Modeling of Speech Localization in a Multitalker Mixture using „Glimpsing“ Models of Binaural Processing

Peter Toth[1], Angela Josupeit[2], Norbert Kopco[3], Volker Hohmann[2]

[1]Charles University Prague, Czech Republic, [2]Medical Physics Section, Department of Medical Physics and Acoustics and Cluster of Excellence "Hearing4all", Oldenburg University, Germany, [3]Safarik University Kosice, Slovakia,

Introduction

A recent study measured the human ability to localize a speech target masked by a mixture of four talkers in a room (Kopco et al., 2010, fig.1.). The presence of maskers resulted in increases in localization errors that depended on the spatial distribution of maskers, the target-to-masker energy ratio (TMR), and the listener's knowledge of the maskers' locations.

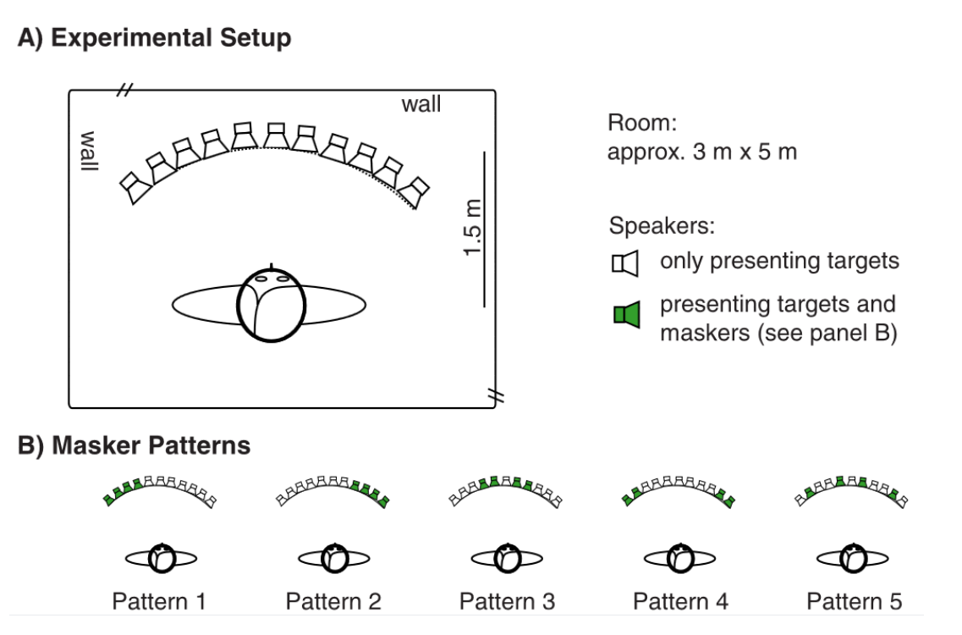

Fig.1. Experimental setup. From (Kopco et al., 2010)

In this study we investigate the performance of two binaural auditory "glimpsing" models in experiments comparable to Kopco et al.,2010:

1. A model (Faller and Merimaa, 2004) based on short-term cross-correlation
2. A model (Dietz et al., 2011) based on interaural phase difference (IPD) extraction

The models were tested under the assumption that optimal information about the TMR in individual spectro-temporal glimpses is available, quantifying the ability of the models to encode spatial properties of complex acoustic scenes (Cooke, 2006).

Materials & Methods

The following framework was used to test the two models:

A: multitalker signal generation

To create sound inputs close to those in the experiment, the stimuli were simulated by convolving speech tokens from the experiment with binaural room impulse responses recorded in a reverberant space similar to the experimental room.

B: auditory peripheral preprocessing

Standard auditory preprocessing involving gammatone filterbank was used. Specific parametrization was adopted from (Dietz et al., 2011).

Twenty-three frequency bands ranging from 200 Hz to 5000 Hz were divided into twelve lower (fc < 1400 Hz) bands containing fine structure information and eleven higher (fc > 1400 Hz) bands containing information processed with modulation filters (creating „envelope").

C: binaural processing

Models use different methods to extract binaural cues – while Faller and Merimaa based their binaural processing on computing IC in range of time shifts, Dietz computes interaural phase differences (IPD) directly. Both models require defined exponential-decaying time window upon which they do computations.

D: selection of glimpses

Both tested models claim to have an ability to select spectro-temporal bins dominated by energy from only one source ("glimpses"). Binaural cues contained in selected bins should point to location of that source.

Faller and Merimaa use glimpsing based on interaural coherence (IC) – high short-term correlation between the left and the right signal indicates that one source is dominant in particular time-frequency window.

Dietz used interaural vector strength (IVS) of interaural phase differences to determine glimpses – i.e., dominance of one source is indicated by low fluctuation of IPD values.

E: target-related information selection

As the main focus of this work is to explore the best possible performance of the glimpsing models, we used ideal binary mask (IBM, Cooke, 2006) to relate individual glimpses to target source and maskers sources. IBM was computed as a ratio between target energy and maskers energy above a certain threshold (default set to 0 dB).

F: target location estimation

We examined several methods for integration of information from glimpses across time and frequency bands. They relied on ITD information. Methods of location estimation varied from simple (mean/median of ITD of selected samples, maximum of histogram) to more complicated which used information from both selected and non-selected glimpses (combining Gaussian kernel-based pdf functions for each group of glimpses).

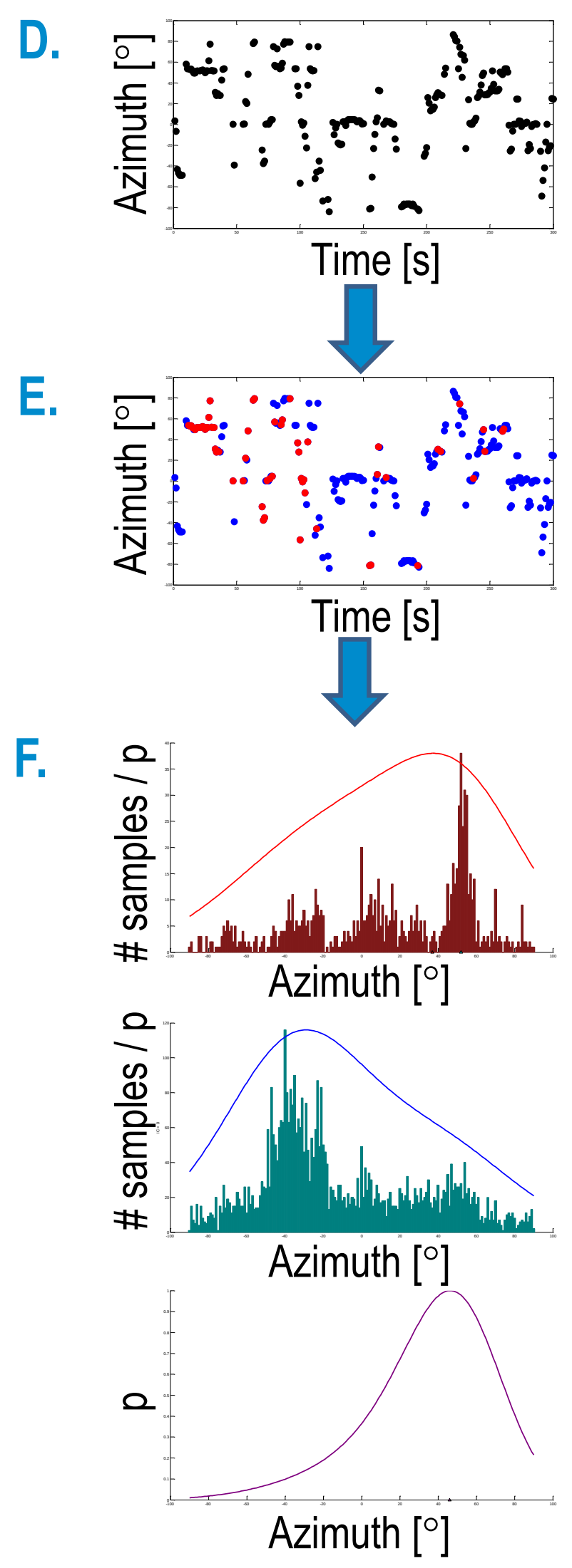

Fig.2. Illustration of glimpsing selection (D., E.) and target location estimation (F.). Each dot represents glimpse and points to the certain angle. Selected glimpses are marked with red, non-selected with blue color. Final estimation (F, purple) is based on division of selected glimpses (F, red) by non-selected glimpses (F, blue).

Results

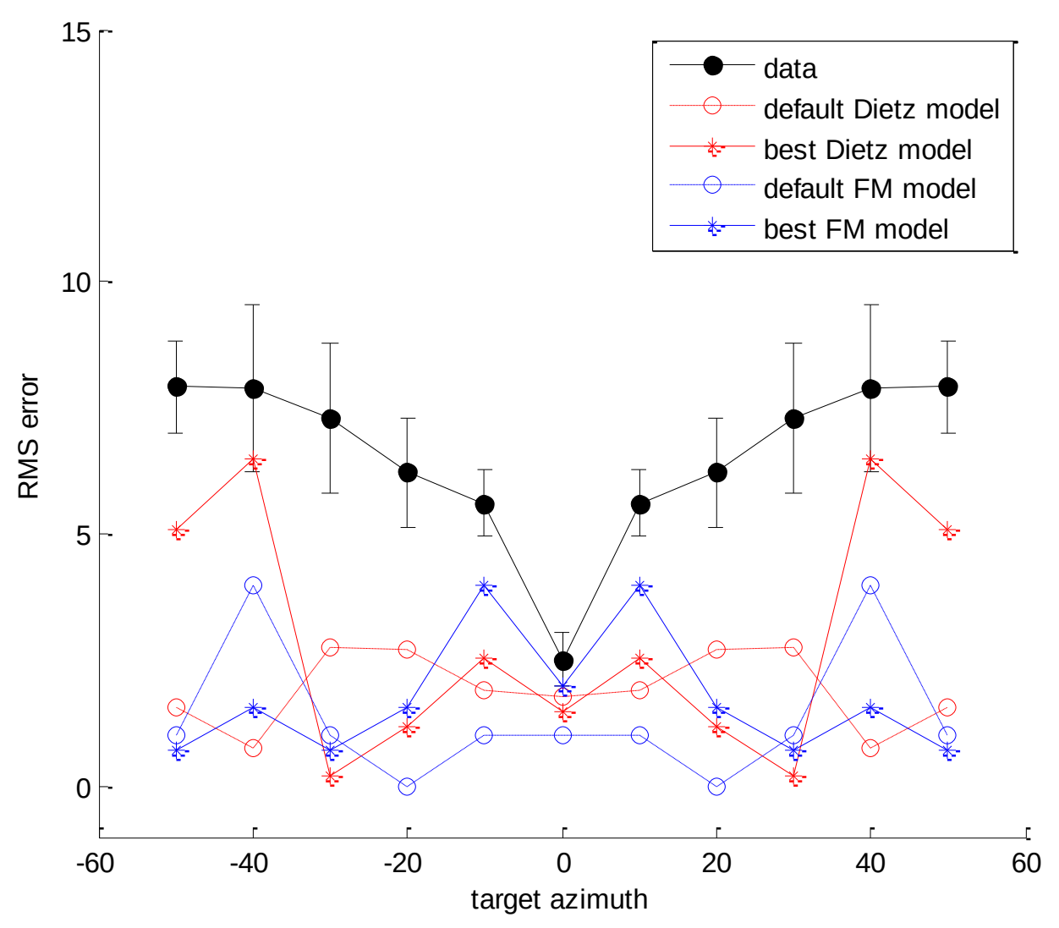

Fig.3. RMS errors for control runs (only target presented)

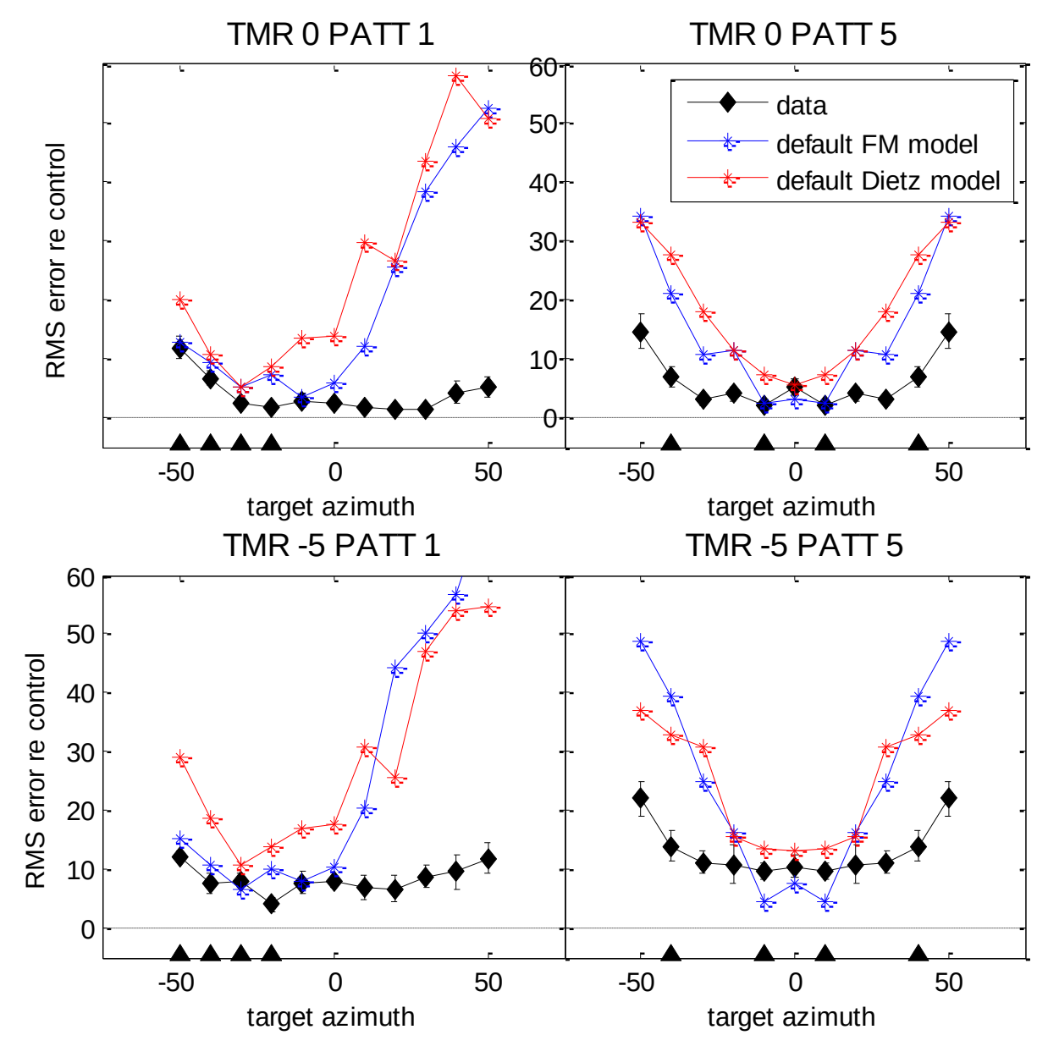

Fig.4. RMS masker-induced error across target positions and masker patterns 1 and 5.

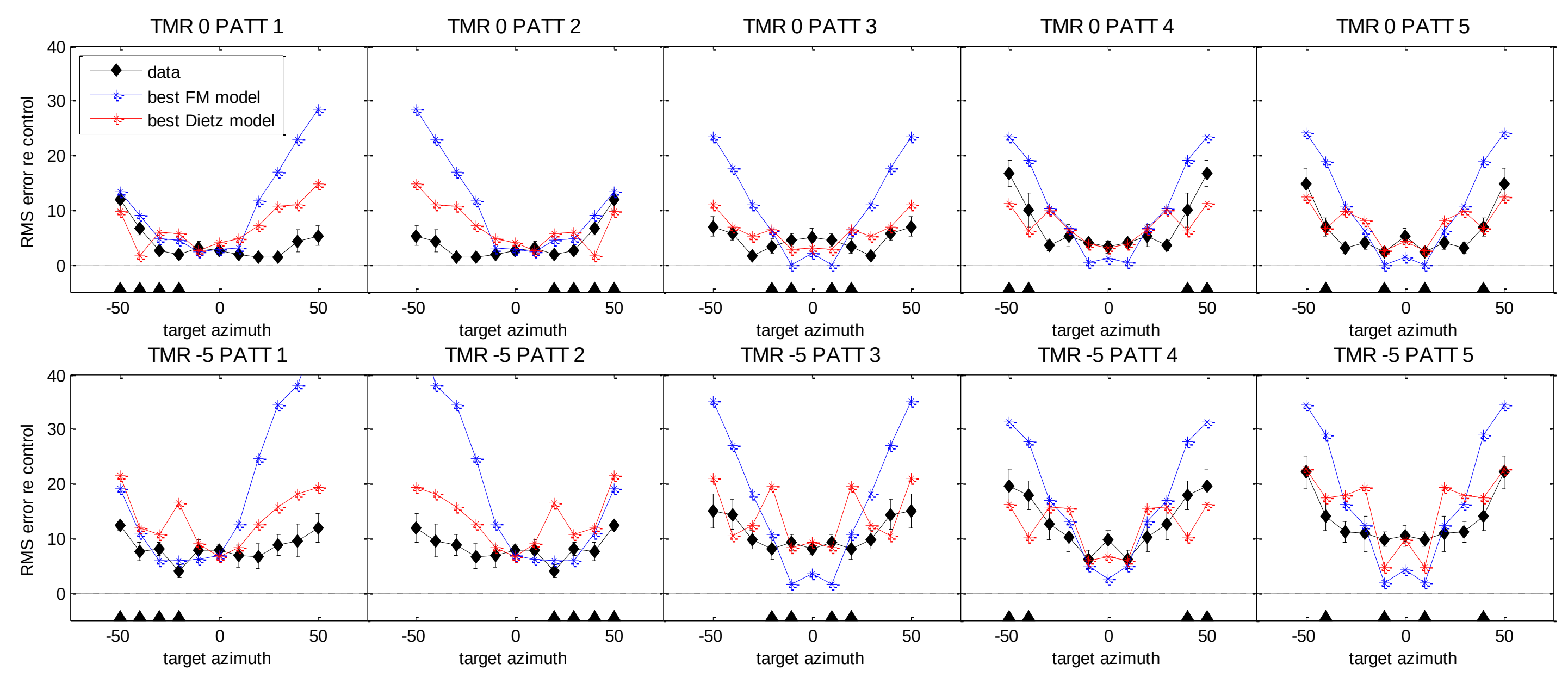

Fig.5. Increase in RMS error due to the presence of the maskers. Predictions of the Dietz and Faller and Merima models with optimized parameters and the Kopco et al. (2010) data. All plots assume left-right symmetry and data were mirror-flipped before plotting.

Model predictions were compared to three measures of human performance: the increase of RMS errors due to maskers, the biases due to maskers, and the miss rates/false alarm rates.

Lowering glimpsing threshold

Performance of glimpsing models was examined by different settings of glimpsing thresholds (gt). Default thresholds for both models was 0.98. We observed that with lowering thresholds results get better. Figures show default (gt = 0.98) and the best (gt = 0) parametrizations for each model.

Lowering binaural processor time constant

Lower time constant means shorter integration window for processing binaural cues and more emphasis on actual states of cues. We have observed improved performance in localization with lowering time constant. In fact, to match humans' performance we had to set time constant to very low values (ca. 1-5 times fundamental period).

On-masker and off-masker target locations

Psychoacoustic data shows that human localization performance is better when target and maskers are spatially separated. Both glimpsing models (in every parametrization) show opposite behavior: the farther the target from the maskers, the worse localization performance is.

Target location estimation

A ways to evaluate information in selected glimpses to make final target location estimation have a big influence on the model results. Although simple methods like mean ITD could have low overall RMS errors, methods based on suppressing non-selected glimpses are closer to experimental data.

High frequency bands

The models can perform very well based only on high frequency channels despite high ITD ambiguity. The best results were achieved when both low and high frequency channels were processed together.

Miss rates and false alarm rates

The subjects in the experiments were asked to indicate when they did not hear the target. Inability to detect targets in normal trials is measured by miss rate while inability to indicate no presence in catch trials by false alarm rate. Default versions of the models have miss rates quantitatively comparable to humans. However, we can observe the same on-masker/off-masker disparity as in the RMS data. With lower glimpsing thresholds miss rates dropped to zero. False alarm rates were unrealistically high in all cases.

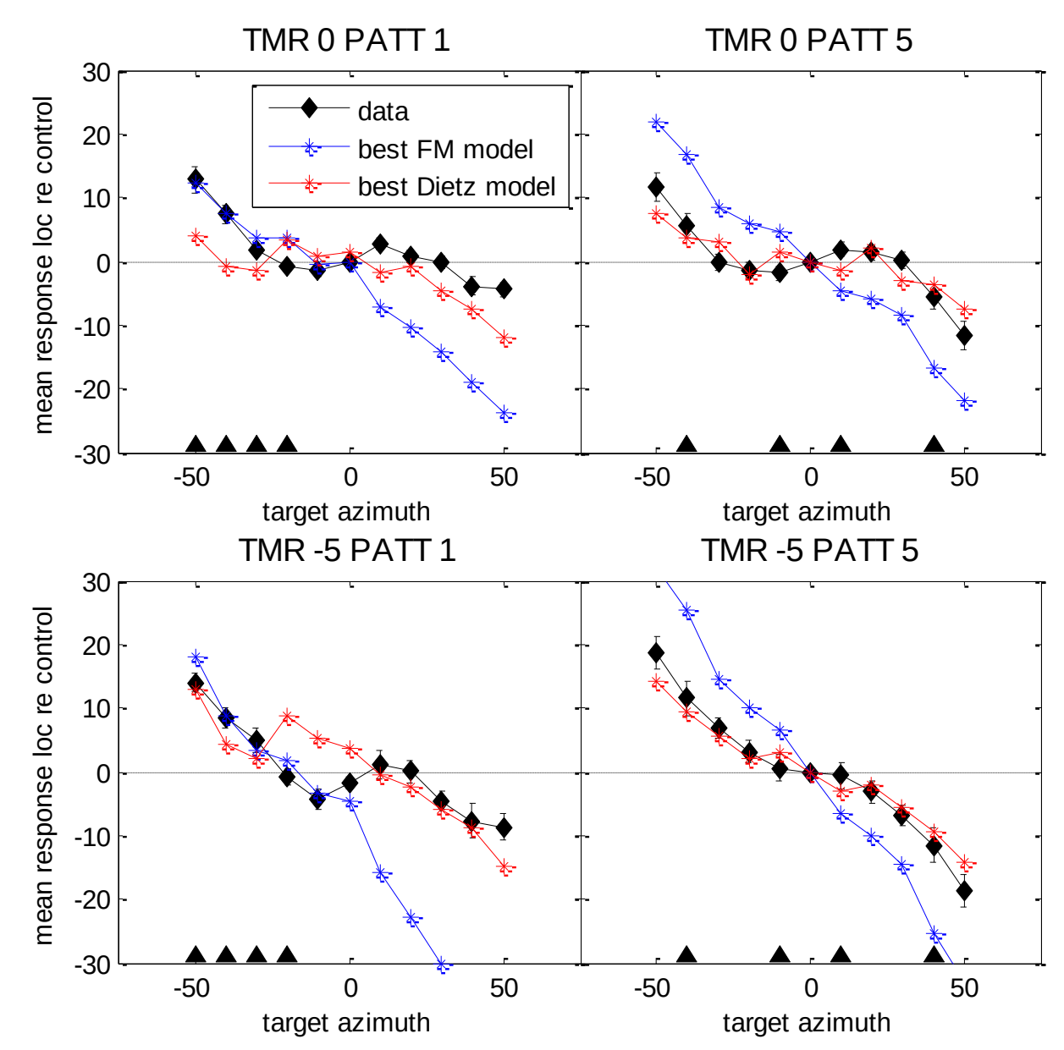

Fig.6. Mean bias in responses re. control condition. Results of the two models and Kopco et al. (2010) data are shown for patterns 1 and 5.

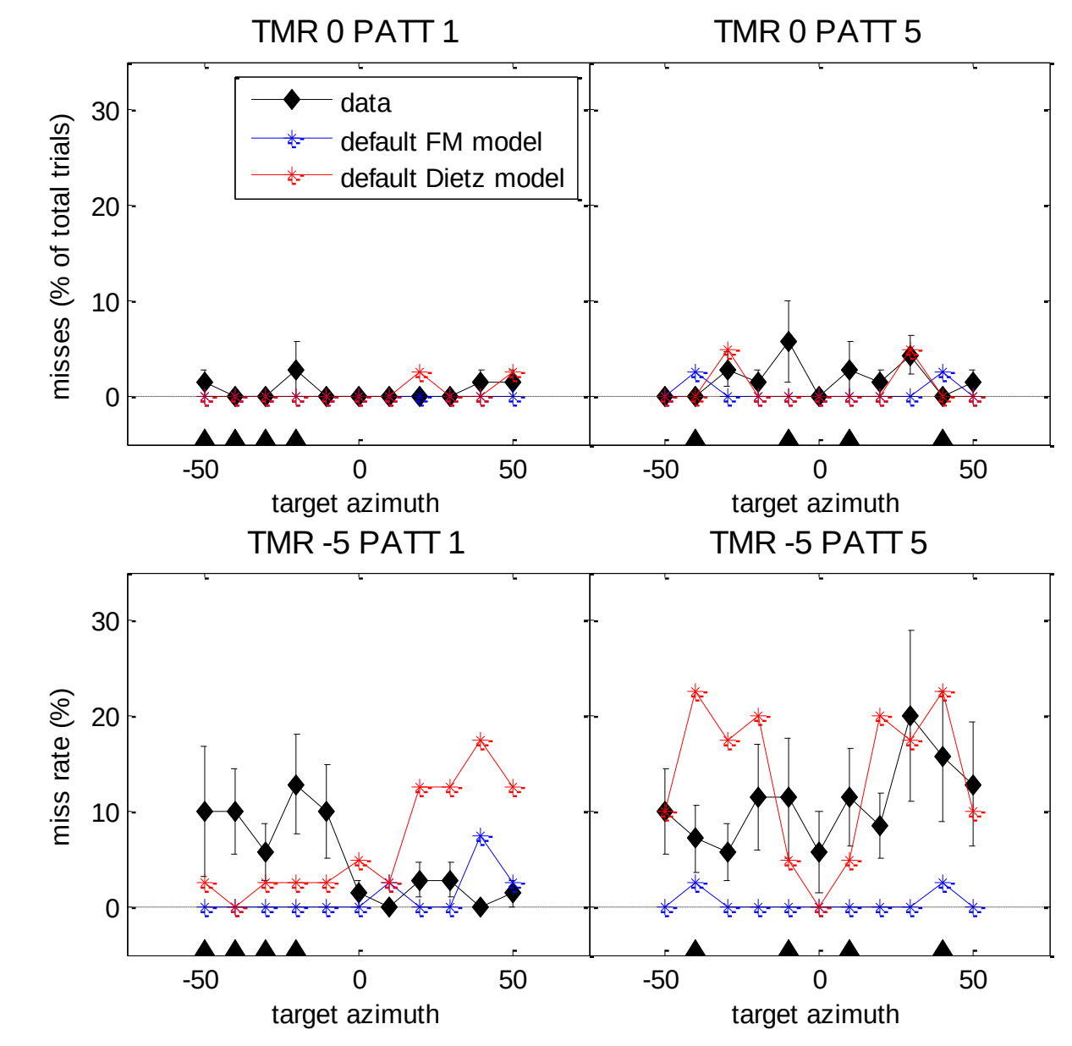

Fig.7. Miss rates across target positions and masker patterns 1 and 5.

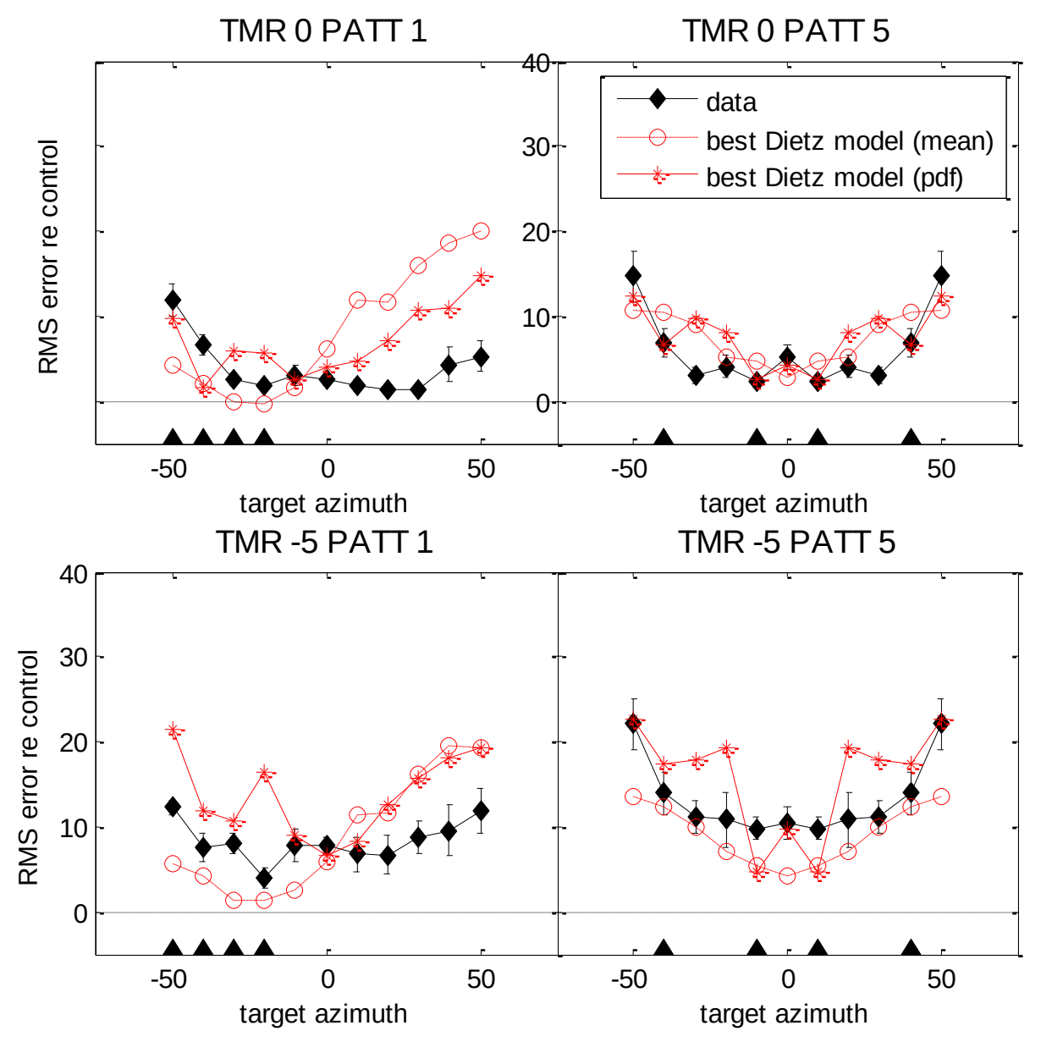

Fig.8. Dietz model predictions with two different methods of target location estimation: mean of the predictions of the target glimpses vs. the ratio of the estimates of the target glimpse pdf and masker glimpse pdf.

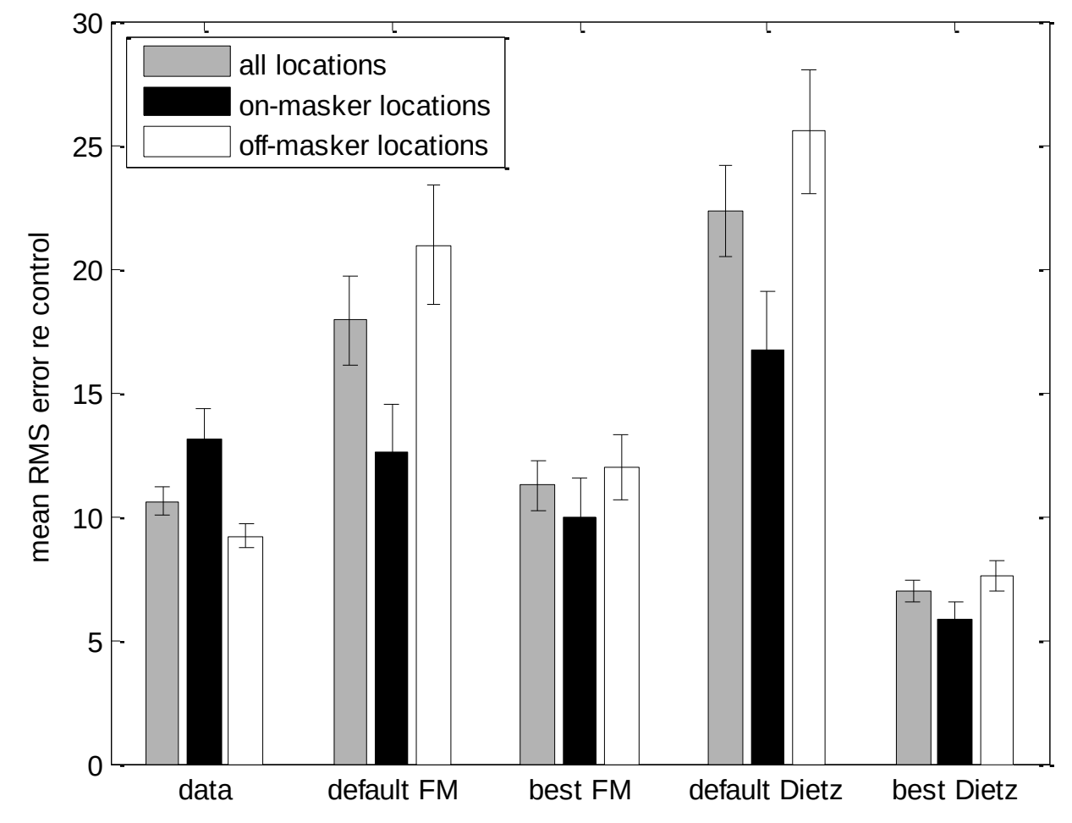

Fig.9. Overall and on-masker/off-masker locations mean RMS errors.

Conclusions

The tested binaural models were able to capture several characteristics of human performance - the mean responses for lateral target locations were medially biased, the RMS errors were smallest for central target locations, and the overall performance varied with TMR. Even though each model extracts binaural information in a different way, the model predictions were comparable, suggesting that the extracted features are equivalent and integrated in similar ways.

Contrary to the expected behavior of glimpsing models, high thresholds do not yield better results. It seems that higher numbers of selected samples are beneficial for further processing in more complex scenarios, where information tends to be noisy so more samples are needed for better estimation. ITD/phase ambiguity can be another reason for having more samples being an advantage.

It seems that on-masker/off-masker disparity is a direct consequence of the way the glimpsing model works. Target glimpses can contain also masker information that causes shifts in estimation – this shift is bigger when maskers are more distant (humans don't make this kind of errors) .

The differences between the model predictions and human performance might be due to differences in interaural level difference processing, across-channel feature integration, the assumed method of combination of target and masker glimpses and due to the usage of IBM.

Kopco et al (2010) also examined the effect of providing a priori information about masker locations on performance. The current models need to be extended to be make use of this kind of information.

References

Cooke, M. (2006). "A glimpsing model of speech perception in noise. " J. Acoust. Soc. Am. 119(3)
Dietz M. et al. (2011). "Auditory model based direction estimation of concurrent speakers from binaural signals" Speech Comm 53(5)
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)
Kopco N. et al. (2010). "Speech localization in a multitalker mixture." J. Acoust. Soc. Am. 127(3)

Acknowledgements

Work supported by EU NIH P41RR14075 FP7-247543, VEGA-1/0492/12, the DFG SFB/TRR 31 "The Active Auditory System", NIH P41RR14075, and the PhD program "Hearing".