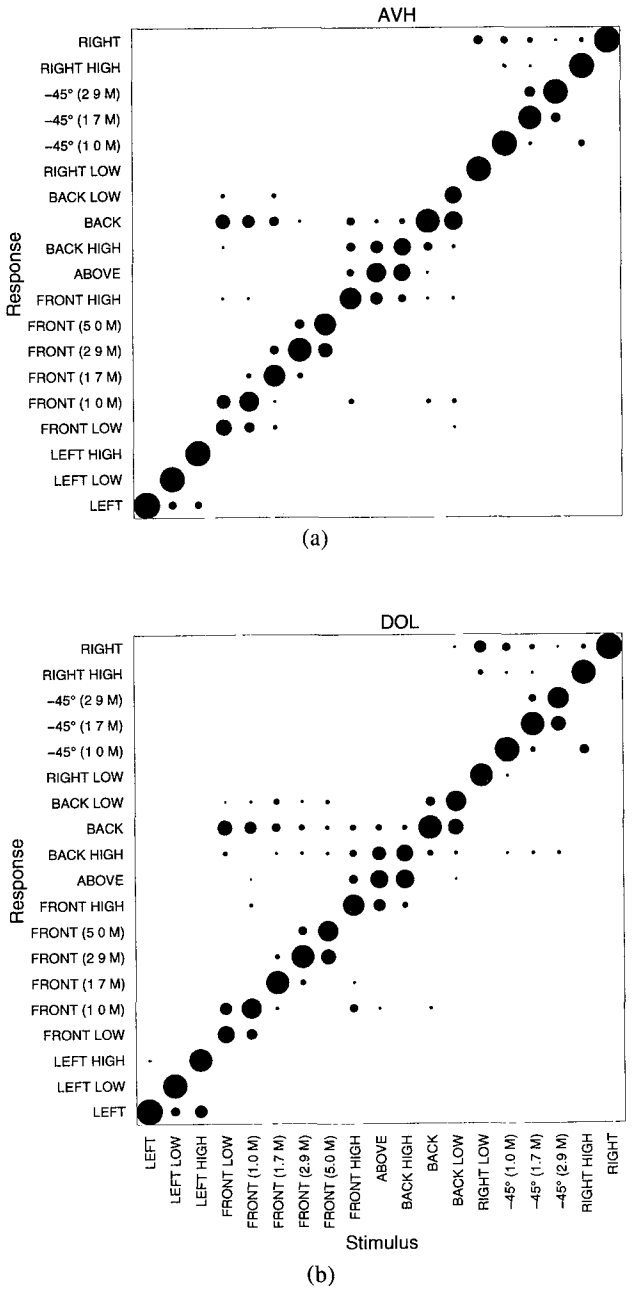


Fig. 6. Main experiment: AVH and DOL (3800 stimuli in each frame).

2.4 Learning

The answers collected during all five real-life sessions in the learning experiment are shown in Fig. 8. The same kind of errors are seen as for the real-life condition in the main experiment. However, there are slightly more errors in the learning experiment. This is most likely explained by the fact that the learning session was always the first of the five sessions on a day.

The distribution of errors between categories is shown in Table 3. In particular the median-plane errors show differences between days, and a graph of these errors is given in Fig. 9(a). There is a general decrease in errors as time progresses, although the decrease is not completely monotonic.

Since a balanced design was used in the main experiment, the errors made on individual days can be accumulated to inspect the effect of learning. This is shown in Fig. 9(b) for the median-plane errors. Clearly the number of errors decreases with time also here. Furthermore,

Table 1. Main experiment: errors in number and percentage of potential number of errors (in parentheses at the bottom of each category) *

	Out of Cone	Within Cone	Median Plane	Distance	Overall
Real life	0 3% 11	0.1% 2	9.7% 193	9.9% 138	9.1% 344
KEMAR 2	2.8% 105	4.4% 62	50 5% 1011	14.2% 199	36.2% 1377
KU100	5.9% 224	11.7% 164	43.0% 860	13 6% 190	37.8% 1438
4100	3.5% 133	6.6% 93	52.4% 1047	13 3% 186	38.4% 1459
HMSII	3.4% 131	4.8% 67	53 7% 1074	12.6% 176	38.1% 1448
MK1	3.2% 120	5.9% 83	50 3% 1006	12.6% 177	36.5% 1386
ITA	4 2% 158	3.1% 43	40.0% 800	17.2% 241	32.7% 1242
VALDEMAR	3 5% 134	2.4% 34	38.1% 762	13 9% 195	29.6% 1125
AVH	2 2% 85	1.3% 18	33.0% 660	13 1% 183	24 9% 946
DOL	4.3% 162	3.0% 42	33.2% 663	13.6% 190	27.8% 1057
	(3800)	(1400)	(2000)	(1400)	(3800)

* Results from all recording heads are significantly different from real life (one-sided Fisher–Irwin tests, 0.1% significance level for all directional errors and at least 5% level for distance errors).

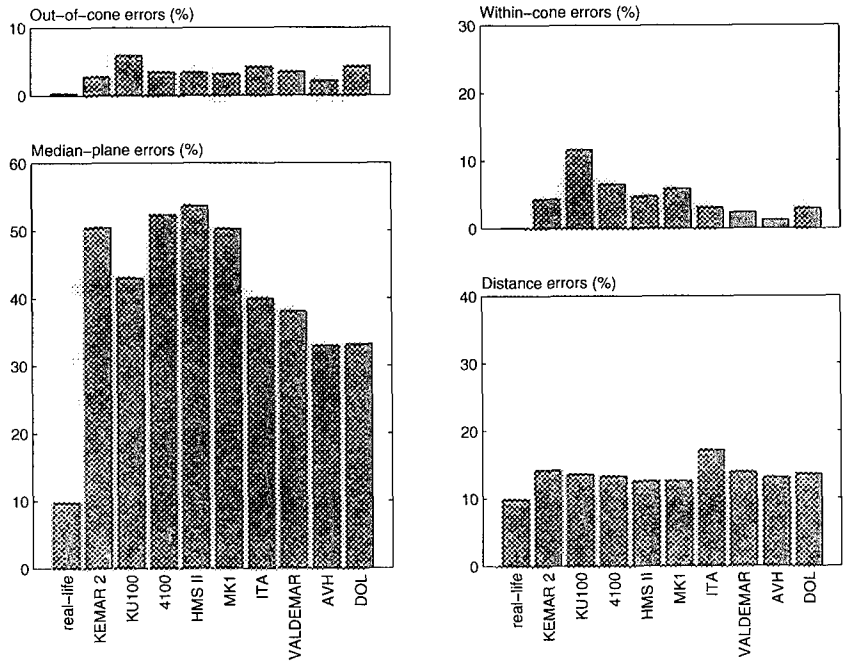


Fig. 7. Main experiment: errors in real life and with binaural recordings divided into four error categories (data as in Table 1).

the learning curves do not seem to level off during the five days of the experiment. The presence of such a strong learning effect emphasizes the need for a balanced design (as used here) in order to make a valid comparison of conditions.

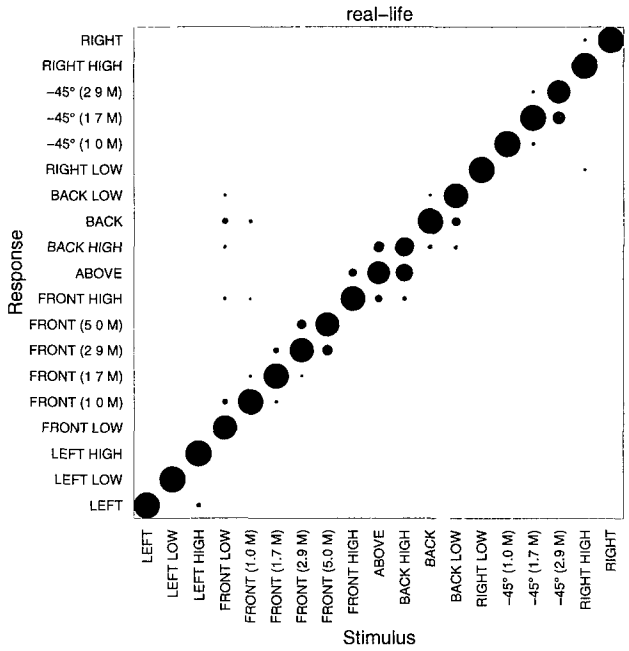


Fig 8. Learning experiment: all 5 real-life sessions pooled (9500 stimuli)

3 COMPARISON OF FIVE EXPERIMENTS

As described in the Introduction, the experiment reported in this paper is one in a series of experiments employing the same loudspeaker setup and general procedure. Although small procedural differences exist and a different number of listeners participated in the individual experiments, it is justified to compare the results of five experiments.

3.1 Experiment A

This experiment was reported in Møller et al. [3] as experiment B. Recordings were made with the build-in microphones of eight artificial heads, and eight listeners participated. The heads were KEMAR 2 A, KU80 A, KU81 A, HMS I A, HMS II A, 5930 A, 4128 A, and TORONTO A. The heads are named here as in the original publication, and the letter A is added to denote experiment A. Please refer to the original publication for a description of the heads and further details.

3.2 Experiment B

This experiment was reported in Møller et al. [3] as experiment C. The 20 listeners included the eight listeners of experiment A and 10 artificial heads were used. Recordings were made with microphones at the blocked entrances to the ear canals of the artificial heads. The heads were KEMAR 1 B, KEMAR 2 B, KEMAR 3 B, KEMAR 4 B, KU80 B, KU81 B, HMS I B, HMS II B,

Table 2. Analysis of variance for percentages of median-plane and distance errors in main experiment (repeated measures, listener as random factor) *

Source	SS	*df*	MS	*F*	*p*
Median-plane errors					
Recording head	10731.1	8	1341.4	21.028	<0.001
Listener	13241.8	19	696.9	10.925	<0.001
Recording head × listener	9696.4	152	63.8		
Total	33669 4	179			
Distance errors					
Recording head	311.8	8	39.0	1 373	0 2124
Listener	5331.2	19	280.6	9.886	<0.001
Recording head × listener	4314.0	152	28.4		
Total	9957 0	179			

* In Bartlett's tests, homogeneity of variance was accepted only for these error categories and rejected for out-of-cone and within-cone errors.

Table 3. Learning experiment: errors in number and percentage of potential number of errors (in parentheses at the bottom of each category).

	Out of Cone	Within Cone	Median Plane	Distance	Overall
Day 1	0.3%	0.1%	15.5%	10.9%	12 5%
	6	1	155	76	238
Day 2	0.5%	0 3%	11.0%	9.0%	9.7%
	9	2	110	63	184
Day 3	0 5%	0 3%	12.4%	7 1%	9.7%
	9	2	124	50	185
Day 4	0 3%	0.7%	9.5%	8.1%	8.5%
	5	5	95	57	162
Day 5	0.1%	0 1%	9.6%	8 4%	8.3%
	1	1	96	59	157
	(1900)	(700)	(1000)	(700)	(1900)

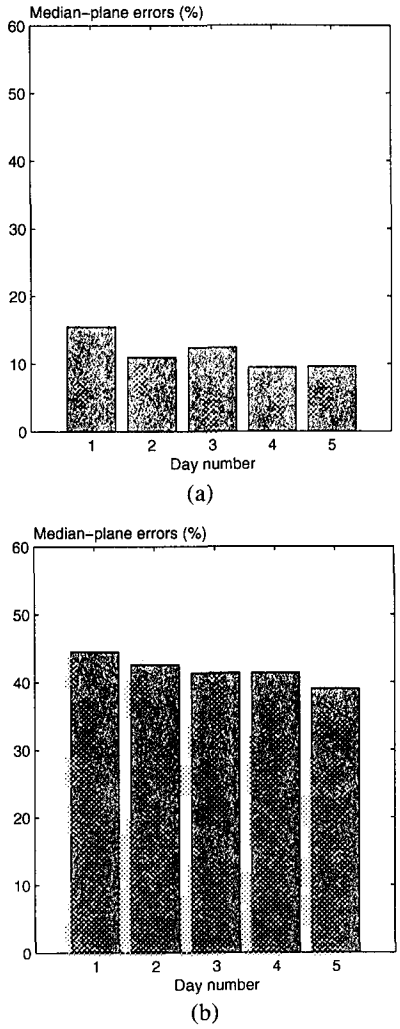


Fig.9. Errors as a function of time. (a) Five real-life sessions in learning experiment (1900 stimuli per day). (b) All sessions in main experiment (7600 stimuli per day)

4128 B, and TORONTO B. As before, the last letter, here B, denotes the experiment.

3.3 Experiment C

Experiment C was done with a new group of eight listeners and included two artificial heads, KU100 C and 4100 C. The KU100 artificial head was placed on a stand, and the 4100 head was used with its jacket and the neck ring in the "forward" position. In both cases built-in microphones were employed. This experiment was carried out by C. B. Jensen and not publicly reported before, but the procedures used were identical to those described in Møller et al. [3].

3.4 Experiment D

This experiment was reported by Sandvad et al. [19] as experiment B. Recordings were made with the built-in microphones of the four artificial heads KU100 D, VALDEMAR D, ITA D, and MK1 D. The experiment was done with 12 listeners. The listeners in experiment D had not participated in any of the previous experiments.

3.5 Experiment E

Experiment E refers to the main experiment reported in this paper. The artificial heads are named as follows: KEMAR 2 E, KU100 E, HMS II E, 4100 E, VALDEMAR E, ITA E, and MK1 E. The 20 listeners in this experiment were again newly recruited.

3.6 Results and Discussion

Median-plan errors for the five experiments are shown in Fig. 10 grouped by recording head. The full bars indicate the observed means, and the (small I-shaped) error bars indicate 84% confidence intervals calculated for each experiment.

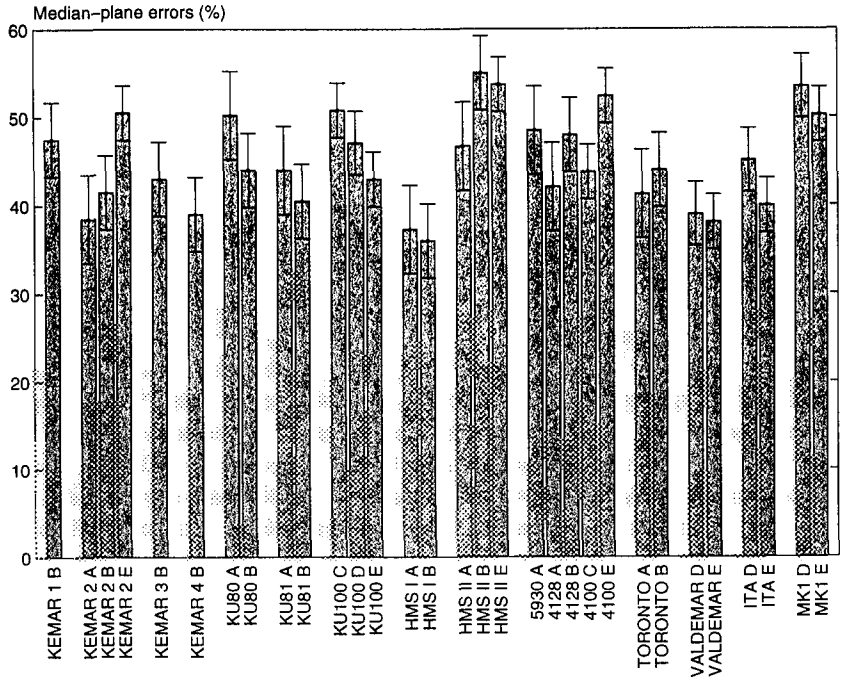


Fig. 10. Comparison of artificial heads in five listening experiments Bars—observed means; error bars—84% confidence intervals. The means of two heads with nonoverlapping confidence intervals are significantly different at the 5% level (approximately) in a *t*-test.

The confidence intervals were calculated from the common variances in the analyses of variance, including the random listener factor (corresponding to the residual variance in a one-way analysis), thus making the confidence intervals valid for a population.

The "odd" size of the confidence interval has been chosen so that the means of two artificial heads with nonoverlapping confidence intervals are significantly different at the 5% level in a *t*-test. (To be exact, this requires a few preconditions, such as that the confidence intervals be of equal size of the conditions compared. The true significance level, though, will only deviate marginally from 5% for the small violation of the preconditions in the present material.)

Systematic differences between the five experiments were inevitably introduced by changing details of the psychometric procedure, equipment, experimenter, group of listeners, group experience, and so on. Despite this, the confidence intervals generally overlap for a head in different experiments. Only five within-head comparisons out of 26 do not overlap, which is only slightly more than chance.

The results suggest that the localization performance with binaural recordings is primarily controlled by properties of the recording heads (provided proper equalization is made, as in all of these experiments). Furthermore, in quite a number of cases the confidence intervals do not overlap between recording heads, indicating that the heads differ significantly from one another.

4 CONCLUSIONS

The experiment reported in this paper is the culmination of a series of experiments on sound localization in a listening room. These experiments have shown that listeners generally do not localize all loudspeakers consistently, even in real life. Furthermore, localization is preserved with individual binaural recordings. Localization is much poorer with nonindividual recordings. This performance is improved if the nonindividual recordings are made with a human head which is selected carefully through subjective testing. However, localization is best with recordings made in our own ears. Artificial heads generally perform poorer than human heads. (In a comparison, 60% of the human heads were better than even the best artificial head.)

Many of the results of earlier experiments were confirmed in the current experiment. In addition to these findings no significant difference was found between the results of speech and white noise stimuli. Localization with binaural recordings was generally much poorer than in real life. Two selected human heads produced less median-plane errors than any of the artificial heads.

An analysis of variance revealed large differences in localization between listeners, and it showed that differences between recording heads were statistically significant for median-plane errors. There were no significant differences between heads for distance errors. An unusually large number of out-of-cone and within-cone errors were seen for one artificial head without shoulders. A very clear learning effect was seen during the listening sessions stretching over five days. This emphasizes the need for a balanced design of conditions (as used in this experiment).

Comparisons of conditions in different experiments must be made with consideration of the confounding effects of learning and systematic differences between experiments. Such a comparison of artificial heads in five experiments showed median-plane errors for which confidence intervals of the means were relatively small and usually overlapping within heads. Quite a number of combinations of two heads did not have overlapping confidence intervals, thus indicating a significant difference between the heads.

Binaural recording is a powerful recording technique with the unique ability of capturing the full spatial information available to a person by only two audio channels The listening experience is, however, limited to the extent that the recording head is fixed in the recorded sound field. It may be argued that some principal localization cues are absent since the changes introduced to the ear signals when moving the head are not represented.

However, the methods for comparing recording heads used in this series of studies are most appropriate, since proper reactions to head movements are by definition not possible with binaural recordings. Therefore the relative performance of the recording heads in the experiments reported are valid in general for applications employing binaural recordings. In this context current artificial heads are not as good as a well-selected human head.

5 ACKNOWLEDGMENT

The authors wish to thank the following persons, manufacturers, and institutions for making the artificial heads available for this study: Benny Glumsø Fredericiaskolen, Denmark (Knowles Electronics KEMAR), Peter Munk at Kinovox A/S, Denmark (Georg Neumann KU100), Poul Ladegård at Brüel & Kjær A/S, Denmark (Brüel & Kjær 4100), Per Sjösten at Lindholmen Utveckling, Sweden (HEAD Acoustics HMS II), Peter Daniel at Cortex Electronic, Germany (Cortex Electronic MK1), and Alfred Schmitz at Aachen University, Germany (Aachen University ITA). Sincere thanks to AVH and DOL for lending us their "human heads" for recording and to Dorte Hammershøi for providing the anechoic speech signal used as stimulus. We are very grateful to the 20 listeners who participated in the experiment; who each came to the laboratory on six different days enabling us to collect 47 500 answers. A great thank you is due to Palle Rye for collecting a substantial amount of the data during the listening experiment and to Clemen Boje Jensen, Michael Friis Sørensen, Dorte Hammershøi, and Jesper Sandvad for making available their data for comparison between experiments. Finally we want to thank Claus Vestergaard Skipper for the competent assistance in the laboratory and Jan Plogsties for the help during the preparation of the manuscript. This work was funded by the Danish National Centre for IT Research.

6 REFERENCES

[1] H. Møller, M. F. Sørensen, D. Hammershøi, and C. B. Jensen, "Head-Related Transfer Functions of Human Subjects," *J. Audio Eng. Soc.*, vol. 43, pp. 300–321 (1995 May).

[2] H. Møller, M. F. Sørensen, C. B. Jensen, and D. Hammershøi, "Binaural Technique: Do We Need Individual Recordings?," *J. Audio Eng. Soc.*, vol. 44, pp. 451–469 (1996 June).

[3] H. Møller, D. Hammershøi, C. B. Jensen, and M. F. Sørensen, "Evaluation of Artificial Heads in Listening Tests," *J. Audio Eng. Soc.*, vol. 47, pp. 83–100 (1999 Mar.).

[4] H. Møller, C. B. Jensen, D. Hammershøi, and M. F. Sørensen, "Using a Typical Human Subject for Binaural Recordings," presented at the 100th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 44, p. 632 (1996 July/Aug.), preprint 4157.

[5] IEC 711, "Occluded-Ear Simulator for the Measurement of Earphones Coupled to the Ear by Ear Inserts," 1st ed., International Electrotechnical Commission, Geneva, Switzerland (1981).

[6] ANSI S3.25-1979 (R1986), "American National Standard for an Occluded Ear Simulator," American National Standards Institute, New York (1980).

[7] IEC 959, "Provisional Head and Torso Simulator for Acoustic Measurements on Air Conduction Hearing Aids," 1st ed., Tech. Rep., International Electrotechnical Commission, Geneva, Switzerland (1990).

[8] ITU-T P.58, "Head and Torso Simulator for Telephonometry," Recommunication International Telecommunications Union, Geneva, Switzerland (1996, Aug.).

[9] F. Christensen, C. B. Jensen, and H. Møller, "Design of the Artificial Head VALDEMAR," in preparation.

[10] F. Christensen, C. B. Jensen, and H. Møller, "Measuring Components of HRTFs," in preparation.

[11] IEC 268-13, "Sound System Equipment, pt. 13, Listening Tests and Loudspeakers," 1st ed., International Electrotechnical Commission, Geneva, Switzerland (1985).

[12] S. H. Nielsen, "Distance Perception in Hearing," Ph.D. dissertation, ISBN 87-7307-447-0 (Aalborg University Press, Aalborg, Denmark, 1991).

[13] S. H. Nielsen, "Auditory Distance Perception in Different Rooms," *J. Audio Eng. Soc.*, vol. 41, pp. 755–770 (1993 Oct.).

[14] H. Møller, "Fundamentals of Binaural Technology," *Appl. Acoust.*, vol. 36, pp. 171–218 (1992).

[15] B. Friedlander and B. Porat, "The Modified Yule–Walker Method of ARMA Spectral Estimation," *IEEE Trans. Aerosp. Electron. Sys.*, vol. AES-20, pp. 158–173 (1984 Mar.).

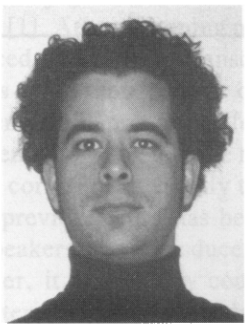
[16] D. R. Cox, *Planning of Experiments* (Wiley, New York, 1958), pp. 269–278.

[17] S. M. Ross, *Introduction to Probability and Statistics for Engineers and Scientists* (Wiley, New York, 1987), pp. 230–234.

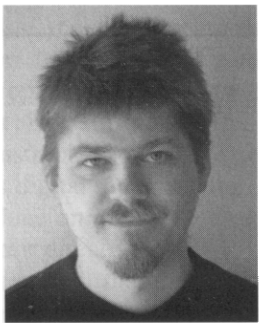
[18] R. E. Glaser in *Encyclopedia of Statistical Sciences*, vol. 1, S. Kotz and N. L. Johnson, Eds. (Wiley, New York, 1982), pp. 189–191.

[19] J. Sandvad, F. Christensen, S. K. Olesen, and H. Møller, "Localization with Artificial Head Recordings," presented at the 134th Meeting of the Acoustical Society of America (San Diego, CA, 1997 Dec.), Abstract in *J. Acoust. Soc. Am.*, vol. 102, no. 5, pt. 2, p. 3116 (1997 Nov.), (paper 2pSP4).

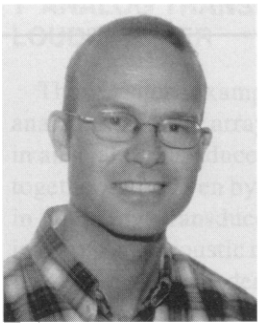
THE AUTHORS



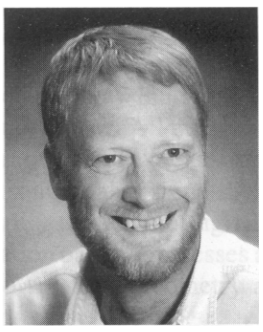
P. Minnaar



S. K. Olesen



F. Christensen



H. Møller

Pauli Minnaar was born in South Africa in 1972 He studied at the University of Pretoria in South Africa, where he received a B Eng degree in mechanical engineering in 1994. He then studied acoustics at the Institute of Sound and Vibration Research (ISVR), University of Southampton in England In 1996 he obtained the M.Sc. degree in sound and vibration, after which he worked at the ISVR as research assistant on the European Union project SOQCRATES that dealt with sound quality in cars Since October 1997 he has been at the Department of Acoustics, Aalborg University in Denmark, where he studied binaural technology He hopes to complete a Ph.D. in the near future. Mr. Minnaar's main research interests are in binaural recording and synthesis, psychoacoustics, and audio digital signal processing.

•

Søren Krarup Olesen was born in 1966 in Denmark. In 1985 he began his studies in electrical engineering and received a M Sc E.E degree with specialization in acoustics from Aalborg University, Denmark, in 1991.

In 1994 he began a Ph.D. study at the Department of Acoustics, Aalborg University, and was later employed as technical amanuensis and research assistant in a project concerning 3-D audio in virtual rooms Mr. Olesen's areas of interest are mainly binaural synthesis, design of computer-controlled psychoacoustical listening experiments and acoustical room simulation and modeling systems for virtual reality purposes He holds membership in the Danish Acoustical Society and Danish Engineering Society.

•

Flemming Christensen was born in 1969 He studied electrical engineering at Aalborg University, where he received his M.Sc E.E. degree in 1993 specializing in acoustics Since then he has been involved in a large number of research projects at the Department of Acoustics within the fields of binaural auralization, design and construction of artificial heads, design of measurement systems, and audiometric measurement techniques For a period of two years he was manager of a research project concerning 3-D audio for multimedia purposes In addition to his involvement in research and education, Mr Christensen maintains a professional career as a musician, playing guitars and keyboards, and occasionally works as a sound studio engineer and acoustics consultant. He is a member of the Danish Engineering Society, the Audio Engineering Society, and the Danish Musician Society.

•

Henrik Møller was born in Århus in 1951. He studied electrical engineering (Danish Engineering Academy) and received a B.Sc. degree in 1974 He worked as a development engineer for Brüel & Kjær from 1974 to 1976. Since then he has been at Aalborg University. He became associate professor in 1980, received a Ph D degree in 1984, and was appointed reader (Danish: docent) in 1988 and professor in 1996. During the period 1991–94 he was partly on leave from the university to work as a director of Perceptive Acoustics A/S, a research subsidiary company of Brüel & Kjær

Dr. Møller's previous and current research reflect his long-time experience with sound, its influence on humans, acoustical measurement techniques, signal processing, hearing, and psychometric methods. His research areas include effects of infrasound and low-frequency noise on humans, investigations of hearing thresholds and loudness assessment, and exploitation of binaural techniques He is the author of numerous scientific publications and invited as well as contributed conference papers

When new high-quality acoustical laboratories were built at Aalborg University in 1987, Dr. Møller was responsible for the design as well as control of the work. As head of the Department of Acoustics, he is now the manager of research and education in a wide range of areas such as human sound perception, audiology, psychometry, electroacoustics, recording and playback techniques, auralization in acoustic room modeling and virtual reality, acoustical measurement techniques, electronics, and signal processing In 1998 the department inaugurated a new section with laboratories for virtual reality, and a further expansion with laboratories for a new research unit for sound quality will be ready in 2001.

Dr Møller has organized conferences on Low Frequency Noise and Hearing (Aalborg 1980), general acoustics (Nordic Acoustical Meeting, Aalborg 1986), and Low Frequency Noise and Vibration (Aalborg 2000). He is convener of ISO Technical Committee 43: "Acoustics," Working Group 1: "Thresholds of Hearing," a member of Working Group 6: "Determination of Noise Immissions from Sound Sources Placed Close to the Ears," and of the editorial board of *Journal of Low Frequency Noise & Vibration* In addition, he serves as reviewer for a number of international journals and national and international research foundations. He holds membership in the Danish Engineering Society, Audio Engineering Society, Acoustical Society of America, IEEE, Danish Acoustical Society, Danish Technical-Audiological Society, Danish Society for Applied Signal Processing, Danish Virtual Reality Society, and Danish Standardisation Organisation (board of Acoustics and Working Group of Audiometry and Hearing).

Dr. Møller spends hours off work (too few) by playing big band music on his baritone saxophone or by keeping his classic British cars in good shape Now and then, he also drives them