

COVER PAGE

Document downloaded by @DAEL

Thu Nov 26 19:42:58 2026
For personal use

When automatic English translation is provided, only the original document is authentic.
The EAA cannot be held responsible of any translation error

Bibliographical reference

Statistical Evaluation of Localization Performance and its Application in an Internet-Based Self-Screening Test Using a Virtual Environment, C. Borß, R. Martin and J. Lyzenga, Acta Acustica **vol. 96** (Number 5), 2010, pp. 936-946

DOI
https://doi.org/10.3813/AAA.918352

ACTA ACUSTICA UNITED WITH ACUSTICA
Vol. 96 (2010) 936 – 946
DOI 10.3813/AAA.918352

Statistical Evaluation of Localization Performance and its Application in an Internet-Based Self-Screening Test Using a Virtual Environment

C. Borß[1], R. Martin[1]*, J. Lyzenga[2]
[1] Institute of Communication Acoustics, Ruhr-Universität Bochum, Bochum, Germany. christian.borss@rub.de
[2] VU University Medical Center, Amsterdam, The Netherlands

Summary
This paper presents a method for scoring forced choice hearing tests with more than two alternatives. The proposed scoring procedure classifies the observed sequence of choices in a maximum likelihood sense. The classifier exploits *a priory* knowledge about the choice statistics estimated from a limited set of observed data. The proposed method was implemented and evaluated in conjunction with a newly developed Internet-based self-screening test. The implemented test measures localization skills in an interactive auditory virtual environment (AVE) while including both acoustical and behavioral aspects of everyday localization situations. By means of cross-validation, it is shown that the test results of listeners with normal localization skills and the test results of listeners with impaired localization skills can be reliably separated using the introduced scoring method. It is further shown that good classification results are obtained when a uniform *a priori* density model for the choices of the localization-impaired reference group is applied.

PACS no. 43.66.Yw, 43.66.Qp, 43.66.Pn

1. Introduction

Spatial hearing and the skill to localize sounds is an important aspect of a subject's auditory performance, since an impaired localization skill significantly reduces a person's spatial orientation. The individual localization blur can be quantitatively evaluated by measuring the minimum audible angle (MAA) [1] where the detection threshold is determined e.g., by matching the parameters of the psychometric function [2] or by applying an adaptive procedure to find a specific point on the psychometric function [3, 4]. This is an established method which is also used for the assessment of other aspects of hearing like e.g., the speech reception threshold in noise [5, 6, 7, 8]. While an MAA test under ideal lab conditions leads to valid and reliable results, it does not directly predict the subject's performance in everyday situations. Natural environments include reverberation and interfering sound sources and are perceived by exploiting more than one modality [9]. In the localization test proposed by Goverts several interfering sound sources generated by a setup of eight loudspeakers create a more complex auditory scene [10]. In this test, the subject has to identify the loudspeaker which plays back the target signal and simple *correct scoring* is used to evaluate the localization performance. However, much information resides in the details of incorrect responses which is not taken into account by simple correct scoring. It makes a difference if the confused sound source is in the proximity of the actual sound source or not. According to information theory, an event which occurs with a lower probability has a higher information content [11]. Not taking the probability for the observed response into account thus means not fully exploiting the available information. Therefore, this paper presents a method for scoring forced choice tests which also considers the confusion likelihoods.

The proposed evaluation method is applied to a newly developed self-screening test which is suited to be performed unsupervised at home via the Internet. The intention of a self-screening test is to indicate a specific hearing deficiency reliably, or to raise the awareness of a specific hearing problem, while it does not replace the well-grounded diagnosis of a professional audiologist. To a certain degree, our self-screening test also takes a subject's behavior into account. To achieve this, binaural signals are synthesized according to the current position of the avatar's head in a virtual test room. In this paper, we make a distinction between the *subject* who sits in front of the computer and the *avatar* as that subject's representation in the virtual environment. The binaural signal which is pre-

Received 20 January 2010,
accepted 19 June 2010.

* This work was funded in part by grants from the European Union FP6, Project 004171 HEARCOM (http://www.hearcom.eu/).

© S. Hirzel Verlag · EAA

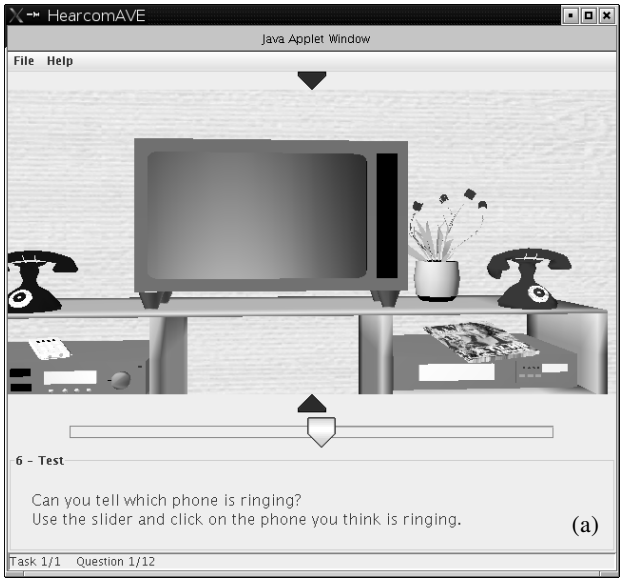


Figure 1. (a): Screenshot of the implemented localization test. (b): Virtual test environment.

sented to the subject dichotically via headphones creates an interactive auditory virtual environment (AVE) [12]. Scoring of localization performance in reverberant rooms is a complex task, especially in a virtual environment, as front-back confusions and localization blur will cause different kinds of errors [13]. Therefore, to deal with the complexity of all the influencing factors, the proposed statistical approach was developed for scoring the localization performance in virtual environments.

Self-screening tests often assign test outcomes to one of the three categories, "good", "moderate", or "insufficient" [14]. Therefore, the following section 2 (Scoring Procedure) explains the statistical considerations for the classification of subjects into those categories with respect to their performance in a localization test. Then, in section 3 (Setup of the Internet-Based Self-Screening Test), we describe our newly developed self-screening test, present classification results for subjects with normal and impaired localization, and validate them using a cross-validation method.

2. Scoring Procedure

In order to make the theoretical considerations for scoring forced choice hearing tests more illustrative, we first give a brief description of our newly developed self-screening test. Details about this test are given in section 3.

2.1. Probabilistic Model

In our implemented self-screening test, the avatar is placed in a virtual living room with $N = 6$ telephones of different colors. Figure 1a depicts a screenshot of the user interface which allows the subject to interact with the virtual environment shown in Figure 1b. The ringing of the spatially distributed telephones is used as stimulus. Every stimulus is presented twice to the subject resulting in $K = 12$ stimuli in random order. Since the subject has to select the assumed sound source for each trial, our self-screening test is a forced choice test.

The stimuli are elements of a stimulus set $\mathbb{S} = \{s_1, \ldots, s_N\}$, i.e. the black, pink, blue, red, yellow, and green telephone. The stimulus sequence $X = \{x_1, \ldots, x_K\}$ specifies a random sequence of elements from the stimulus set $\mathbb{S}$. Whenever the subject clicks on one of the telephones, he chooses an element from the answer set $\mathbb{A} = \{a_1, \ldots, a_N\}$. We denote the sequence of choices as answer sequence $Y = \{y_1, \ldots, y_K\}$. Rating a subject as either having normal localization skills (NL) or having impaired localization skills (LI) is a classification problem, where the answer sequence Y is the observation. The likelihood for class membership P_{NL} and P_{LI} respectively can be determined on the basis of the probability density for the observed answer sequence Y under the condition of the stimulus sequence X and the class membership,

$$P_{\mathrm{NL}} = P_{Y|X,\mathrm{NL}}\,(Y|X, \mathrm{NL}) \tag{1}$$

$$P_{\mathrm{LI}} = P_{Y|X,\mathrm{LI}}\,(Y|X, \mathrm{LI})\,. \tag{2}$$

Based on the likelihood ratio, a subject can be classified in the maximum likelihood sense: if the likelihood ratio $\frac{P_{Y|X,\mathrm{NL}}(Y|X,\mathrm{NL})}{P_{Y|X,\mathrm{LI}}(Y|X,\mathrm{LI})}$ is greater than or equal to 1, the subject will be classified as having normal localization skills or else as having impaired localization skills.

Since the avatar is brought back to a default position after each trial i.e., facing the television in the virtual living room, we can assume that the trials are statistically independent. Thus, the joint probabilities $P_{Y|X,\mathrm{NL}}\,(Y|X, \mathrm{NL})$ and $P_{Y|X,\mathrm{LI}}\,(Y|X, \mathrm{LI})$ can be determined by the products

$$P_{Y|X,\mathrm{NL}}\,(Y|X, \mathrm{NL}) = \prod_{k=1}^{K} P_{a|s,\mathrm{NL}}(y_k|x_k, \mathrm{NL}) \tag{3}$$

$$P_{Y|X,\mathrm{LI}}\,(Y|X, \mathrm{LI}) = \prod_{k=1}^{K} P_{a|s,\mathrm{LI}}(y_k|x_k, \mathrm{LI}), \tag{4}$$

where $P_{a|s,\mathrm{NL}}(y_k|x_k, \mathrm{NL})$ and $P_{a|s,\mathrm{LI}}(y_k|x_k, \mathrm{LI})$ denote the previous estimated probabilities of the individual choices. The corresponding log-likelihoods are:

$$L_{Y|X,\mathrm{NL}}(Y|X, \mathrm{NL}) = \sum_{k=1}^{K} L_{a|s,\mathrm{NL}}(y_k|x_k, \mathrm{NL}) \quad (5)$$

$$L_{Y|X,\mathrm{LI}}(Y|X, \mathrm{LI}) = \sum_{k=1}^{K} L_{a|s,\mathrm{LI}}(y_k|x_k, \mathrm{LI}). \quad (6)$$

Based on these log-likelihoods, the classification decision can be formulated alternatively by the inequality $L_{Y|X,\mathrm{NL}}(Y|X, \mathrm{NL}) - L_{Y|X,\mathrm{LI}}(Y|X, \mathrm{LI}) \geq 0$.

As a consequence of the sum over all stimuli k, the joint log-likelihoods are linear dependent on the number of played stimuli. A division by K eliminates this linear dependence and yields our final test score ϑ,

$$\vartheta = \frac{L_{Y|X,\mathrm{NL}}(Y|X, \mathrm{NL}) - L_{Y|X,\mathrm{LI}}(Y|X, \mathrm{LI})}{K}. \quad (7)$$

This normalization has the advantage that changing the number of played stimuli would not change the range of the test score or the range of the thresholds which we will derive in section 2.4. A positive test score results from $L_{Y|X,\mathrm{NL}}(Y|X, \mathrm{NL}) > L_{Y|X,\mathrm{LI}}(Y|X, \mathrm{LI})$, which is equivalent to $P_{\mathrm{NL}} > P_{\mathrm{LI}}$. Thus, a positive score value indicates that the observed answer sequence originates from a subject with a "good" localization performance. Accordingly, a negative value means that the observed answer sequence is more likely to originate from a subject with an "insufficient" localization performance.

2.2. Computation of the Test Score

The test score ϑ can be directly computed from the joint log-likelihoods specified in equations (5) and (6). In this subsection we will use a matrix notation to yield a more compact form.

The observed sequence of choices can be represented by a confusion matrix $\mathbf{C}_{Y|X}$,

$$\mathbf{C}_{Y|X} = \begin{bmatrix} c_{11} & \cdots & c_{N1} \\ \vdots & \ddots & \vdots \\ c_{1N} & \cdots & c_{NN} \end{bmatrix},$$

where the elements c_{ij} (positive integers which sum up to K) specify how often the subject chose a_i when the stimulus s_j was presented. The *a priori* knowledge about classification likelihoods, which is derived during the training of the classifier as described in the following section 2.3, can be represented by classification matrices $\mathbf{M}_{\mathrm{NL}}$ and $\mathbf{M}_{\mathrm{LI}}$,

$$\mathbf{M}_{\mathrm{NL}} = \begin{bmatrix} L_{a|s,\mathrm{NL}}(a_1|s_1, \mathrm{NL}) & \cdots & L_{a|s,\mathrm{NL}}(a_N|s_1, \mathrm{NL}) \\ \vdots & \ddots & \vdots \\ L_{a|s,\mathrm{NL}}(a_1|s_N, \mathrm{NL}) & \cdots & L_{a|s,\mathrm{NL}}(a_N|s_N, \mathrm{NL}) \end{bmatrix}$$

$$\mathbf{M}_{\mathrm{LI}} = \begin{bmatrix} L_{a|s,\mathrm{LI}}(a_1|s_1, \mathrm{LI}) & \cdots & L_{a|s,\mathrm{LI}}(a_N|s_1, \mathrm{LI}) \\ \vdots & \ddots & \vdots \\ L_{a|s,\mathrm{LI}}(a_1|s_N, \mathrm{LI}) & \cdots & L_{a|s,\mathrm{LI}}(a_N|s_N, \mathrm{LI}) \end{bmatrix}.$$

The matrices $\mathbf{M}_{\mathrm{NL}}$ and $\mathbf{M}_{\mathrm{LI}}$ store the likelihoods for all possible combinations of a stimulus and an observed answer. The difference of these matrices results in what we denote as our scoring matrix $\mathbf{M}_{\mathrm{score}}$,

$$\mathbf{M}_{\mathrm{score}} = \mathbf{M}_{\mathrm{NL}} - \mathbf{M}_{\mathrm{LI}}. \quad (8)$$

Since the elements m_{ij} of the scoring matrix $\mathbf{M}_{\mathrm{score}}$ are the log-likelihood ratios $L_{a|s,\mathrm{NL}}(a_i|s_j, \mathrm{NL}) - L_{a|s,\mathrm{LI}}(a_i|s_j, \mathrm{LI})$, we can compute the test score ϑ by the normalized scalar inner product,

$$\vartheta = \frac{\left\langle \mathbf{M}_{\mathrm{score}}, \mathbf{C}_{Y|X} \right\rangle}{K} \quad (9)$$

where we define the scalar inner product for two matrices $\mathbf{M}$ and $\mathbf{C}$ with elements m_{ij} and c_{ij}, respectively, as follows:

$$\langle \mathbf{M}, \mathbf{C} \rangle = \sum_{i=1}^{N} \sum_{j=1}^{N} m_{ij} c_{ij}. \quad (10)$$

Thus, $\mathbf{C}_{Y|X}$ selects those entries of $\mathbf{M}_{\mathrm{score}}$ which correspond to a particular stimulus-answer sequence.

2.3. Determination of the Classification Matrices

The likelihood of confusing two sound sources in a natural environment depends on a multitude of factors. Reflections at the room boundaries are superposed with the direct sound and cause a degradation of the localization accuracy [15, 16]. The amount of degradation depends on the room geometry, the reflection properties of the walls, the resulting reverberation time, the position of the resulting image sources, and the distance of the sound sources [17]. In addition to these physical factors, subjective effects which help listeners to cope with reverberation can also be observed in binaural hearing [18]. For example, listeners with an initial poor localization performance in an unknown environment can improve their localization accuracy by learning the acoustic characteristics of the room [19]. In a virtual environment, the localization accuracy is further reduced by auralization artifacts caused by e.g., non-individual head-related transfer functions (HRTFs) or by the lack of low-latency head-tracking [13].

Due to the multitude of factors, it is difficult to predict the confusion likelihoods $L_{a|s,\mathrm{NL}}(a_i|s_j, \mathrm{NL})$ and $L_{a|s,\mathrm{LI}}(a_i|s_j, \mathrm{LI})$ in adverse acoustical conditions. These likelihoods need to be estimated from the relative frequency of observed answer/stimulus combinations $[a_i, s_j]$ of two reference groups with normal and with impaired localization. This estimation constitutes the training / learning step in a pattern recognition task and can be realized by summing up (and later normalizing) the observed confusion matrices for all subjects to cumulative confusion matrices $\mathbf{G}_{\mathrm{NL}}$ and $\mathbf{G}_{\mathrm{LI}}$. In order to avoid an underestimation of probabilities for combinations which were not observed, we then adjust the elements g_{ij} of the cumulative

confusion matrices by limiting them to a lower bound ε,

$$\widehat{\mathbf{G}} = \begin{bmatrix} \widehat{g}_{11} & \cdots & \widehat{g}_{N1} \\ \vdots & \ddots & \vdots \\ \widehat{g}_{1N} & \cdots & \widehat{g}_{NN} \end{bmatrix} = \begin{bmatrix} \max(\varepsilon, g_{11}) & \cdots & \max(\varepsilon, g_{N1}) \\ \vdots & \ddots & \vdots \\ \max(\varepsilon, g_{1N}) & \cdots & \max(\varepsilon, g_{NN}) \end{bmatrix}. \quad (11)$$

Note that for readability reasons the indices NL and LI are omitted in equation (11) and that this limitation step is applied to both cumulative confusion matrices (this notational simplification will be used in the following equations wherever possible). As relative frequency is always greater than or equal to 0 and as the combinations for which we apply the limitation have been observed less than 1 time, it is reasonable to choose the mean value $\varepsilon = 0.5$ as lower bound. For comparison: estimating the elements $\widehat{g}_{ij}$ from 52 normally-localizing answer sequences and from 24 localization-impaired answer sequences, as described in section 3.1, results in $\widehat{g}_{ij} \in [0, 104]$ and $\widehat{g}_{ij} \in [0, 48]$ respectively.

Dividing each row of $\widehat{\mathbf{G}}_{\mathrm{NL}}$ and $\widehat{\mathbf{G}}_{\mathrm{LI}}$ by its sum normalizes the elements of $\widehat{\mathbf{G}}_{\mathrm{NL}}$ and $\widehat{\mathbf{G}}_{\mathrm{LI}}$ and yields our estimated probabilities $P_{a|s,\mathrm{NL}}(a_i|s_j, \mathrm{NL})$ and $P_{a|s,\mathrm{LI}}(a_i|s_j, \mathrm{LI})$. Thus, the log-likelihood values of the classification matrices $\mathbf{M}_{\mathrm{NL}}$ and $\mathbf{M}_{\mathrm{LI}}$ can be obtained as follows:

$$\mathbf{M} = \begin{bmatrix} \ln(\hat{g}_{11} / \sum_{i=1}^{N} \hat{g}_{i1}) & \cdots & \ln(\hat{g}_{N1} / \sum_{i=1}^{N} \hat{g}_{i1}) \\ \vdots & \ddots & \vdots \\ \ln(\hat{g}_{1N} / \sum_{i=1}^{N} \hat{g}_{iN}) & \cdots & \ln(\hat{g}_{NN} / \sum_{i=1}^{N} \hat{g}_{iN}) \end{bmatrix}$$

The choice of proper training sets has a major impact on the classification performance. Choosing a normally-localizing reference group for the training set by selecting normally-hearing subjects with a self-reported good localization performance is relatively straightforward. However, the choice of the localization-impaired reference group is more complicated. This requires prior rating of the localization performance and a selection of subjects with a significantly worse rating as compared to the rating of the normally-hearing reference group. Hence, the accuracy of the final system depends on this selection. This dependency can be reduced by testing the likelihood of normally-localizing subjects against a uniform distribution,

$$\mathbf{M}_{\mathrm{unif}} = \begin{bmatrix} \ln\frac{1}{N} & \cdots & \ln\frac{1}{N} \\ \vdots & \ddots & \vdots \\ \ln\frac{1}{N} & \cdots & \ln\frac{1}{N} \end{bmatrix}, \quad (12)$$

which in fact makes the least assumptions on the performance of localization-impaired subjects. This yields the alternative test score

$$\vartheta_{\mathrm{unif}} = \frac{\left\langle \mathbf{M}_{\mathrm{NL}} - \mathbf{M}_{\mathrm{unif}}, \mathbf{C}_{Y|X} \right\rangle}{K}. \quad (13)$$

Table I. Probabilities of classifier decisions.

classifier decision	H_0 is true	H_0 is false
accept H_0	$1-\lambda$	β
no decision	$\lambda-\alpha$	$\mu-\beta$
reject H_0	α	$1-\mu$

The likelihood values in (12) can be expected from subjects who are not able to extract any spatial information from the binaural signal, giving answers that are indistinguishable from purely random (uniformly distributed) answers. Subjects with a moderate localization impairment are able to extract at least some spatial information from a binaural signal. It is more likely that the observed answers of these subjects correspond to $\mathbf{M}_{\mathrm{LI}}$ coefficients which are larger than the $\mathbf{M}_{\mathrm{unif}}$ coefficients. Consequently, the alternative score $\vartheta_{\mathrm{unif}}$ based on $\mathbf{M}_{\mathrm{unif}}$ will be larger than ϑ i.e., $\vartheta_{\mathrm{unif}}$ is an overestimated (normalized) likelihood ratio.

2.4. Threshold Determination

The proposed self-screening test is intended to distinguish between "good" and "insufficient" localization performance as well as an intermediate range ("moderate" localization performance). This is a classification task where the test score is classified as belonging to one of these three classes. We define the hypothesis that the subject is a normally-localizing subject as the null-hypothesis H_0. This hypothesis is either accepted ("good" localization performance), rejected ("insufficient" localization performance), or no decision is made concerning H_0 ("moderate" localization performance). We define α as the probability that the localization performance of a normally-localizing subject is rated as "insufficient" and β as the probability that the localization performance of a localization-impaired subject is rated as "good". We further define λ as the probability that the localization performance of a normally-localizing subject is rated as either "moderate" or "insufficient" and μ as the probability that the localization performance of a localization-impaired subject is rated as either "good" or "moderate". Table 1 gives an overview of the probabilities for the different decisions.

In order to map the test score to the aforementioned ordinal scale ("insufficient", "moderate", or "good"), two thresholds are required. If the test score is below the lower threshold Θ_{l}, then H_0 will be rejected. If the test score is above the upper threshold Θ_{u}, then H_0 will be accepted. If the cumulative distribution function (CDF) $F_{\mathrm{NL}}(\vartheta)$ for the score of normally-localizing subjects is known, Θ_{l} can be selected to yield the desired error probability α. If the CDF $F_{\mathrm{LI}}(\vartheta)$ for the score of localization-impaired subjects is known, Θ_{u} can be selected to yield a desired error probability β (the probabilities λ and μ result from the selected thresholds). For example, if the distribution is known to be Gaussian, the threshold Θ_{l} could be chosen as one, two, three, or four standard deviations below the mean of the probability density function (PDF)

$f_{\mathrm{NL}}(\vartheta)$ to yield $\alpha = 15.9\%$, $\alpha = 2.28\%$, $\alpha = 0.135\%$, or $\alpha = 0.00317\%$ respectively (Θ_{u} can be chosen accordingly above the mean of $f_{\mathrm{LI}}(\vartheta)$ to produce the corresponding β value). For the probability distribution of our test scores we have no evidence for the assumption of a specific distribution like a Gaussian distribution. Thus, $F_{\mathrm{NL}}(\vartheta)$ and $F_{\mathrm{LI}}(\vartheta)$ have to be estimated. This is achieved by the integration over $f_{\mathrm{NL}}(\vartheta)$ and $f_{\mathrm{LI}}(\vartheta)$ which we estimate from the training data.

Estimating the statistics of a group, i.e. estimating $f_{\mathrm{NL}}(\vartheta)$ and $f_{\mathrm{LI}}(\vartheta)$, does not mean that all members of a group share the same response characteristics. In fact, a reliable test should always result in a test score within a small individual margin (in an ideal test, the variance of each subject's score should be zero i.e., the test will always result in the same score for the same subject). To assume that all members of a group share the same response characteristics would clearly make no sense, because it would imply that all members have the same localization performance. If the individual response characteristics are combined to a group response characteristic, this does not imply any homogeneity assumption within the group. The PDFs $f_{\mathrm{NL}}(\vartheta)$ and $f_{\mathrm{LI}}(\vartheta)$ only specify the *average* PDF of all group members which is given by the mean of the individual PDFs. If we assume that the individual scores are normally distributed, we can apply a Gaussian mixture model,

$$f(\vartheta) = \frac{1}{I}\sum_{i=1}^{I} \phi\left(\vartheta, \mu_i, {\sigma_i}^2\right), \tag{14}$$

where i denotes the subject, I denotes the number of group members, $\phi\left(\cdot, \mu, \sigma^2\right)$ denotes the normal PDF, and μ_i and ${\sigma_i}^2$ specify the means and variances of the individual test scores.

For the estimation of the means and variances of the individual test scores, we apply a bootstrap method which is well suited for small data sets [20]. Instead of using the observed answer sequence Y and the actual stimulus sequence X, we construct a large number of bootstrap versions $\hat{Y}$ and $\hat{X}$ by random sampling with replacement (each in the same order) e.g., $\hat{Y} = [y_6, y_{10}, y_9, y_{11}, y_{11}, y_5, y_9, y_3, y_1, y_9, y_7, y_6]$ and $\hat{X} = [x_6, x_{10}, x_9, x_{11}, x_{11}, x_5, x_9, x_3, x_1, x_9, x_7, x_6]$. By Calculating the mean and the variance of the test scores which result from these bootstrap versions, we can estimate μ_i and ${\sigma_i}^2$. This bootstrap procedure needs to be applied for each member of both reference groups to determine the coefficients of the Gaussian mixture models for $f_{\mathrm{NL}}(\vartheta)$ and $f_{\mathrm{LI}}(\vartheta)$.

The estimated CDF $F_{\mathrm{NL}}(\vartheta)$ is finally obtained from the sum over the corresponding normal CDFs $\Phi\left(\cdot, \mu, \sigma^2\right)$,

$$F_{\mathrm{NL}}(\vartheta) = \frac{1}{I}\sum_{i=1}^{I} \Phi\left(\vartheta, \mu_i, {\sigma_i}^2\right). \tag{15}$$

For practical reasons, $F_{\mathrm{LI}}(\vartheta)$ is obtained likewise from the sum over the corresponding complementary normal CDFs. The advantage of using the sum over the CDFs instead of a numerical integration over $f_{\mathrm{NL}}(\vartheta)$ and $f_{\mathrm{LI}}(\vartheta)$ is that it is less prone to numerical errors. Based on these CDFs, the thresholds Θ_{l} and Θ_{u} can then be selected to yield the desired classification error probabilities α and β.

The choice of the thresholds always results in a compromise between the error probabilities α, β, λ, and μ. In principle, if the impact of the different error types can be quantified by linear weights c_α, c_μ, c_β, and c_λ, then an optimal compromise can be found by applying the Bayes-Risk decision rule [21],

$$\Theta_{\mathrm{l}} = \underset{\Theta_{\mathrm{l}}}{\operatorname{argmin}}\left(c_\alpha \alpha(\Theta_{\mathrm{l}}) + c_\mu \mu(\Theta_{\mathrm{l}})\right) \tag{16}$$

$$\Theta_{\mathrm{u}} = \underset{\Theta_{\mathrm{u}}}{\operatorname{argmin}}\left(c_\beta \beta(\Theta_{\mathrm{u}}) + c_\lambda \lambda(\Theta_{\mathrm{u}})\right), \tag{17}$$

where the error probabilities are obtained from the probability functions,

$$\begin{aligned}
\alpha(\Theta_{\mathrm{l}}) &= F_{\mathrm{NL}}(\Theta_{\mathrm{l}}) \\
\mu(\Theta_{\mathrm{l}}) &= F_{\mathrm{LI}}(\Theta_{\mathrm{l}}) \\
\beta(\Theta_{\mathrm{u}}) &= F_{\mathrm{LI}}(\Theta_{\mathrm{u}}) \\
\lambda(\Theta_{\mathrm{u}}) &= F_{\mathrm{NL}}(\Theta_{\mathrm{u}}).
\end{aligned}$$

This optimization eliminates the only non-automatic parameter of our methodology. However, for our implemented self-screening test, which is described in the following section 3, we cannot use this automatic threshold selection due to the lack of a proper criterion for the error impacts.

3. Setup of the Internet-Based Self-Screening Test

The proposed scoring procedure was implemented and evaluated for a newly developed Internet-based self-screening test which aims at assessing the localization performance of subjects in everyday life situations [22]. As an adverse acoustical "everyday life" situation, we have chosen a reverberant living room with a diffuse background noise, i.e. music with low dynamic variations. The task of the subject is to localize ringing telephones which are spatially distributed around the subject's avatar. Under lab conditions, a physical setup with real telephones or a multi-speaker setup can be used to create the above mentioned everyday life situation. Unfortunately, the end-users who conduct our implemented test at home do not have such a sophisticated setup. While some of them may have a surround sound setup, which could be used for a self-screening test, it is questionable whether they will use a standard compliant speaker arrangement [23]. A possible solution for this dilemma is to use headphones to create an auditory virtual environment. An additional advantage of using headphones instead of loudspeakers is that the influence of the room acoustics at the end-user side is eliminated. This leads to more controllable conditions under which the test is conducted. Using this method, a

self-screening localization test can be realized which is not only valid but also reliable [10].

In our implementation, each stimulus contains 6 ring tones of 2 s length and is presented to the subject twice in random order. A mechanical bell ringing tone with a broad frequency spectrum and a long duration is used, because the localization of broadband signals of long duration is not much affected by the uncalibrated playback level [24]. The level of the music which we use as interference produces a mean segmental signal-to-noise ratio $\overline{SNR_{seg}} = -10.9$ dB (mean over both ears, six directions, and different playback starting times; only segments which contain ringing and background noise were taken into account).

The binaural signals which are fed to the headphones are rendered by the auralization software IKA-SIM [12]. This real-time software simulates specular reflections up to the 3rd order and uses an artificial reverberator for the synthesis of the late reverberation [25]. By omitting the direct path of the music sound source, we remove the main localization cues for the interferer and obtain a diffuse background noise. The virtual test room depicted in Figure 1b has a rectangular shape and a size of 5.50 x 4.50 x 2.70 m^3. The only reflecting surfaces which are modeled by the auralization software are the walls, the ceiling, and the floor. Figure 2 outlines the positions of the six omnidirectional sound sources and the avatar, all are placed at an equal level, 1.20 m above the floor. In order to avoid potential front-back confusions, all sound sources are positioned in the frontal half-plane and the avatar is reset to a default position facing the television before each test trial. The virtual living room has a measured reverberation time of $T_{60} = 0.26$ s. For our test, we neither vary this reverberation time nor the level of the interferer, because varying the level of the background noise over a significant range would shift the emphasis towards a signal-in-noise test and varying the reverberation time over a significant range is not possible without the loss of plausibility (a common living room does not have a reverberation time of 1 s or longer). And as plausibility/naturalness is an important design criterion for an audiological self-screening test which takes behavioral aspects of everyday situations into account, this is not an option.

According to Blauert, "in all cases in which localization is at first imprecise or anomalous, exploratory head movements take on great importance" [26, p.44]. In everyday situations, head movements reduce spatial ambiguity and can to some extent compensate for hearing deficiencies. Therefore, incorporating behavioral aspects into audiological test methods provides additional information on the relevance of an individual's hearing deficiencies in everyday life and leads to improved insights into the perception of everyday sounds [27]. Two kinds of exploratory head movements can be observed: unconscious, spontaneous movements and conscious searching and orienting movements [26, p.178ff]. For the synthesis of a virtual environment, the usage of a head tracker would be highly beneficial, because it allows to take both kinds of interactions into account, but unfortunately, such equipment cannot be used for a self-screening test. A self-screening test conducted at home is restricted to the keyboard and the mouse as exclusive input devices. The same is true for 3D computer games where this constraint proved to be only a minor restriction. Due to this kind of entertainment software, a significant amount of people is familiar with using a mouse or a keyboard to interact with a virtual environment and to conduct conscious, systematic motions. For our test we use an audio-visual display similar to what is used in 3D computer games. But as the interaction is limited to turning the avatar's head, our test has a lower demand on the eye-hand motor coordination. The user interface shown in Figure 1a includes the visual display, a marker for the frontal position, and an enlarged horizontal slider. A short tutorial at the beginning of the test explains the principle of the user interface and allows the subject to become familiar with it. The spoken instructions of the tutorial are detailed in Figure 3.

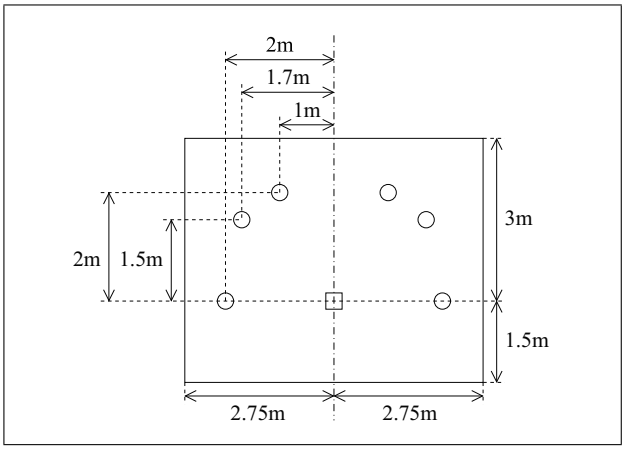


Figure 2. Layout of the virtual test environment (square: avatar position; circles: sound source positions).

To avoid the necessity of prior software installation which might be a barrier for elderly people, we implemented the self-screening test as a web service [28]. The client software is executed in the context of a standard web-browser using a Java applet. This applet remote controls an auralization server running IKA-SIM. The binaural signals are rendered in real-time on the server, streamed over the Internet, and played back by the applet via headphones. For an AVE-based localization test which is provided via the Internet, the system latency is of crucial importance. Previous studies have shown that the influence of the system latency on localization performance for long stimuli is not apparent for latencies below 500 ms [29, 30, 31]. An interactive localization test can be conducted even with large system latencies of up to 1500 ms without significant deviations of the test results [29]. However, such high latencies reduce the user's quality of experience. To avoid user annoyance, the overall delay between user interaction and audio output should be less than 600 ms for the given use case [29]. With a system

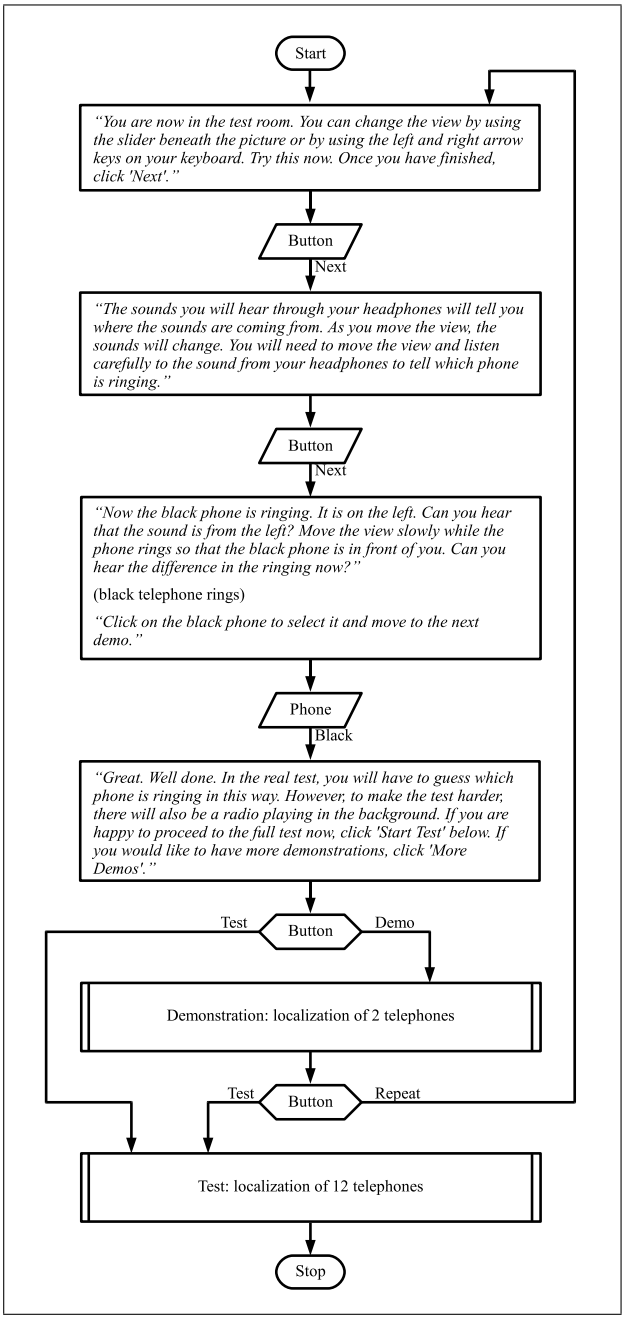


Figure 3. The spoken instructions of the tutorial.

latency of ∼400 ms, our implemented Internet-based self-screening test fullfills this requirement.

The set of head-related transfer functions (HRTFs) which is used to synthesize the binaural signals was selected among several sets of human HRTFs on the basis of informal listening tests [32, section 6.2]. The selected set which resulted in a good localization performance for most of the test participants was measured with 15° azimuth and 10° elevation resolution, interpolated to a full grid with 5° resolution, and post-processed [33]. For practical reasons, a self-screening localization test that is provided via the Internet and addresses the general public can only be implemented with non-individual HRTFs. Furthermore, the test is conducted with uncalibrated end-user equipment. Mismatching HRTFs as well as a deviation of the headphones' frequency characteristics from the assumed diffuse field calibration can result in displaced auditory events or in the worst case in intra-cranial auditory events (inside-the-head localization) whereas the localization test degenerates into a lateralization test [13, 34]. Allowing the subject to turn his virtual head does not only incorporate behavioral aspects but also reduces the front-back ambiguity [35]. While a mismatching headphone equalization may result in auditory events displaced along the cone of confusion[1], mismatching HRTFs may also result in mismatching inter-aural time delays (ITDs) and inter-aural level differences (ILDs) which are the main localization cues [36, 37]. Hence, the accuracy of a virtual localization test is limited by the non-individual HRTFs. For the frontal position, however, a significant lateral displacement will not occur, because ITD and ILD are zero. Hence, by instructing the subject to turn the avatar's head until he faces the sound source, this limitation can be compensated to some extend. As the subject has to click on the telephone in the frontal position, strictly speaking this is not a localization task with a head pointing input method [38], but rather a six-alternative forced choice (6-AFC) closed-loop localization/lateralization task [34, 39]. Furthermore, because the reference groups are subject to the same limitation, the reduced accuracy is indirectly taken into account by the estimated choice statistics.

Instructing the subject to turn the avatar's head until he faces the sound source has an additional advantage in an audio-visual environment where a visual display is used in combination with an auditory display. To render the visual environment on the screen, a certain angle of view is used. However, this angle of view is not necessarily identical with the angle of view seen from the eyes of the subject [40]. Thus, it is quite likely that the visual virtual environment is not congruent with the auditory virtual environment except for the frontal position. Requesting the subject to select the source in the frontal region reduces deviations due to non-congruent audio-visual coordinate systems.

3.1. Test Results

The observed answer sequences as well as the stimulus sequences of the two reference groups were stored in a database to estimate the confusion likelihoods as described in section 2.3. The current training set comprises 24 datasets from 12 localization-impaired subjects and 52 datasets from 26 normally-localizing subjects. The data for the localization-impaired training set originates from the VU Medical Centre in Amsterdam (The Netherlands) which selected 12 subjects with an asymmetrical hearing loss. In contrast to a moderate symmetrical hearing loss of peripheral origin, an asymmetrical hearing loss significantly increases the localization blur [26]. The normally-localizing

[1] If the frequency characteristic of the headphones' left speaker differs from the right speaker or in the case of a misadjusted balance settings, also a lateral displacement may be observed, but this is less likely to happen than a uniform deviation from the assumed diffuse field equalization.

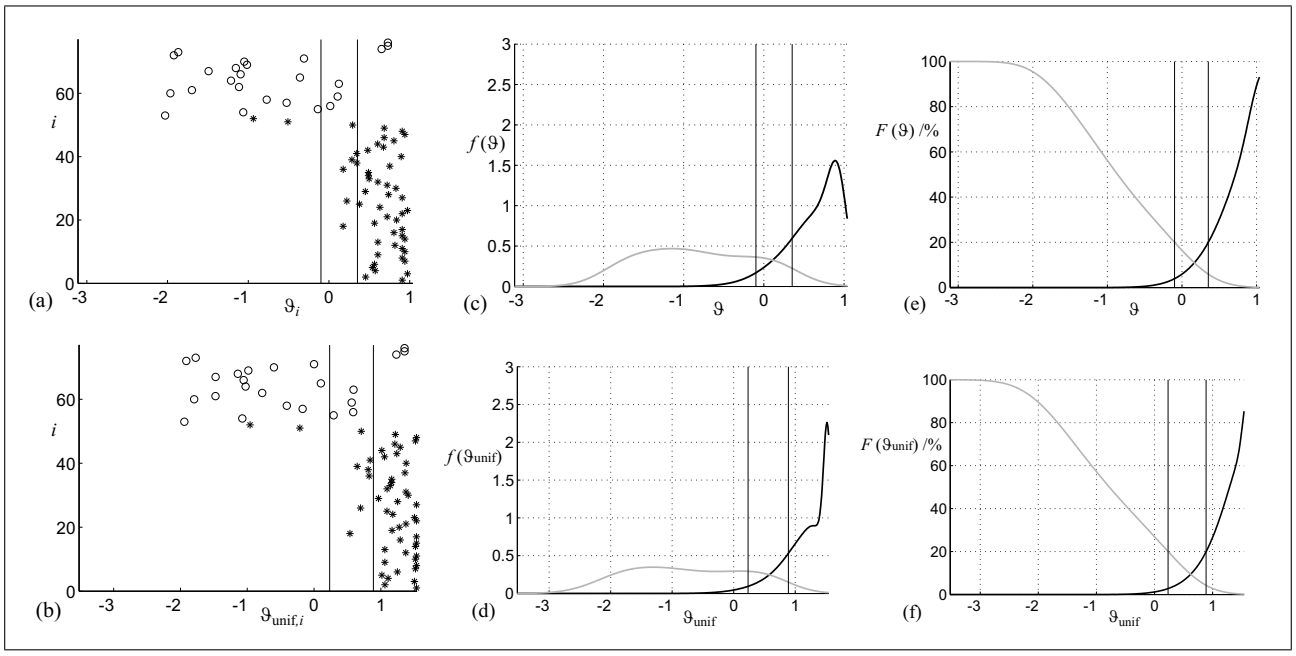


Figure 4. (a) Training set scores (the data set contains two results from each subject; horizontal axis: training set scores ϑ_i; vertical axis: data set number i; circles: localization-impaired set; asterisks: normally-localizing set); (c) estimated PDFs for the training set scores (black: $f_{\mathrm{NL}}(\vartheta)$; gray: $f_{\mathrm{LI}}(\vartheta)$); (e) estimated (complementary) CDFs (vertical lines: derived thresholds); (b), (d), and (f) show the corresponding results for a localization-impaired classification matrix with uniform *a priori* density.

training set comprises 9 subjects tested at the VU Medical Centre and 17 subjects tested at the Ruhr-Universität Bochum (Germany). All subjects conducted our AVE-based localization test twice.

Figure 4a shows the resulting test scores for the complete training set. The vertical lines mark the thresholds derived from the estimated CDFs. High scores indicate a good localization performance and low scores indicate an insufficient localization performance. The classification matrices were determined by omitting two data-sets from the normally-localizing reference group and three data-sets from the localization-impaired reference group (data-sets $i = \{51, 52\}$ and $i = \{74, 75, 76\}$), because they are right in the middle of the score range of the other group and are thus indistinguishable from those results[2]. Figure 4c shows the probability density functions $f_{\mathrm{NL}}(\vartheta)$ and $f_{\mathrm{LI}}(\vartheta)$ estimated from these scores as described in section 2.4. Figure 4e shows the obtained CDF $F_{\mathrm{NL}}(\vartheta) = \int_{-\infty}^{\vartheta} f_{\mathrm{NL}}(s)\,\mathrm{d}s$ and the complementary CDF $F_{\mathrm{LI}}(\vartheta) = \int_{\vartheta}^{\infty} f_{\mathrm{LI}}(s)\,\mathrm{d}s$. The thresholds Θ_{l} and Θ_{u} shown in these figures were selected to yield 80% correctly classified normally-localizing subjects ($\lambda = 20\%$) and 80% correctly classified localization-impaired subjects ($\mu = 20\%$). This selection results in $\alpha = 3.7\%$ and $\beta = 5.8\%$ for the score based on the estimated *a priory* density and $\alpha = 2.7\%$ and $\beta = 4.1\%$ for the score based on the uniform *a priory* density. Figures 4b, 4d, and 4f show the corresponding results for the score $\vartheta_{\mathrm{unif}}$ assuming uniformly distributed conditional probabilities for the classification matrix $\mathbf{M}_{\mathrm{LI}}$.

For the score based on the estimated *a priory* density, the test and retest scores correlate with $\varphi_{\mathrm{tr}} = 0.86$. The same value (rounded to two decimals) can be observed for the score based on the uniformly distributed conditional probabilities. If the subjects with the test score outliers are not taken into account, then the correlation coefficient increases to $\varphi_{\mathrm{tr}} = 0.91$ and $\varphi_{\mathrm{tr}} = 0.92$, respectively.

3.2. Cross-Validation

The results shown in Figure 4 are obtained from classification matrices which were estimated from the same data. Since the classification matrices not only fit the data statistics but also the data itself, they are not appropriate for a validation of the test procedure. This should ideally be done by separate training and validation data-sets. Because of the limited amount of available data, there is no option to split the data into two parts for training and evaluation. A method which is better suited for small data-sets is the leave-one-out cross-validation. In this method, the validation set contains only one single element of the whole data-set, all other elements are used as a training set, and this split-up is repeated with all elements.

Figure 5a shows the resulting test scores for the complete training set using the leave-one-out method. The data-sets $i = \{51, 52\}$ from the normally-localizing set and the data-sets $i = \{74, 75, 76\}$ from the localization-impaired set were again omitted for the determination

[2] If the observed answers of a subject with an assumed localization impairment is indistinguishable from those of the NL group, this indicates that the assigned label was wrong - a consequence of the difficulty to select proper reference groups as stated in section 2.3. Leaving out this data during classification training has the effect that the resulting wrong classification will become even more evident.

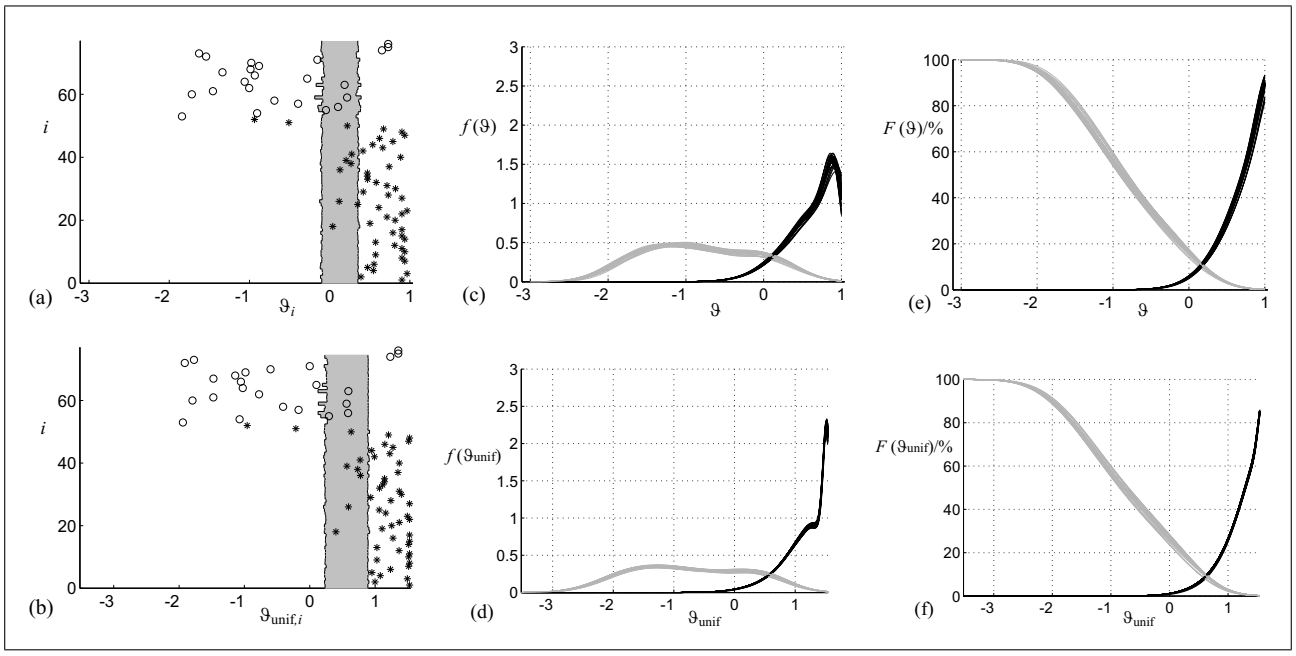


Figure 5. (a) Cross-validation: training set scores (horizontal axis: training set scores ϑ_i; vertical axis: data set number i; circles: localization-impaired set; asterisks: normally-localizing set; left boundary of the gray area: lower threshold Θ_l; right boundary of the gray area: upper threshold Θ_u); (c) estimated PDFs for the training set scores (black: $f_{\mathrm{NL}}(\vartheta)$; gray: $f_{\mathrm{LI}}(\vartheta)$); (e) estimated (complementary) CDFs; (b), (d), and (f) show the corresponding results for a localization-impaired classification matrix with uniform *a priori* density.

of the classification matrices. The left boundary of the gray area marks for $\lambda = 20\%$ the lower threshold Θ_l which varies due to the leave-one-out method. The right boundary of the gray area marks the upper threshold Θ_u for $\mu = 20\%$. Figure 5c shows the resulting set of probability density functions $f_{\mathrm{NL}}(\vartheta)$ and $f_{\mathrm{LI}}(\vartheta)$ which were also obtained using the leave-one-out method. Figure 5e shows the resulting set of (complementary) probability functions $F_{\mathrm{NL}}(\vartheta) = \int_{-\infty}^{\vartheta} f_{\mathrm{NL}}(s)\,\mathrm{d}s$ and $F_{\mathrm{LI}}(\vartheta) = \int_{\vartheta}^{\infty} f_{\mathrm{LI}}(s)\,\mathrm{d}s$. Figures 5b, 5d, and 5f show the corresponding cross-validation results for the score $\vartheta_{\mathrm{unif}}$ based on a localization-impaired classification matrix with uniform *a priori* density.

Compared to the results shown in Figure 4, some variation can be observed for the leave-one-out method. However, the variation is small and good separability is preserved.

3.3. Discussion

In section 2.3 two scoring methods were described. For the more comprehensive first method which resulted in the test score ϑ, both classification matrices, $\mathbf{M}_{\mathrm{NL}}$ and $\mathbf{M}_{\mathrm{LI}}$, were determined by the statistics of the reference groups. For the second method, which resulted in the alternative test score $\vartheta_{\mathrm{unif}}$, a simplified model based on a localization-impaired classification matrix with uniform *a priori* density was used. As stated in section 2.3, the alternative score will overestimate the likelihood ratio. This is clearly visible by means of the shifted values and the shifted thresholds in Figure 4b. If we compare the test scores in Figures 4a and 4b, we will observe similar patterns and similar good separability. The alternative score $\vartheta_{\mathrm{unif}}$ requires less *a priori* knowledge about the localization-impaired population and, more important, does not require a prior proper selection (labelling) of localization-impaired subjects. This means less dependency on other (physical) localization tests. Hence, this is the preferred scoring method.

The thresholds Θ_l and Θ_u for the cross-validation scores in Figure 5a show some variation across the localization-impaired data-set. This is also visible for the scoring procedure based on uniform distributed probabilities (Figure 5b). The greatest deviation can be observed for scores which are close to one of the thresholds. This is because the PDFs are estimated by a Gaussian mixture model and leaving out one of the mixtures results in the greatest deviation at it's mean value i.e., the omitted score. Ideally, the variation should be very small, if the residual data-sets with one element left-out share the same statistics. Larger variations indicate insufficient amount of training data. Because of this, we did not derive the thresholds Θ_l and Θ_u from the error probabilities α and β. Instead, the thresholds Θ_l and Θ_u shown in Figure 4 were selected to yield 80% correctly classified normally-localizing subjects ($\lambda = 20\%$) and 80% correctly classified localization-impaired subjects ($\mu = 20\%$). Due to the steeper slopes of $F_{\mathrm{NL}}(\Theta_\mathrm{u})$ and $F_{\mathrm{LI}}(\Theta_\mathrm{l})$ compared to the slopes of $F_{\mathrm{NL}}(\Theta_\mathrm{l})$ and $F_{\mathrm{LI}}(\Theta_\mathrm{u})$, the confidence is significantly higher (for a database which comprises 24 localization-impaired scores, $\mu = 20\%$ means that this threshold relies on no more than 5 scores which support it). Besides these limitations, we observe that the normally-

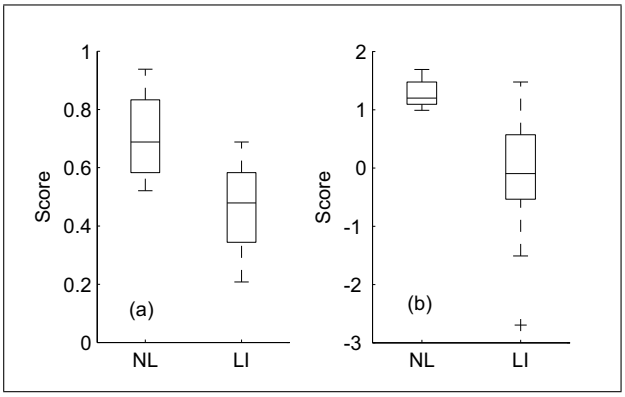


Figure 6. Median values and quartile ranges of the 8-speaker localization test results; (a) legacy method i.e., simple correct scoring; (b) same results evaluated with the proposed method.

localizing and localization-impaired scores are well separable. Both scoring procedures seem to be appropriate to discriminate the localization performance. The same conclusion can be made, if the validity is assessed by a stricter form of the leave-one-out cross-validation method where not only the individual scores but both scores, the test as well as the retest score, of each subject are left out. If both validation results are compared, only marginally score shifts can be observed which are less evident than the difference between Figure 4 and Figure 5. Leaving out two instead of only one score increases the variation of the thresholds, but it does not change any classification result.

The proposed scoring procedure was developed for the presented Internet-based self-screening test, but it is broadly applicable to other N-AFC tests. In [41], the authors apply the proposed scoring procedure to an 8-speaker localization test and compare it to the outcome of the legacy method i.e., simple correct scoring. Figure 6 depicts the median values (central horizontal mark), the quartile ranges (boxes), and the limits (upper/lower horizontal marks) of the published data. While the quartile ranges of the legacy scores adjoin each other, they do not for the proposed method which results in a better separability. Using a single threshold two quartile ranges below the median value of the "normal" scores would yield 100% correctly classified NL subjects and more than 75% correctly classified LI subjects. This is a clear evidence for the assumption that superior classification results can be obtained if the additional information in the different wrong choices is exploited.

4. Summary and Conclusions

A scoring procedure for psychophysical experiments based on forced choices has been developed which classifies the observed sequence of choices in the maximum likelihood sense. The proposed evaluation method is superior to established evaluation methods in terms of exploiting the information contained in the observed sequence of choices, because it takes *a priori* knowledge into account.

The presented method uses the cumulative distribution functions of the test scores as criterion for the selection of the classification thresholds. The proposed algorithm yields the optimal classification thresholds for a desired compromise with regard to the resulting error probabilities α, β, λ, and μ (false-alarm and missed-hit rates).

Furthermore, a novel Internet-based self-screening test for the assessment of localization performance using an interactive auditory virtual environment has been presented and a cross-validation of test results obtained from normally-localizing and localization-impaired subjects has been conducted. The two proposed scores, ϑ and $\vartheta_{\mathrm{unif}}$, the latter being based on a localization-impaired classification matrix with uniform *a priori* density, result in similar separability. As the score $\vartheta_{\mathrm{unif}}$ based on a uniform distribution requires less detailed *a priori* knowledge about the localization-impaired population, we consider this to be the favorable scoring procedure.

Acknowledgement

The authors would like to thank Prof. Torben Poulsen, Department of Electrical Engineering, Technical University of Denmark (DTU), and the anonymous reviewers for their valuable comments and suggestions which helped to improve this article.

References

[1] A. W. Mills: On the Minimum Audible Angle. J. Acoust. Soc. Am. **30** (1958) 237–246.

[2] J. L. Hall: Maximum-Likelihood Procedure for Estimation of Psychometric Functions. J. Acoust. Soc. Am. **44** (1968) 370.

[3] H. Levitt: Transformed Up-Down Methods in Psychoacoustics. J. Acoust. Soc. Am. **49** (1971) 467–477.

[4] D. M. Green: Stimulus Selection in Adaptive Psychophysical Procedures. J. Acoust. Soc. Am. **87** (1990) 2662–2674.

[5] N. J. Versfeld et al.: Method for the Selection of Sentence Material for Efficient Measurement of the Speech Reception Threshold. J. Acoust. Soc. Am. **107** (2000) 1671–1684.

[6] B. Kollmeier and M. Wesselkamp: Development and Evaluation of a German Sentence Test for Objective and Subjective Speech Intelligibility Assessment. J. Acoust. Soc. Am. **102** (1997) 2412–2421.

[7] M. Hällgren et al.: Speech Understanding in Quiet and Noise, With and Without Hearing Aids. Int. J. Audiol. **44** (2005) 574–583.

[8] J. Bench and A. Kowal and J. Bamford: The BKB (Bamford-Kowal-Bench) Sentence Lists for Partially-Hearing Children. Br. J. Audiol. **13** (1979) 108–112.

[9] A. Kohlrausch and S. van de Par: Audio-Visual Interaction in the Context of Multi-Media Applications. – In: Communication Acoustics. J. Blauert (ed.). Springer Verlag, Berlin-Heidelberg-New York, 2005.

[10] T. S. Goverts: Assessment of Spatial and Binaural Hearing in Hearing Impaired Listeners. Dissertation. VU University, Amsterdam, The Netherlands, 2004.

[11] C. E. Shannon: A Mathematical Theory of Communication. The Bell System Technical Journal **27** (Jul., Oct. 1948) 379–423, 623–656.

[12] A. Silzle, P. Novo, H. Strauss: IKA-SIM: A System to Generate Auditory Virtual Environments. 116th AES Convention, Berlin, Germany, 2004.

[13] D. R. Begault, E. M. Wenzel, L. S. Stone, M. R. Anderson: Direct Comparison of the Impact of Head Tracking, Reverberation, and Individualized Head-Related Transfer Functions on the Spatial Perception of a Virtual Speech Source. 108th AES Convention, Paris, France, Feb. 2000.

[14] C. Smits and P. Merkus and T. Houtgast: How We Do It: The Dutch Functional Hearing-Screening Tests by Telephone and Internet. Clinical Otolaryngology **31** (2006) 436–440.

[15] W. M. Hartmann: Localization of Sound in Rooms. J. Acoust. Soc. Am. **74** (1983) 1380–1391.

[16] B. Shinn-Cunningham: Localizing Sound in Rooms. ACM SIGGRAPH and EUROGRAPHICS Campfire: Acoustic Rendering for Virtual Environments, Snowbird, UT, USA, 2001.

[17] B. Shinn-Cunningham and N. Kopco and T. J. Martin: Localizing Nearby Sound Sources in a Classroom: Binaural Room Impulse Responses. J. Acoust. Soc. Am. **117** (2005) 3100–3115.

[18] W. Koenig: Subjective Effects in Binaural Hearing. J. Acoust. Soc. Am. **22** (1950) 61–62.

[19] B. Shinn-Cunningham: Learning Reverberation: Considerations for Spatial Auditory Displays. 6th International Conference on Auditory Display, Atlanta, GA, USA, 2000.

[20] B. Efron and R. J. Tibshirani: An Introduction to the Bootstrap. Chapman & Hall/CRC, Boca Raton, FL, 1993.

[21] A. Wald: Contributions to the Theory of Statistical Estimation and Testing Hypotheses. Ann. Math. Statist. **10** (1939) 299–326.

[22] Internet-Based Self-Screening Test Using an Interactive Auditory Virtual Environment. `http://hearcom.eu/main/Checkingyourhearing/testarea/DE-RUB/testexplain.html`.

[23] ITU-R: Rec. ITU-R BS.775-1, Multichannel Stereophonic Sound System with and without Accompanying Picture. 2006.

[24] W. M. Hartmann and B. Rakerd: Auditory Spectral Discrimination and the Localization of Clicks in the Sagittal Plane. J. Acoust. Soc. Am. **94** (Oct. 1993) 2083–2092.

[25] C. Borß and R. Martin: D-3-7: Improved Internet-based AVE for Demonstration and Self-Screening Purposes. Sep. 2007. Public Report, FP6-004171 HearCom, `http://hearcom.eu/about/DisseminationandExploitation/deliverables/HearCom_D-3-7_V2.0.pdf`.

[26] J. Blauert: Spatial Hearing, The Psychophysics of Human Sound Localization. second ed. MIT Press, Cambridge, MA, 1997.

[27] R. Plomp: The Intelligent Ear: On the Nature of Sound Perception. Lawrence Erlbaum Associates, London, UK, 2002.

[28] C. Borß and A. Silzle and R. Martin: Internet-Based Interactive Auditory Virtual Environment Generators. 14th International Conference on Auditory Display, Paris, France, Jun. 2008. `http://www2.ika.rub.de/publications/2008/borss08internet.pdf`.

[29] A. Silzle and C. Borß: Latency Annoyance of Interactive Auditory Virtual Environments. Proc. Annual Meeting of the German Acoust. Society (DAGA), Stuttgart, Germany, Mar. 2007. `http://www2.ika.rub.de/publications/2007/silzle07latency.pdf`.

[30] Y. Suzuki and S. Yairi and Y. Iwaya: Effect of Large System Latency of Virtual Auditory Display on Listener's Head Movement in Sound Localization Task. 13th Int. Conf. on Auditory Display, Montreal, Canada, June 2007.

[31] E. M. Wenzel: Effect of Increasing System Latency on Localization of Virtual Sounds with Short and Long Duration. International Conference on Auditory Display, Espoo, Finland, Aug. 2001.

[32] A. Silzle: Generation of Quality Taxonomies for Auditory Virtual Environments by Means of Systematic Expert Survey. Dissertation. Institute of Communication Acoustics, Ruhr-Universität Bochum, Bochum, Germany, 2007.

[33] A. Silzle: Quality of Head-Related Transfer Functions - Some Practical Remarks. First ISCA Tutorial & Research Workshop on Auditory Quality of Systems, Akademie Mont-Cenis, Germany, 2003.

[34] G. Plenge: On the Differences Between Localization and Lateralization. J. Acoust. Soc. Am. **56** (1974) 944–951.

[35] E. M. Wenzel et al.: Localization Using Nonindividualized Head-Related Transfer Functions. J. Acoust. Soc. Am. **94** (1993) 111–123.

[36] E. A. Macpherson and J. C. Middlebrooks: Listener Weighting of Cues for Lateral Angle: the Duplex Theory of Sound Localization Revisited. J. Acoust. Soc. Am. **111** (2002) 2219–2236.

[37] M. A. Akeroyd: The Psychoacoustics of Binaural Hearing. Int. J. Audiol. **45** (2006) 25–33.

[38] J. C. Makous and J. C. Middlebrooks: Two-Dimensional Sound Localization by Human Listeners. J. Acoust. Soc. Am. **87** (1990) 2188–2200.

[39] A. W. Bronkhorst: Localization of Real and Virtual Sound Sources. J. Acoust. Soc. Am. **98** (1995) 2542–2553.

[40] W. P. J. de Bruijn and M. M. Boone: Subjective Experiments on the Effects of Combining Spatialized Audio and 2D Video Projection in Audio-Visual Systems. 112th AES Convention, Munich, Germany, May 2002.

[41] J. Lyzenga and J. Koopman: D-11-6: Validation of the Screening Localisation Test and an Inter Comparison of all HearCom Screening Tests. 2008. Public Report, FP6-004171 HearCom, `http://hearcom.eu/about/DisseminationandExploitation/deliverables/HearCom_D11-6_V2p.pdf`.