Spectro-temporal factors in two-dimensional human sound localization

Paul M. Hofman[a] and A. John Van Opstal
University of Nijmegen, Department of Medical Physics and Biophysics, Geert Grooteplein 21, 6525 EZ Nijmegen, The Netherlands

(Received 6 June 1997; revised 5 December 1997; accepted 29 January 1998)

This paper describes the effect of spectro-temporal factors on human sound localization performance in two dimensions (2D). Subjects responded with saccadic eye movements to acoustic stimuli presented in the frontal hemisphere. Both the horizontal (azimuth) and vertical (elevation) stimulus location were varied randomly. Three types of stimuli were used, having different spectro-temporal patterns, but identically shaped broadband averaged power spectra: noise bursts, frequency-modulated tones, and trains of short noise bursts. In all subjects, the elevation components of the saccadic responses varied systematically with the different temporal parameters, whereas the azimuth response components remained equally accurate for all stimulus conditions. The data show that the auditory system does not calculate a final elevation estimate from a long-term (order 100 ms) integration of sensory input. Instead, the results suggest that the auditory system may apply a ''multiple-look'' strategy in which the final estimate is calculated from consecutive short-term (order few ms) estimates. These findings are incorporated in a conceptual model that accounts for the data and proposes a scheme for the temporal processing of spectral sensory information into a dynamic estimate of sound elevation. [S0001-4966(98)03005-7]

PACS numbers: 43.66.Qp, 43.66.Ba, 43.66.Mk [RHD]

INTRODUCTION

Human auditory localization depends on implicit binaural and monaural acoustic cues. Binaural cues arise through interaural differences in sound level (ILD) and timing (ITD) that relate in a simple way to the horizontal component of sound direction relative to the head. The auditory system, however, cannot distinguish, from these cues, between all source positions with the same horizontal component that lie on the so-called ''cone of confusion.'' Due to the front–back symmetry of the ITD and ILD cues, a sound azimuth (α) estimate provided by these binaural cues is therefore ambiguous (e.g., Blauert, 1996; Wightman and Kistler, 1992).

Monaural cues consist of direction-dependent linear spectral filtering caused by the torso, head, and pinnae. Incident waveforms are reflected and diffracted in a complex and direction-dependent way, which typically gives rise to strong enhancement and attenuation at particular frequency bands (Hebrank and Wright, 1974; Shaw, 1974; Mehrgardt and Mellert, 1977; Middlebrooks *et al.*, 1989; Lopez-Poveda and Meddis, 1996). These cues are essential for elevation (ϵ) localization and in resolving front–back ambiguities (e.g., Batteau, 1967; Musicant and Butler, 1984; Blauert, 1996).

The relation between the pinna geometry and the direction-dependent filtering has been shown in both experimental and theoretical studies (Batteau, 1967; Teranishi and Shaw, 1968; Han, 1994; Lopez-Poveda and Meddis, 1996). The actual importance of the filtering in two-dimensional localization has been underlined by various behavioral experiments. In some of these studies, the geometrical structure of the pinna was altered artificially, resulting in a degraded localization performance (Gardner and Gardner, 1973; Oldfield and Parker, 1984b). In other studies, manipulation of the sound spectra and the resulting effects on elevation localization could be related to spectral features in the pinna filters (e.g., Middlebrooks, 1992; Butler and Musicant, 1993).

The features in the spectral pinna filter, or head-related transfer function (HRTF), are believed to carry the information about sound location (Kistler and Wightman, 1992). Yet, a fundamental problem arises in the processing of these features into estimates of sound location (Zakarouskas and Cynader, 1993; Hofman and Van Opstal, 1997). At the level of the eardrums, the available spectrum, $y(\omega;\mathbf{r}_S)$, associated with the source position, $\mathbf{r}_S=(\alpha_S,\epsilon_S)$, results from the source spectrum, $x(\omega)$, filtered by the particular direction-dependent HRTF, $h(\omega;\mathbf{r}_S)$:

$$y(\omega;\mathbf{r}_S)=h(\omega;\mathbf{r}_S)\cdot x(\omega) \tag{1}$$

with ω the frequency in octaves. In principle, the auditory system has no knowledge about the relative contributions of neither $h(\omega;\mathbf{r}_S)$ nor $x(\omega)$. Yet, it is necessary to minimize the source influences in order to recognize $h(\omega;\mathbf{r}_S)$ and thus determine the position, $\mathbf{r}_S$ (mainly elevation and front–back angle, as left–right angle is determined predominantly from binaural cues).

So far, only a few computational models have been proposed, which suggest how this could be done. A possible solution to this problem would be to make a priori assumptions about the sound source spectrum. It would enable the auditory system to directly compare the incoming spectrum to the (stored) HRTFs (e.g., Neti *et al.*, 1992; Zakarouskas

[a]Electronic mail: paul@mbfys.kun.nl

and Cynader, 1993; Middlebrooks, 1992). Neti *et al.* (1992) showed that a feed-forward neural network could be trained to accurately extract the location of broadband noise on the basis of the spectral filter properties of the cat's pinna. Since the model was only trained with white noise, however, it is not clear how it withstands spectral variations.

Zakarouskas and Cynader (1993) proposed a model in which the comparison between the cochlear spectrum and the HRTFs is based on spectral derivatives of first and second order (i.e., $dy/d\omega$, $d^2y/d\omega^2$). For a sound source spectrum that is locally constant, or has a locally constant slope, the model was shown to recognize essential spectral features of the underlying HRTF (specifically relevant peaks and notches), and thus correctly extract the direction of the sound.

In an alternative model, proposed by Middlebrooks (1992), the sensory spectrum is compared with the HRTFs by computing a spectral correlation coefficient. This scheme suggested that localization is accurate if the source spectrum is broadband and sufficiently flat, such that the sensory input maintains maximal correlation with the underlying HRTF of the associated source position.

In the latter two models, accurate localization relies on specific spectral constraints on the source spectrum. This assumption is supported by recent experimental results. On one hand, the auditory system appears to tolerate random variations within the broadband sound spectrum (e.g., Wightman and Kistler, 1989). On the other hand, if relative variations in the source spectrum become too large, localization can be disturbed dramatically (e.g., Middlebrooks, 1992). It therefore remains unclear what the actual spectral constraints are.

So far, the majority of localization studies have applied stimuli with stationary spectral properties. Yet, natural sounds possess a high degree of nonstationarity. Although it is commonly accepted that spectral shape cues play an essential role in elevation detection, it is as yet unknown how the auditory system applies the spectral analysis to nonstationary sensory information. In addition, due to the fundamental relation between the temporal and spectral domains, a sufficiently high spectral resolution requires a minimal time window over which the spectral estimation is integrated: $\Delta f \cdot \Delta T = \text{constant}$. The temporal and spectral resolutions needed for adequate sound localization, however, are not well known.

One possibility is that the sensory information is integrated over a time scale of order, say, 100 ms to obtain an average spectrum on which a spectral (shape) analysis can be applied. If true, sounds with the same average power spectrum on that time scale would be localized equally well: It would allow a considerable amount of freedom for the phase spectrum on that time scale.

An alternative possibility could be that the auditory system applies a ''multiple-look'' strategy, in which elevation estimation is based on multiple, consecutive short-term (say, of order a few ms) spectral analyses of the ongoing sensory information. This latter scheme would imply that the spectral-temporal behavior of the stimulus on that short time scale is also important.

In the present study, we specifically focused on these spectro-temporal aspects for sound localization in two dimensions. To our knowledge, such data are not available in the current literature. In two experiments, the sensitivity of the localization process to short-term spectro-temporal variations was of interest. Localization to frequency-modulated tones (''FM sweeps'') and trains of short-duration broadband bursts was measured. These stimuli had similar broadband average power spectra, but fundamentally different spectro-temporal behaviors. In a third experiment, we estimated the minimal time needed by the localization process to complete elevation estimation. To that means, localization to broadband noise of various durations (3–80 ms) was measured. In a recent pilot study by Frens and Van Opstal (1995), it was shown that localization performance systematically deteriorates in elevation, but not in azimuth, when stimulus duration of broadband noise bursts is shorter than 10 ms.

Saccadic eye movements were used to quantify the response accuracy. It enabled accurate measurements (within 1 deg) of a very early spatial percept (below 200 ms) for stimuli presented within the oculomotor range (35-deg eccentricity range in all directions). Subjects were tested under entirely open-loop conditions, i.e., neither acoustic, nor visual or verbal feedback was provided.

The results of our experiments show a consistent and systematic influence of spectro-temporal factors of the stimulus on the elevation component of the localization response in all subjects tested. These findings will be discussed in terms of a conceptual spectro-temporal model of human auditory localization in two dimensions.

I. METHODS

A. Subjects

Seven male subjects participated in the localization experiments. Their ages ranged from 23 to 39 years. Subjects were employees and students of the department. Three of the subjects (PH, JO, and JG) were experienced in sound localization experiments, whereas the other subjects had no such previous experience and were kept naive as to the purpose of this investigation. Inexperienced subjects were given one or two practice sessions to get acquainted with the setup and localization paradigm and to gain stable performance. All subjects reported to have no hearing problems of any kind.

B. Apparatus

Experiments were conducted in a completely dark and sound-attenuated room with dimensions $3\times3\times3$ m. The walls, floor, and ceiling were covered with acoustic foam, that effectively absorbed reflections above 500 Hz. The room had an A-weighted ambient background noise level of 30 dB SPL.

The orientation of the subject's right eye was measured with the scleral search coil technique (Collewijn *et al.*, 1975; Frens and Van Opstal, 1995). The oscillating magnetic fields that are needed by this method were generated by two orthogonal pairs of square 3×3 m coils, attached to the room's edges: one pair of coils on the left and right walls generated a horizontal magnetic field (30 kHz) and a second pair on the ceiling and floor created a vertical magnetic field (40 kHz).

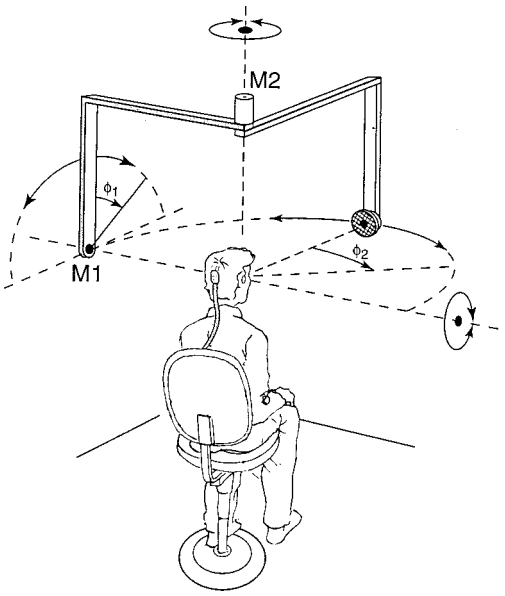

FIG. 1. Experimental setup for delivering acoustic stimuli at various spatial locations. Two stepping motors, M1 and M2, independently control the rotation angles ϕ_1 and ϕ_2, respectively. This construction ensures a fixed distance from the speaker to the center of the subject's eyes (0.9 m) for any stimulus direction (ϕ_1, ϕ_2).

An acoustically transparent frontal hemisphere (consisting of a thin wire framework, covered with black silk cloth) with 85 red light-emitting diodes (LEDs) was used for calibration of the eye-coil measurements and for providing a fixation light at the start of each localization trial. LED coordinates are defined in a two-dimensional polar coordinate system with the origin at the straight-ahead gaze direction. Target eccentricity, $R \in [0,35]$ deg, is measured as the gaze angle with respect to the straight-ahead fixation position, whereas target direction, $\phi \in [0,360]$ deg, is measured in relation to the horizontal meridian. For example, $R=0$ (for any ϕ) corresponds to ''straight ahead,'' and $\phi=0$, 90, 180, and 270 deg (for $R>0$) correspond to ''right,'' ''up,'' ''left,'' and ''down'' positions, respectively. LEDs were mounted at a distance of 85 cm from the subject's eye, at directions $\phi=0,30,60,...,330$ deg and at eccentricities $R=0,2,5,9,14,20,27,35$ deg.

Sound stimuli were delivered through a broad-range lightweight speaker (Philips AD-44725) mounted on a two-link robot (see Fig. 1). The robot consisted of a base with two nested L-shaped arms, each arm driven by a separate stepping motor (Berger Lahr VRDM5) with an angular resolution of 0.4 deg. It enabled rapid (within 2 s) and accurate positioning of the speaker at practically any point on a frontal hemisphere with a radius of 90 cm, the center of which was aligned with the LED hemisphere's center. To prevent spurious echoes, the robot was entirely coated with acoustic foam.

The robot's stepping engines produced some sound while moving which, at first sight, might be suspected to provide additional cues with regard to the speaker position. However, the first engine (M1) always remained in place, at the left of the subject. The sound of the second engine (M2), that moved in the midsaggital plane above the subject, appeared to provide no localizable stimulus cues. This was experimentally verified with two subjects in a previous study (Frens and van Opstal, 1995).

A second potential source for response biases was the speaker displacement between consecutive trials. Especially if a new stimulus position is close to the position of the previous trial, the subject might conclude a small displacement of the speaker from the short duration of motor movements. This potential problem was effectively resolved by incorporating a random movement with a minimal displacement of 20° for each engine and prior to each trial.

Two PCs controlled the experiment: a PC-386 and PC-486 that acted as a master and slave computer, respectively. The master PC was equipped with hardware for data acquisition, stimulus timing and control of the LED hemisphere: eye position signals were sampled with an AD-board (Metrabyte DAS16) at a sampling rate of 500 Hz, stimulus timing was controlled by a digital IO board (Data Translation DT2817) and the LEDs were controlled through a second digital IO board (Philips I2C).

The slave PC controlled the robot and generated the auditory stimuli. It received commands from the master PC through its parallel port. Stimulus generation was done by storing a stimulus in the slave's RAM before a trial and, after receiving a trigger from the timing board in the master PC, passing it through a DA converter (Data Translation DT2821) at a sampling output rate of 50 kHz. The output of the board was fed into a band pass filter (Krohn–Hite 3343) with a pass band between 0.2 kHz and 20 kHz, amplified (Luxman A-331), and passed to the speaker.

C. Sound stimuli

Gaussian white noise (GWN), recorded from a function generator (Hewlett–Packard HO1-3722A) and passed through a band pass filter (Krohn–Hite KH 3343 with pass band 0.2–20 kHz flat within 1 dB) was used as a basis for the noise stimuli. The speaker characteristic was flat within 12 dB between 2 and 15 kHz and was not corrected for. In all experiments (also in each session), the same broadband noise burst with a duration of 500 ms was included as the control stimulus. All stimuli had 1-ms sine-squared onset–offset ramps.

In experiment I (subjects BB, JO, PH), the test stimulus set consisted of broadband noise of various durations, $D=3$, 5, 10, 20, 40, and 80 ms [see Fig. 2(a)]. Each stimulus that was presented within one session was drawn randomly from the recorded noise.

In experiment II (subjects JG, JR, PH), the test stimuli were trains of 3-ms bursts with various duty cycles, $\Delta T=3$, 10, 20, 40, 80 ms [see Fig. 2(b)]. A duty cycle of ΔT ms means that the onsets of consecutive bursts were ΔT ms apart. Each 3-ms burst of the train had been drawn randomly from the recorded noise. A single 3-ms burst stimulus (i.e., same as in experiment I with $D=3$ ms) was included as the limit case $\Delta T=\infty$. The total duration of each burst train was about 500 ms.

In experiment III (subjects KH, PH, VC), the test stimulus set consisted of sweeps of various periods [see Fig. 2(c)]. Inverse Fourier transforms ($N=64,128,...,2048$ points) were

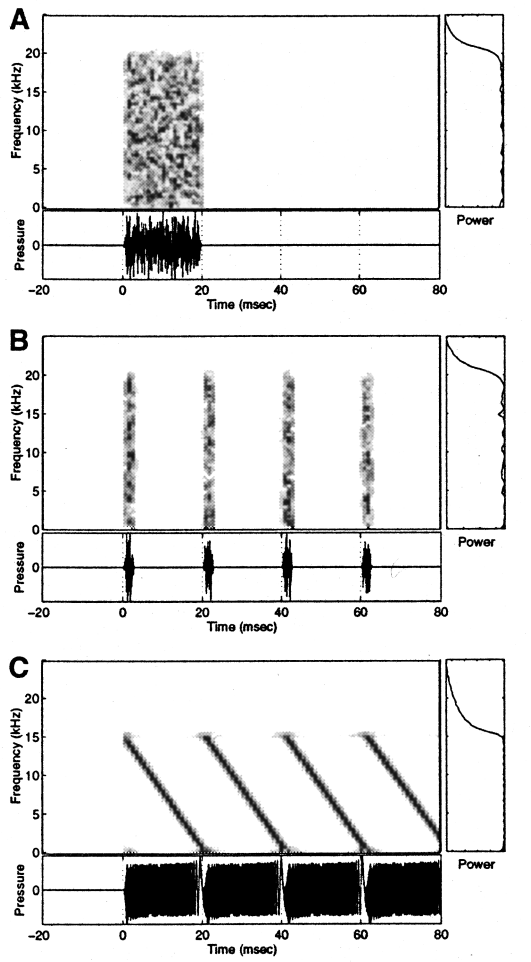

FIG. 2. Examples of the three stimulus types used in the localization experiments: a noise burst with $D=20$ ms (a), a pulse train with $\Delta T=20$ ms (b) and an FM sweep with $T=20$ ms (c). The graphs show the stimuli from 20 ms before stimulus onset until 80 ms after stimulus onset. The large panels show the sonograms, which describe the spectro-temporal behavior of the power spectrum; spectral power is coded by the gray scale, where bright corresponds to low power, dark to high power. The panels on the right contain the time-integrated power spectra (all on the same scale). Finally, the lower panels show the stimulus waveforms.

used to transform an amplitude spectrum, that was flat up to 16 kHz with a corresponding phase spectrum calculated according to Schröder's algorithm, in to the time domain (Schröder, 1970; Wightman and Kistler, 1989). This procedure yields FM sweeps with fixed periods $T=1.28$, 2.56, 5.12, 10.24, 20.48, or 40.96 ms. The instantaneous characteristics of this stimulus are narrow band with a center frequency that (repetitively) traverses the entire frequency range downwards with a constant velocity of $16/T$ kHz/s. The total stimulus duration was 500 ms.

Examples of the spectograms of synthesized stimuli that were used in the experiments (before being passed through the speaker) are shown in Fig. 2: a noise burst with duration $D=20$ ms, a burst train with duty cycle $\Delta T=20$ ms, and a sweep with period $T=20$ ms. As can be seen, the noise burst contains spectral power over the entire frequency range and over the entire stimulus duration. In a burst train, the total stimulus duration is long and the stimulus is broadband, but power is only present during short 3-ms intervals. Like the noise burst, the sweep contains no silence periods throughout the whole stimulus duration. The sweep can be considered ''broadband'' if one regards a whole period, yet narrow band on smaller time scales. Thus all stimuli have flat, broadband time-averaged power spectra, whereas the spectro-temporal behavior is fundamentally different for the three stimulus types.

The control stimulus and all test stimuli had equal rms values. For the burst trains this rms value refers to the non-silence periods. The A-weighted sound level at which the stimuli were delivered was 70 dB, measured at the subject's head position (measuring amplifier Brüel & Kjær BK2610 and microphone Brüel & Kjær BK4144).

D. Stimulus positions

In this paper, the coordinates of both the oculomotor response and the sound source position are described in a double-pole coordinate system, in which the origin coincides with the center of the head. The horizontal component, azimuth α, is defined by the stimulus direction relative to the vertical median plane, whereas the vertical component, elevation ϵ, is defined by the stimulus direction relative to the horizontal plane.

The stimulus positions were confined to 25 ''boxes'' centered at azimuths $\alpha=0,\pm 13,\pm 26$ deg, and elevations $\epsilon=0,\pm 13,\pm 26$ deg [see, e.g., Fig. 5(a)]. The dimensions of each box were 8 deg$\times$8 deg, limiting the total stimulus range in both azimuth and elevation to $[-30,30]$ deg. Sets of 25 stimulus positions were composed by randomly selecting a position within each box. Already for one set, this selection procedure ensured a high degree of uncertainty in the stimulus position while maintaining a homogeneous distribution over the oculomotor range. Importantly, the number of stimulus positions could thus be limited for each stimulus condition, which was highly desirable as each experiment had seven different conditions.

E. Paradigm

The eye position in a head-fixed reference frame was used as an indicator of the perceived sound location (see also Frens and Van Opstal, 1995). In order to calibrate the eye coil, each session started with a run in which all 84 peripheral LEDs on the hemisphere were presented in a random order. Subjects were instructed to generate an accurate saccade from the central fixation LED at 0-deg eccentricity to the peripheral target, and to maintain fixation as long as the target was visible. After calibration, the eye position was known with an absolute accuracy of 3% or better over the full oculomotor range.

In the subsequent runs, sound stimuli were presented. A trial always started with the central visual fixation stimulus. Then, after a random period of 0.4–0.8 s, the LED was switched off and the sound was presented at some peripheral location. The subject's task was to direct the eyes as fast and as accurately as possible toward the apparent sound location without moving the head. A firm head rest enabled the subject to stabilize his head position throughout the session.

A typical run with sound stimuli consisted of the control stimulus and six other sound stimulus conditions. During the 175 consecutive trials, each stimulus was presented once at a set of 25 pseudo-randomly drawn positions (see above). The order of stimulus conditions and positions throughout a session was randomized.

Each subject participated in four sessions on four different days. Hence, each subject traversed a total number of 700 localization trials. Each of the seven temporally defined stimuli was presented 100 times in total four times within each of the 25 stimulus boxes. Subject PH participated in all three experiments (2100 trials).

F. Data analysis

Eye positions were calibrated on the basis of responses to 85 visual stimuli in the first run of the session. From this run, sets of raw eye position signals (AD values of the horizontal and vertical position channel) and the corresponding LED positions (in azimuth and elevation) were obtained. LED azimuth, α, and elevation, ϵ, were calculated from the polar coordinates, (R, ϕ), of the LEDs by:

$$\alpha = \arcsin(\sin R \cos \phi),$$
$$\epsilon = \arcsin(\sin R \sin \phi). \tag{2}$$

These data were used to train a three-layer backpropagation neural network that mapped the raw data signals to calibrated eye position signals. In addition, the network also corrected for small inhomogeneities of the magnetic fields and a slight crosstalk between the horizontal and vertical channels.

A custom-made PC program was applied to identify saccades in the calibrated eye-position signals on the basis of preset velocity criteria for saccade onset and offset, respectively. The program enabled interactive correction of the detection markings. The endpoint of the first saccade after stimulus onset was defined as the response position (see also Sec. II). If saccade latency re. stimulus onset was less than 80 ms or exceeded 500 ms, the saccade was discarded from further analysis. Earlier or later stimulus-evoked responses are highly unlikely for a normal, attentive subject, although the precise values of the boundaries are somewhat arbitrary. Earlier responses are generally assumed to be predictive and are very inaccurate, even for visually evoked saccades. Later responses are considered to be caused by inattention (see also Sec. II).

Response positions versus stimulus positions were fitted for the respective components with a linear fit procedure that minimizes the summed absolute deviation (Press *et al.*, 1992). This method is less sensitive to outliers than the more common least-squares method. For the same reason, the correlation between response- and stimulus positions was quantified by the nonparametric rank correlation coefficient, rather than by Pearson's linear correlation coefficient.

Confidence levels for both the linear fit parameters and the correlation coefficients were obtained through the bootstrap method (Press *et al.*, 1992), since explicit expressions for the confidence levels in the methods described above are not available. In the bootstrap method, one creates N synthetic data sets by randomly selecting, with return, data points from the original set (N typically about 100). A synthetic set has the same size as the original set, so that a given data point from the original set can occur more than once in the synthetic set. The parameter of interest is then computed for each synthetic data set and the variance in the resulting N parameter values is taken as the confidence level.

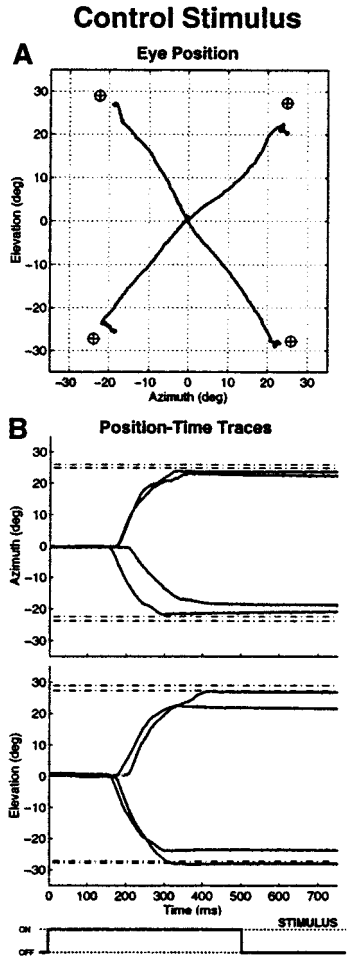

FIG. 3. Typical saccadic eye movement responses towards the control stimulus (GWN, $D = 500$ ms) at four different eccentric positions. (a): Eye-position trajectories during the full duration of the trial. (b): Corresponding position-time traces for both azimuth and elevation during the first 750 ms after stimulus onset. Stimulus positions are indicated by the visor symbols (a) and by the dot-dashed lines (b). Stimulus timing is indicated at the bottom. Data from naive subject BB.

II. RESULTS

A. Control condition

Characteristics of typical responses to four eccentric stimulus positions in the control condition are plotted in Fig. 3. It shows the eye-position trajectory in space and the separate eye-position components (i.e., azimuth and elevation) versus time. One can see that the offset position of the eye (i.e., the response position) corresponds closely to the stimulus position. The response is accurate for both azimuth and elevation components.

It can be seen from Fig. 3(b) that the responses follow shortly after stimulus onset. In these examples, primary sac-

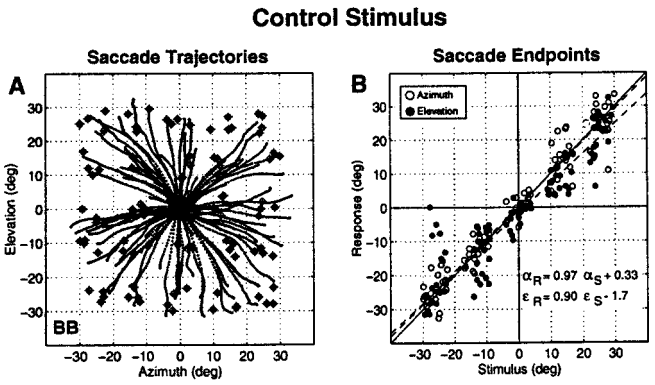

FIG. 4. Saccadic responses to the control stimulus. (a) Trajectories of primary saccades (thick dotted lines, sampled at 2-ms intervals) to control stimuli presented throughout the entire stimulus range. Stimulus positions are indicated by the visor symbols. (b) Endpoint positions of the same primary saccades as in (a) versus stimulus positions for both azimuth (○) and elevation (●). Also the linear-fit results for response positions versus stimulus position are provided (azimuth: α_S vs α_R; elevation: ϵ_S vs ϵ_R). Note large slopes (0.97 and 0.90, respectively; $N=98$) and small offsets (within 2 deg). Subject BB.

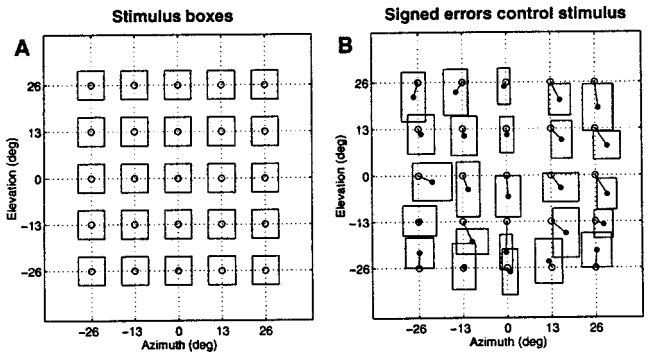

FIG. 5. (a) Stimulus positions were presented within the square boxes, with dimensions 8×8 deg. Sets of 25 stimulus positions were composed by selecting a position at random within each box. (b) The lines from the open symbols to the closed symbols correspond to local mean signed errors. For each stimulus box, the mean signed error is computed by averaging the individual mean signed errors of all seven subjects. The width and height of a box correspond to twice the standard deviation in the signed errors for azimuth and elevation, respectively. The center of each stimulus box is indicated by the open circle [in both (a) and (b)]. Note the clear separation of almost all response boxes, and the larger response scatter in the elevation components as compared to the azimuth components.

cades are initiated with a latency of approximately 200 ms, and are completed within 400 ms after stimulus onset (i.e., before stimulus offset time).

All subjects in this study made accurate localization responses to the control stimulus (500-ms GWN) in all directions (azimuth and elevation). Figure 4 shows all saccade trajectories [Fig. 4(a)] and saccade endpoints [Fig. 4(b)] for all four sessions of one of the subjects (BB). Note how both the stimulus positions and saccade trajectories are distributed over the entire stimulus range. Note also, that the accuracy of response elevation is both quantitatively and qualitatively similar to azimuth localization. A summary of the results for all subjects is listed in Table I.

Figure 5(b) gives an overall impression of the local localization accuracy by showing the averaged signed errors of the saccadic responses for each stimulus box. For each subject and each stimulus box, the mean signed error was computed for all responses to stimuli presented within that box (typically 4–8 responses for each box and each subject). Then, a final mean signed error for each stimulus box was obtained by averaging the subject-mean signed errors over all 7 subjects.

Note that, for the majority of the error boxes in Fig. 5(b), the height is larger than the width. Thus the scatter in the responses is somewhat larger for elevation than for azimuth. These scatter properties underline the fact that azimuth and elevation localization are dissociated processes (see the Introduction). They clearly contrast with the scatter properties of visually evoked saccadic responses, that betray the polar organization of the visuomotor system (e.g., Van Opstal and Van Gisbergen, 1989).

One may observe that the highest accuracy for azimuth localization is reached near the median plane, whereas elevation accuracy is approximately homogeneous over the entire oculomotor field. The largest signed errors were found for

TABLE I. Parameters of the azimuth (α) and elevation (ϵ) response components to the control stimulus (500-ms white noise) for each subject. Columns 2–3: rank correlation coefficient ρ between response position and stimulus position. Note that correlation coefficients $\rho_\epsilon > 0.90$ for elevation and $\rho_\alpha > 0.95$ for azimuth. Columns 4–5: slope, or gain, G, of a straight-line fit for response versus stimulus position. Columns 6–7: standard deviation Δ_{FIT} of the difference between the actual response and the response predicted by the fit (in degrees). Columns 8–9: average absolute localization error Δ_{RESP} (in degrees). Column 10: median response latency (in ms). The two bottom rows present, for each column, the mean and standard deviation, respectively, pooled for all subjects.

	Corr ρ		Gain G		Δ_{FIT}		Δ_{RESP}		
Subject	α	ϵ	α	ϵ	α	ϵ	α	ϵ	Lat.
JO	0.97	0.91	1.1	0.67	4.3	4.9	5.1	6.3	166
BB	0.97	0.94	0.97	0.90	4.5	6.2	3.3	4.9	198
VC	0.96	0.91	0.96	0.81	5.0	6.8	3.8	8.3	226
KH	0.96	0.92	1.0	0.88	5.7	5.5	5.2	4.6	244
JR	0.97	0.95	1.1	1.1	5.0	6.9	4.3	5.9	148
JG	0.97	0.93	1.3	1.0	4.6	6.4	5.4	7.8	156
PH	0.98	0.95	1.1	1.1	3.9	5.9	4.4	6.4	166
mean	0.97	0.93	1.06	0.91	4.6	6.0	4.5	6.3	186
s.d.	0.01	0.02	0.11	0.15	0.7	0.8	0.9	1.5	38

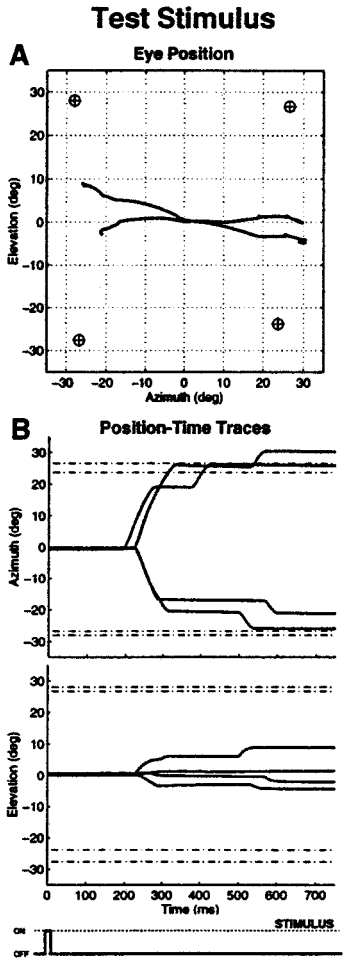

FIG. 6. Typical responses toward short-duration noise bursts ($D=3$ ms) at four different eccentric positions. Note marked undershoot of the elevation components, whereas the azimuth components remain accurate (cf. Fig. 3). Note further that also small secondary saccades are made after approximately 400 ms. See legend of Fig. 3 for further details.

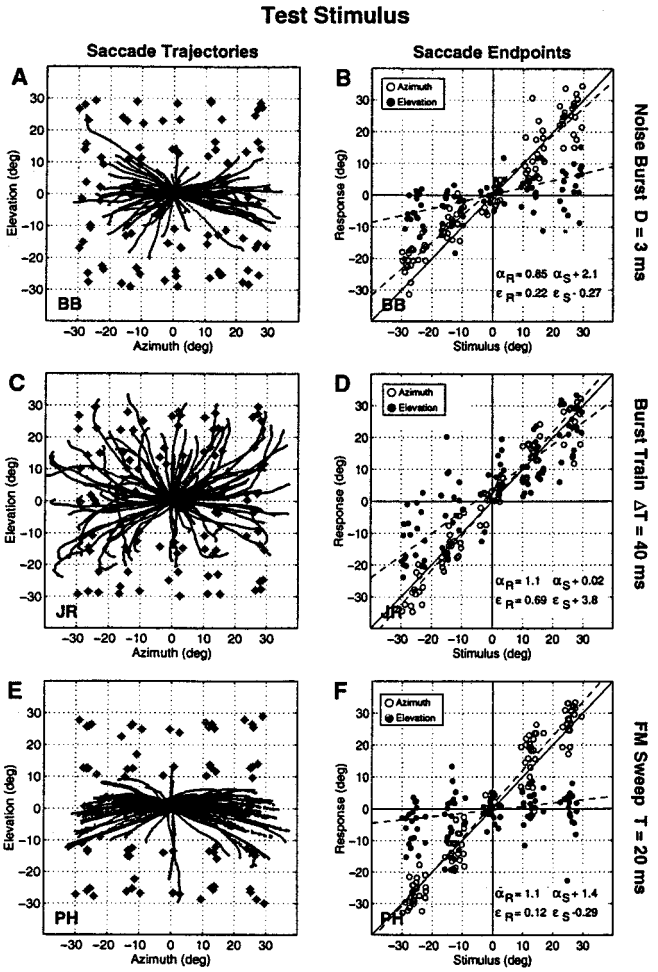

FIG. 7. Saccadic responses to test stimuli. (a) and (b) responses of subject BB to noise bursts of duration $D=3$ ms (correlations for azimuth and elevation: $\rho_\alpha=0.97$, $\rho_\epsilon=0.69$; $N=95$). (c) and (d) responses of subject JR to pulse trains with duty cycle $\Delta T=40$ ms ($\rho_\alpha=0.97$, $\rho_\epsilon=0.86$; $N=97$). (e) and (f) responses of subject PH to FM sweeps with period $T=20$ ms ($\rho_\alpha=0.96$, $\rho_\epsilon=0.38$; $N=143$). Correlations for response position versus stimulus position are listed in Tables II, III, and IV. See legend of Fig. 4 for further details.

stimulus positions within the box at $(\alpha,\epsilon)=(0,-13)$ deg, where response elevations were, on average, about 8 ± 5 deg lower than the actual stimulus elevations.

B. Test conditions

Typical responses to four eccentric stimulus positions in one of the test conditions (noise burst, duration $D=3$ ms) are plotted in Fig. 6. It is clear that azimuth localization is accurate. However, in contrast to the control condition, where elevation detection was also accurate, the responses now exhibit large undershoots in elevation.

Typical responses for the three different stimulus types employed in this study are shown in Fig. 7 (data from subjects BB, JR, and PH, respectively). The most obvious feature of these data, obtained for all three test conditions, is that the saccade trajectories cover only part of the vertical stimulus range [compare Fig. 7(a) with Fig. 4(a)]. Compared to the control condition (see Fig. 4), localization accuracy has clearly deteriorated for elevation, whereas it has remained the same for azimuth [Fig. 7(b), open symbols]. However, although elevation accuracy has deteriorated in these test conditions, the correlation between stimulus elevation and response elevation is still highly significant. This can also be seen in Tables II, III, and IV, which summarize the regression results for all subjects and the test conditions.

In summary, we found for all spectro-temporal stimuli that, although response elevation could be incorrect, it was quite consistent and correlated highly with the real stimulus elevation: $\rho_\epsilon \geqslant 0.4$ for all conditions. This is also expressed by the difference, Δ_{FIT}, between the responses and the straight line fits (dashed lines in Fig. 4 and Fig. 7): Δ_{FIT} remained within 4–8 deg for elevation and within 3–6 deg for azimuth for all stimulus conditions and subjects. Hence, the variance in the responses did not increase substantially with respect to the control condition (compare with Table I).

C. Response gain

Next, we compared the responses for each subject for the different temporal parameters D, ΔT, and T with the results from the control stimulus ($D=500$ ms). The straight

TABLE II. Parameters of the responses to noise-burst stimuli with various durations D (in ms) for three subjects. Note that elevation gain increases systematically with increasing D, but that the variability in the data (Δ_{FIT}) is independent of D. Azimuth gain, however, is not affected by D. Note also, that the median latencies tend to decrease as D increases. See Table I for further explanation of the columns.

		Corr ρ		Gain G		Δ_{FIT}		Δ_{RESP}		
Subject	D	α	ϵ	α	ϵ	α	ϵ	α	ϵ	Lat.
BB	3	0.97	0.69	0.85	0.22	4.9	5.9	4.9	13	210
	5	0.96	0.71	0.81	0.33	5.5	6.4	5.1	11	213
	10	0.96	0.79	0.86	0.30	5.1	4.8	4.7	11	207
	20	0.97	0.89	0.87	0.50	5.4	5.6	4.4	9.1	191
	40	0.97	0.94	0.91	0.63	4.6	5.3	3.7	6.7	188
	80	0.95	0.94	0.90	0.76	6.4	5.8	5.0	5.7	190
JO	3	0.98	0.80	0.99	0.28	3.4	4.1	4.7	12	182
	5	0.98	0.76	1.0	0.35	3.4	5.1	4.1	11	180
	10	0.97	0.80	1.0	0.43	3.8	5.0	4.4	10	173
	20	0.97	0.88	1.0	0.42	4.1	4.6	4.3	9.8	164
	40	0.96	0.92	1.0	0.56	4.3	4.4	4.8	7.2	164
	80	0.97	0.92	1.1	0.61	4.0	4.8	4.7	7.0	163
PH	3	0.97	0.85	1.1	0.67	4.4	8.0	4.4	10	188
	5	0.97	0.87	1.1	0.70	4.2	7.1	4.4	9.4	180
	10	0.97	0.87	1.1	0.82	4.6	7.5	4.7	10	172
	20	0.96	0.89	1.1	0.81	4.4	7.1	4.9	8.3	164
	40	0.98	0.93	1.1	0.99	3.3	6.2	3.7	6.7	164
	80	0.97	0.94	1.1	1.0	4.1	6.0	4.2	6.1	162

line appeared to yield a reasonable description of the relation between response versus stimulus position, and the slope of the line turned out to be a characteristic parameter for the responses in a given condition. One can immediately see from Fig. 8 that the response gains for elevation varied systematically with all three temporal parameters (bottom panels). The azimuth component of the saccades was unaffected by the stimulus parameters (top panels).

For the noise burst, the same systematic variation with D is observed for all three subjects: the response elevation gain increases gradually with stimulus duration D, from $D=3$ ms up to $D=80$ ms. Although there is some intersubject variability as to the absolute values of these gains, all subjects followed a similar trend. Note that similar quantitative inter-subject differences were also obtained for the control stimuli. Furthermore, as the control stimulus is a noise burst with $D=500$ ms, the results suggest that for broadband noise bursts responses stabilize at roughly $D=80$ ms. Note that the

TABLE III. Parameters of the responses to the burst-train stimuli with various duty cycles ΔT (in ms). Elevation gain, but not azimuth gain, depends systematically on ΔT. $\Delta T=\infty$ refers to the single 3-ms burst stimulus. Median latencies tend to decrease with ΔT for all three subjects. See Table I for explanation of the columns.

		Corr ρ		Gain G		Δ_{FIT}		Δ_{RESP}		
Subject	ΔT	α	ϵ	α	ϵ	α	ϵ	α	ϵ	Lat.
JR	3	0.98	0.94	1.1	1.1	4.4	6.9	3.8	6.4	148
	10	0.98	0.93	1.1	1.0	4.4	7.3	3.9	5.7	154
	20	0.98	0.90	1.2	0.86	4.2	7.7	4.0	7.0	160
	40	0.97	0.86	1.1	0.69	5.0	7.7	4.2	8.6	166
	80	0.97	0.85	1.1	0.61	4.9	6.7	4.0	8.9	168
	∞	0.96	0.74	1.1	0.49	5.1	7.8	4.2	10	166
JG	3	0.98	0.94	1.2	1.0	4.4	5.9	5.2	6.6	156
	10	0.98	0.87	1.2	0.82	4.1	7.7	4.9	8.0	158
	20	0.96	0.93	1.2	0.73	5.5	4.7	5.6	7.1	166
	40	0.97	0.92	1.2	0.77	4.8	5.8	5.5	7.1	170
	80	0.97	0.92	1.3	0.77	4.9	6.2	5.7	6.1	182
	∞	0.97	0.83	1.3	0.55	5.7	7.9	6.8	12	176
PH	3	0.98	0.92	1.1	1.0	4.4	7.5	4.2	6.1	164
	10	0.97	0.96	1.1	0.97	4.4	5.1	4.1	5.8	178
	20	0.97	0.91	1.1	0.83	4.3	6.5	4.6	9.0	170
	40	0.97	0.91	1.1	0.71	4.3	6.0	4.7	10	182
	80	0.97	0.82	1.1	0.59	4.0	6.9	4.7	11	190
	∞	0.97	0.84	1.1	0.76	4.4	8.1	4.5	11	182

TABLE IV. Parameters of the responses to sweep stimuli with various repetition periods T (in ms). Elevation gain, but not azimuth gain, depends systematically on T. Median latencies tend to increase with T for subject PH, but not for subjects VC and KH. See Table I for explanation of the columns.

		Corr ρ		Gain G		Δ_{FIT}		Δ_{RESP}		
Subject	T	α	ϵ	α	ϵ	α	ϵ	α	ϵ	Lat.
VC	1.3	0.95	0.79	0.92	0.58	5.4	8.5	4.4	13	244
	2.6	0.96	0.87	0.84	0.63	4.2	7.2	4.5	9.8	232
	5.1	0.96	0.70	0.86	0.27	5.0	6.9	5.2	12	236
	10	0.95	0.55	0.87	0.15	5.4	6.8	4.8	13	237
	20	0.96	0.43	0.94	0.14	5.4	5.8	4.6	14	234
	41	0.95	0.47	0.97	0.15	5.6	7.2	4.7	14	240
PH	1.3	0.97	0.91	1.1	0.91	4.3	7.2	4.0	10	164
	2.6	0.97	0.91	1.1	0.81	5.6	7.6	4.7	6.5	163
	5.1	0.96	0.86	1.0	0.64	4.8	6.9	4.0	7.9	170
	10	0.97	0.43	1.2	0.15	5.3	8.1	5.5	14	168
	20	0.96	0.38	1.1	0.12	4.8	6.5	4.3	14	188
	41	0.96	0.63	1.2	0.20	6.2	4.1	7.1	14	184
HK	1.3	0.94	0.84	0.95	0.55	5.7	7.1	4.9	8.3	251
	2.6	0.95	0.84	0.81	0.66	5.3	7.4	5.0	8.4	246
	5.1	0.96	0.85	0.94	0.43	4.7	5.0	4.4	12	269
	10	0.95	0.57	0.87	0.29	4.8	7.5	4.3	15	243
	20	0.93	0.41	0.79	0.13	4.7	5.9	5.3	20	258
	41	0.97	0.60	0.93	0.19	4.6	5.3	3.9	18	239

short bursts already account for relatively large gains: for example, the gains at $D=10$ ms are already about 40%–80% of the final gains obtained for $D=80$ ms.

For the burst-train stimulus, response gain decreases monotonously with the duty cycle ΔT for subjects PH and JR. The response gain of subject JG, although displaying a similar overall trend, does not vary significantly for the intermediate values of $\Delta T=10,20,40,80$ ms. In this experiment, the intersubject variability for the gains is small for both the burst trains and the control condition. For the shortest duty cycles applied ($\Delta T=3$ ms), the gain is similar as for the control condition (indicated by C). In addition, for subject PH who also participated in experiment I, the gain for $\Delta T=80$ ms is very similar to the gain observed for the single noise burst at $D=3$ ms (condition $D3$ in Fig. 8). This result could be expected: the latency of this subject's response lies around 150 ms (see Table III), so that only the first burst (i.e., at $t=0$ ms) and maybe the second one (at $t=80$ ms) could actually have been processed by the auditory system for generating the first saccade.

For the sweeps, the elevation gain decreases when the period T increases. In contrast to the noise burst and the burst train, the change with the temporal parameter is more sudden. Approximately $T=5$ ms seems to be a critical value, where the response elevation changes most rapidly with T. From $T=10$ ms (subject KH) or $T=20$ ms (subjects PH and VC) there is little change in the gain which lies between 0.1 and 0.2. Although the gains for $T\geqslant 10$ ms are relatively low, correlations for the sweep data were still 0.38 or higher.

D. Response latency

The latencies of primary saccades are typically well below 300 ms. Figure 9(a) shows a latency distribution for saccades to the control stimulus (subject PH). One can see that latencies peak near 170 ms and remain within the [100,300]-ms interval. This interval was typical for all subjects, which can be further appreciated from the cumulative latency distributions in Fig. 9(b). Data from three different subjects for the control condition are presented: latencies that were relatively short (subject JR), long (VC), and intermediate (PH). Note that, also for subject VC, more than 90% of the latencies remained below 300 ms.

It may be observed that in Fig. 9(b) the curves are nearly linear and roughly parallel to each other (see Sec. III). This is most obvious for the distributions of subjects PH and VC, which are well-defined, since they consist of a relatively high number of saccades. For other subjects, fewer saccades were available for each condition, but the distributions exhibited roughly similar characteristics [e.g., subject JR, Fig. 9(b)].

For subject PH, also cumulative distributions are shown that resulted from the several short-noise burst experiments. One can see that latencies systematically increase as the duration of the noise burst decreases. The same trend was observed for the other subjects (BB, JO) who participated in this experiment (see Table II). Also in the other experiments, a systematic shift of the (reciprocal) latency distribution as function of the temporal stimulus parameter was observed as well (except for subjects VC and KH in response to FM sweeps). Median latencies also show this trend (see Tables II, III, and IV). For different stimulus conditions, the reciprocal latency distributions differed in their offset, but retained their shape.

E. Primary and secondary saccades

The endpoint of the primary saccade was accepted as a valid response when its onset latency fell in the interval [80,500] ms. For the average subject, a valid response was measured in $97\%\pm 2\%$ of the trials. In the same time inter-

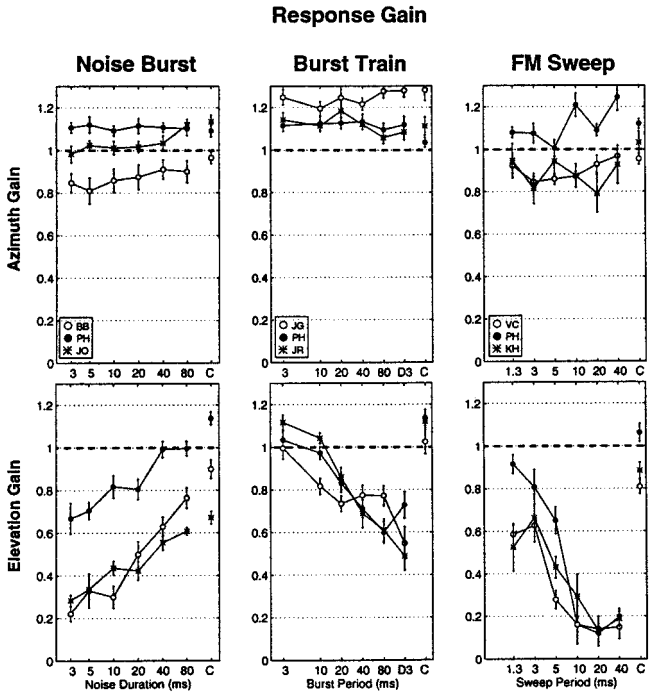

FIG. 8. Response gains for all subjects and all conditions. The gain is defined as the slope of the straight line fitted through response component versus stimulus component (see also Fig. 4 and Fig. 7). Top panels: azimuth gains. Lower panels: elevation gains. Gains are plotted as function of the temporal stimulus parameters D, ΔT, and T. The control condition is indicated by C. The single-burst condition in the burst-train experiments is labeled by D3. Optimal agreement of response and stimulus position is associated with gain 1.0 (dashed lines). Lowest correlations ρ_ϵ obtained for noise bursts, burst trains and sweeps were 0.58, 0.83, and 0.38, respectively, but still highly significant (see Tables II, III, and IV).

val, a secondary saccade was observed in $23\% \pm 9\%$ of the trials, and a third saccade in less than 4% of the trials.

To test whether the secondary saccade was corrective (which is known to be the case for visually evoked saccades), the unsigned errors after the primary saccade, and after the secondary saccade were compared for each subject. This was done for saccade azimuth α, elevation ϵ, amplitude R and direction ϕ. The analyses revealed that incorporating the second saccade did not significantly change (i.e. neither improve nor deteriorate) response accuracy (data not shown).

To further check for a possible relation of the secondary saccade with the stimulus position, an additional analysis was performed by comparing the directions of the primary saccade, the secondary saccade and the stimulus. First, the difference in direction, $\Delta\phi_{12}$ between the primary and the secondary saccade was computed. For all subjects pooled, it was found that $\Delta\phi_{12} = 4 \pm 48$ deg. The individual results for each subject were similar. Thus the secondary saccade generally proceeds in the same direction as the primary saccade. Next, $\Delta\phi_{12}$ was compared to the difference in direction, $\Delta\phi_{01}$, between the primary saccade and the stimulus position (i.e., the direction error). No significant correlation between $\Delta\phi_{12}$ and $\Delta\phi_{01}$ was found. Thus the secondary saccade does not correct for a residual direction error after the primary saccade either.

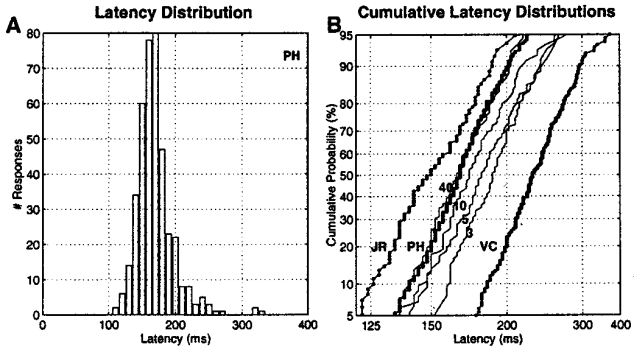

FIG. 9. (a) Response latency distribution for the control condition. Subject PH. (b) Cumulative response latency distributions. The latency axis has a reciprocal scale, whereas the abscissa is on probit scale. In this format, a Gaussian distribution of reciprocal latency results in a straight line. The thick dotted lines refer to the control conditions for subject JR, VC, and PH; the distributions consist of 99, 219, and 398 saccades, respectively. The thin lines show the cumulative distributions of responses to short noise bursts with $D = 3$, 5, 10, 40 ms (subject PH). For each condition D, about 116 saccades are included.

III. DISCUSSION

A. General findings

In the present study, human localization experiments were performed using a wide range of acoustic stimuli. The time-averaged power spectra of the stimuli were always broadband and identical in shape, but the spectro-temporal behavior on a millisecond time scale was fundamentally different. Localization performance varied systematically with the experimental parameters T, ΔT, and D. Whereas elevation detection appeared to be very sensitive to the spectro-temporal stimulus behavior, azimuth localization remained unaffected and was equally accurate for all conditions.

Our findings provide new insights into the spectro-temporal processing of acoustic sensory information. The data suggest specific temporal constraints for accurate acoustic localization of stimulus elevation. Moreover, these results underline the presence of separate dynamical processes underlying the analysis of the different acoustic cues for the detection of azimuth (ITD, IID) and elevation (spectral shape cues).

B. Saccadic eye movements

1. Orienting to sounds through saccadic eye movements

Saccadic eye movements were used to measure the perceived sound direction. The results show that this method yields a highly reproducible and accurate measure of the acoustic localization percept for stimuli presented within the oculomotor range (approximately 35 deg in all directions). Correlations for both the horizontal and vertical components of responses to control stimuli exceeded 0.9 in experienced as well as in naive subjects. The oculomotor response forms an important part of the natural repertoire of stimulus-evoked orienting (including head and body movements) and does not require any specific training of the subjects. Moreover, the response is fast (latencies remain well below 300 ms; see also Frens and Van Opstal, 1995).

In previous localization studies with human subjects, different response methods have also been used to quantify the localization percept: arm pointing (e.g., Oldfield and Parker, 1984a), the naming of learned coordinates (e.g., Wightman and Kistler, 1989), and head pointing (e.g., Makous and Middlebrooks, 1990). The first two methods are substantially slower than the eye-movement method and may thus be assumed to measure a later acoustic percept, possibly also incorporating cognitive aspects. The head-pointing method is potentially faster. Although latencies of head movements can be similar to eye movement latencies, head-movements dynamically alter the acoustic input for long-duration stimuli, which was deemed to be an undesirable factor for the purpose of this study.

2. Potential artefacts

We are confident that our main results cannot be ascribed to peculiar properties of the oculomotor system. First, the sound stimuli were presented well within the oculomotor range and stimulus types and stimulus positions were always presented in a randomized order. Therefore, the results cannot be due to a (conscious or subconscious) strategy adopted by the subjects.

A second argument for believing that the results reflect properties of the auditory system, rather than of the visuomotor system, is that quite different behaviors were obtained for the azimuth components of the responses, than for the elevation components. Such a behavior is never encountered in visually evoked eye movements toward stimuli at similar locations.

We also believe that the auditory stimuli were always presented well above the detection threshold, since similar temporal-dependent response behaviors were obtained for all three acoustic stimulus types. The finding that the azimuth components of the responses remained unaffected by the stimulus parameters, indicates that the stimuli were well-perceived by the auditory system, despite the fact that the overall acoustic energy of the stimuli varied greatly. This is also supported by the fact that the standard deviations of the latency distributions were hardly affected [see Fig. 9(b), and below].

Finally, also the result that response variability did not change systematically with the temporal stimulus parameters (Tables II, IV, and III) argues against the possibility that the auditory stimuli may have been close to the detection threshold for the shortest stimulus durations D or longest duty cycles ΔT.

3. Latency characteristics

There are some interesting aspects regarding the auditory-evoked saccadic latency distributions. First, the cumulative latency probabilities in Fig. 9(b) are plotted on a so-called probit scale, which is the inverse of the Error function. If the cumulative distribution on probit scale yields a straight line, the variable follows a Gaussian distribution. This linear feature of reciprocal latency distributions has been shown to be characteristic for visually evoked saccades (Carpenter, 1995). In the present study, straight lines were also obtained for auditory saccades, despite the different origin and encoding format of acoustic sensory information.

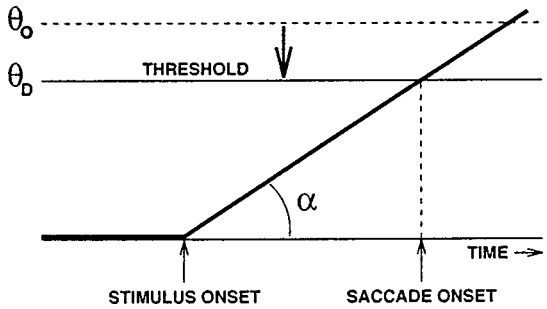

FIG. 10. A simple model that accounts for the observed characteristics of the latency distribution (see also Carpenter, 1995). At stimulus onset, a decision signal rises with a constant rate, α, and upon exceeding a fixed threshold, at θ_D, a saccade is initiated. In each trial, the increase rate, α, is drawn from a gaussian distribution. In order to incorporate the condition-dependent modulations on the latency in this study, it is proposed that the threshold is not constant, but instead decreases during stimulus presentation, from the initial value θ_0 to the final value θ_D.

Carpenter (1995) proposed a simple model for visual saccade initiation that accounts for this characteristic property (Fig. 10). The model assumes that a decision signal starts to increase after stimulus onset and, subsequently, initiates the saccade upon exceeding a fixed threshold (θ_0). Consequently, the time-to-threshold (i.e., the saccade latency) is inversely proportional to the increase rate of the decision signal. The increase rate (slope α) is assumed to vary from trial to trial, but remains constant during a trial. If this rate is distributed as a gaussian over trials, then the reciprocal latency is necessarily distributed likewise. Our data suggest that a similar mechanism could underlie the initiation of auditory-evoked saccades.

Yet, the question remains how stimulus-related shifts in the distributions could arise. For example, Fig. 9(b) shows that latencies become larger for shorter stimulus durations. The simple model described above has only a few parameters that can cause a shift in the distribution: the initial level of the decision signal, its mean increase rate, and the threshold level. The first two parameters have preset values and are assumed to depend on expectation about the stimulus. Since the different stimuli were presented in random order, expectation is not likely to play a role in the present situation. The shift in the median latencies is qualitatively explained, however, if it is assumed that the threshold decreases as more sensory information comes in (θ_D). For short burst durations, the threshold would then remain systematically higher (yielding longer latencies) than for longer stimulus durations (see also Fig. 10).

4. Secondary saccades

The typical response pattern of an auditory orienting trial usually consists of a large primary saccade followed by a smaller secondary (and occasionally a tertiary) saccade. A similar pattern is also typical for visually elicited responses. The primary saccade carries the eye over roughly 90% of the required trajectory (undershoot), whereas the secondary saccade corrects for the remaining retinal error. It has been proposed (Harris, 1995) that such a motor strategy minimizes

(in a statistical sense) the time needed to fixate the target, by taking into account both the effects of a longer duration needed to complete larger saccades, and the additional time needed to program a corrective saccade in the opposite direction (due to an overshoot). Here, the question was what the underlying strategy for secondary saccades could be in the case of a sound stimulus.

Analysis of the auditory responses indicated that the secondary saccades were not corrective. Note that, in contrast to visual stimulus conditions, the oculomotor system is not provided with any feedback concerning its performance after completing the primary auditory saccade. Notwithstanding, the occurrence of a small secondary saccade in roughly the same direction as the primary saccade was quite typical in all subjects tested. We propose that this phenomenon reflects a property of the programming mechanisms underlying the oculomotor response, rather than auditory processing. The data suggest that the secondary saccade is a pre-programmed movement that in this case, however, does not (cannot) correct for a residual error.

C. Temporal aspects of sound localization

1. Effect of stimulus duration

It should be realized, that the programming of an auditory-evoked saccade consists of at least two main stages: a target localization stage, in which the acoustic information is transformed into an estimate of target location, and a response initiation stage in which the estimated target coordinates are transformed into the appropriate motor commands. We believe that the data obtained from the noise-burst experiments may be interpreted in the light of these two, partly separate, stages.

First, it was found that the response elevation gain increases systematically as function of stimulus duration, and that it reaches a plateau for stimulus durations exceeding 80 ms (Fig. 8). From these data we infer that the auditory localization system needs roughly 80 ms of Gaussian broadband input to reach a stable estimate of target elevation.

Second, the latency data [Fig. 9(b) and Table II] may provide further insights into the processing time of the response initiation stage. Although it takes less time for the short-duration bursts to complete, the associated latencies were about 20 ms longer than for the longest stimuli (see Table II). This suggests that the movement initiation stage takes, at least, about 20 ms longer to initiate a saccade toward the shortest burst. We have no simple explanation for this apparent, yet consistent, discrepancy.

In contrast to sound elevation, sound azimuth can apparently be determined accurately on the basis of only a few milliseconds of sensory information. Even for the shortest stimuli ($D=3$ ms), the accuracy was about the same as for the control condition ($D=500$ ms). In this sense, azimuth localization can be considered as a much ''faster'' process than elevation localization. From our data it is not possible to conclude whether this property is due to mechanisms processing either the interaural phase or intensity differences, since the broadband bursts provided both cues simultaneously.

2. Effect of nonstationarity of the short-term spectrum

The results obtained from the FM sweeps provide some further interesting suggestions regarding the temporal processing of sensory information in sound localization. In contrast to the noise bursts, the duration of the sweeps was kept constant (at 500 ms), but the spectro-temporal behavior was varied by means of the cycle variable T. Response characteristics were similar as in the case of the noise bursts. Although elevation localization accuracy varied with T, responses remained consistent: high correlation coefficients of response versus stimulus position were obtained for all conditions. Because elevation localization performance was most accurate for $T<5$ ms and relatively inaccurate for $T>5$ ms, it is suggested that the auditory localization system discriminates spectro-temporal patterns at a temporal resolution down to about 5 ms.

A possible explanation for this ability is that the auditory system applies a so-called *multiple-look* strategy (e.g., Viemeister and Wakefield, 1991). In such a mechanism, the input spectrum is measured over consecutive short time windows, each lasting only a few milliseconds. Each short-term spectrum is processed into a position estimate, which, at a higher level, is integrated with the earlier estimates into a final estimate (see below, Fig. 12). This explanation is also in line with the interpretation of the burst-duration experiments (see above).

As an example, consider the computation of a short-term spectrum over a 5-ms time window somewhere in the cycle of an FM sweep with a long period, e.g., $T=20$ ms. The outcome will vary substantially for different window positions within the cycle. Elevation updates computed from consecutive (but very different) short-term spectra would then be inconsistent throughout the cycle and prevent the dynamic elevation estimate to stabilize at the actual target elevation. Note, that such a 5-ms time-averaged spectrum in a fast sweep ($T<5$ ms) would be broadband, whereas it would be narrow band for very slow sweeps ($T\gg 5$ ms).

3. Effect of silence gaps in the stimulus

The data obtained from the burst-train stimuli suggest further constraints on the dynamics of the spectral analysis in the localization process. For the shortest burst intervals, ΔT, near-optimal localization performance was obtained, although each burst was only 3 ms long.

Interestingly, the elevation gain was observed to decrease significantly when the individual bursts of the train were all chosen to be identical, rather than randomly drawn as in the experiments presented here, even for $\Delta T=3$ ms (data from subject PH only, not shown). Thus acoustic input presented at later times indeed contributes to the improvement of the final elevation estimate.

The low gains at long intervals are not explained by the fact that, within the latency period of roughly 180 ms, only two to four 3-ms bursts may have contributed to the programming of the first saccade, because the final estimate of the oculomotor responses (even after 500 ms) was not systematically better than the estimate recorded after the primary saccade. Rather, the decrease of elevation gain with

increasing ΔT indicates that the subsequent elevation estimates, based on each 3-ms sound burst, are not kept in acoustic memory forever. Possibly, the integrative mechanism that combines subsequent elevation estimates (see above) is leaky. If so, the data suggest a time constant for this integrator in the order of a few tens of ms.

D. Toward a spectro-temporal model of sound localization

In this section we study the properties of a monaural localization model that relies on spectral correlations in order to estimate sound elevation, ϵ_S (see also Middlebrooks, 1992). In addition, a biologically plausible mechanism for the dynamic implementation of these correlations, based on our data, will be briefly described.

1. Spectral correlations

In the proposed spectral correlation approach, it is assumed that the auditory system bases its comparison between the sensory signal and the HRTF on the log power of these spectral functions. In addition, a logarithmic scaling of the frequency domain, reminiscent to the tonotopic neural representation of sound frequency throughout the auditory system, is applied. By applying the logarithm to the power of the sensory signal measured at the eardrum, Eq. (1) can be rewritten as:

$$Y(\omega;\epsilon_S)=H(\omega;\epsilon_S)+X(\omega), \tag{3}$$

where ω is in octaves, and ϵ_S is the source elevation. The capitals indicate the logarithmic power spectra [e.g., $X(\omega) \equiv \log|x(\omega)|$]. Note, that elevation, ϵ_S, rather than position, $\mathbf{r}_S$, is used in Eq. (3), as it is assumed that azimuth, α_S, is already extracted from binaural cues.

A quantitative scalar measure of similarity for two spectral functions can be given by the spectral correlation coefficient. Therefore, the mean, $\overline{F}$, and variance, σ_F^2, for an arbitrary spectral function, $F(\omega)$, are first introduced:

$$\begin{aligned}\overline{F}&=\langle F(\omega)\rangle\equiv\int_0^\infty d\omega\ p(\omega)F(\omega),\\ \sigma_F^2&=\langle(F(\omega)-\overline{F})^2\rangle,\end{aligned} \tag{4}$$

with $p(\omega)$ a normalized weighting function that is nonzero in the (broad) frequency band of interest. Here $p(\omega)$ was chosen to be uniform in the [2, 16] kHz range, and zero elsewhere. The spectral correlation, $\mathsf{C}(F(\omega),G(\omega))$, between two functions, $F(\omega)$, and, $G(\omega)$, is then defined as

$$\mathsf{C}(F(\omega),G(\omega))\equiv\left\langle\left(\frac{F(\omega)-\overline{F}}{\sigma_F}\right)\left(\frac{G(\omega)-\overline{G}}{\sigma_G}\right)\right\rangle. \tag{5}$$

The outcome varies in $[-1,1]$, where $\mathsf{C}(\cdot)=1$ corresponds to ''maximal similarity,'' and $\mathsf{C}(\cdot)\leqslant 0$ means ''no similarity.''

In the present discussion, the spectral correlation compares the sensory signal, $Y(\omega;\epsilon_S)$, with (neurally stored) HRTFs, $H(\omega;\epsilon)$, for all sound elevations ϵ. By using Eq. (3) and by taking the mean (where $\overline{Y}=\overline{H_S}+\overline{X}$), this comparison reads:

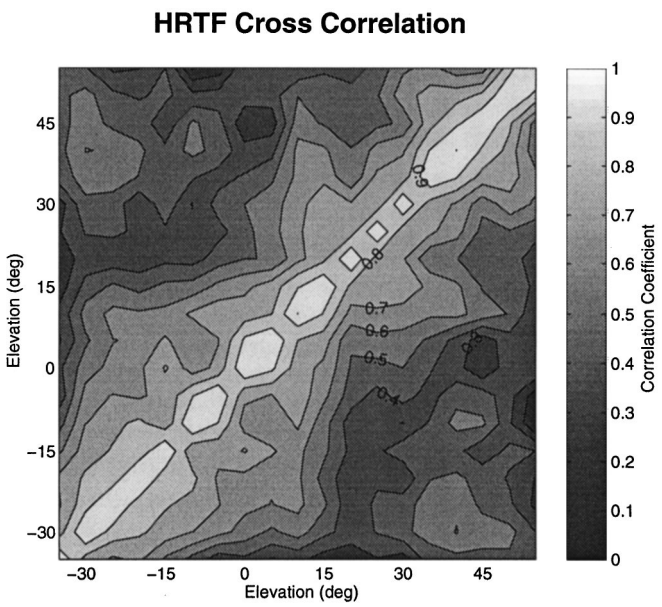

FIG. 11. Spectral correlations of human HRTFs, measured in the midsaggital plane at elevations $\{-35,-30,...,+50,+55\}$ deg. Subject PH. The strength of the correlation is coded by brightness. Markers label the contours at different values. Note that HRTFs correlate strongest with themselves and with HRTFs at neighboring elevations. Correlations were computed in the frequency range [2,16] kHz and HRTFs were sampled at 1/20 octave intervals.

$$\begin{aligned}C_Y(\epsilon;\epsilon_S)&\equiv\mathsf{C}(Y(\omega;\epsilon_S),H(\omega;\epsilon))\\ &=\left(\frac{\sigma_{H_S}}{\sigma_Y}\right)\mathsf{C}(H(\omega;\epsilon_S),H(\omega;\epsilon))\\ &\quad+\left(\frac{\sigma_X}{\sigma_Y}\right)\mathsf{C}(X(\omega),H(\omega;\epsilon))\\ &\equiv\left(\frac{\sigma_{H_S}}{\sigma_Y}\right)C_H(\epsilon;\epsilon_S)+\left(\frac{\sigma_X}{\sigma_Y}\right)C_X(\epsilon),\end{aligned} \tag{6}$$

with σ_Y, σ_{H_S}, σ_X the standard deviations of the respective quantities in Eq. (3). Thus Eq. (6) quantifies the spectral correlation of each stored HRTF (associated with sound elevation, ϵ) with the measured spectrum at the eardrum (resulting from a sound source at elevation, ϵ_S). It proposed that the perception of sound elevation essentially consists of selecting the unique position $\epsilon=\epsilon_P$ at which the global maximum, $C_Y(\epsilon_P;\epsilon_S)$, is attained.

In its decomposed form, Eq. (6) can be readily interpreted. The first term on the right-hand side, $C_H(\epsilon;\epsilon_S)$, is the spectral correlation between the HRTFs $H(\omega;\epsilon)$ and $H(\omega;\epsilon_S)$. Under the assumption that HRTFs are unique with respect to position, $C_H(\epsilon;\epsilon_S)$ reaches a maximum in the neighborhood of ϵ_S, i.e., at $\epsilon\approx\epsilon_S$.

Figure 11 shows that this is indeed a reasonable assumption. In this figure, measured HRTF functions in the median plane of one of our subjects have been correlated with each other. Note that only along the principal diagonal a high correlation is obtained, indicating that, indeed, HRTFs only resemble themselves. Therefore, these functions contain unique information about target elevation, when all frequency bands are allowed to contribute to the computational

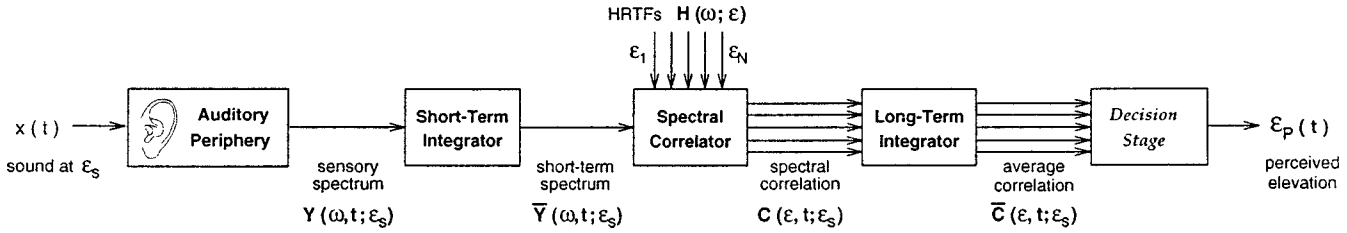

FIG. 12. Conceptual spectro-temporal correlation model of human sound localization. The first stage represents the auditory periphery, where a sound, $x(t)$, at elevation, ϵ_S, is processed from the external ear up to the cochlea. In the second stage, a short-term (order a few ms) averaged spectrum, $\bar{Y}(\omega,t;\epsilon_S)$, is determined around the current time, t, from the ongoing sensory spectrum, $Y(\omega,t;\epsilon_S)$. Then, $\bar{Y}(\omega,t;\epsilon_S)$ is correlated with the neurally stored HRTF representations, $H(\omega;\epsilon)$, for all possible elevations ϵ, i.e., $\epsilon_1,\ldots,\epsilon_N$, yielding short-term correlations $C(\epsilon,t;\epsilon_S)$. Next, integration is done over a longer time span (order several tens of ms), resulting in a long-term correlation up to current time t, $\bar{C}(\epsilon,t;\epsilon_S)$. This integration process may be leaky. In the decision stage, the average correlation at time t yields a perceived elevation, $\epsilon_P(t)$. It is proposed that this decision depends on the maximum of the average correlation, but also on the consistency of that average and on the initial elevation percept.

process. We have verified that this conclusion also holds for the entire elevation domain ($\epsilon \in [-60,90]$ deg, front and back), although the region of maximum correlation broadens appreciably close to zenith positions ($\epsilon \approx +90$ deg; data not shown). It is therefore expected, that stimuli presented within that range are not well localized and discriminated.

The second term in Eq. (6), $C_X(\epsilon)$, expresses the resemblance of the source spectrum $X(\omega)$ with each particular HRTF $H(\omega;\epsilon)$. If the source spectrum does not correlate significantly with any of the stored HRTFs, then $C_X(\epsilon) \approx 0$ for all ϵ [for example, this would be true if $X(\omega)$ is flat and broadband]. In that case, the maximum correlation $C_Y(\epsilon_P;\epsilon_S)$ will be reached at $\epsilon_P = \epsilon_S$, thus the sound would be accurately localized. It is expected that there will be a broad range of naturally occurring acoustic stimuli for which there is little or no resemblance with the stored HRTFs, i.e., $C_X(\epsilon) \approx 0$.

However, the second term does come into play, if the source spectrum $X(\omega)$ does correlate well with a given HRTF, say $H(\omega;\epsilon_*)$. This occurs if $X(\omega)$ contains prominent features that are characteristic for $H(\omega;\epsilon_*)$. Then, $C_X(\epsilon_*) \gg 0$, meaning that $C_Y(\epsilon;\epsilon_S)$ contains a second local maximum at $\epsilon = \epsilon_*$. This could even be the global maximum in which case the perceived position, ϵ_P, would be, ϵ_*, rather than the actual position, ϵ_S, thus $\epsilon_P = \epsilon_* \neq \epsilon_S$. This might occur, for example, when $X(\omega)$ is narrow band (i.e., peaked) with its center frequency at a characteristic peak of $H(\omega;\epsilon_*)$. In the study of Middlebrooks (1992), consistent mislocalizations were attributed to such spurious correlations between the narrow-band source spectrum and one of the HRTFs.

2. Conceptual spectro-temporal model

The data presented in this study clearly indicate that the spectral estimation performed by the auditory localization system is also a temporal process. Therefore, if the auditory system bases its spatial estimation on spectral correlations, the underlying computational mechanisms have to be incorporated within a temporal scheme. Figure 12 provides a conceptual model of the successive stages in this spectral-temporal process (see legend for specific details). The model accounts for the different aspects that have been derived from the data for each of the spectro-temporal stimulus patterns (see above, Sec. III C).

(1) The sweep data suggest that the elevation localization system first measures spectra on a short time scale of about 5 ms (''multiple looks''). Accurate localization requires a broadband, short-term spectrum (fast sweeps, noise bursts). Inaccurate localization results if the short-term spectrum is narrow band (slow sweep).
(2) The noise-burst data suggest that acoustic information needs to be delivered over a (longer) time scale of roughly 80 ms. If short-term estimates are consistent over time (fast sweeps, noise) averaging will enhance the final estimate. Yet, if the estimates vary strongly (e.g., for slow sweeps) the estimates may cancel each other on average.
(3) The burst-train data suggest a power-dependent gating mechanism and leakiness of the long-term integration process. Estimates at different time windows (e.g., in burst trains with a long duty cycle) are not heavily suppressed by the intervening silence periods, thus gating is plausible. Leakiness is inferred from the finding that the elevation gain drops as the silence period increases.

In this scheme, the current estimate, $\epsilon_P(t)$, of target position, ϵ_S, smoothly develops over time as more acoustic information enters the system.

The experimental data further suggest that, in the absence of sufficient spectral processing the auditory localization system stays close to its default initial estimate of elevation, typically near the horizontal plane. An interesting question that remains to be investigated, is whether the system's default elevation estimate depends on the initial gaze direction, or on head orientation. In the present study, initial eye position, the horizontal plane of the head, and the earth-fixed horizon always coincided. This problem could be studied by systematically changing these different frames of reference with respect to each other under similar stimulus conditions as applied in this study.

ACKNOWLEDGMENTS

The authors are indebted to Jeroen Goossens and Jos Van Riswick for their participation in the experiments. We thank Hans Kleijnen and Ton Van Dreumel for valuable technical assistance. BB, KH, and VC are acknowledged for volunteering as experimental subjects. The authors also thank both referees for their useful and constructive com-

ments. This research was supported by the Netherlands Foundation of the Life Sciences (SLW; PH) and the University of Nijmegen (AJVO).

Batteau, D. W. (**1967**). ''The role of pinna in human localization,'' Proc. R. Soc. London, Ser. B **168**, 158–180.

Blauert, J. (**1996**). *Spatial Hearing: The Psychophysics of Human Sound Localization* (MIT, Cambridge, MA).

Butler, R. A., and Musicant, A. D. (**1993**). ''Binaural localization: Influence of stimulus frequency and the linkage to covert peak areas,'' Hearing Res. **67**, 220–229.

Carpenter, R. H. S., and Williams, M. L. L. (**1995**). ''Neural computation of log likelihood in control of saccadic eye movements,'' Nature (London) **377**, 59–62.

Collewijn, H., Van der Mark, F., and Jansen, T. C. (**1975**). ''Precise recording of human eye movements,'' Vision Res. **15**, 447–450.

Frens, M. A., and Van Opstal, A. J. (**1995**). ''A quantitative study of auditory-evoked saccadic eye movements in two dimensions,'' Exp. Brain Res. **107**, 103–117.

Gardner, M. B., and Gardner, R. S. (**1973**). ''Problem of localization in the median plane: Effect of pinnae occlusion,'' J. Acoust. Soc. Am. **53**, 400–408.

Han, H. L. (**1994**). ''Measuring a dummy head in search of pinna cues,'' J. Audio Eng. Soc. **42**, 15–37.

Harris, C. M. (**1995**). ''Does saccadic undershoot minimize saccadic flight-time? A Monte-Carlo study,'' Vision Res. **35**, 691–701.

Hebrank, J., and Wright, D. (**1974**). ''Spectral cues used in the localization of sound sources on the median plane,'' J. Acoust. Soc. Am. **56**, 1829–1234.

Hofman, P. M., and Van Opstal, A. J. (**1997**). ''Identification of spectral features as sound localization cues in the external ear acoustics,'' in *Biological and Artificial Computation: From Neuroscience to Technology*, edited by J. Mira, R. Moreno-Díaz, and J. Cabestany (Springer-Verlag, Berlin).

Kistler, D. J., and Wightman, F. L. (**1992**). ''A model of head-related transfer functions based on principal component analysis and minimum-phase reconstruction,'' J. Acoust. Soc. Am. **91**, 1637–1647.

Lopez-Poveda, E. A., and Meddis, R. (**1996**). ''A physical model of sound diffraction and reflections in the human concha,'' J. Acoust. Soc. Am. **100**, 3248–3259.

Makous, J. C., and Middlebrooks, J. C. (**1990**). ''Two-dimensional sound localization by human listeners,'' J. Acoust. Soc. Am. **87**, 2188–2200.

Mehrgardt, S., and Mellert, V. (**1977**). ''Transformation characteristics of the external ear,'' J. Acoust. Soc. Am. **61**, 1567–1576.

Middlebrooks, J. C. (**1992**). ''Narrow-band sound localization related to external ear acoustics,'' J. Acoust. Soc. Am. **92**, 2607–2624.

Middlebrooks, J. C., Makous, J. C., and Green, D. M. (**1989**). ''Directional sensitivity of sound-pressure levels in the human ear canal,'' J. Acoust. Soc. Am. **86**, 89–108.

Musicant, A. D., and Butler, R. A. (**1984**). ''The influence of pinnae-based spectral cues on sound localization,'' J. Acoust. Soc. Am. **75**, 1195–1200.

Neti, C., Young, E. D., and Schneider, M. H. (**1992**). ''Neural network models of sound localization based on directional filtering by the pinnae,'' J. Acoust. Soc. Am. **92**, 3141–3155.

Oldfield, S. R., and Parker, S. P. (**1984a**). ''Acuity of sound localisation: a topography of auditory space. I. Normal hearing conditions,'' Perception **13**, 581–600.

Oldfield, S. R., and Parker, S. P. (**1984b**). ''Acuity of sound localisation: a topography of auditory space. II. Pinna cues absent,'' Perception **13**, 601–617.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (**1992**). *Numerical Recipes in C* (Cambridge U.P., Cambridge, MA), 2nd ed.

Schröder, M. R. (**1970**). ''Synthesis of low-peak-factor signals and binary sequences with low autocorrelation,'' IEEE Trans. Inf. Theory **16**, 85–89.

Shaw, E. A. G. (**1974**). ''Transformation of sound pressure level from free field to eardrum in the horizontal plane,'' J. Acoust. Soc. Am. **56**, 1848–1861.

Teranishi, R., and Shaw, E. A. G. (**1968**). ''External-ear acoustic models with simple geometry,'' J. Acoust. Soc. Am. **44**, 257–263.

Van Opstal, A. J., and Van Gisbergen, J. A. M. (**1989**). ''Scatter in the metrics of saccades and properties of the collicular motor map,'' Vision Res. **29**, 1183–1196.

Viemeister, N. F., and Wakefield, G. H. (**1991**). ''Temporal integration and multiple looks,'' J. Acoust. Soc. Am. **90**, 858–865.

Wightman, F. L., and Kistler, D. J. (**1989**). ''Headphone simulation of free field listening I: stimulus synthesis,'' J. Acoust. Soc. Am. **85**, 858–867.

Wightman, F. L., and Kistler, D. J. (**1992**). ''The dominant role of low-frequency interaural time differences in sound localization,'' J. Acoust. Soc. Am. **91**, 1648–1661.

Zakarouskas, P., and Cynader, M. S. (**1993**). ''A computational theory of spectral cue localization,'' J. Acoust. Soc. Am. **94**, 1323–1331.