Modeling localization and word recognition in a multitalker setting

Angela Josupeit, Volker Hohmann

CvO Universität Oldenburg, Medizinizsche Physik, Cluster of Excellence Hearing4all, Email: angela.josupeit@uni-oldenburg.de

Introduction and objective

The human auditory system is able to attend to one particular sound source, even if other sources (background noise, competing talkers, reverberation) are present. In particular, this holds for situations with multiple talkers, which is why this phenomenon is termed the “Cocktail Party Effect”. The task of the listener can be divided into certain subtasks. This includes the identification of the talker of interest (“target talker”), where we can imagine situations where we hear a key word, e.g. our own name, and then try to focus on this talker. Once identified, it is important to localize or track the target talker. Finally, the task is to understand what this person is saying, even of other talkers are present at the same time.

The present study proposes a model framework for solving these tasks. To evaluate the model, we test it in an experimental procedure based on a psychoacoustical study [1]. In this procedure, two to four speech streams are presented to the listener. The speech streams are male talkers, each uttering a sentence from the “Coordinate Response Measure” (CRM) corpus [2], starting simultaneously. The corpus consists of sentences following a common structure: “Ready (call-sign) got to (color) (number) now”. The task of the listener was to recognize the color and number word of the talker that uttered the call-sign “Baron”. Speech streams were presented in different spatial configurations: two talkers close ($[-5°, 5°]$), two talkers far ($[-60°, 60°]$), three talkers close ($[-15°, 0°, 15°]$). three talkers far ($[-60°, 0°, 60°]$), and four talkers ($[-60°, -20°, 20°, 60°]$). The task was done in an anechoic setting. The corpus contained 8 call-signs, 4 color words, and 8 number words.

Model

The model consists of 2 steps: (1) Identification and localization the target talker, based on the word “Baron”. (2) Recognizing the color and number word uttered by the target talker, based on the previously identified identity and location of the target talker.

The first step is similar to previous studies of localizing a target talker in a multitalker setting on the basis of a known target utterance [5, 6, 7]. In that studies, a harmonicity template matching approach was used. Because the corpus used in the present study is also limited to a few words and talkers, we use a similar harmonicity template matching approach for both steps, including the same methods for feature extraction. In the following, the model implementation for the two steps is described in detail.

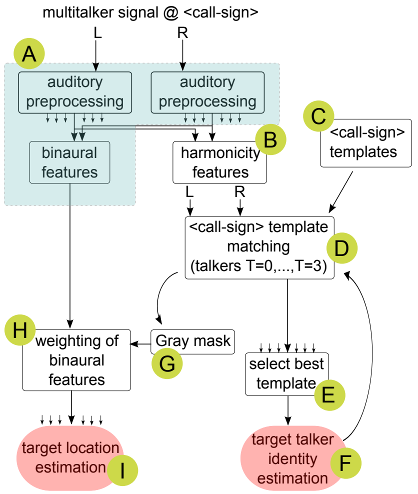
multitalker signal @ <call-sign>
L
R
A
auditory preprocessing
auditory preprocessing
binaural features
harmonicity features
B
C
<call-sign> templates
L
R
<call-sign> template matching (talkers T=0,...,T=3)
D
H
weighting of binaural features
Gray mask
G
select best template
E
target location estimation
I
target talker identity estimation
F

Fig. 1: Model outline for step 1: Target talker identification and localization.

Step 1: Target identification and localization

Fig. 1 shows the model outline for the first step. First, the multitalker signal was preprocessed and binaural features are extracted in each frequency channel f_c using a binaural model [3] (A). Harmonicity features were extracted from the same preprocessed signals using a “synchrogram” method [4] (B). In this extraction, energy criteria were applied to assure that only robust features were extracted, i.e. those features that likely belong to a single focused sound source.

The procedure for the generation of call-sign (“Baron”) templates (C) is illustrated in fig. 2: First, harmonicity features were extracted from all “Baron” words in the speech corpus for one specific talker (32 words), see left panel of fig. 2. Second, for each time step, histograms were calculated based on the harmonicity feature values, see right panel of fig. 2. The histogram bin width was chosen as $P_c/10$ where $P_c = 1/f_c$ is the center period of the auditory filter. These histograms form the templates and are termed as $H(P, t, f_c, T)$ in the following, where P, t, f_c, and T denote period, time, center frequency of the auditory channel and target talker identity, respectively.

The template matching procedure (D) compares the templates with the extracted harmonicity features of the multitalker input signal (B). As a measure “how good” a template fits to the multitalker harmonicity features, the histogram height at the harmonicity values $P_k(t, f_c)$ of the multitalker signal was used as a measure. The procedure is illustrated in fig. 3. First, for the left and right channel, the mean value across harmonicity values was calculated:

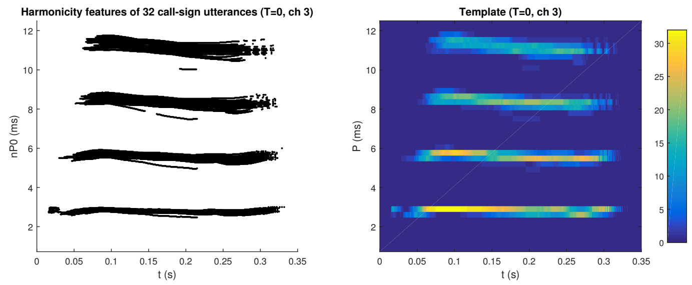

Fig. 2: Generation of call-sign ("Baron") templates, shown here for talker $T = 0$ and frequency channel $f_c = 348.4$ Hz. Left: Extracted harmonicity features of all uttered clean call-signs (32 total) for this T and f_c. Right: Generated template $H(P, t, f_c = 348.4 \text{ Hz}, T = 0)$.

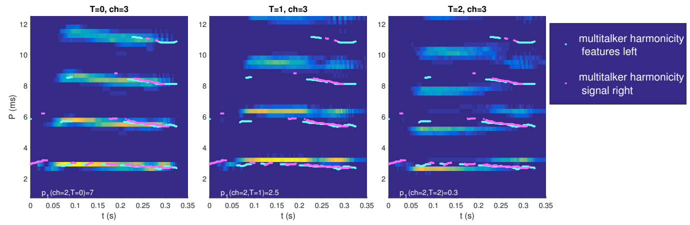

Fig. 3: Template matching procedure. The numbers in the lower left corners indicate the mean of $p_1(t, f_c, T)$ across time t, for frequency channel $f_c = 348.4$ Hz and the respective talkers T. Note that $T = 0$ (left plot) is the real target talker in this scene.

$$p_0(t, f_c, T) = \frac{1}{N_P(t, f_c)} \sum_{k=1}^{N_P(t,f_c)} H(P_k(t, f_c), t, f_c, T),$$

with $N_P(t, f_c)$: number of extracted harmonicity feature values in a t-f_c bin. The procedure actually integrates different time shifts between template and multitalker features; due to simplicity reasons, this step is not described in detail here. Subsequently, left and right channel were combined:

$$p_1(t, f_c, T) = \sqrt{p_{0,L} \cdot p_{0,R}},$$

and finally as a measure of "how good" a template of a certain talker T fits:

$$p(T) = \frac{1}{N_t} \frac{1}{N_{f_c}} \sum_{i=1}^{N_t} \sum_{j=1}^{N_{f_c}} p_1\left(t_i, f_{c,i}, T\right).$$

The estimated target talker (F) was than calculated as the best fitting template (E):

$$T_{\text{est}} = \underset{T}{\operatorname{argmax}} \left(p(T)\right).$$

For the location estimation, the binaural features of the multitalker signal were weighted using a gray mask (G) that was defined by

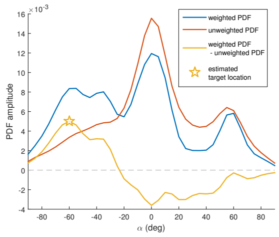

Fig. 4: Estimation of target location for one sample run. The target talker is located at −60°, and the masker talkers at 0° and 60°, respectively. Motivated by previous localization modeling studies [6, 7], the target location is estimated based on the difference function of the weighted PDF and the unweighted PDF. This reduces the influence of masker locations and strengthens the distinctness of the target location.

$$\text{GM}_{\text{est},1}(t, f_c) = p_1(t, f_c, T_{\text{est}}).$$

The binaural features were then integrated across time and frequency using a Gaussian kernel based probability density function (PDF) estimate. PDFs for weighted and unweighted binaural features are shown in fig. 4 for one sample run. It is visible that the target position at −60° is more pronounced for the weighted PDF than at the unweighted PDF. To account for this difference between the weighted and unweighted PDF, the difference function

$$f(\alpha) = f_{\text{weighted}}(\alpha) - f_{\text{unweighted}}(\alpha)$$

was used as a basis for the estimation of target location [6, 7]

$$\alpha_{\text{est}} = \underset{\alpha}{\operatorname{argmax}} \left(f(\alpha)\right).$$

Step 2: Recognition of color and number word

The model outline for step 2 is illustrated in fig. 5, shown here for the color word recognition as an example. For this task, the following a priori information is available:

1. Target talker location, resp. the function $f(\alpha)$
2. Target talker identity: T_{est}

A gray mask was estimated on the basis of "how good" the multitalker binaural features fit to the previously estimated location (J):

$$\text{GM}_{\text{est},2}(t, f_c) = f^* \left(\alpha(t, f_c)\right).$$

f^* denotes a scaled version of f so that the minimum is zero instead of a negative value. The gray mask was used to weigh the harmonicity features (K, in the actual implementation, this step is done after the template matching (L)).
The template generation (C) and template matching procedure (L) in step 2 was similar to the one in step

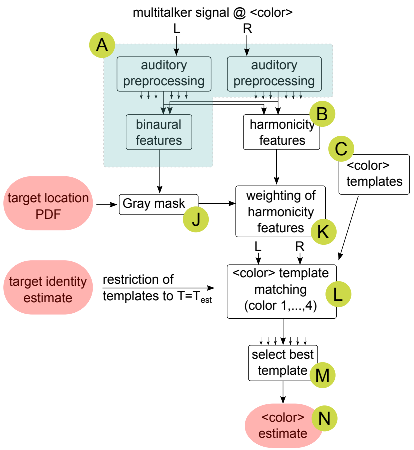

Fig. 5: Model outline for step 2: Color word recognition.

1. However, here we only used the color word templates for talker T_{est}, resulting in a set of templates differing in color (index c): $H(P, t, f_c, c)$. The measures $p_0(t, f_c, c)$ and $p_1(t, f_c, c)$ were calculated analogue to step 1 (D). The calculation of $p(c)$ is done as follows:

$$p(c) = \frac{1}{N_t} \frac{1}{N_{f_c}} \sum_{i=1}^{N_t} \sum_{j=1}^{N_{f_c}} p_1\left(t_i, f_{c,i}, c\right) \cdot \mathrm{GM}_{\text{est},2}(t, f_c).$$

The multiplication with $\mathrm{GM}_{\text{est},2}(t, f_c)$ makes sure that target-related t-f_c bins are strengthened. It is an equivalent for weighting the harmonicity features (K).

The color estimation (N) is based on the “best fitting” template (M):

$$c_{\text{est}} = \underset{c}{\operatorname{argmax}} \left(p(c)\right).$$

Results

The model performed 100 runs for each spatial configuration. Fig. 6 shows the results for step 1, target talker identification and localization. Green bars show the performance of the default model version which contained no ideal *a priori* knowledge other than the four “Baron” templates. The light blue bars show the results when optimal knowledge about the target talker location resp. identity was used. That is, for target talker location estimation, we used the actual target talker T_{real} instead of T_{est} for the estimation of the gray mask $\mathrm{GM}_{\text{est},1}(t, f_c)$. For target talker identification, we applied an additional weighting to the template matching procedure which was calculated similar to the gray mask in step 2 (J), based on the real target location. Dark blue bars show the results for using Ideal Gray Masks (IGMs), calculated as the SNR in each t-f_c band based on the spectral energy signals of target and masker alone. For target talker location estimation, the IGM was used instead of $\mathrm{GM}_{\text{est},1}(t, f_c)$ to weigh binaural features. For target talker identification, the additional weighting of the template matching procedure was done using the IGM.

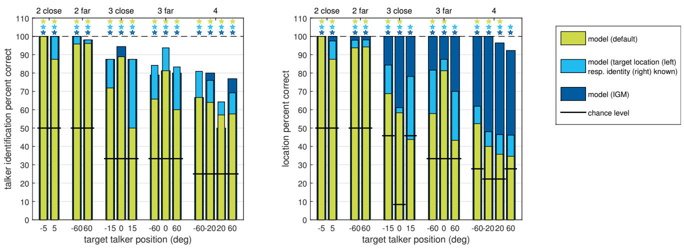

Fig. 6: Results for step 1: Percent correct scores for target identification (left panel) and localization (right panel; a location is considered to be correct if it is closer to the target position than to any masker position). Different colors indicate model results for different amounts of *a priori* knowledge. Green: Results for the default model version as described earlier (no ideal *a priori* knowledge). Light blue: For target identity estimation (left), optimal knowledge about the target location was used; for target location estimation (right), optimal knowledge about the target identity was used. Dark blue: Ideal Information about the target and masker in isolation, i.e. Ideal Gray Masks (IGMs) were used as a priori knowledge. Horizontal black lines show the chance levels; star symbols on top of the figure indicate whether the model results are significantly higher than chance level (binomial test with $\alpha = 0.05$ significance level).

Results show that the target identification using the default model is always significantly above chance level. Performance is degraded with increasing number of talkers, and is generally better for center positions in the three talker conditions. Using an additional loop with a weighting of harmonicity features according to the correct location further improves the results. Interestingly, almost no difference is observed whether this weighting is done using the estimated GM or the IGM.

The target location estimation for the default model is very accurate for the two talker conditions. As seen before, performance degrades for an increasing number of talkers. For the center positions in the three talker conditions, the performance is clearly above chance level; however, for the flanking positions performance degrades, especially for positive angles. The same is seen in the four talker condition. If optimal knowledge about the target talker identity is available, the model performance increases; in this case, all estimations are above chance level. If the weighting of binaural features is done using the IGM, the performance is nearly perfect for all conditions, with some slight degradations in the four talker condition.

Fig. 7 shows the results for step 2, here only color word recognition. Green bars show the performance of the default model version using the previously estimated target identity and location as priors. Light blue bars show the results if the correct target talker and the correct location function is used. Dark blue bars show the results if the IGM is used to weigh the harmonicity features.

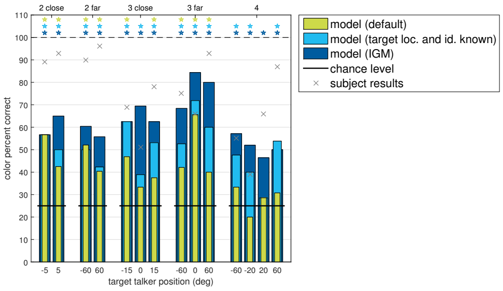

Fig. 7: Results for step 2: Percent correct scores for color recognition. Different colors show results for different amounts of a priori knowledge. Green: default model version. Light blue: Optimal knowledge about the target talker identity and its location. Dark blue: Weighting of harmonicity features using the IGM. Gray crosses identify the subject data for color *and* number both correct [1], serving as a rough comparison.

Results show that the performance of the default model is above chance level for the two and, with one exception, three talker conditions, and at chance level for the four talker condition. Even though the model performance is above chance level for some conditions, the percent correct scores are, in most cases, not as good as the subject scores in the original psychoacoustic study [1]. Optimal knowledge about the target talker identity and its location improves performance, especially for the three and four talker conditions. Presumably, this difference might be due to the fact that in these conditions the target talker identification and localization was not accurate initially (see results of step 1). Using the IGM for weighting of harmonicity features further improves the results, especially for the three talker conditions. Generally, the model performance is still lower than the subject performance, especially for the two talker conditions. However, for the center position in the three talker condition, and for the left positions in the four talker conditions, the performance is in the range of the subject performance. It needs to be noted that the illustrated subject performance refers to the color *and* number correct scores.

Summary and conclusions

1. The model is able to identify the target talker identity using a harmonicity template matching procedure based on the call-sign utterances for every talker. This implies that presumably harmonicity features are crucial for the ability to differentiate between different talkers, even in a difficult situation where only male talkers have to be distinguished. The identification of the target talker improves when the correct direction is known.

2. The model is able to localize the target talker for two talker scenarios, and to a certain extend for conditions with more talkers. Although some errors arise due to wrong information about the target talker identity, the larger amount of the errors can be explained by differences of the estimated GM from the IGM. Given that the localization performance is nearly perfect for the binaural feature selection based on the IGM, it can be assumed that the binaural features are not a primary source of error.

3. The color and number word recognition performance of the model is generally above chance level for two and three talker scenarios; however, especially for the two talker conditions it is still clearly below subject performance. Optimal knowledge about target identity and location, or the usage of the IGM, improves the results to a certain extend, but not enough to reach subject performance in all conditions. This implies that the harmonicity template matching procedure is not sufficient to predict psychoacoustic results. It can be assumed that further features are needed that allow for the differentiation of other speech features such as resonances of the vocal tract.

Acknowledgements

This work was supported by DFG SFB/TRR 31 “The Active Auditory System”

Literatur

[1] D. S. Brungart and B. D. Simpson, “Cocktail party listening in a dynamic multitalker environment,” *Perception & Psychophysics*, vol. 69(1), pp. 79–91, 2007.

[2] R. S. Bolia, W. T. Nelson, M. A. Ericson, and B. D. Simpson, “A speech corpus for multitalker communications research,” *The Journal of the Acoustical Society of America*, vol. 107(2), pp. 1065–1066, 2000.

[3] M. Dietz, S. D. Ewert, and V. Hohmann, “Auditory model based direction estimation of concurrent speakers from binaural signals,” *Speech Communication*, vol. 53(5), pp. 592–605, 2011.

[4] S. D. Ewert, C. Iben, and V. Hohmann, “Robust fundamental frequency estimation in an auditory model,” *Proceedings of the International Conference on Acoustics AIA-DAGA 2013, Deutsche Gesellschaft für Akustik e.V., Berlin*, pp. 271–274, 2013.

[5] A. Josupeit, S. van de Par, N. Kopco, V. Hohmann, “Modeling of speech localization in a multitalker environment using binaural and harmonic cues,” *Proceedings of the International Conference on Acoustics AIA-DAGA 2013, Deutsche Gesellschaft für Akustik e.V., Berlin*, pp. 724-727, 2013.

[6] A. Josupeit and V. Hohmann, “Sound localization in complex multitalker conditions by harmonic template matching,” *Fortschritte der Akustik - DAGA 2014, DEGA e.V., Berlin, Oldenburg*, pp. 353-354, 2014.

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