Sound localization in complex multitalker conditions by harmonic template matching

Angela Josupeit, Volker Hohmann

Medical Physics Section, Department of Medical Physics and Acoustics and Cluster of Excellence "Hearing4all",
Oldenburg University, Email: angela.josupeit@uni-oldenburg.de

Introduction

A recent psychoacoustic study found that normal hearing listeners can localize a target talker in a complex multitalker condition with high accuracy and low variability [6]. The task of the subject was to localize a constant female target token presented alongside four male masker tokens concurrently uttering a random monosyllabic word, each longer than the target word. Each target and masker utterance had the same signal energy. Maskers were presented in different spatial patterns which were random across trials. The experimental room was slightly reverberant.

A previous modeling study [7] showed that "short-time" binaural localization models [3, 1] can successfully model the psychoacoustic results if full *a priori* knowledge about the SNR in each time-frequency (t-f) bin is available. Therefore we use one of these models [1] for the extraction of localization features. Another previous study attempted to blindly estimate target-dominant t-f bins based on harmonic template matching – using the target harmonicity as a prior. This approach, however, did not lead to results comparable to the psychoacoustic results [5]. In the present study, among further analysis of this aforementioned procedure we introduce and analyze a template matching procedure based on spectral energy and a combination of both harmonic and energy features.

Methods

The model framework proposed in this study consists of three major steps that are explained in the following.

Step 1: Binaural feature extraction

For binaural feature extraction we used the localization model of Dietz et al. [1] that includes an auditory preprocessing stage (with frequency channels f_c), a fine structure and envelope processing, and the extraction of interaural phase differences. For binaural resolution, a rather low time constant of $\tau = 1/f_c$ is used. Furthermore, only those binaural information is used where the interaural vector strength $\mathrm{IVS}(t, f_c) > 0.9$. Based on the binaural model alone, a segregation of target and masker localization information is not possible because most of the binaural information is determined by the masker background in the acoustic condition investigated here. A strong selection of target-related binaural information is requried to solve the task.

Step 2: Selection of target-related binaural information

The selection of target-related binaural information is based on template matching procedures of different auditory features: (A) harmonicity, (B) spectral energy, and (C) a combination of both features. Basically, the template matching procedure estimates binary masks (A) BM_{F0}, (B) BM_E, and (C) $\mathrm{BM}_{E,F0}$=BM_{F0}·BM_E of target-dominant t-f bins which serve as a basis to read out target-related binaural information (see fig. 2).

The extraction of harmonic features (A) is based on the calculation of synchrograms ([4, 2]) in the different auditory channels (preprocessing like in Step 1). Synchrograms represent the proportion of the harmonic energy of a signal with fundamental period P compared to the overall energy, for every point t, and every period P. Synchrogram maxima with high values (e.g. > 0.9) correspond to the dominant fundamental period P_0 and its multiples nP_0. Thus, those maxima (possibly multiple per t and f_c!) function as harmonicity features, referred to as $nP_0(t, f_c)$. Spectral energy features $E(t, f_c)$ (B) are calculated as the energy at the output of each preprocessed frequency band using 10 ms windows and a sampling frequency of 1 kHz.

The target template was calculated as the average of all isolated target utterances (2 channels × 11 directions). For template matching, the features extracted from the multitalker utterance were compared to the target templates. For harmonicity template matching the following rules were used: First, the number of extracted maxima for each t and f_c must not differ by more than 2. Second, the difference in period between corresponding maxima must not exceed 0.1 ms. For the spectral energy approach, a t-f_c bin is considered as target-related if the absolute energy difference is not greater than 2.5 dB. These rules have to apply for both left and right channel.

Step 3: Estimation of target location

The target location $\hat{\alpha}$ is estimated as follows:

$$\hat{\alpha} = \mathrm{argmax}_{\alpha} \left(\mathrm{PDF}_{\mathrm{sel}}(\alpha)^b / \mathrm{PDF}_{\mathrm{nsel}}(\alpha)\right).$$

$\mathrm{PDF}_{\mathrm{sel}}(\alpha)$ and $\mathrm{PDF}_{\mathrm{nsel}}(\alpha)$ are gaussian kernel based probability density functions (PDFs) calculated from the selected and the not-selected binaural information, respectively. The division by the PDF of not-selected (i.e. masker-related) binaural information leads to a highlighting of target-related binaural information outlying the main distribution. The function of the exponent b is to influence the individual proportion of the two PDFs.

Results

The subject and model localization performance in terms of bias and standard deviation (STD) are shown in Fig. 1. The IBM approach leads to results which are generally in line with the psychoacoustic findings - with some uncertainties in STD at off-masker positions.
The harmonicity approach leads to proper results mostly for on-masker target positions. At peripheral and off-masker target positions, the approach fails both in terms of bias and STD. This performance degradation is also reflected in the low congruence of BM$_{F0}$ and IBM (see f. 2, plots 3 and 1): We observe a relatively low hit rate, in particular at the onset and in higher frequency bands.
The performance of the energy model approach is better than the previous approach; however, the performance is not as good as the subjects' performance. The congruence between BM$_E$ and IBM is higher than observed in the harmonicity approach (see f. 2, plots 2 and 1).
The combination of harmonicity and energy approach leads to proper results for the bias, but to higher STDs than any other model version. This might be caused by uncertainties due to the very low hit rate and high sparseness of the BM$_{E,F0}$ compared to the IBM.

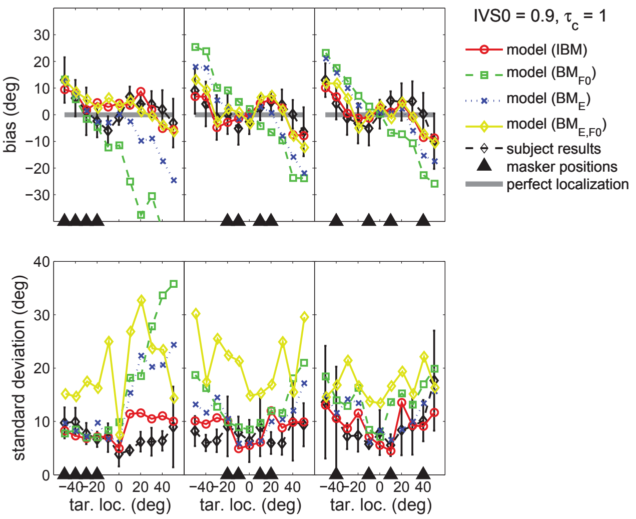

Figure 1: Model (colored lines) and subject (black) localization performance in a multitalker mixture for three example masker patterns. Top: Biases from real target location as a function of target location. Bottom: Standard deviation across runs. Different colors identify different procedures of target glimpse selection in the models. Error bars in subject data identify standard deviations across subjects.

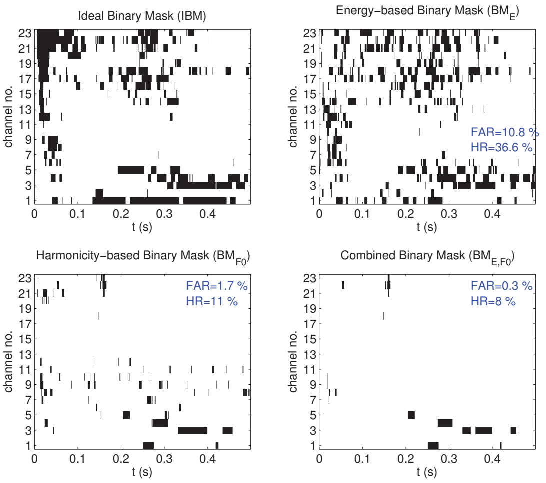

Figure 2: Examples of binary masks obtained with different template matching methods compared to the Ideal Binary Mask (IBM). False alarm rates (FAR) and hit rates (HR) are calculated with reference to the IBM.

Conclusions

(1) The binaural model [1] - embedded in the described framework - is basically able to predict the psychoacoustic results on localization in a multitalker mixture. This is reflected in the good results obtained with the Ideal Binary Mask (IBM) approach, i.e., by using full *a priori* knowledge about the input signal SNR.
(2) Harmonicity as extracted using the synchrogram alone is not a sufficient prior to read-out target-related "glimpses" in the multitalker mixture. That is both reflected in the poor localization performance as well as in the poor congruence between the obtained binary mask and the IBM.
(3) The energy contour of the target in different frequency bands might be a relatively strong cue for the read-out of target-related "glimpses" in the multitalker mixture.
(4) A combination of both harmonicity and energy-based approaches seems promising if hit rates could be further enhanced.

Acknowledgements

This work was supported by the DFG Collaborative Research Centre SFB/TRR 31.

References

[1] M. Dietz, S. D. Ewert, V. Hohmann. Auditory model based direction estimation of concurrent speakers from binaural signals. Sp. Comm., 53(5):592-605, 2011.

[2] S. D. Ewert, C. Iben, V. Hohmann. Robust fundamental frequency estimation in an auditory model. AIA-DAGA 2013, pp. 271-274, Berlin, 2013.

[3] C. Faller, J. Merimaa. Source Localization in Complex Listening Situations: Selection of Binaural Cues Based on Interaural Coherence. J. Acoust. Soc. Am., 116(5):3075–89, 2004.

[4] V. Hohmann. Verfahren zur Extraktion periodischer Signalkomponenten und Vorrichtung hierzu., Patent 2006.

[5] A. Josupeit, S. van de Par, N. Kopco, V. Hohmann. Modeling of speech localization in a multitalker environment using binaural and harmonic cues. AIA-DAGA 2013, pp. 724-727, Berlin, 2013.

[6] N. Kopco, V. Best, S. Carlile. Speech localization in a multitalker mixture. J. Acoust. Soc. Am., 127(3):1450-7, 2010.

[7] P. Toth, A. Josupeit, N. Kopco, V. Hohmann. Modeling of speech localization in a multitalker mixture using "glimpsing" models of binaural processing. ARO abstracts, vol. 37, pp. 94(A), 2014.