Modeling of speech localization in a multi-talker mixture using periodicity and energy-based auditory features

Angela Josupeit[a)]
Medizinische Physik, Cluster of Excellence Hearing4all, Universität Oldenburg, 26111 Oldenburg, Germany

Norbert Kopčo
Institute of Computer Science, Faculty of Science, P. J. Šafárik University, Jesenná 5, 04001 Košice, Slovakia

Volker Hohmann
Medizinische Physik, Cluster of Excellence Hearing4all, Universität Oldenburg, 26111 Oldenburg, Germany

(Received 20 October 2015; accepted 3 May 2016; published online 20 May 2016)

A recent study showed that human listeners are able to localize a short speech target simultaneously masked by four speech tokens in reverberation [Kopčo, Best, and Carlile (2010). J. Acoust. Soc. Am. **127**, 1450–1457]. Here, an auditory model for solving this task is introduced. The model has three processing stages: (1) extraction of the instantaneous interaural time difference (ITD) information, (2) selection of target-related ITD information ("glimpses") using a template-matching procedure based on periodicity, spectral energy, or both, and (3) target location estimation. The model performance was compared to the human data, and to the performance of a modified model using an ideal binary mask (IBM) at stage (2). The IBM-based model performed similarly to the subjects, indicating that the binaural model is able to accurately estimate source locations. Template matching using spectral energy and using a combination of spectral energy and periodicity achieved good results, while using periodicity alone led to poor results. Particularly, the glimpses extracted from the initial portion of the signal were critical for good performance. Simulation data show that the auditory features investigated here are sufficient to explain human performance in this challenging listening condition and thus may be used in models of auditory scene analysis.

© *2016 Acoustical Society of America*. [http://dx.doi.org/10.1121/1.4950699]

[JFL] Pages: 2911–2923

I. INTRODUCTION

Human listeners are able to attend to and understand one specific talker in complex acoustic settings, such as reverberant rooms in which multiple talkers speak at the same time (e.g., Bronkhorst, 2000). One aspect of this ability is the localization of the attended target talker in a multitalker environment. How the monaural signal-related auditory features, used for the discrimination of the target against the maskers, are combined with binaural features to identify the location of the target is still largely unknown (e.g., see Shamma and Fritz, 2014). This study examines this question by simulating a localization task in a multi-talker setting using an auditory model, and comparing it to human data. Periodicity and spectral energy were investigated as monaural features; interaural time differences were used as binaural features. One important characteristic of the proposed model is that it uses *a priori* information about the target speech token, similar to the optimal detector approaches established in psychoacoustic detection models (e.g., Dau *et al.*, 1996). This way, information about the target is used optimally, making it possible to assess the relative salience of the features and their interaction in solving the task, which is the main purpose of this study.

The ability of human listeners to localize speech in complex listening scenarios depends on a number of factors. It has been shown that frontal azimuth localization performance degrades with decreasing SNR (Kopčo *et al.*, 2010), increasing number of maskers (Langendijk *et al.*, 2001), masker uncertainty (Kopčo *et al.*, 2010) and reverberation (Giguère and Abel, 1993). Experimental results suggest that, primarily, the binaural features at the onset of a sound (Houtgast and Aoki, 1994; Freyman *et al.*, 1997), or at rising segments of the signal envelope (Dietz *et al.*, 2013) are used for localization.

Auditory modeling of frontal azimuthal localization has been done using physiologically inspired models based on normalized cross-correlation (Faller and Merimaa, 2004; Roman *et al.*, 2003) or on the extraction of instantaneous interaural phase differences (IPDs; Dietz *et al.*, 2011). In some cases, these models contain a measure for identifying robust binaural information: Only binaural information with a high interaural correlation (Faller and Merimaa, 2004), or a high interaural vector strength (IVS; Dietz *et al.*, 2011) is taken into account to estimate locations. Using these measures is especially important for scenarios that include reverberation and multiple sound sources. It has been shown that these models can accurately estimate the locations of multiple talkers. However, the models alone are not able to determine which of the segregated sources is the target and which are the maskers. Furthermore, the models' performance was not previously compared to human data.

To identify a speech target in a multi-talker mixture, further features are needed. It has been shown that periodicity

[a)]Electronic mail: angela.josupeit@uni-oldenburg.de is an important cue for distinguishing between different talkers (e.g., Darwin, 1981; Alain *et al.*, 2005). Another important cue is the spectral profile (Gockel and Colonius, 1997; Gockel, 1998).

In a typical auditory scene analysis task, several features need to be integrated to distinct auditory objects. One principle that guides this integration is that temporally and spectrally coherent features are bound to the same object (Elhilali *et al.*, 2009; Shamma *et al.*, 2011; Teki *et al.*, 2013).

The present study introduces an auditory model that simulates the task of localizing a female speech token presented simultaneously with four male speech tokens arranged in different spatial configurations (Kopčo *et al.*, 2010). The scene is complex in the sense that it has a high number of maskers, a relatively low SNR of −6 dB, a simultaneous onset of target and maskers, a complete temporal overlap of the target word by the masker words, short utterances (mostly < 300 ms), a slightly reverberant environment, unknown masker words, and an unknown spatial masker configuration.

This study investigated three different aspects of the simulated localization task. First, it examined whether an auditory binaural model (Dietz *et al.*, 2011) is suitable for modeling human localization performance in this challenging condition when optimal selection of target-related binaural information is assumed. For the optimal selection of target-related features, an ideal binary mask (IBM) was used (Wang, 2005; Barker and Cooke, 2007). Second, it investigated how target-related features can be selected using *a priori* knowledge about the unmasked target utterance. For this, we adopted the optimal detector method developed to predict human detection performance (e.g., Dau *et al.*, 1996). In particular, a template was generated that consisted of the extracted monaural features of the unmasked target. Then, a template-matching procedure compared the template with the respective features from the multi-talker input signal and selected the matching time-frequency bins. Under the assumption that auditory features occurring in the same time-frequency bin belong to the same source (Shamma *et al.*, 2011), the target-related binaural information was read out from the selected bins, whereas the binaural information from the remaining bins was assigned to the maskers. As monaural features, periodicity, spectral energy, and a combination of both features were compared to investigate the relative salience of these features. Finally, the present study investigated the importance of early vs late portions of the signal in the localization task. This was done by analyzing the localization accuracy of the model for the early and late signal portions, and by manipulating the selection of target-related information using mixtures of template-matching procedures and optimal IBM-based selection.

II. MODEL DESCRIPTION

Figure 1 shows the outline of the model. The input signal was a multi-talker signal, as used by Kopčo *et al.* (2010). The task was to estimate the location of the target—the word "two" uttered by a female talker—presented simultaneously with four different male masker speech tokens originating from four different locations. A detailed description of the stimuli is given in Sec. III A. First, the left and right channels of the input signal were preprocessed by a model of auditory periphery, and auditory features were extracted from the preprocessed signals (Fig. 1, model part A). Binaural features were calculated using a slightly modified version of the binaural model of Dietz *et al.* (2011); monaural features were periodicity (Chen and Hohmann, 2015) and spectral energy. Second, target-related binaural features were selected using a binary mask (BM) (Fig. 1, model part B). This BM could either be an IBM (which replaces the stages in the dashed-dotted box), derived by analyzing the target and masker signals separately, or BMs based on a template-matching procedure that compared the monaural features derived from the target alone with those derived from the target and masker mixture signal. Third, the final target location was estimated based on the distributions of selected and not-selected binaural features across the whole utterance (Fig. 1, model part C). It is important to note here that the binaural and periodicity features were pre-selected according to a robustness measure. It is thus assumed that each pre-selected feature value mainly represents a single sound source and that a binary decision as implemented here is sufficient to separate target- and background-related feature values. A detailed description of the three model parts is given in the following.

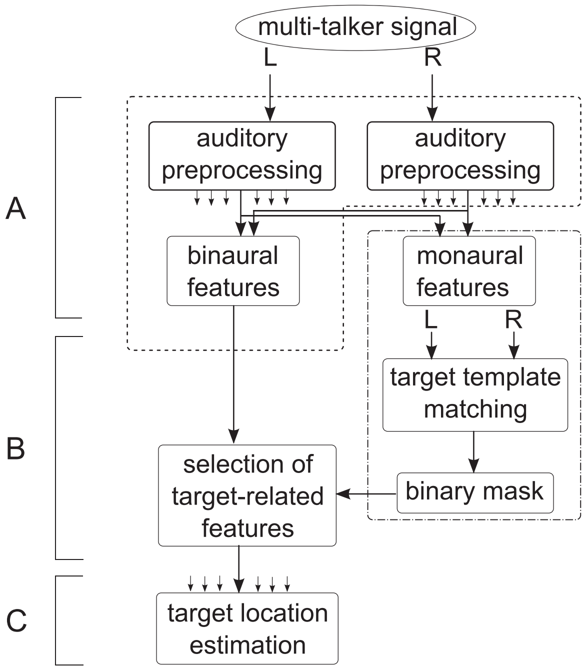

FIG. 1. Model outline. Part A: The left and right ear signals are first preprocessed by a peripheral model. After that, binaural and monaural features are extracted; the dashed box identifies the processing steps used in the binaural model adapted from Dietz *et al.* (2011); the extracted monaural features are periodicity and spectral energy, derived from the left and right channel signals individually. Part B: Based on the monaural features, a template-matching procedure is applied from which BMs are estimated. The selection of target-related binaural features is based on these BMs, or on the IBM (in which case the stages enclosed in the dash-dotted box are replaced by IBM extraction). Part C: The target location is estimated based on the binaural information selected as belonging to the target as well as on the maskers information in all frequency channels.

A. Feature extraction

1. Auditory preprocessing

The left and right multi-talker input signals passed a preprocessing stage based on a model of the auditory periphery as used by Dietz *et al.* (2011). In brief, this model includes a middle ear band-pass filter, a gammatone filter bank with 23 filters ranging from approximately $f_c = 200$ Hz to $f_c = 5$ kHz, followed by an instantaneous compression, half wave rectification and a low-pass filter. In addition to the original model by Dietz *et al.* (2011), a differentiator was implemented after the low-pass filter to remove the DC component before extracting periodicity features as described in Sec. II A 3. The signals then passed a fine structure filter (for $f_c < 1400$ Hz) or envelope filter (for $f_c > 1400$ Hz). In line with the original model, the fine structure filters were set to the respective center frequency f_c of each filter and a bandwidth of $f_c/3$. The envelope filter was set to a center frequency of $f_m = 250$ Hz and a bandwidth of 250 Hz to cover the full range of the target talker fundamental frequency (ca. 170–270 Hz).

2. Binaural features

Binaural features were extracted as described by Dietz *et al.* (2011). This model computes the IPDs as a function of time t in each frequency band f_c. ITDs are calculated from IPDs and the sub-band instantaneous frequency. Interaural level differences (ILDs) are extracted from the preprocessed signals before the differentiation stage; the sign of the ILD is used to resolve IPD ambiguities in the fine structure filters. ITDs are low-pass filtered using a time constant τ, which defines the binaural temporal resolution. As a measure for the robustness of the binaural features, the IVS was calculated. Only those binaural feature values are further processed whose corresponding IVS values exceed a threshold IVS_0. A second measure for robustness is the "rising flanks" criterion. That is, only those features are further processed where the derivative of the IVS time signal is positive. As a binaural time constant, we chose $\tau = 1/f_c$ for the fine structure channels and $\tau = 1/f_m$ for the envelope channels; as a threshold for robust information, we chose $\mathrm{IVS}_0 = 0.9$. In the original Dietz *et al.* (2011) study, these parameters were set to $\tau = 5/f_c$ resp. $\tau = 5/f_m$ and $\mathrm{IVS}_0 = 0.98$. The parameters were changed in this study to achieve a sufficiently high number of robust features during the short duration of the target utterance.

ITDs were mapped to azimuth angles $\alpha_1(t,f_c)$ using a fitting function that was calculated similarly to the one described by Dietz *et al.* (2011): First, we generated speech signals based on the speech corpus employed (Kidd *et al.*, 2008); each signal consisted of one random word uttered by one random talker of the experiment. This utterance was convolved with the binaural room impulse response (BRIR) for a specific direction ranging from −60° to 60° in 10° steps; the same BRIRs were used to generate the input signals in the simulations (see Sec. III A). The final 0.8 s of each signal were discarded because they tended to be dominated by reverberant energy. Second, we extracted ITDs and ILDs of these signals, as described earlier; as parameters, $\tau = 2.5/f_c$ or $\tau = 2.5/f_m$ and $\mathrm{IVS}_0 = 0.98$ were chosen. Third, we calculated one ITD for each azimuth direction α as the median of the ITDs across time. For each azimuth, 25 iterations with random words and random talkers were done and the ITD was found as the median across these iterations, resulting in values $\mathrm{ITD}(\alpha)$ for each f_c. Fourth, a linear fitting function was applied to the inverse values $\alpha(\mathrm{ITD})$ for each f_c. The parameters τ and IVS_0 for the calculation of the lookup table were chosen to select robust binaural information for the single-source reference signal and differed from the parameters used for the extraction of binaural features in the simulations. The target-localization in quiet (cf. Fig. 2), however, was not influenced by the change in the parameter values.

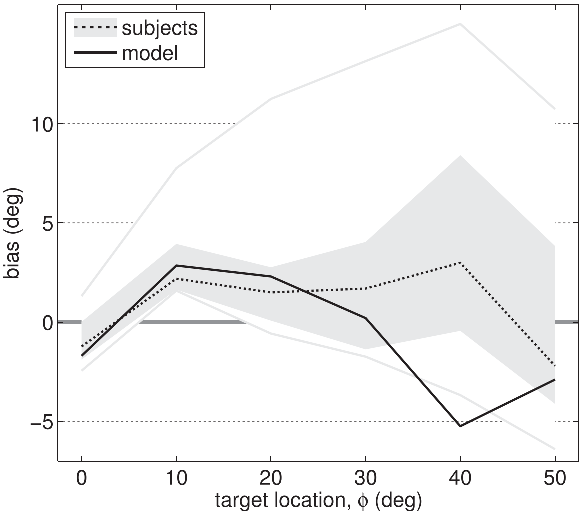

FIG. 2. Median target localization bias from actual target location as a function of target location without maskers (simulation A, control condition). The dashed line, the gray filled area and the thin gray lines indicate the median, the upper and lower quartiles, and the minimum and maximum of the subjects' individual median biases, respectively (data from Kopčo *et al.*, 2010). The solid line represents the model results.

Azimuth signals $\alpha_1(t,f_c)$ were then downsampled from $\mathrm{fs}_1 = 44.1$ kHz to $\mathrm{fs} = 1$ kHz in order to reduce storage usage and to provide better temporal alignment with the periodicity and spectral energy features, both of which were extracted with a sampling frequency of 1 kHz. The downsampling algorithm calculated the mean value of IVS-selected binaural information every 1 ms, resulting in a sampling frequency of $\mathrm{fs} = 1$ kHz. The resulting signal is referred to as $\alpha(t,f_c)$.

3. Periodicity features

Periodicity features were extracted from the preprocessed signals. They were based on the extraction of the normalized "synchrogram" $S(t,f_c,P)$ (Chen and Hohmann, 2015). The normalized synchrogram $S(t,f_c,P)$ is the ratio of the harmonic signal energy for the period P and the total signal energy in the same time window for a $[t, f_c]$ bin, computed for a number of tested candidate periods P'. If $S(t,f_c,P) = 1$, the signal is fully harmonic with a period P; if $S(t,f_c,P) = 0$, there is no harmonic energy at the period P.

It was therefore assumed that the locations of local maxima with high peak values of the synchrogram function across candidate periods P' correspond to the dominating fundamental period $P0$ and its multiples. The extraction of periodicity features is explained in detail in the following.

The set of all tested periods was given by

$$P' = \{P'_{\min}, P'_{\min} + \Delta P, \ldots, P'_{\max} - \Delta P, P'_{\max}\}, \tag{1}$$

with $P'_{\min} = 1/(1400\,\mathrm{Hz}), P'_{\max} = 1/(80\,\mathrm{Hz})$ and $\Delta P = 1/\mathrm{fs}_1$, with $\mathrm{fs}_1 = 44.1\,\mathrm{kHz}$.

The set of all local maxima of the synchrogram for one given $[t, f_c]$ bin was given by

$$P_{\max}(t,f_c) = \{P \in P' \,|\, S(t,f_c,P)\ \text{is a local maximum}\}. \tag{2}$$

A local maximum was defined as a value which is larger than its two neighboring values. The periodicity features were chosen from the set of local maxima $P_{\max}(t,f_c)$ if they fulfill certain energy requirements: The largest local maximum had to exceed a value of P_1, making sure that there is enough harmonic energy in the signal. If this requirement was fulfilled, all local maxima exceeding a certain threshold P_2 were chosen as periodicity features,

$$P0(t,f_c) = \{P \in P_{\max}(t,f_c) \,|\, \max_{P^* \in P_{\max}(t,f_c)} (S(t,f_c,P^*)) > P_1 \wedge S(t,f_c,P) > P_2\}. \tag{3}$$

P_1 and P_2 were set to 0.9 and 0.8, respectively, in fine structure bands ($f_c < 1400$ Hz) or 0.5 and 0.4 in envelope bands ($f_c > 1400\,\mathrm{Hz}$), respectively. These values were chosen to make sure that periodicity features represent sub-band signal sections with a salient predominant periodicity, similar to the coherence-based selection of binaural features. The periodicity features were determined separately for the left and right channels and in the following, they are referred to as $P0_l(t,f_c)$ and $P0_r(t,f_c)$, respectively.

4. Spectral energy features

Spectral energy features $E_l(t,f_c)$ and $E_r(t,f_c)$ were calculated from the preprocessed signals. They were calculated every 1 ms as the mean signal power in a 10-ms rectangular moving window.

B. Selection of target-related binaural information

The binaural signal $\alpha(t,f_c)$ contains azimuth information of target and masker stimuli. The selection of target-related information was based on BMs that label the target-dominant $[t, f_c]$ bins, or "glimpses," with a value of 1 and all other bins with 0,

$$\alpha_{\mathrm{sel}}(t,f_c) = \{\alpha(t,f_c) | \mathrm{BM}(t,f_c) = 1\}. \tag{4}$$

The selection mechanism was restricted to the initial portion of the stimulus ($t \in [0, 300]$ ms), which contained direct-sound energy of the target. In this study, BMs were estimated in several different ways: First, as an IBM; second, via a template-matching procedure of periodicity features ($\mathrm{BM}_{\mathrm{P0}}$), spectral energy features (BM_{E}), or a combination of both ($\mathrm{BM}_{\mathrm{E,P0}}$); third, using a combination of IBM for the early signal portions and BM for the late signal portions, or vice versa. Each of these procedures is explained in the following.

1. IBM

The IBM is defined as

$$\mathrm{IBM}(t,f_c) = \begin{cases} 1 & \text{if } \mathrm{SNR}(t,f_c) > 0\,\mathrm{dB} \\ 0 & \text{if } \mathrm{SNR}(t,f_c) < 0\,\mathrm{dB}, \end{cases} \tag{5}$$

where $\mathrm{SNR}(t,f_c)$ is the ratio of target signal energy and masker signal energy; these energies were calculated analogously to the multi-talker energy described in Sec. II A 4. For this approach, full *a priori* knowledge about the separated target and masker signals is needed.

2. Template matching

The selection of target-related binaural information was based on a template-matching procedure that used the monaural features of the target alone as a template. This is in line with the Kopčo *et al.* (2010) experiment, in which the subjects had the opportunity to create a template, as the experiment included a target-alone control condition prior to the main experiments (see also simulation A in Secs. III and IV). To create the BMs $\mathrm{BM}_{\mathrm{P0}}$, BM_{E}, and $\mathrm{BM}_{\mathrm{E,P0}}$ in each simulated experimental trial, the template of the target's periodicity and/or spectral energy was matched with the corresponding features extracted from the multi-talker mixture. The derivation of the templates and the computation of the BMs is described in detail in the following.

a. Periodicity template matching. To calculate the periodicity template $P0_{\mathrm{tar}}(t,f_c)$, periodicity features were extracted as described in Sec. II A 3. for all possible unmasked target utterances (11 locations with 2 channels each), referred to as the sets $P0_{\mathrm{tar},i}(t,f_c)$, $i = 1, \ldots, 22$. Second, a probability density function (PDF) $\mathrm{PDF}(t,f_c,P)$ across all sets $P0_{\mathrm{tar},i}(t,f_c)$ was calculated as follows:

$$\mathrm{PDF}(t,f_c,P) = C \sum_{i=1}^{22} \left(\sum_{P0' \in P0_{\mathrm{tar},i}(t,f_c)} \mathcal{N}(P0', 10^{-4}\ \mathrm{s}) \right), \tag{6}$$

where $\mathcal{N}(\mu,\sigma)$ denotes a Gaussian function with an expected value μ and standard deviation σ. The factor C was chosen so that the integral of the PDF was one. The resulting PDF was usually a multi-peak function with peaks at multiples of the fundamental period. Third, the peak positions of the PDF were chosen as the possible candidates for the template. A candidate contributed to the template $P0_{\mathrm{tar}}(t,f_c)$ if a minimum number of period values from the original sets $P0_{\mathrm{tar},i}(t,f_c)$ lay $\pm 10^{-4}$ s from a candidate. These minimum numbers were set to 12 for the fine structure filters and 6 for the modulation filters.

In the template-matching procedure, a given multi-talker input's periodicity, $P0(t,f_c)$, was evaluated against the periodicity template, $P0_{\mathrm{tar}}(t,f_c)$, separately for each $[t,\ f_c]$ bin. Two criteria had to be fulfilled for the input at each ear to consider it a match to the template: (1) the number of periodicity values had to be similar between $P0(t,f_c)$ and $P0_{\mathrm{tar}}(t,f_c)$, and (2) the periodicity values found in the input had to be similar to the periodicity values in the template. Specifically, the two criteria were defined as follows.

Criterion 1: The difference of the number of periodicity values in one $[t,f_c]$ bin should not exceed a threshold of 2,

$$A(t,f_c) = \begin{cases} 1 & \text{if } |\#P0(t,f_c) - \#P0_{\mathrm{tar}}(t,f_c)| \leq 2 \\ 0 & \text{else.} \end{cases} \tag{7}$$

The symbol # defines the number of elements in a set. The rule is applied to both the left and right channel periodicity features separately; the corresponding variables are termed $A_l(t,f_c)$ and $A_r(t,f_c)$.

Criterion 2: This criterion had two versions depending on whether there were fewer values in the multi-talker input or in the template. If the number of values in the multi-talker input was lower than in the template, then for each periodicity value in the multi-talker input there had to be a value in the template that did not differ by more than 0.1 ms. If the number of values in the template was lower than in the multi-talker input, for each periodicity value in the template there had to be a value in the multi-talker input that did not differ by more than 0.1 ms. Formally,

$$B(t,f_c) = \begin{cases} 1 & \text{if } \forall P0 \in P0_x(t,f_c) : \min_{P0' \in P0_y(t,f_c)}(P0 - P0') < 0.1\,\mathrm{ms} \\ 0 & \text{else,} \end{cases} \tag{8}$$

with

$$P0_x(t,f_c) = \begin{cases} P0(t,f_c) & \text{if } \#P0(t,f_c) < \#P0_{\mathrm{tar}}(t,f_c) \\ P0_{\mathrm{tar}}(t,f_c) & \text{else,} \end{cases} \tag{9}$$

$$P0_y(t,f_c) = \begin{cases} P0_{\mathrm{tar}}(t,f_c) & \text{if } \#P0(t,f_c) < \#P0_{\mathrm{tar}}(t,f_c) \\ P0(t,f_c) & \text{else.} \end{cases} \tag{10}$$

The second rule was also implemented for the left and right channel features individually; the corresponding variables are termed $B_l(t,f_c)$ and $B_r(t,f_c)$. The $\mathrm{BM_{P0}}$ was estimated on the basis of the aforementioned rules that had to apply for both the left and the right channel,

$$\mathrm{BM_{P0}}(t,f_c) = \begin{cases} 1 & \text{if } A_l(t,f_c) = A_r(t,f_c) = B_l(t,f_c) = B_r(t,f_c) = 1 \\ 0 & \text{else.} \end{cases} \tag{11}$$

b. Spectral energy template matching. The spectral energy template was calculated as the mean of all spectral energy features of the 22 unmasked target utterances. BM estimation based on energy template matching was based on the absolute difference between target template and left and right multi-talker signal $\Delta E_l(t,f_c) = |E_l(t,f_c) - E_{\mathrm{tar}}(t,f_c)|$ and $\Delta E_r(t,f_c) = |E_r(t,f_c) - E_{\mathrm{tar}}(t,f_c)|$,

$$\mathrm{BM_E}(t,f_c) = \begin{cases} 1 & \text{if } \Delta E_l(t,f_c) < 2.5\,\mathrm{dB} \wedge \Delta E_r(t,f_c) < 2.5\,\mathrm{dB} \\ 0 & \text{else.} \end{cases} \tag{12}$$

c. Combination of periodicity and spectral energy. The BM for the combination of periodicity and spectral energy features was calculated as the product

$$\mathrm{BM_{E,P0}}(t,f_c) = \mathrm{BM_{P0}}(t,f_c) \cdot \mathrm{BM_E}(t,f_c). \tag{13}$$

That means that the $\mathrm{BM_{E,P0}}$ is only one in the $[t,f_c]$ bins in which both the periodicity features and the spectral energy features matched the template.

3. BMs based on early vs late signal portions

To examine how different temporal portions of the signals contribute to the BMs, an additional analysis was performed in which BMs of the early portion of the signal ($t \leq 100\,\mathrm{ms}$) were treated separately from the BMs of the late portion of the signal ($t > 100\,\mathrm{ms}$). These BMs are referred to as $\mathrm{BM^{early}}$ and $\mathrm{BM^{late}}$, respectively. Combinations of different BM types were denoted as additions of $\mathrm{BM^{early}}$ and $\mathrm{BM^{late}}$, e.g., the combination of IBM in the onset and $\mathrm{BM_{P0}}$ in the offset was termed $\mathrm{IBM^{early}} + \mathrm{BM^{late}_{P0}}$.

C. Estimation of target location

To estimate the target location, two PDFs of the location estimates were generated, one based on the selected bins, $\mathrm{PDF_{sel}}(\alpha)$, and one based on the not-selected bins, $\mathrm{PDF_{nsel}}(\alpha)$. The PDFs were generated by summing up Gaussian kernels centered at the selected or not-selected estimated locations at each $[t,f_c]$ bin,

TABLE I. Overview of performed simulations.

Simulation	Description
A	Control condition (target alone)
B	Multi-talker condition with IBM-based selection
C	Multi-talker condition with selection based on template matching using periodicity ($\mathrm{BM_{P0}}$), spectral energy ($\mathrm{BM_E}$) and a combination of both features ($\mathrm{BM_{E,P0}}$)
D	Influence of early and late portions of the signal

$$\mathrm{PDF_{sel}}(\alpha) = C_1 \cdot \sum_t \sum_{f_c} \mathcal{N}(\alpha_{\mathrm{sel}}(t,f_c),\sigma), \quad (14)$$

$$\mathrm{PDF_{nsel}}(\alpha) = C_2 \cdot \sum_t \sum_{f_c} \mathcal{N}(\alpha_{\mathrm{nsel}}(t,f_c),\sigma). \quad (15)$$

C_1 and C_2 were chosen so that the PDF integrals were one. The target location estimate, $\hat{\alpha}$, was then defined as

$$\hat{\alpha} = \underset{\alpha}{\mathrm{argmax}}(b \cdot \mathrm{PDF_{sel}}(\alpha) - \mathrm{PDF_{nsel}}(\alpha)). \quad (16)$$

The factor b controls the relative influence of selected and not-selected azimuth values for the decision. On the basis of pilot experiments, we set $b = 3$ and the standard deviation of the Gaussian kernels $\sigma = 30°$. This relatively large standard deviation was chosen because it generates smooth PDFs and thus leads to robust predictions. The subtraction of $\mathrm{PDF_{nsel}}(\alpha)$ suppresses the remaining masker-related information in the target-related PDF and resembles the active suppression of masker positions (Dong *et al.*, 2013).

III. METHODS

A. Stimuli

The speech material used here was the same as that used in Kopčo *et al.* (2010; speech corpus of Kidd *et al.*, 2008). The target to be localized was a female voice uttering the word "two," which was kept constant throughout the experiment. The target azimuthal location was between −50° and 50° in 10° steps. The maskers were four male voices uttering a random monosyllabic word which completely overlapped the target word. Each target and masker utterance had approximately the same energy, so that the target-to-masker ratio was 0 dB, as stipulated by Kopčo *et al.* (2010). The resulting SNR was approximately −6 dB. The male talkers were the same throughout the experiment with the same left-to-right order. Five masker location patterns were used: $[-50,-40,-30,-20]^\circ$, $[20,30,40,50]^\circ$, $[-20,-10,10,20]^\circ$, $[-50,-40,40,50]^\circ$S, and $[-40,-10,10,40]^\circ$.

The input signals were generated using virtual acoustics. Clean speech tokens were set to a root-mean-square (RMS) of 1 before convolution with a BRIR for the respective angle. BRIRs were measured in the ears of a human listener in a slightly reverberant room (Kopčo and Shinn-Cunningham, 2011). The distance between head and sound sources was 1 m and the azimuth spacing was 10°. All other methods for measuring BRIRs were the same as described by Shinn-Cunningham *et al.* (2005). In our study, we used only the BRIRs from the left hemisphere and switched left and right channels for the other hemisphere.

B. Simulations

Table I shows an overview of all of the simulations in this study. To assess the model performance for the localization of the unmasked target, a control condition (simulation A) was simulated in accordance with the psychoacoustic study of Kopčo *et al.* (2010). In this simulation, no selection mechanism was implemented, so that all extracted azimuth angles $\alpha(t,f_c)$ contributed to the estimated target location [cf. Sec. II C, Eq. (16)],

$$\hat{\alpha} = \underset{\alpha}{\mathrm{argmax}}(\mathrm{PDF}(\alpha)). \quad (17)$$

Consistent with the computations used for the masked localization simulations, the PDF was calculated based on Gaussian kernels with a standard deviation of $\sigma = 30°$. Only one model run was performed for each target location, because the target utterance was kept constant, in line with Kopčo *et al.* (2010). The model did not simulate any of the localization inaccuracies that presumably occur in the psychoacoustic experiment, e.g., due to the head tracking procedure.

Simulation B investigated the model with the selection of target-related binaural information based on the IBM (see Sec. II B 1). The IBM selection requires full *a priori* knowledge of the target and masker signals in isolation. The simulation can be seen as an investigation of the performance of the binaural model as well as the performance of the location estimation mechanism.

Simulation C investigated the model using BMs based on template matching using periodicity, spectral energy, or a combination of both features ($\mathrm{BM_{P0}}$, $\mathrm{BM_E}$, or $\mathrm{BM_{E,P0}}$, see Sec. II B 2).

Simulation D investigated how the model performance depends on information in the early (first 100 ms of the input) vs late portions of the signal (rest of the input). For this, the BMs from simulations B and C were combined such that either the IBM (from simulation B) was used for early signal portions and $\mathrm{BM_{P0}}$, $\mathrm{BM_E}$, or $\mathrm{BM_{E,P0}}$ (from simulation C) for late signal portions, or vice versa (see Sec. II B 3).

Furthermore, the BMs $\mathrm{BM_{P0}}$, $\mathrm{BM_E}$, and $\mathrm{BM_{E,P0}}$ estimated using these procedures were compared to the IBM in terms of positive predictive values (PPVs), negative predictive values (NPVs), accuracy (ACCs), and glimpse proportions (GPs). The PPV was defined as the total number of true positives, i.e., bins for which both the given BM and the

IBM is one, divided by the number of bins with a value of one in the BM. Thus, it is basically a measure of how many of the selected glimpses are actually target-related, as defined by the IBM, serving as the "gold standard." The NPV is defined as the total number of true negatives, i.e., bins for which both the given BM and the IBM are zero, divided by the total number of bins with a value of zero in the BM. Analogous to the PPV, the NPV is a measure of how many of the not-selected glimpses are actually not target-related, as defined by the IBM. The ACC is defined as the sum of true positives and true negatives divided by the total number of bins. The GP is defined as the number of ones in a BM divided by the total number of bins.

C. Descriptive statistics

In the experiment of Kopčo *et al.* (2010), seven subjects S_i participated, each of them performing ten runs per target position ϕ (11 total) and masker pattern p (5 total). Masker words and masker patterns were randomized across (ϕ, p) conditions and subjects. For the model simulations, 50 runs were performed for each ϕ and p with randomized masker words.

For illustration and descriptive statistics, results for the spatially symmetric (masker) conditions, $p=3$, 4, and 5 were merged across hemispheres. Furthermore, masker patterns $p=1$ and $p=2$ are spatially anti-symmetric, and their results were merged after mirroring the data of pattern 2. That is, for each target location the number of runs was doubled by adding the target location estimations of the respective mirrored location, so that 20 runs were examined instead of 10 (subjects) or 100 instead of 50 (model), except for $\phi = 0°$ in masker patterns $p=3$ through $p=5$. The same merging was done for the control condition with the difference that the subjects performed 20 runs per target location so that 40 runs were examined for the mirrored data. This procedure reduced the influences of the sequence of maskers and room asymmetries on the results.

Model and subject data were compared with regard to the median bias and interquartile range (IQR) across runs within a (ϕ, p) condition. The median bias is a measure of the deviation from perfect localization, referred to here as $\Delta_{S_i}(\phi, p)$ for the subject S_i and $\Delta_M(\phi, p)$ for the model. The IQR was used as a measure for the variation across different runs for a given (ϕ, p) condition, referred to as $\mathrm{IQR}_{S_i}(\phi, p)$ for the subject S_i and $\mathrm{IQR}_M(\phi, p)$ for the model.

As a measure for the similarity between model and subject performance, global and local root-mean-square errors (RMSEs) were used. These RMSEs were always calculated with reference to the medians across individual $\Delta_{S_i}(\phi, p)$ and $\mathrm{IQR}_{S_i}(\phi, p)$, referred to as $\Delta_S(\phi, p)$ and $\mathrm{IQR}_S(\phi, p)$, respectively. The global bias RMSE was used to assess overall performance averaged across location and pattern. It was defined as

$$\mathrm{RMSE}_{\Delta,\mathrm{global},X} = \sqrt{\frac{1}{N}\sum_{\phi=-50°}^{50°}\sum_{p=1}^{5}\left(\Delta_S(\phi,p) - \Delta_X(\phi,p)\right)^2}, \tag{18}$$

where $X = S_i$ for the subject and $X = M$ for the model. The local bias RMSEs were used to assess performance separately for each combination of pattern and target location. They were defined as

$$\mathrm{RMSE}_{\Delta,\mathrm{local},X}(\phi,p) = |\Delta_S(\phi,p) - \Delta_X(\phi,p)|, \tag{19}$$

with the variable X as used in Eq. (18). The calculation of global and local IQR RMSEs was done analogously.

The reference measure for the comparison of model and subject results was the mean and standard deviation of the individual subject's global and local RMSE. The performance of the model was considered similar to human subject performance if the model RMSE lay within two standard deviations of the across-subject RMSE mean. For a rough statement about whether human and model performance were comparable or not, and how large the deviation was, the global RMSEs were used. To make a statement about the difference between human and model performance for individual (ϕ, p) conditions, the local RMSEs were used.

IV. RESULTS

A. Simulation of control condition

Model and subject median biases for the control condition are shown in Fig. 2. The results for the $-50°$ to $-10°$ locations were mirror-flipped and combined with the 50° to 10° location results. As the same target utterance was used for all runs at all locations, the model did not show any variance across runs. Hence, the results for model and subject IQRs across runs are not shown here. The model results were in good agreement with the subject results. This was reflected in the model's global RMSE of 3.5°, compared to the mean subject global RMSE of $3.6° \pm 2.9°$; that is, the global RMSE of the model lay within 0.06 subject standard deviations of their global RMSE. Analysis of the local RMSE revealed that the model was in good agreement with the subject data for all target locations, as can be seen in Fig. 2. At the 40° location, the model and subject median biases differed considerably. At this location, the mean local RMSE of the subjects was $4.68° \pm 4.27°$ compared to a local RMSE of 8.24° for the model; due to the large variability across subject estimates, the model still fulfilled the criterion of not differing by more than two standard deviations from the mean subject RMSE.

B. Simulation using the IBM

Figure 3 shows the model and subject median biases and IQRs across runs for the masked localization data. The human bias data showed similar localization estimates across patterns (top row in Fig. 3). The main feature across the patterns was that the most lateral sources tended to be biased medially. This effect was strongest in masker pattern 1 for the lateral target locations near the distractors (50°) and weakest for masker pattern 1 for the target locations far from the distractor ($-50°$). The model captured the general trend considerably well. However, it did not show the asymmetry between the data at $\phi = -50°$ and $\phi = 50°$ for masker pattern 1. For the IQRs, human data showed a similar behavior

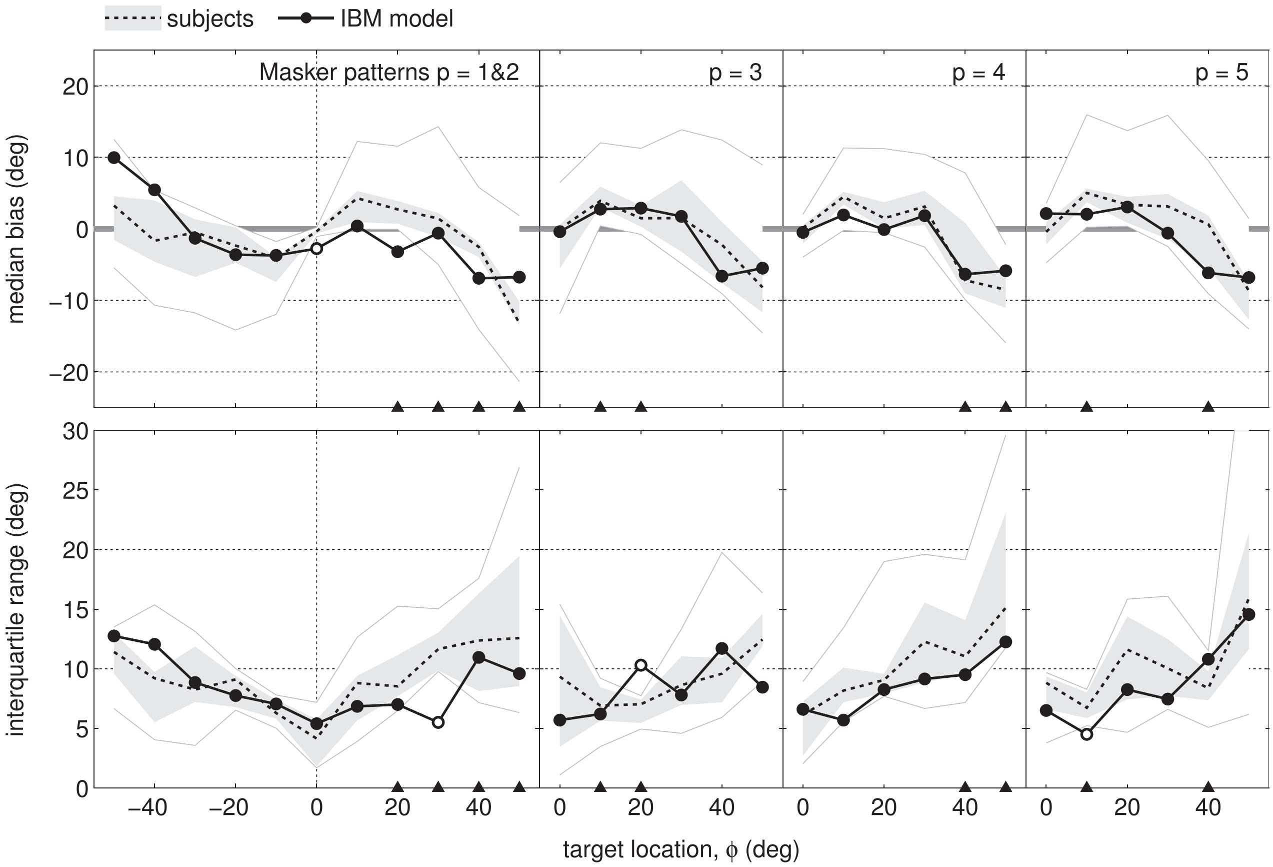

FIG. 3. Masked target localization modeled using the IBM as a selector for target-related bins (simulation B). Median biases (top row) and IQRs across runs (bottom row) are shown as a function of target positions ϕ for subjects and model. Each panel shows the results for a specific spatial masker pattern p, indicated by the black triangles on the abscissa. Results were merged based on spatial masker pattern symmetry. Medians, lower and upper quartiles and minimum and maximum of individual median biases and IQRs are shown for the subjects as dashed lines, filled areas and thin gray lines, respectively (data from Kopčo *et al.*, 2010). Black circles indicate the median biases and IQRs of the model. Open circles indicate that the local model RMSE was more than two standard deviations away from the mean local RMSE of the subjects.

across masker patterns (bottom row in Fig. 3). In particular, the IQRs for lateral target positions tended to be larger than for medial target positions. In most cases, this trend was observable in the model data. For masker pattern 1, the IQR in the human data was higher than the model IQR for the target locations near the distractors and lower than the model IQR for the target locations far from the distractors. A similar trend, although weaker, can be seen for patterns 4 and 5. However, there the model predictions seem to be less stable. So the model did not capture well the difference between human IQRs near vs far from the maskers. Generally, the model showed lower IQRs than the subjects, which might be probably due to the fact that the model incorporated idealized knowledge about the target-dominant $[t, f_c]$ bins that the subjects did not have.

The global RMSEs for the median biases were 4.2°±2.9° for the subjects and 3.5° for the model (see also Table II). For the IQRs the global RMSEs were 3.6°±1.7° for the subjects and 2.5° for the model. That is, for both the biases and the IQRs, the global RMSEs of the model were within two standard deviations of the mean global RMSEs of the subjects. This indicates that the overall model predictions did not differ significantly from the subject data and that the

TABLE II. Global model RMSEs for the bias and the IQR for the different model versions. RMSEs were calculated relative to the median subject data. Model data were obtained using different types of BMs and combinations of BMs to test the influence of selection of early vs late portions of the signal on localization performance. Z values identify how many standard deviations the global RMSEs of the model differed from the mean global RMSEs of the subjects.

	Bias		IQR	
BM type	RMSE (deg)	z	RMSE (deg)	z
IBM	3.5	−0.2	2.5	−0.6
$\text{IBM}^{\text{early}}$	4.2	−0.0	4.6	0.6
IBM^{late}	5.4	0.4	11.2	4.4
BM_{P0}	26.2	7.6	12.9	5.4
$\text{IBM}^{\text{early}}+\text{BM}^{\text{late}}_{\text{P0}}$	7.6	1.2	9.5	3.4
$\text{BM}^{\text{early}}_{\text{P0}}+\text{IBM}^{\text{late}}$	6.7	0.8	10.8	4.1
BM_{E}	6.3	0.7	6.9	1.9
$\text{IBM}^{\text{early}}+\text{BM}^{\text{late}}_{\text{E}}$	3.8	−0.2	3.0	−0.3
$\text{BM}^{\text{early}}_{\text{E}}+\text{IBM}^{\text{late}}$	5.2	0.3	3.8	0.2
$\text{BM}_{\text{E,P0}}$	3.2	−0.4	5.4	1.0
$\text{IBM}^{\text{early}}+\text{BM}^{\text{late}}_{\text{E,P0}}$	3.5	−0.3	2.8	−0.4
$\text{BM}^{\text{early}}_{\text{E,P0}}+\text{IBM}^{\text{late}}$	4.1	−0.1	5.1	0.9

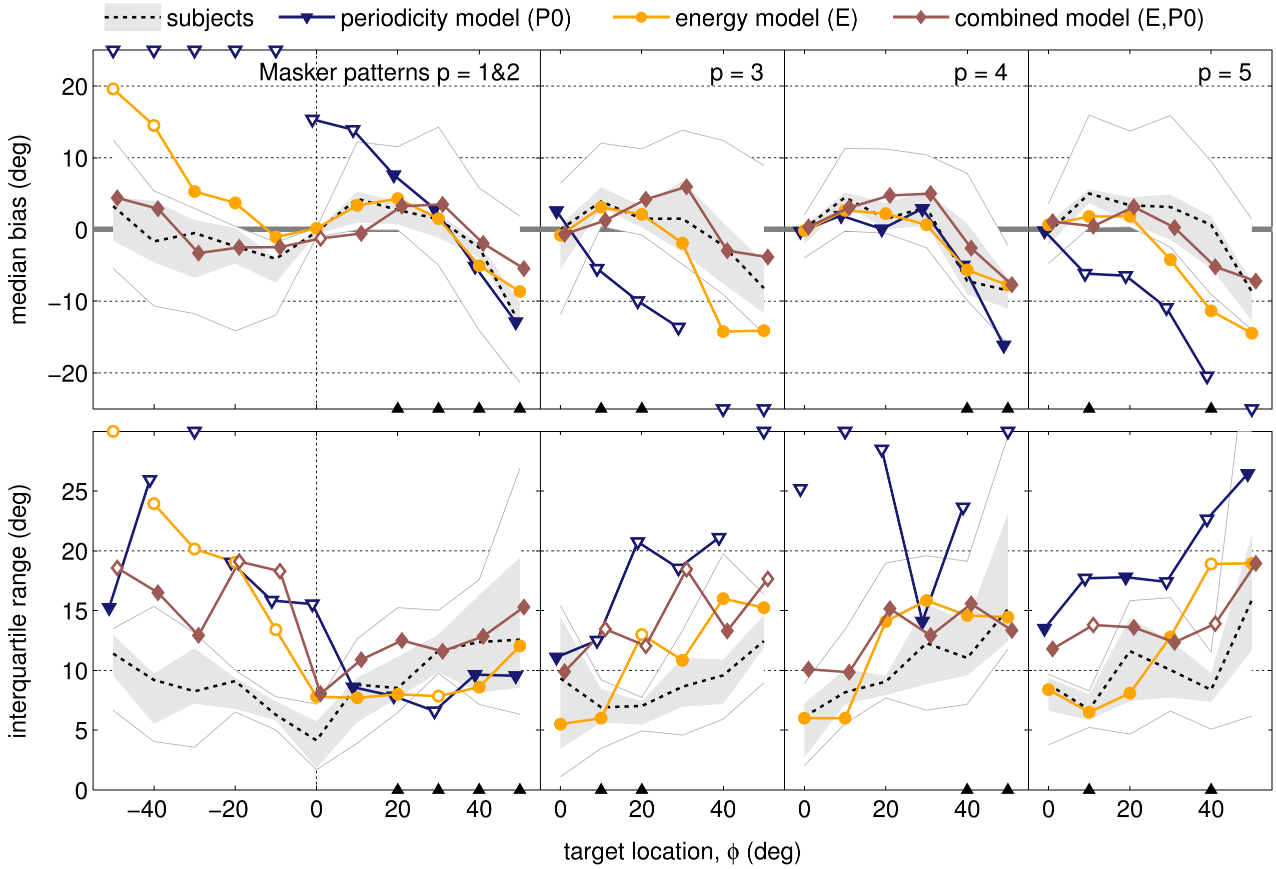

FIG. 4. (Color online) Masked target localization biases (top row) and IQRs (bottom row) as a function of target position ϕ for the template-matching models (simulation C). The layout of the figure and the human data are identical to Fig. 3. Different colors and symbols represent the model variations using different monaural features (triangles: periodicity; circles: spectral energy; diamonds: combination of both monaural features). The open symbols indicate that the local model RMSE was more than two standard deviations away from the mean local RMSEs of the subjects. Values that fall outside the plot range are plotted along the plot edges and not connected to the other data points.

IBM-based model can be used as a reference for evaluating models that do not use optimal *a priori* information for selection of target-related $[t, f_c]$ bins. Investigating the local RMSEs, it was observed that for most (ϕ, p) conditions, the model did not significantly differ from the subject data (open circles in Fig. 3 indicate where it did).

C. Simulations using template matching

Figure 4 shows the simulation results using the models based on template matching, using a layout similar to Fig. 3. The periodicity model shows a global RMSE of 26.2° for the biases and 12.9° for the IQR (triangles in Fig. 4, see also Table II). Both of these model RMSEs lay outside two standard deviations of the mean global RMSE of the subjects. Particularly, large differences were observed for the left hemisphere of masker pattern 1, in which the model responses showed a very strong bias toward the middle and even toward the masker positions. Also the IQRs were very large for these conditions. However, there were similarities to both the subject biases and IQRs in terms of local RMSE in masker pattern 1 for the on-masker positions. In masker patterns 3 and 5 a good performance was found for the center target positions ($\phi = 0°$) in terms of both bias and IQR. This performance degraded for the more lateral positions, where the model estimates strongly differed from the subject biases and IQRs. For masker pattern 4, the bias estimates were close to the subject biases; however, with the exception at $\phi = 30°$, the IQRs of the estimates were considerably higher than observed in the subject data.

For the energy model (circles), the global RMSE was 6.3° for the bias and 6.9° for the IQRs. Both of these values lay inside two standard deviations of the mean global RMSEs of the subjects. For most (ϕ, p) conditions, the model was in good agreement with the subject results as analyzed with the local RMSEs. The model generally captured the trends of a medial localization bias and the increment of IQRs for lateral positions (masker patterns 3–5). These trends tended to be more distinct in the model than in the subject data. In masker pattern 1, the performance strongly degraded for positions far from the masker positions.

The combined model results (diamonds) had a global RMSE of 3.2° for the biases and 5.4° for the IQRs. Both of these values lay inside two standard deviations from the mean global RMSE of the subjects. The biases were in good agreement with the subject results for all target positions and all masker patterns. The IQRs generally seemed to be higher than the subject IQRs. Significant differences were found within the off-masker locations in masker pattern 1 and for some of the locations in masker patterns 3 and 5. Generally, the results for the combined model nearly approached the performance of the IBM model.

D. Influence of early vs late signal portions

Table II shows the global RMSEs of median bias and IQR for the model using different BMs and BM combinations as a selector for target-related binaural features. BMs were combined using the IBM in the early portions of the signal, and $\mathrm{BM}_{\mathrm{P0}}$, BM_{E}, or $\mathrm{BM}_{\mathrm{E,P0}}$ in the late portions of the signal; or vice versa (see Sec. II B 3).

Using only the early or only the late signal portions, selected with the IBM, generally increased the bias and IQR RMSEs compared to using the whole signal. This increment was stronger when using only the late signal portions, especially for the IQR. These findings suggest than an accurate selection seems to be more important for the early portions of the signal than for the late portions. Still, the most accurate results were found when the complete IBM was used.

Results for the mixed BMs showed that replacing the selection at early and late signal portions by an optimal selection lowered the RMSEs. This effect was strongest for the periodicity model. The ideal selection in early signal portions generally led to slightly lower bias and IQR RMSEs than the ideal selection in late signal portions.

Comparing the results for $\mathrm{IBM}^{\mathrm{early}}$ and the combination of $\mathrm{IBM}^{\mathrm{early}}$ and $\mathrm{BM}^{\mathrm{late}}$ showed diverse results for the different features: While the combinations of $\mathrm{IBM}^{\mathrm{early}}$ with $\mathrm{BM}_{\mathrm{E}}^{\mathrm{late}}$ and $\mathrm{BM}_{\mathrm{P0,E}}^{\mathrm{late}}$ led to a decrease in RMSE compared to $\mathrm{IBM}^{\mathrm{early}}$ alone, the combination of $\mathrm{IBM}^{\mathrm{early}}$ and $\mathrm{BM}_{\mathrm{P0}}^{\mathrm{late}}$ led to a relatively strong RMSE increase compared to $\mathrm{IBM}^{\mathrm{early}}$ alone. These findings suggest that the selection in the late signal portions of BM_{E} and $\mathrm{BM}_{\mathrm{P0,E}}$ is similar to the contribution of $\mathrm{IBM}^{\mathrm{late}}$, while the selection in the late signal portions of $\mathrm{BM}_{\mathrm{P0}}$ possibly contains false positives and false negatives.

E. Comparison of BMs

The results of simulation B showed that the model performance using the IBM as a selector for target-related binaural features was very similar to the subject results. Using the BMs based on template matching showed different results depending on the monaural features used. It is therefore of interest to compare those BMs to the IBM. Figure 5 shows BMs obtained with the four different approaches for one sample run, alongside the individual PPVs and NPVs, ACCs, and GPs. Table III shows the average measures across all conditions and runs. For the IBM (top left panel in Fig. 5) glimpses were found during the first 50 ms in almost all frequency bands. This was also reflected in a relatively high average GP of $(20.3\pm11.0)\%$ within the early portion of the signal (0–100 ms; see Table III). There were also distinct glimpses observable during the late portions (>130 ms) in the modulation channels and the channels with $f_c = 236\,\mathrm{Hz}$, $f_c = 414\,\mathrm{Hz}$ and $f_c = 488\,\mathrm{Hz}$. There were no glimpses found after approximately 50 ms in the central frequency channels with $f_c = 569\,\mathrm{Hz}$ to $f_c = 1470\,\mathrm{Hz}$. This pattern generalized to other sample runs as well, resulting in a much higher average GP during the early portions than during the late portions in the IBM model (Table III).

The GP for the energy model (BM_{E}) was similar to the GP of the IBM. The pattern of selected glimpses was also similar in the two models: both had a higher GP in the early than in the late portion, and both lacked glimpses in the late portions for frequency bands between $f_c = 569\,\mathrm{Hz}$ and $f_c = 1470\,\mathrm{Hz}$. In contrast to the IBM and the BM_{E}, $\mathrm{BM}_{\mathrm{P0}}$ was very sparse. The $\mathrm{BM}_{\mathrm{E,P0}}$ was the intersection of $\mathrm{BM}_{\mathrm{P0}}$ and BM_{E}, and was therefore also very sparse.

It is notable that the very few glimpses in the $\mathrm{BM}_{\mathrm{E,P0}}$ were very accurate estimates of the actual target-related glimpses defined by the IBM, as seen in a PPV of $67.1\%\pm20.9\%$. Compared to that, the $\mathrm{BM}_{\mathrm{P0}}$ showed a relatively low PPV of $22.5\%\pm12.7\%$, indicating that only a

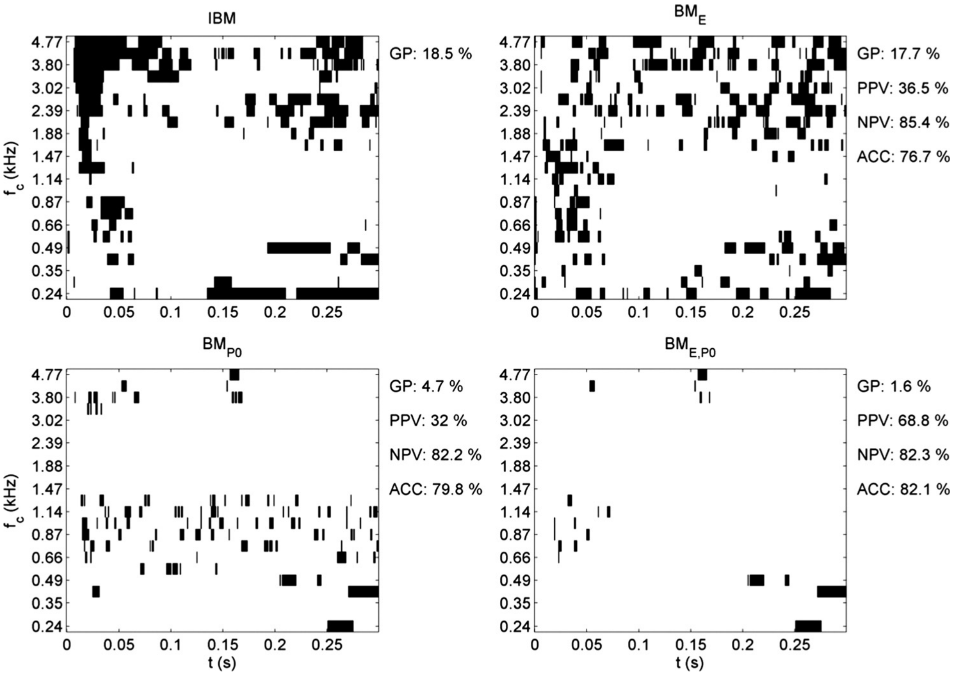

FIG. 5. Comparison of BMs, which serve as the basis for the selection of target-related binaural information, for one sample run (top left: IBM, bottom left: $\mathrm{BM}_{\mathrm{P0}}$, top right: BM_{E}, bottom right: $\mathrm{BM}_{\mathrm{E,P0}}$). Black areas identify the estimated target-dominant time-frequency bins ("glimpses"). The calculation of PPVs, NPVs, and ACCs (shown on the right, next to each panel) of the template-based BMs was done with reference to the IBM. Furthermore, the GPs are shown. Vertical dashed lines indicate the separation between early portions (<100 ms) and late portions (>100 ms) of the signal.

TABLE III. PPVs, NPVs, and ACCs of the different template-matching BMs with respect to the IBM, and GPs for IBM and BMs. The table shows the measures for the whole BMs, and for the early and late portions individually.

	PPV (%)	NPV (%)	ACC (%)	GP (%)
IBM				
Whole				13.0 ± 5.5
Early				20.3 ± 11.0
Late				9.3 ± 4.1
BM$_{\mathrm{P0}}$				
Whole	22.5 ± 12.7	87.5 ± 5.2	84.8 ± 4.9	4.2 ± 0.9
Early	30.3 ± 22.5	80.4 ± 10.4	78.4 ± 9.4	4.6 ± 1.8
Late	16.2 ± 10.2	91.0 ± 4.0	88.0 ± 3.8	3.9 ± 0.9
BM$_{\mathrm{E}}$				
Whole	33.3 ± 12.6	90.7 ± 4.0	82.6 ± 3.8	14.7 ± 3.8
Early	39.2 ± 17.5	84.7 ± 8.8	77.1 ± 7.5	19.4 ± 8.2
Late	27.4 ± 10.2	93.2 ± 3.1	85.3 ± 3.2	12.3 ± 2.7
BM$_{\mathrm{E,P0}}$				
Whole	67.1 ± 20.9	87.5 ± 5.2	87.4 ± 5.1	0.9 ± 0.6
Early	61.3±29.9[a]	80.4 ± 10.5	80.4 ± 10.1	1.4 ± 1.4
Late	66.1±27.1[a]	91.0 ± 3.9	90.9 ± 3.9	0.6 ± 0.5

[a] Not included are all runs with GP = 0, i.e., 127 runs for BM$^{\mathrm{early}}_{\mathrm{E,P0}}$ and 32 runs for BM$^{\mathrm{late}}_{\mathrm{E,P0}}$.

small part of the already few selected glimpses actually identified target-related bins. As seen in Fig. 5, there was a large number of mis-selections during the late portions for frequency channels from $f_c = 569$ Hz to $f_c = 1296$ Hz. This was reflected in a low PPV of BM$_{\mathrm{P0}}$ in the late portions (16.2%±10.5%). Interestingly, the PPV of BM$_{\mathrm{E}}$ was also relatively low (33.3%±12.6%), although the overall congruence with the IBM seemed to be relatively high (see Fig. 5). These results revealed that the false selections arising from the periodicity and spectral energy features in isolation could be largely removed if the two features were combined.

While the PPVs differed between the BMs, their NPVs were relatively similar. NPVs generally had high values around 90%, meaning that the not-selected $[t, f_c]$ bins were generally correctly identified as not target-related. The accuracy was also similar between the different BMs and showed a relatively high value. This may have been due to the dominance of the number of correct negatives in this measure.

Higher PPVs were found in the early portions of BM$_{\mathrm{P0}}$ and BM$_{\mathrm{E}}$, while PPVs were similar for the early and late portions of BM$_{\mathrm{E,P0}}$. However, for all BMs, the NPV was approximately 10% higher in the early portions than in the late portions, which was also reflected in the accuracy.

V. DISCUSSION

The present study introduced an auditory model for localization of target speech in a complex acoustic environment. The model was evaluated on experimental data in which the target, a female voice uttering the word “two,” was masked by four spatially separated male voices (Kopčo *et al.*, 2010). Notable properties of the acoustic scene were a relatively low SNR, short utterances, a full temporal overlap of the target by the maskers, temporally aligned onsets, reverberation, and previously unknown masker words whose spatial configuration was also unknown. The model extracted ITD-based binaural features from the multi-talker scene, selected the target-related binaural features based on a BM, and estimated the target location by combining information from the selected and not-selected binaural features. As a selector, we evaluated the IBM and BMs based on a template-matching procedure using periodicity, spectral energy, and a combination of both features. Additionally, the contribution of BMs was examined separately in the early and late portions of the signals.

When the binaural feature selection was based on the IBM, the model performance was in good agreement with the subject performance. However, to create an IBM one needs to have the signal containing the maskers without the target, a requirement that is not fulfilled in regular localization tasks. The model performance using the template-matching BMs depended on the monaural features applied: Using periodicity, the overall model performance was worse than the subjects’ performance. Using spectral energy, the performance was only slightly worse than the subject performance. Using both features combined led to subject-like performance in terms of bias and a slight performance degradation in terms of IQR. Replacing these BM-based selections by an optimal IBM-based selection in the early or late portions of the signal led to an improved model performance.

A. Differences between simulation design and experimental setup

Although the task was the same for the model and the subjects, there were some differences between the experiment and the simulation, which may have influenced the general comparability of model and subject performance.

First, the room used to record the BRIRs for the model simulations in this study was similar to, but not the same as, that used to collect the subject data. The differences in room geometry and materials of walls, ceilings and floor may have caused some differences in reverberation and therefore a difference in performance. However, given the good match between the subject and IBM-model performance, this was unlikely a big factor.

Another difference is that the model does not incorporate any individualization to account for differences in behavior between subjects. First, subjects may show some characteristic variabilities arising from the head-tracking procedure in the experiment. Applying an “internal noise” to the model location estimate would account for these variabilities; our present model version, however, does not do this. Second, each individual may have a characteristic response behavior, e.g., a certain azimuth offset. Our model did not account for these kinds of differences. This could be incorporated by modeling each subject individually.

B. Differences between model and subject performance

Several factors may have caused the observed differences between model and subject performance. (1) There may be a difference in how binaural information was extracted and combined, as the model used primarily ITDs while the subjects may have also based their localization on ILDs. (2) There may be differences in the selection of target-related time-frequency bins (assuming that the humans use such a selection at all). (3) There may be a mismatch between the model and the subjects in the binding process that links the binaural and monaural information related to the target and the maskers to estimate the target location. The data using the IBM as a selector for target-related binaural features showed absolute biases and IQRs comparable to or even lower than the subject data. We can thus assume that the binaural features (stage 1), and the location estimation procedure used here (stage 3) accurately simulate human performance. Thus, larger absolute biases and IQRs in specific models most likely occurred due to inaccuracies in the selection of target-related time-frequency bins. These inaccuracies can arise from both incorrect selection of masker-dominated bins as target bins (false positives) and omission of target bins (false negatives).

The combined BM yielded results close to the subject performance and IBM, although it was very sparse and therefore missed many target-related time-frequency bins. On the other hand, the relative number of false positives (1-PPV) was rather low. This finding suggests that misses are not necessarily a drawback, as long as the few selected bins are accurately estimated. As seen in the results for the periodicity BM, too many false positives can have a large negative effect on the model performance.

False positives occur whenever, by coincidence, the template and the multi-talker mixture differ in feature values by less than the chosen minimum difference threshold. For periodicity, false alarms were observed in the fine structure filters with center frequencies of approximately 600–1400 Hz. This may be due to an overlap of the high harmonics of target and masker signals. Therefore, voiced masker-dominated bins might easily be classified as target bins. For spectral energy, false positives occur whenever the mixture and the target template have a similar energy, while the target is not active in the mixture.

We attempted to reduce the influence of false positives on target location estimation by subtracting the PDF of not-selected binaural features from the PDF of selected binaural features before estimating the target location. Because the NPVs were generally high for all BMs, the estimation of the background is considered to be relatively accurate. However, especially for the periodicity model, this method was not sufficient to exclude the influence of false positives. One way to potentially improve the results would be to optimize the parameter b in Eq. (16), which determines the relative influence of the PDFs of selected and not-selected binaural features. In the present model, b was optimized for IBM results and was not changed for the other BMs. It is possible that location estimates that are remote from the masker locations would become more accurate if b were decreased; however, this would come at the cost of more inaccuracies for positions close to the maskers.

The influence of false positives and false negatives was especially prominent for masker patterns 1 and 2, in which all of the maskers were in one hemisphere, and the targets were in the other hemisphere (azimuths less than $-10°$). Here, masker-related binaural features were all in the range of 20° to 50°, so there was a large difference between them and the target positions. If the number of false positives and false negatives was high, the resulting PDF had a maximum either between target and masker positions or at the masker positions, resulting in large biases from the actual target position, and a wider possible spread of location estimates across runs. However, the subjects did not seem to have a problem localizing the target in these conditions.

C. Influence of early vs late portions of the signal

Several studies have shown that binaural information is primarily read out at the signal onset or at rising segments of the signal envelope (Houtgast and Aoki 1994; Freyman *et al.*, 1997; Dietz *et al.*, 2013). The present simulations support these findings. In particular, in simulation D, IBM-based selection led to better results when used in the early signal portions than in the late signal portions. This implies that binaural features in the early portions of the signal are more accurate than in the late portions of the signal; this was expected, since reverberation has a smaller influence in the early portions.

In this experiment, all target and masker tokens started synchronously, so no onset features or temporal order features were available to segregate the talkers. Therefore, correct selection of target-related time-frequency bins was important, especially in the early signal portions when the binaural features are more reliable. Simulation D showed that the combination of IBM-based selection in the early portions and template-matching BM-based selection in the late portions can distinctly improve the results compared to using the template-matching BMs in the early and late portions. Furthermore, the analysis of BMs revealed a generally lower accuracy and NPV of BMs for the late than for the early signal portions: In the early portions, the proportion of misses (false negatives) of all not-selected $[t, f_c]$ bins was higher. On the other hand, PPVs in the early portions were higher than or equal to PPVs in the late portions. These results show that the tested template-matching procedures are not able to bring out enough target-related binaural features available in the early portions of the signal.

VI. CONCLUSIONS

(1) The binaural model of Dietz *et al.* (2011) is capable of extracting a sufficient amount of ITD information to model localization of speech in a multi-talker masking speech mixture. Together with a location estimation back-end that is based on both target-related and background-related features, the model performance is comparable to the subject performance. However, this requires optimal selection of target-related "glimpses" in the time-frequency plane, e.g., using the IBM; the target localization cannot be achieved based on the binaural model alone. It requires in addition a sophisticated method to separate the target-related glimpses from the masker ones.

(2) Segregation based on target-alone template matching, while more realistic than the IBM-based segregation, could not predict the human data as accurately as the IBM approach when using either periodicity features or spectral energy features alone. However, while periodicity features alone led to a strong performance degradation, spectral energy features were still reasonably accurate. Combining the two features improved the model performance so that it approached subject performance.

(3) Extracting binaural information from the target-dominated time-frequency bins during the early portions of the signal seems to be important for performing the task in reverberant environments. This is likely because reverberant energy is initially low, and does not affect binaural information. However, neither of the template-matching features was capable of extracting enough of the critically important target-related information during the signal onset.

(4) The failure of the template-based BMs to extract the target-related information during the early signal portions indicates a more complex selection process, possibly involving temporal integration and across-frequency integration of correlative extracted features, which was not considered in this study. Alternatively, it is possible that the listeners combined ITDs and ILDs to estimate the target, an option not considered in the binaural model used in this study.

(5) Binaural and periodicity features were selected based on a salience measure with a rather strict criterion. It was then assumed that each selected feature either belongs to the target or the background. This means that binaural unmasking as implemented, e.g., in equalization-cancellation models of binaural processing, was excluded. Still, the model performed as well as human listeners. This suggests that explicit modeling of target-masker superposition may not be needed for modeling human sound localization.

ACKNOWLEDGMENTS

A.J. and V.H. were supported by the DFG (SFB/TRR 31 "The active auditory system") and the Cluster of Excellence EXC 1077/1 "Hearing4all." N.K. was supported by the Slovak Research and Development Agency Contract No. APVV-0452-12 and by the TECHNICOM project, ITMS-26220220182, of the European Union Research and Development Fund. The authors would like to thank Peter Tóth, who contributed to the implementation of the model and simulations.

Alain, C., Reinke, K., He, Y., Wang, C., and Lobaugh, N. (**2005**). "Hearing two things at once: Neurophysiological indices of speech segregation and identification," J. Cognit. Neurosci. **17**(5), 811–818.

Barker, J., and Cooke, M. (**2007**). "Modelling speaker intelligibility in noise," Speech Commun. **49**(5), 402–417.

Bronkhorst, A. W. (**2000**). "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," Acta Acust. Acust. **86**(1), 117–128.

Chen, Z., and Hohmann, V. (**2015**). "Online monaural speech enhancement based on periodicity analysis and a priori SNR estimation," IEEE/ACM Trans. Audio Speech Language Processing **23**(11), 1904–1916.

Darwin, C. J. (**1981**). "Perceptual grouping of speech components differing in fundamental frequency and onset-time," Q. J. Exp. Psychol. **33**(2), 185–207.

Dau, T., Püschel, D., and Kohlrausch, A. (**1996**). "A quantitative model of the 'effective' signal processing in the auditory system. I. Model structure," J. Acoust. Soc. Am. **99**(6), 3615–3622.

Dietz, M., Ewert, S. D., and Hohmann, V. (**2011**). "Auditory model based direction estimation of concurrent speakers from binaural signals," Speech Commun. **53**(5), 592–605.

Dietz, M., Marquardt, T., Salminen, N. H., and McAlpine, D. (**2013**). "Emphasis of spatial cues in the temporal fine structure during the rising segments of amplitude-modulated sounds," Proc. Natl. Acad. Sci. **110**(37), 15151–15156.

Dong, J., Colburn, H. S., and Sen, K. (**2013**). "A computational model of spatial tuning in the auditory cortex in response to competing sound sources," Proc. Meet. Acoust. **19**(1), p. 050105.

Elhilali, M., Ma, L., Micheyl, C., Oxenham, A. J., and Shamma, S. A. (**2009**). "Temporal coherence in the perceptual organization and cortical representation of auditory scenes," Neuron **61**(2), 317–329.

Faller, C., and Merimaa, J. (**2004**). "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," J. Acoust. Soc. Am. **116**(5), 3075–3089.

Freyman, R. L., Zurek, P. M., Balakrishnan, U., and Chiang, Y. C. (**1997**). "Onset dominance in lateralization," J. Acoust. Soc. Am. **101**(3), 1649–1659.

Giguère, C., and Abel, S. M. (**1993**). "Sound localization: Effects of reverberation time, speaker array, stimulus frequency, and stimulus rise/decay," J. Acoust. Soc. Am. **94**(2), 769–776.

Gockel, H. (**1998**). "On possible cues in profile analysis: Identification of the incremented component," J. Acoust. Soc. Am. **103**(1), 542–552.

Gockel, H., and Colonius, H. (**1997**). "Auditory profile analysis: Is there perceptual constancy for spectral shape for stimuli roved in frequency?," J. Acoust. Soc. Am. **102**(4), 2311–2315.

Houtgast, T., and Aoki, S. (**1994**). "Stimulus-onset dominance in the perception of binaural information," Hear. Res. **72**(1), 29–36.

Kidd, G., Jr., Best, V., and Mason, C. R. (**2008**). "Listening to every other word: Examining the strength of linkage variables in forming streams of speech," J. Acoust. Soc. Am. **124**(6), 3793–3802.

Kopčo, N., Best, V., and Carlile, S. (**2010**). "Speech localization in a multitalker mixture," J. Acoust. Soc. Am. **127**(3), 1450–1457.

Kopčo, N., and Shinn-Cunningham, B. G. (**2011**). "Effect of stimulus spectrum on distance perception for nearby sources," J. Acoust. Soc. Am. **130**(3), 1530–1541.

Langendijk, E. H., Kistler, D. J., and Wightman, F. L. (**2001**). "Sound localization in the presence of one or two distracters," J. Acoust. Soc. Am. **109**(5), 2123–2134.

Roman, N., Wang, D., and Brown, G. J. (**2003**). "Speech segregation based on sound localization," J. Acoust. Soc. Am. **114**(4), 2236–2252.

Shamma, S., and Fritz, J. (**2014**). "Adaptive auditory computations," Curr. Opin. Neurobiol. **25**, 164–168.

Shamma, S. A., Elhilali, M., and Micheyl, C. (**2011**). "Temporal coherence and attention in auditory scene analysis," Trends Neurosci. **34**(3), 114–123.

Shinn-Cunningham, B. G., Kopčo, N., and Martin, T. J. (**2005**). "Localizing nearby sound sources in a classroom: Binaural room impulse responses," J. Acoust. Soc. Am. **117**(5), 3100–3115.

Teki, S., Chait, M., Kumar, S., Shamma, S., and Griffiths, T. D. (**2013**). "Segregation of complex acoustic scenes based on temporal coherence," eLife **2**, e00699.

Wang, D. (**2005**). "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines* (Springer, New York), pp. 181–197.