Modeling of Speech Localization in a Multitalker Environment using Binaural and Harmonic Cues

Angela Josupeit[1*], Steven van de Par[1], Norbert Kopčo[2], Volker Hohmann[1]
[1] *Carl von Ossietzky University Oldenburg, Germany*
[2] *Safarik University Košice, Slovakia*
[*] *Email: angela.josupeit@uni-oldenburg.de*

Introduction

In typical social situations we can be confronted with multiple speakers, noise and reverberation. Despite the complexity of such a situation we are still able to follow the speaker we want to attend to. One important ability that we use to solve such a difficult auditory scene is the ability to localize the target speaker. Kopčo et al. (2010) measured localization performance of normal-hearing subjects in a task in which the target speaker was fully temporally overlapped by masker speakers and was presented with slight reverberation. In their experiment, subjects were placed in front of a loudspeaker array with 11 loudspeakers at positions of $-50°$ to $50°$ with $10°$ spacing. The task of the subjects was to localize a female target token, pronouncing the word "two", played randomly from one of the loudspeakers. The target utterance was constant throughout the experiment. Alongside the target token, four maskers were presented each with the same signal intensity as the target, leading to a target-to-masker ratio (TMR) of 0 dB. The maskers were presented in four different loudspeakers with different spatial patterns which were randomized throughout the experiment. Masker tokens were four different male talkers, uttering randomly selected monosyllabic words which each were longer than the target word. This led to a full temporal overlap of the target by the maskers. Although the psychoacoustic study examined different spatial masker patterns, in this contribution only one pattern is examined with masker positions at $-50°$, $-40°$, $-30°$, and $-20°$.

Due to the full temporal overlap as well as the presence of reverberation, solving the task is difficult and requires a high temporal resolution of the analysis to perceive "glimpses" of the target. To get a first insight into the possible processing strategy, we propose and analyze an auditory model framework based on established peripheral auditory processing models. The general approach of this framework is to integrate binaural and harmonic features, where binaural features deliver the required location information and harmonic features are basically used to separate the target voice from the masker voices. Key to the approach is the application of the concept of "glimpsing" (Cooke, 2006), which assumes that, after peripheral auditory frequency analysis and exploiting its high temporal resolution, reliable cues are selected for further processing, leading to a sparse spectro-temporal representation of the input. We hypothesize that this concept allows extracting all features that are necessary to perform the task of localizing a fully temporally masked female target talker in a slightly reverberant condition. One important assumption is the availability of *a priori* information, the harmonic features of the target alone, referred to as "target template". Similar *a priori* information was also available for the subjects of the psychoacoustic study, because they performed localization experiments of the target alone before performing the main experiment with the masked target. Therefore, they might have had a clear impression of the harmonicity content of the constant target utterance. Finally, the integration of binaural and harmonic features is based on the assumption that temporally coinciding glimpses belong to the same source.

Model description

Step 1: Generation of binaural signals

For generating the binaural input signals, binaural room impulse responses (BRIRs) were used that were recorded in a room similar to the one in which the psychoacoustic experiment was performed. The speech corpus used in the proposed model framework was the same as the one used in the psychoacoustic study.

In common with the psychoacoustic experiment, the female target token, the word "two" stayed constant throughout the experiment while its position was randomized from trial to trial; the masker tokens varied both in their monosyllabic word and in their loudspeaker positions while the talker identities and their spatial order stayed constant. All tokens are adjusted so that they have the same RMS values and the same lengths. Then, all individual tokens were convolved with the appropriate BRIRs according to their spatial position. The simulated utterances were then summed to form the binaural input signal.

Step 2: Auditory preprocessing

The auditory preprocessing used in this model was the same as used in the binaural model of Dietz et al. (2011). It consisted of

1. middle ear band pass filtering,
2. a gammatone filterbank ranging from ca. $f_c > 200$ Hz to $f_c < 1400$ Hz center frequencies with 1 ERB spacing,
3. half wave rectification,
4. compression,

5. only for harmonicity model: low pass filtering with 770 Hz cutoff frequency

Step 3: Binaural feature extraction

For the binaural feature extraction, the binaural model of Dietz et al. (2011) was used which uses the auditory preprocessing as described in step 2. The outputs of this auditory preprocessing stage are processed through a fine structure filter (gammatone filter with center frequency equal to the center frequency f_c of the respective frequency band). The modulation filters as described in Dietz et al. (2011) were not taken into account here.

For each fine structure output, the instantaneous interaural phase difference IPD(n) as a function of time$n$ is calculated. This IPD signal is smoothed using a low pass filter with a relatively short time constant of $\tau_{f_c} = \frac{2.5}{f_c}$ (e.g. for $f_c = 1000$ Hz: $\tau_{1000Hz} = 2.5$ ms).

To determine the reliability of the extracted IPDs, the stability of IPD values over a short time window is investigated. The measure for this stability is referred to as "interaural vector strength" (IVS). If the IVS is high, the extracted IPDs are considered to be reliable, i.e. these IPDs are "glimpses". Azimuth glimpses $\alpha(n)$ are calculated using the instantaneous frequency of the signal, a lookup table providing a function ITD(α), and the interaural level differences ILD(n) to avoid the ambiguities that occur in ITD-to-α mapping.

After the binaural model processing, the extracted azimuth features were sampled down to a sampling frequency of $fs = 1000$ Hz. This step is not included in the original binaural model.

The extracted localization glimpses of all frequency bands are shown in the left panel of Fig. 2. In this plot, only the selected glimpses (as described in step 6) are highlighted in different colors, each presenting one frequency band (center frequencies f_c in Hz are identified in the legend on top of the figure). The glimpses that were not selected are shown in gray color independent of the frequency band.

Step 4: Harmonic feature extraction

The extraction of harmonic features is based on a method described by Hohmann (2006) and further developed by Ewert et al. (2013). It is performed on the auditory preprocessed signals of the "better ear". For each output channel, the normalized synchrogram $S(n, P)$ is calculated. Each point of this synchrogram identifies the proportion of a harmonic signal with the period P compared to the total signal energy of the investigated time window at time n. A value of $S(n', P') = 1$ would mean that at the time n' the harmonic signal with the period P' can explain all the signal energy we observe, i.e. it is the only component in the signal. Therefore, high values of $S(n, P)$ are considered as robust and form the basis of glimpse selection which is done in two steps for each time step n:

1. The highest peak value of the signal must exceed 0.9. Thus, it is made sure that the signal has enough harmonic content. If this requirement is not fulfilled, no glimpse is generated at time n.
2. If the first requirement is fulfilled, all peaks are selected that exceed the value of 0.8. In doing so, we extract the fundamental period $P0$ and multiples of the fundamental period with a high probability. The output signal of the harmonic feature extraction stage is therefore referred to as $nP0(n)$.

The harmonic features $nP0_{\mathrm{mult}}(n)$ extracted from a typical multitalker signal appearing in this task are shown in Fig. 1 (blue dots) for the frequency band with the center frequency $f_c = 414.2$ Hz.

Step 5: Target template

For the calculation of the target template, the harmonic features of all target configurations are extracted from presentations of the target alone (2 channels x 11 positions). The target template then consists of the set of harmonic glimpses appearing in all of these configurations. The target template $nP0_{\mathrm{tar}}(n)$ for the frequency band with $f_c = 414.2$ Hz is shown in Fig. 1 (black dots).

Step 6: Selection of target-related localization glimpses

For the selection of target-related localization glimpses, a pattern matching of the multitalker harmonic features $nP0_{\mathrm{mult}}(n)$ and the target template $nP0_{\mathrm{tar}}(n)$ is implemented. Both signals are shown in Fig. 1 for a sample simulation with the target at 30° and maskers at -50, -40, -30, and $-20°$ (blue: multitalker harmonic features; black: target template). The harmonic glimpses for each time instance appear as several values $nP0(n)$, where the smallest value usually identifies the fundamental period and the higher values identify multiples of the fundamental period. For each time instance n it is determined if the two patterns match or not: The following requirements have to be applied:

1. Both signals $nP0_{\mathrm{mult}}(n)$ and $nP0_{\mathrm{tar}}(n)$ must not be empty at time n.
2. If both signals $nP0_{\mathrm{mult}}(n)$ and $nP0_{\mathrm{tar}}(n)$ have the same amount of values, all of these values have to coincide in a certain range, i.e. they must not differ by more than 0.12 ms each. In Fig. 1, this is for example the case for times of appr. $0.21 - 0.25$ s.
3. If the signals $P0_{\mathrm{mult}}(n)$ and $nP0_{\mathrm{tar}}(n)$ have a different amount of values, all of the values of the shorter signal have to coincide with values of the longer signal. In Fig. 1, this is the case for times of appr. $0.30 - 0.35$ s.

The time instances where these requirements are applied are identified as gray squares on the time axis of Fig. 1. It is assumed that at these times the target is the dominant source in the multitalker mixture in this particular frequency band.

Finally, the localization glimpses appearing at time instances that the pattern matching procedure identified as target-dominated are selected. The procedure described here is done for each frequency band. The hereby selected localization glimpses are shown in Fig. 2, left panel, as

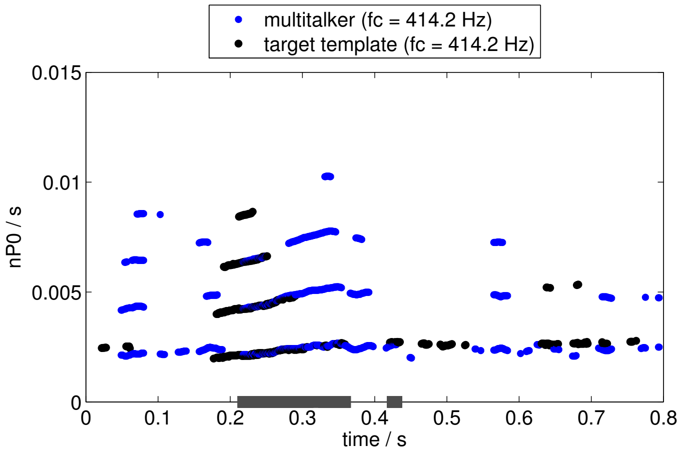

Figure 1: Harmonic features of the multitalker signal $nP0_{\mathrm{mult}}(n)$ (blue) and the target template $nP0_{\mathrm{tar}}(n)$ (black) for the frequency band with center frequency $f_c = 414.2$ Hz. Gray squares on the time axis identify time instances where the two patterns match. It is assumed that at these time instances the target is the dominant source in the multitalker mixture.

highlighted colored glimpses. Different colors identify different frequency bands.

Step 7: Target location estimation

The first step in estimating the target location is the gaussian-kernel-based estimation of the probability density function $p_{\mathrm{sel}}(\alpha)$ across all localization glimpses selected as target-related by the harmonic pattern matching procedure described in step 6. This function is shown by the black curve in the right panel of Fig. 2. If the selection of target related glimpses would have been perfect we would observe a maximum of this function at the target location. However, the selection process generally is not perfect, so a considerable amount of selected glimpses coincide with the masker locations, leading to a maximum of $p_{\mathrm{sel}}(\alpha)$ at the masker locations in many cases. To reduce the dominance of masker locations on the distribution, the probability density function $p_{\mathrm{sel}}(\alpha)$ is divided by the probability density function $p_{\mathrm{nsel}}(\alpha)$ of all not selected glimpses. To be able to vary the contribution of both functions, an exponent ϵ is added to the formula:

$$p(\alpha) = \frac{p^{\epsilon}_{\mathrm{sel}}(\alpha)}{p_{\mathrm{nsel}}(\alpha)} \tag{1}$$

This function, using a value of $\epsilon \approx 1.8$, is shown as a red curve in the right panel of Fig. 2. To avoid an overestimation at peripheral azimuth angles, this function is lowered for angles exceeding 60°. The blue curve in the right panel of Fig. 2 shows the resulting distribution $p_w(\alpha)$. The maximum position of this distribution is selected as the final target location.

Results and discussion

The results of ten model runs for each target location is shown in Fig. 3 for the masker locations [−50° −40° −30° −20°] as magenta circles. Results from the psychoacoustic experiment are also shown in this figure identified by black crosses. The model data for each location vary due to different masker utterances.

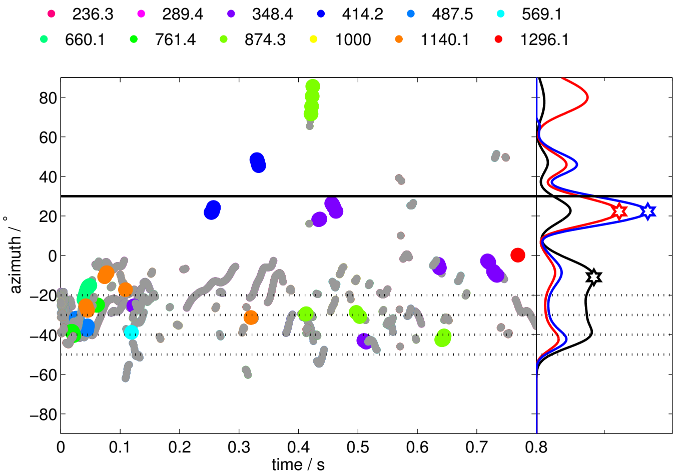

Figure 2: Left panel: Output of the localization model for a multitalker input signal for all frequency bands used in the analysis (center frequencies in Hz identified in the top legend). The horizontal black lines identify target (solid) and masker (dashed) locations. Glimpses selected by the harmonic pattern matching procedure are highlighted in color; not-selected glimpses are shown as small gray dots. Right panel: Different distributions as basis for selecting the target location (identified by the star symbol). The black curve is the probability density function (pdf) $p_{\mathrm{sel}}(\alpha)$ calculated over all selected glimpses. The red curve is the same function raisd to the ϵ-th power, divided by the pdf of all not selected glimpses: $p(\alpha) = \frac{p^{\epsilon}_{\mathrm{sel}}(\alpha)}{p_{\mathrm{nsel}}(\alpha)}$. The blue curve is the same function as $p(\alpha)$, but with application of a function decay at high values of α, leading to the weighted function $p_w(\alpha)$. Target locations are extracted as the position of the maximum of the function $p_w(\alpha)$.

For on-masker positions, the localization ability of the model is similar to that of the subjects. This is due to the fact that the target location distribution is part of the masker location distribution in these trials. Therefore, it does not matter if masker-related glimpses are accidentally selected as target-related. Good localization ability is also observed up to target locations of 20°, and for a target location of 40° where only one of the ten simulations produced a large error.

Variations between model and subjects data can be explained by the fact that BRIRs of a different room were used for the simulations. Also the measurement procedure in the psychoacoustic task, head tracking, could lead to the slightly different values, especially to higher variances in the subjects data. Finally, also inter-subject differences exist for which the model framework does not account for.

For a target location of 30° only four runs, and for 50° only two runs can be considered as successful. At these positions, probably too many glimpses at masker locations are selected as target-related. The contribution of these incorrectly selected glimpses is even so high that the reduction of masker dominance as formulated in Eq. 1 does not yield suitable results.

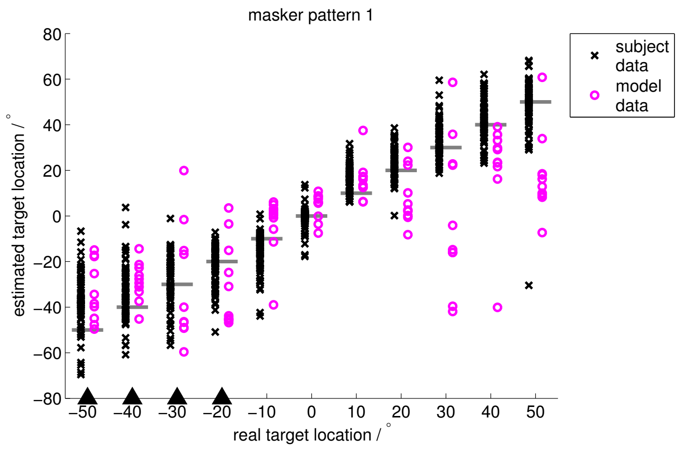

Figure 3: Results of the psychoacoustic experiment (each black cross represents one response of one subject for a given target location; 7 subjects participated in the study, each performing 10 trials per target location and masker pattern) and the model runs (magenta circles) for all runs. Black triangles identify the positions of the maskers. Short gray horizontal lines identify the case of perfect localization.

Conclusions

A model framework was proposed that was partially able to solve the task of localizing a female target talker presented along four male talkers which temporally fully overlapped the female target. Despite the difficulty of this task, the model results coincide with the psychoacoustic results in most of the cases. In particular, the following conclusions can be made:

- The proposed binaural and harmonic auditory models deliver the glimpses that are necessary to perform the task. The concept of "glimpsing" is crucial here to avoid using non-robust information.
- Since some glimpses only last for a few milliseconds, it can be assumed that a high temporal resolution is necessary to extract the glimpses related to the target.
- The processing in different frequency bands is necessary, because on a purely temporal domain, the target talker is completely temporally overlapped by maskers. However, the analysis showed that there seem to be robust target-related spectro-temporal glimpses.
- Since there are still many masker-related localization glimpses identified as target-related, the proposed method to identify the target-dominant time-frequency regions is not yet totally successful. This reflects the limited resolution of the harmonicity analysis, which leads to errors in selecting target-related spectro-temporal glimpses. A possible refinement of the method by including explicit speech formant tracking could possibly alleviate this problem.

The results shown here are obtained for just one of the masker patterns used in the psychoacoustic task. In order to get stronger conclusions more extended simulations are needed.

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, SFB/TRR 31) and the PhD program "Hearing".

References

[1] Cooke, M.: A glimpsing model of speech perception in noise.
The Journal of the Acoustical Society of America **119** (2006), pp. 1562–1573

[2] Dietz, M., Ewert, S., and Hohmann, V.: Auditory model based direction estimation of concurrent speakers from binaural signals.
Speech Communication **53** (2011), pp. 592–605

[3] Ewert, S. D., Iben, C. T., and Hohmann, V.: Robust fundamental frequency estimation in an auditory model.
AIA-DAGA 2013, Meran, 2013

[4] Hohmann, V.: Verfahren zur Extraktion periodischer Signalkomponenten und Vorrichtung hierzu.
Patent (2006) Germany 102004045097.

[5] Kopčo, N., Best, V., and Carlile, S.: Speech localization in a multitalker mixture.
The Journal of the Acoustical Society of America **127** (2010), pp. 1450–1457