# The effect of spectro-temporal integration in a probabilistic model for robust acoustic localization

Tobias May[1], Steven van de Par[2], and Armin Kohlrausch[1,2]

[1] *Eindhoven University of Technology, Department of Human Technology Interaction, P.O. Box 513, NL-5600 MB Eindhoven, The Netherlands*

[2] *Philips Research, High Tech Campus 36, NL-5656 AE Eindhoven, The Netherlands*

A robust acoustic localization model will be presented, which is based on the supervised learning of azimuth-dependent binaural feature maps consisting of interaural time differences (ITD) and interaural level differences (ILD). Motivated by the robust localization performance of the human auditory system, the associated peripheral stage is used in this study as a front-end for binaural cue extraction. Multi-conditional training is performed to take into account the variability of the binaural features which results from the combination of multiple sources, the effect of reverberation and changes in the source/receiver configuration. One way of accumulating evidence of possible sound source locations is to combine information across auditory channels. Alternatively, integrating evidence across groups of time-frequency (T-F) units, so called *fragments*, which are believed to belong to a single source, was reported to significantly improve ITD-based localization performance [Christensen *et al*., Proc. of Interspeech, 2769-2772 (2007)]. Instead of accumulating the localization cue directly, the proposed model combines likelihoods, taking into account the uncertainty which is associated with the azimuth estimate of a particular T-F unit. Various procedures of controlling the spectro-temporal integration will be discussed and the influence on sound source localization will be presented.

## INTRODUCTION

The human auditory system is able to identify and localize acoustic objects in adverse acoustic conditions. Regarding speech perception and localization tasks, the robustness of the human auditory system is superior to computer algorithms that have access to the information also available to the human auditory system (binaural signal). Bregman's auditory scene analysis (ASA) is an approach to describe how the human auditory system derives a description of complex acoustic scenes (Bregman, 1990). First, the auditory input is segregated into fragments, representing groups of T-F units which are dominated by a single source. In the second step, fragments which correspond to the same acoustic source are grouped together to form an acoustic description of the sources present in the scene. In order to perform this higher level analysis of complex acoustic scenes, a front-end is required to partition the T-F plane into groups of coherent T-F units, which are believed to be dominated by one source.

Proceedings of ISAAR 2009: Binaural Processing and Spatial Hearing. 2nd International Symposium on Auditory and Audiological Research. August 2009, Elsinore, Denmark. Edited by J. M. Buchholz, T. Dau, J. Cristensen-Dalsgaard, and T. Poulsen. ISBN: 87-990013-2-2. The Danavox Jubilee Foundation, 2010.

The segregation of multiple sources based on localization cues was shown to lead to high performance in anechoic conditions (Roman *et al*., 2003). However, the accuracy of localization information based on individual T-F units rapidly decreases with increasing reverberation time. In order to increase the reliability of the localization estimate, information could be grouped and integrated across T-F units that are dominated by the same source. Recently, a two-stage model for exploiting spatial cues across groups of T-F units was proposed to improve frame-based localization in reverberation (Christensen *et al*., 2007, 2008, 2009). First, a pitch tracking algorithm was employed to group T-F units together to so called *fragments*, according to common pitch information. Those fragments of consistent pitch information were used in the second stage to integrate the ITD information across the corresponding spectro-temporal regions.

The current paper presents a model to create T-F-based localization maps of multi-source scenarios in reverberation, by integrating evidence of sound source position across groups of T-F units. There is strong evidence that the formation of auditory objects is not primarily driven by spatial cues (Darwin, 2008). Nevertheless, it is interesting to explore whether in realistic acoustic settings with reverberation, there is enough localization information to derive fragments based on common spatial information. Compared to the pitch-based fragment generation, a method based on spatial cues would work for arbitrary acoustic signals (voiced/unvoiced/noise). In contrast to accumulating spatial cues as proposed by Christensen and colleagues, in this study the evidence will be accumulated by integrating likelihoods of sound source locations across fragments. In this way, the uncertainty of localization estimates of individual T-F units is taken into account. With respect to frame-based localization performance, it was shown that a probabilistic integration of likelihoods is superior to integrating the localization cue across frequency (May *et al*., 2009a).

## MODEL ARCHITECTURE

The model that we propose to derive the localization map consists of four major building blocks, which are shown in Fig. 1. The model requires a binaural input, consisting of sources at fixed spatial locations. Given this binaural input, the proposed model first computes the likelihood of sound source azimuth for individual T-F units. Secondly, T-F units are grouped together to fragments based on consistency in localization information. Those fragments are used in the third block to accumulate likelihoods across the corresponding spectro-temporal regions of each fragment. Fourthly, the modified likelihood map is transformed into the final localization map. Each of the four building blocks is explained in detail in the following sections. The description is supported by visualizing outputs of various processing steps involved in estimating the localization map of a two-source mixture in reverberation ($T_{60}$ = 0.39 s).

### Spatial likelihood map

In previous work we have presented a probabilistic model for sound source localization

that can be used for the computation of a spatial likelihood map (May *et al*., 2009a,b). In this model, the acoustic signal is decomposed into 64 frequency channels

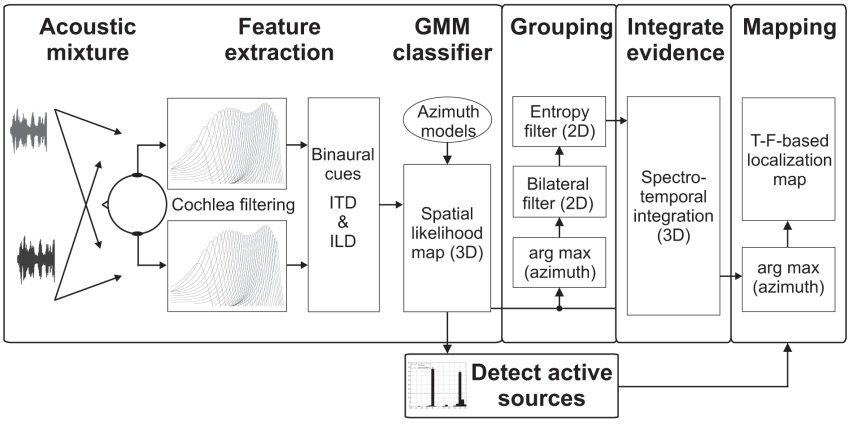


**Fig. 1**: Model architecture for creating a T-F-based localization map. See text for details.

using a fourth-order gammatone filterbank. Spatial ITD and ILD cues are extracted independently for each frequency channel using a time window of 20 ms with a 10 ms frame shift. The joint distribution of both spatial cues resulting from multiple sources and reverberation is learned by a classifier for a range of 21 azimuth positions between ±50° in steps of 5°. Given a binaural mixture, the model computes a 3D likelihood map, which represents the sound source probability for each time instant, frequency and azimuth position.

The aim of the final localization map is to group and label the T-F representation of multi-source mixtures according to azimuth information of the active sources. In order to achieve this, the number and the azimuth of active sources is required. Figure 2 shows the estimated azimuth map before grouping and spectro-temporal integration for an acoustic mixture consisting of two male speakers in reverberation ($T_{60}$ = 0.39 s). In panel (a), the most likely source position for individual T-F units is presented. Although dominant azimuth locations seem to be detectable by visual inspection, an automatic detection of active sources using the T-F-based localization information is a nontrivial task. Integrating evidence across frequency and estimating the most dominant source per frame leads to a good frame-by-frame indicator of the azimuth of the most dominant source (see panel (b)). Using these frame-based localization estimates, an efficient histogram technique can be utilized to detect active sources and the corresponding azimuth positions by applying a threshold criterion (May *et al*., 2009b), as shown in panel (c). Note that the histogram analysis is performed on a sentence basis for each acoustic mixture.

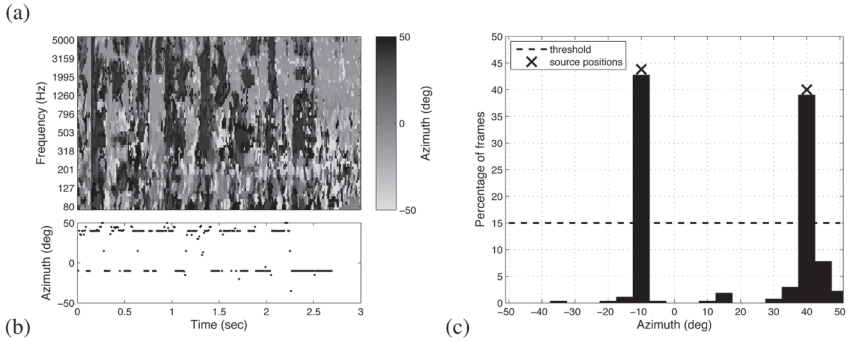


**Fig. 2**: Estimated localization information for an acoustic mixture consisting of 2 male speakers (at 40° and -10° azimuth) in reverberant conditions ($T_{60}$ = 0.39 s) based on: (a) individual T-F units and (b) time frames after integrating evidence across frequency channels. As illustrated in (c), a histogram of the frame-based localization estimates can be used to detect the number of active sources in the mixture and the corresponding azimuth positions.

### Grouping of consistent localization information across T-F units

In this study, consistent localization information is grouped across T-F units leading to fragments. For this purpose, the 3D likelihood map is transformed into a preliminary 2D localization map by applying a leaky integrator to the likelihood map across frames and selecting the most likely source position for each T-F unit. The leaky integrator is intended to reduce fluctuations of the azimuth information. The time constant $\tau$ = 30 ms is adopted from Christensen *et al.* (2007) to match the average length of speech fragments in multi-source scenarios. The preliminary localization map is shown in panel (a) of Fig. 3.

In the next step, localization information is averaged across neighboring T-F units by applying a bilateral filter (Tomasi and Manduchi, 1998). The output of the bilateral filter is a weighted sum of the azimuth of neighboring T-F units, where the weighting function is data-dependent. The weighting function is a multiplication of two Gaussians. The first one decreases with the distance of neighboring T-F units to the center T-F units, putting more emphasis on the T-F units that are in the proximity of the center unit. The second weighting function decreases with increasing azimuth difference to the azimuth of the center T-F unit, using predominantly T-F units with similar azimuth values for the averaging. The latter weight provides the edge preserving property of the bilateral filter. The width of the Gaussian functions are specified by two parameters, $\sigma_{\mathrm{TF}}$ and $\sigma_{\varphi}$, controlling the decrease of the Gaussian weight in the T-F domain and in the azimuth domain, respectively. To ensure that the edge-preserving property of the bilateral filter is maintained for arbitrary source configurations, the decrease of the Gaussian weight along the azimuth domain $\sigma_{\varphi}$ is adjusted by using the estimated source positions described in the previous section.

The minimum spatial distance between all detected sources is taken as the change in azimuth which should be preserved by the bilateral filter. Therefore, $\sigma_\varphi$ is changing linearly as a function of the minimum source distance. In this way, the smoothing of the bilateral filter is stronger if acoustic sources are spatially further apart, whereas less smoothing is applied if sources are spatially close to one another. The effect of the bilateral filter can be seen in Fig. 3, by comparing the preliminary localization map in panel (a) with the output of the bilateral filter in panel (b). Although groups of connected T-F units which correspond to either of the two source positions can be visually identified in panel (a), the azimuth within those areas is quite noisy and fluctuates around the true azimuth position. The localization information after bilateral filtering is consistent across large areas of T-F units and small azimuth deviations are smoothed, whereas boundaries between two different source positions are maintained.

After bilateral filtering, fragments of consistent localization information are extracted to complete the grouping procedure. To achieve this, the variability of azimuth information is computed over a window of adjacent T-F units. More specifically, the variability is measured by calculating the entropy of the azimuth values across a window of 3 x 3 neighboring T-F units. Panel (c) of Fig. 3 shows the entropy map of the localization information presented in panel (b). The entropy map is close to zero if the azimuth is consistent across the analysis window (low variability) and approaches 1 if the azimuth is completely random. Hence, the entropy filter can be used to effectively group T-F units and to extract the boundaries between areas of consistent localization information. An experimentally determined threshold of $\theta_e = 0.3$ is used to transform the entropy map into a binary image, labeling consistent and inconsistent T-F units with 1 and 0, respectively. T-F units labeled with 0 are considered as background, and are therefore not considered for the fragment creation. T-F units labeled with 1 are grouped across time and frequency to obtain the final fragments. A pair of T-F units is considered to belong to the same fragment, if the two spectro-temporal units are connected either horizontally or vertically (4-neighbors connectivity).

**Spectro-temporal integration**

Based on the extracted fragments, the spectro-temporal integration accumulates evidence of source locations across all corresponding T-F units. For each fragment, the likelihoods corresponding to a specific azimuth are multiplied to obtain an estimate of the likelihood that the source is at this specific azimuth. This 2D integration is done for all 21 azimuth positions of the likelihood map independently. Various weighting schemes are discussed in Christensen *et al*. (2009) to control the contribution of each T-F unit to the overall average of the fragment. But because the likelihood map already reflects the uncertainty which is associated with local T-F units, a uniform combination of likelihoods seems to be optimal.

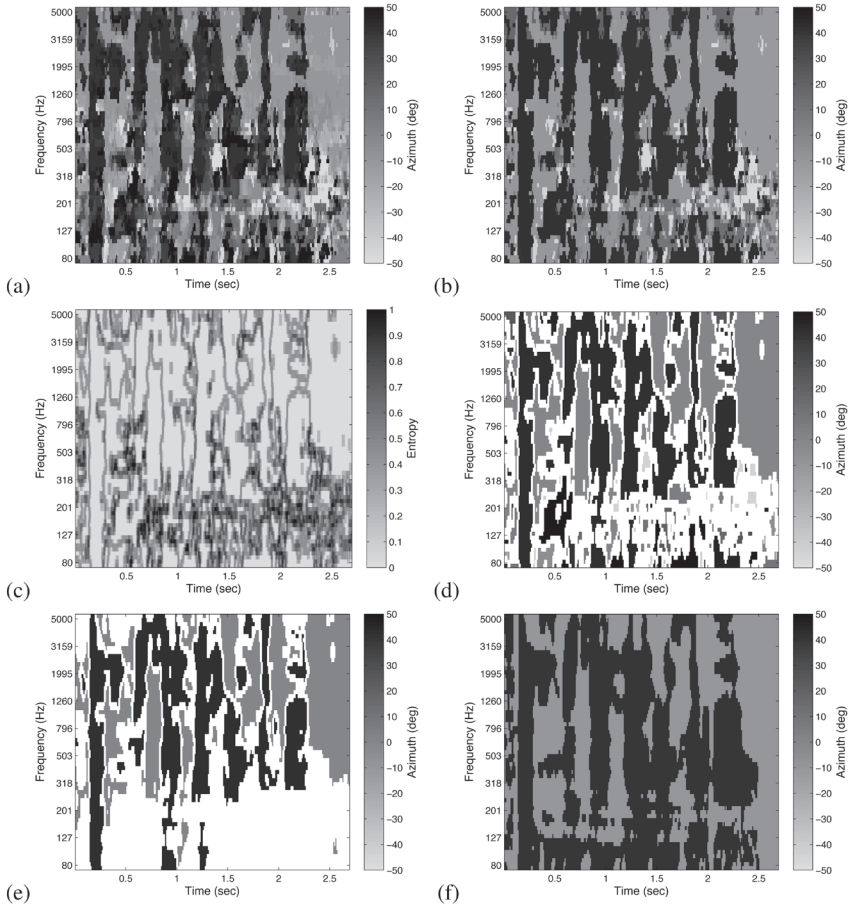


**Fig. 3**: Process of creating a localization map of an acoustic mixture consisting of 2 male speakers (at 40° and -10° azimuth) in reverberant conditions ($T_{60}$ = 0.39 s): (a) Preliminary localization map, (b) Localization map after bilateral filter, (c) Entropy map of filtered localization map, (d) Localizing groups of T-F units, (e) Final localization map and (f) Ideal grouping based on a priori SNR. White color indicates the background.

## Mapping

After performing spectro-temporal integration, a 2D localization map is formed by estimating the most likely source location for each fragment. The azimuth labeling is done only for those T-F units which do correspond to a fragment. The remaining T-F units are labeled as background. This process can be seen in panel (d) of Fig. 3. While the estimated azimuth of most fragments correspond to either of the real sound sources at 40° and -10°, some fragments in lower frequency channels point to different azimuth positions.

After assigning a localization label for each fragment in the T-F plane, a post-processing is performed to remove T-F units from the localization map which do not correspond to one of the active sound source positions. The detection of source activity is based on the histogram technique described in the spatial likelihood map section. The final localization map after postprocessing is depicted in Fig. 3, panel (e). For comparison, the ideal localization map based on *a priori* SNR between both sources is presented in panel (f).

## EVALUATION SETUP

### Acoustic conditions

The performance of the proposed system was evaluated in simulated two-source scenarios. Binaural mixtures were generated by convolving anechoic speech files with simulated binaural room impulse responses (BRIR). The speech files were randomly selected from the TIMIT database (Garofolo *et al*., 1993). The BRIRs were synthesized for a room of dimensions 5.1 x 7.1 x 3 meter by using the room simulation package (Campbell *et al*., 2005). The preset *Acoustic plaster* was selected in the software package to model a frequency-dependent absorption characteristic of the room. Different sets of absorption coefficients were used to realize mild-to-strong reverberation. Each binaural mixture consisted of two, fully overlapping speech files of different speakers which were mixed at a signal-to-noise ratio (SNR) of 0 dB, as defined after spatialization. The probabilistic model was trained to recognize locations between ±50° in steps of 5°, resulting in 21 possible source positions. 21 mixtures consisting of two sources were created by randomly selecting all 21 source positions, but the minimum sources distance was constrained to be at least 10°. All 21 mixtures were created and presented five times, leading to a total of 105 mixtures for each reverberation condition.

### Performance evaluation

The accuracy of the estimated localization map is evaluated by comparing it to the ideal binary mask (IBM) (Wang, 2005), which represents the ideal grouping of T-F units based on the a priori SNR of the target and interfering source. Because this work focuses on competing sources rather then separating a target from interfering sources, the IBM was modified to take the azimuth of the dominant source per T-F unit. A T-F unit is considered to be correct, if the estimated azimuth corresponds to the azimuth of the IBM. Thus, the percentage of localization errors is computed by dividing the number of correctly identified T-F units by the number of T-F units in the estimated localization map.

The estimated localization map is not fully occupied, but only contains information which is believed to belong to one of the detected sources. The remaining T-F units are labeled as background. Hence, an additional metric is used to measure the amount of available information, expressed in percentage of T-F units. Ideally, the percentage of errors should be minimized while maximizing the percentage of

available information. But there is an interdependency between both measures. At a certain amount of information, a further increase is only possible at the cost of increasing the localization errors.

### Algorithms

The performance of the proposed method is compared against two baseline systems. The first system *T-F units* produced localization maps on the basis of individual T-F units without any grouping and integration. The second system *Leaky* is based on the preliminary localization map. Compared to *T-F units*, it explores the time-dependent development of sound source likelihoods by applying a leaky integrator across frames. The third system *Fragments* used the proposed spectro-temporal integration by accumulating evidence across fragments. Note that the post-processing described in the mapping section is applied to the localization maps of all methods.

## EXPERIMENTAL RESULTS

Panel (a) of Fig. 4 shows the error rate of the estimated localization maps for acoustic mixtures consisting of two sources in reverberation. The first method *T-F units*, which computed localization maps based on individual T-F units, achieved an error rate of 5.7% in anechoic conditions. However, performance rapidly deteriorated with reverberation time, showing that the estimated localization information based on individual T-F units is not robust in adverse acoustic conditions. Although significantly lower error rates were accomplished by the *Leaky* system, the general dependency of localization accuracy on reverberation time remained. The proposed method *Fragments*, which performed spectro-temporal integration based on fragments, outperformed both baseline systems in all reverberation conditions. Furthermore, the localization error per T-F unit was almost independent of the reverberation time, ranging from 4.5% in anechoic conditions to 5.5% in strong reverberation ($T_{60} = 0.69$ s). This demonstrates the benefit of grouping and integrating localization information across groups of T-F units.

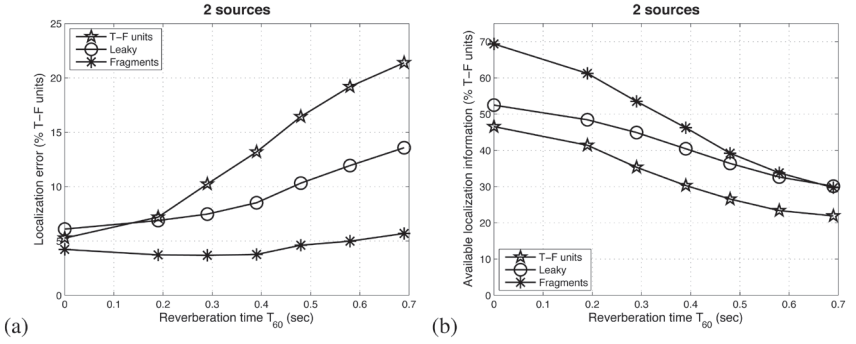


**Fig. 4**: The accuracy of the estimated localization maps for two-source mixtures, expressed in (a) percentage of localization errors and (b) percentage of available localization information per T-F unit.

The percentage of available localization information is presented in panel (b) of Fig. 4. With increasing reverberation time, the estimated localization information becomes more noisy and therefore, the percentage of T-F units which correspond to one of the detected source positions decreases. However, especially for reverberation times below $T_{60} = 0.4$ s, the localization maps produced by the *Fragments* system contained significantly more information compared to the two baseline systems.

## CONCLUSIONS

This paper presented a method to create T-F-based localization maps for multi-source mixtures. By accumulating sound source evidence across groups of T-F units which are believed to be dominated by one source, the estimated localization information was shown to be robust in simulated, adverse acoustic conditions. The new method for extracting fragments which was proposed, is based on common spatial information. Compared to pitch-based grouping, no explicit assumption about the sound source is required, which makes the method generally applicable to arbitrary acoustic scenarios. Future work will investigate the use of primitive grouping principles, e.g. onset and offset analysis (Hu and Wang, 2007), to further enhance the fragment generation process. In addition, the estimated localization map will be used as a front-end for higher order analysis of complex acoustic scenes.

## REFERENCES

Bregman, A. S. (**1990**). *Auditory Scene Analysis* (MIT Press, Cambridge, MA).

Campbell, D. R., Palomäki, K. J., and Brown, G. (**2005**). "A MATLAB simulation of "shoebox" room acoustics for use in research and teaching," Computing and Information Systems **9**, 48 – 51.

Christensen, H., Ma, N., Wrigley, S. N., and Barker, J. (**2007**). "Integrating pitch and localisation cues at a speech fragment level," Proc. Interspeech , 2769–2772.

Christensen, H., Ma, N., Wrigley, S. N., and Barker, J. (**2008**). "Improving source localisation in multi-source, reverberant conditions: exploiting local spectro-temporal location cues," J. Acoust. Soc. Am. **123**, 3294 (A).

Christensen, H., Ma, N., Wrigley, S. N., and Barker, J. (**2009**). "A speech fragment approach to localising multiple speakers in reverberant environments," IEEE Proc. ICASSP, 4593–4596.

Darwin, C. J. (**2008**). "Spatial Hearing and Perceiving Sources," in *Auditory perception of sound sources*, edited by W. A. Yost, R. R. Fay, and A. N. Popper (Springer Verlag), pp. 215–232.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., and Zue, V. (**1993**). "TIMIT Acoustic-Phonetic Continuous Speech Corpus," Technical report NISTIR 4930. National Institute of Standards and Technology, Gaithersburg, MD.

Hu, G. and Wang, D. L. (**2007**). "Auditory Segmentation Based on Onset and Offset Analysis," IEEE Trans. Audio, Speech, Lang. Process. **15**, 396–405.

May, T., van de Par, S., and Kohlrausch, A. (**2009**a). "A Probabilistic Model for Robust Acoustic Localization based on an Auditory Front-end," NAG/DAGA , 254 (A).

May, T., van de Par, S., and Kohlrausch, A. (**2009**b). "A Probabilistic Model for Robust Localization based on a Binaural Auditory Front-end," *submitted* to IEEE Trans. Audio, Speech, Lang. Process. .

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

Tomasi, C. and Manduchi, R. (**1998**). "Bilateral Filtering for Gray and Color Images," Proc. ICCV , 839 – 846.

Wang, D. L. (**2005**). "On ideal binary masks as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, edited by P. Divenyi (Kluwer Academic), pp. 181–197.