esented at DAGA2016 in Aachen, Germa

Probabilistic 2D localization of sound sources using a multichannel bilateral hearing aid

Joachim Thiemann and Steven van de Par

Cluster of Excellence "Hearing4All", CvO Universität Oldenburg, 26129 Oldenburg, Deutschland

Email: Joachim.Thiemann@uni-oldenburg.de

Introduction

Modern hearing aids typically use multiple microphones, which allow for advanced spatial filtering techniques (e.g. beamforming) to be used. To be effective, these algorithms need to be steered intelligently, such that the source of interest is enhanced whereas interfering sources are suppressed. Computational Auditory Scene Analysis (CASA) is the process of determining the acoustic scene by computational analysis of the microphone signals. In the context of hearing aids, the goal is to use the hearing aid microphone signals to localize sources and identify them as important to the hearing aid user.

In this work, we examine some aspects of localizing sources in terms of both azimuth and elevation, in particular how three different approaches compare to each other not only in performance, but also in memory requirements and computational complexity, the latter two being important due to the size and power constraints of digital hearing aids. The three approaches we examine are: a) localization by considering all grid positions as individual locations, b) estimating azimuth and elevation separately in a vertical-polar coordinate system and c) estimating azimuth and elevation separately in an interaural-polar coordinate system.

In our framework, localization is performed using a probabilistic classifier. This has the advantage that no prior assumption about the number of sources in the scene are made. Instead, for every discrete direction, the probability that a sound source is present in that direction is computed. The disadvantage is that for a spherical grid of directions, a large number of points need to be evaluated. Processing such a large localization space requires significant memory and computational resources that are scarce in wearable devices such as hearing aids.

To alleviate this problem, we investigate two localizers that estimate the horizontal and vertical direction of sources separately, and compare these to the grid point localizer. The two localizers differ in the coordinate system used: vertical-polar or interaural-polar. While the former is more natural to humans, the latter matches better to auditory cues.

Experimental Setup

Our comparative experimental setup is based on a conventional vertical-polar setup, using a high-resolution database of head-related impulse responses (HRTFs) [1]. To reduce computation time and highlight issues regarding the geometry at the poles for the interaural-polar setup, we restrict ourselves to an elevation range of $-30°$ to $60°$, and a resolution of $10°$.

The audio signal consists on four channels from a bilateral behind-the-ear (BTE) hearing aid fitted with three microphones per side [2]. In previous experiments [3] we found that four channels were sufficient to obtain good performance, using only the front and back microphones on each hearing aid.

The feature extraction for localization is based on an auditory model system from [4], where the signal is first analyzed using a $F = 32$ channel gammatone filterbank with spacing of center frequencies based on the effective rectangular band (ERB) scale [5], followed by a half-wave rectification and nonlinear compression to emulate the neural transduction.

From the auditory spike-train-like signal, three features are computed on a frame basis, where frames are 20 ms long, with a frame advance of 10 ms. As in [4], we calculate the interaural level difference (ILD) and interaural time difference (ITD) using only the front microphones of the hearing aids. The third feature is a time difference of arrival (TDOA) value calculated from either the left front and back microphone pair or the right front and back microphone pair. On both sides, the crosscorrelation between the front and back microphones is computed. The lag is determined by finding the maximum correlation coefficient, refined to subsample resolution by exponential interpolation. The estimated lag on the side giving the higher maximum correlation coefficient is chosen to be used as feature. We denote the feature vector at time instance t in frequency band f with $\overline{x}_{t,f}$.

Based on the three-element feature vector for each frequency band, a Gaussian mixture model (GMM) classifier is used to perform the probabilistic localization. In each band, each source direction $\varphi_k, k = 1 \ldots K$ is modeled by a frequency-dependent GMM λ_{f,φ_k}. A likelihood map (over the time-frequency plane) can be calculated for all K directions with

$$\mathcal{L}(t, f, k) = \dot{p}(\overline{x}_{t,f}|\lambda_{f,\varphi_k}). \qquad (1)$$

A robust posterior probability is then obtained by integrating and normalizing over frequency and time

$$\mathcal{P}(k|\overline{x}_{t'}) = \frac{\prod_{f,t} m(t,f)\mathcal{L}(t,f,k)}{\sum_{k'} \prod_{f,t} m(t,f)\mathcal{L}(t,f,k')}, \tag{2}$$

where $m(t,f)$ is a binary mask indicating frames where the features are deemed reliable, as determined by a threshold on the maximum ITD crosscorrelation coefficient.

The GMM with $\mathcal{V} = 15$ components was trained using multi-conditional training, to simulate the uncertainty of binaural cues in complex acoustic scenarios. The GMMs used full covariance matrices. We used anechoic HRTF to spatialize sources, mixed with diffuse speech-shaped noise (SSN) at -10, 0, 10, and 20 dB SNR. The cylindrical diffuse noise was generated mixing uncorrelated SSN spatialized at 36 points with 10° spacing around the head at elevation 0°. The target sources were randomly selected male and female speech items from the TIMIT database [6].

Baseline full grid localizer

The baseline system treats each point on the azimuth-elevation grid as a separate location to be evaluated. This results in $K_{\text{FG}} = 360$ directions to be evaluated. While this method has the advantage of being able to handle multiple sources simultaneously without ambiguities, it requires large memory footprint and has a high computational complexity, which makes it unfeasible for mobile and wearable systems. In terms of memory, the classifier needs to store the parameters of $F \cdot K_{\text{FG}} \cdot \mathcal{V} = 172800$ 3-dimensional Gaussians with full covariances.

Vertical-polar localizer

The vertical-polar localizer consists of two localizers that classify sounds in the horizontal (azimuth) and vertical (elevation) directions separately. As shown in Fig. 1, in the vertical-polar coordinate system, points with constant azimuth relative to the head form a plane; points with constant elevation form a cone. Thus, the poles of the spherical grid are above and below the head. This coordinate system is commonly used, as in [7].

The azimuth localizer evaluates $K_{\text{VA}} = 36$ directions, forming a full circle around the head, and requires storing the parameters of $F \cdot K_{\text{VA}} \cdot \mathcal{V} = 17280$ Gaussians. The elevation localizer evaluates $K_{\text{VE}} = 10$ directions, from $-30°$ to $60°$, needing to store the parameters of 4800 Gaussians. The combined directions match exactly to the full grid localizer points, but require storing only a combined 22080 Gaussian parameters.

Interaural-polar classifier

Like the vertical-polar localizer described above, the interaural-polar evaluates azimuth and elevation separately, however the coordinates are arranged such that the poles are aligned with the interaural axis. In this coordinate system, points with constant elevation form a plane, whereas points with constant azimuth form a cone in the direction of the ear, illustrated in Fig. 2.

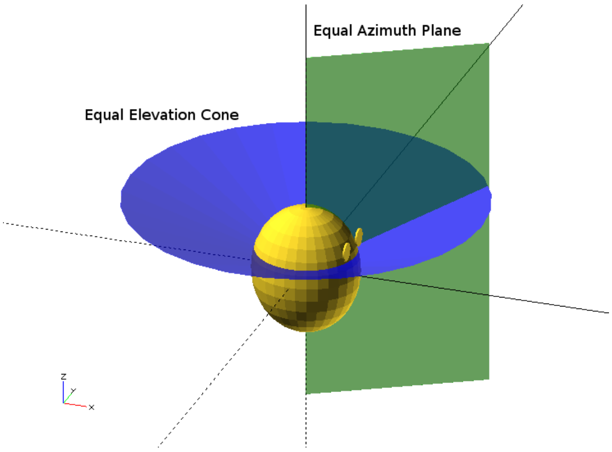

Figure 1: Constant azimuth and constant elevation surfaces for the vertical-polar coordinate system.

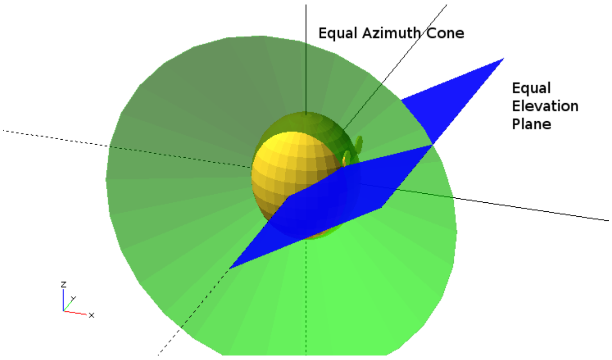

Figure 2: Constant azimuth and constant elevation surfaces for the interaural-polar coordinate system.

The interaural-polar localizer implemented in our experiment also uses a 10° resolution in azimuth and elevation. However, due to the geometry at the poles, the grid points of the source locations no longer line up, introducing a source of error. Furthermore, the range of elevations needs to cover the full range from $-170°$ to $180°$, thus needing $K_{\text{IE}} = 36$ directions. The memory requirements of this localizer are the parameters of 17280 Gaussians. However, the azimuth localizer is now reduced to $K_{\text{IA}} = 19$ directions (from $-90°$ to $90°$), requiring the storage of 9120 Gaussian parameters, for a total of 26400 Gaussian parameters for both localizers. The alignment problem at the right ear pole is illustrated in Fig 3.

Results

We assess localizer performance by testing with stimuli similar to the stimuli used for training. A male and a female speech sample from the TIMIT database (distinct from the training set) was spatialized at each of the grid positions, and mixed at 10 dB SNR with cylindrical diffuse SSN. Localization decisions are made by integrating over 1 s blocks (see Eq. 2) with 50% overlap, and for each block selecting the direction k with highest likelihood.

Localizer performance is assessed using the F-measure as shown in Table 1, and examining the confusion matrices. The F-measure (or F1 score) is the harmonic mean of

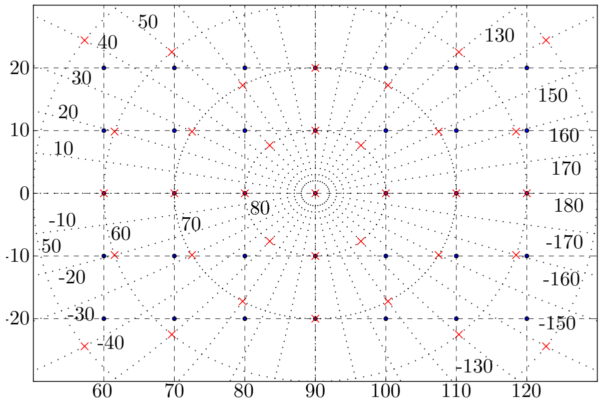

Figure 3: Quantization of vertical polar coordinates near the ear when translated to the interaural-polar grid. Blue dots show the ground truth directions, red crosses the 10° resolution interaural-polar nearest match.

Table 1: Evaluation of localizers using the F-measure.

Localizer	F-Measure	
	Azimuth	Elevation
full grid	(overall) 0.51	
full grid	0.76	0.57
vertical polar	0.91	0.77
interaural polar	0.92	0.59

precision and recall, and has a range from 0 to 1, with a score of 1 indicating perfect recognition without false positives or false negatives. For the full grid localizer, we see from the F-measure that the overall performance is relatively poor. Separating the results into azimuth and elevation shows that errors are mostly made in the vertical direction, which is apparent when examining the confusion matrices in Fig. 4.

If localization is split into dedicated azimuth and elevation estimators, performance increases significantly, as shown by the confusion matrices in Fig. 5. The lower number of classes (the underpinning unit of the directions the localizer evaluates) mean that the localizer can generalize the observed training data better than the full grid classifier. This effect is also clear in the elevation estimate.

Fig. 6 shows the confusion matrices of the localizers using the interaural-polar coordinate system. While it appears that compared to the vertical-polar localizer the performance is similar for the azimuth estimation but worse for the elevation estimation, we note that the mismatch in grid coordinates at the poles, as well as the larger number of vertical directions to be evaluated can explain the lower performance in elevation estimation, and should also have reduced performance of the azimuth estimator.

Discussion

Accurate localization of sound sources relative to the human head is a challenging task, both for humans and for CASA algorithms analyzing hearing aid signals. Unlike humans — which can estimate the horizontal direction of a source very well but are not as good at localizing sources in the vertical direction — hearing aids can be fitted with multiple microphones, which helps with estimating the vertical location of sources (which also resolves the front-back confusion).

Regarding points on some grid on a sphere around the head is conceptually simple but computationally difficult since there are many points to be considered if a reasonable resolution is desired. Thus a separate estimation of azimuth and elevation is a necessity, even if it introduces ambiguities for simultaneously active sources. Splitting the localization reduces the memory requirements and computational complexity by about one order of magnitude.

It then remains to define exactly what is meant by "azimuth" and "elevation", and this article shows that the choice of coordinate system can affect the performance of localization algorithms. While the vertical-polar seems to have better overall performance than the interaural-polar localizer, the difference in performance can be explained by the mismatch of the tested target directions to the quantization of directions of the interaural-polar localizer. Thus, drawing a final conclusion is not possible without redesigning the test to eliminate this bias.

References

[1] J. Thiemann and S. van de Par, "Multiple model high-spatial resolution HRTF measurements," in *Proc. DAGA 2015*, Nürnberg, Germany, Mar. 2015.

[2] H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses," *EURASIP J. on Applied Sig. Proc.*, 2009.

[3] J. Thiemann, S. Doclo, and S. van de Par, "Features for speaker localization in multichannel bilateral hearing aids," in *Proc. EUSIPCO*, Nice, France, Aug 2015, pp. 1266–1270.

[4] T. May, S. van de Par, and A. Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 1–13, 2011.

[5] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, Academic Press, 5th edition, 2003.

[6] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continous speech corpus," Tech. Rep. NISTIR 4930, National Institute of Standards and Technology, Gaithersburg, MD, 1993.

[7] J. Blauert, *Spatial Hearing: The psychophysics of human sound localization*, MIT Press, Cambridge, MA, 1996.

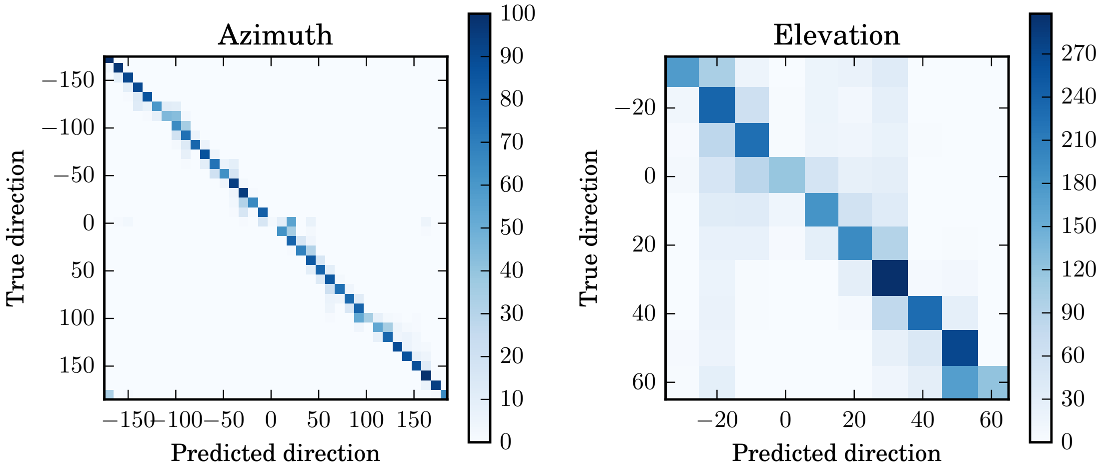

Figure 4: Confusion matrices for the full grid classifier, separated into azimuth and elevation accuracy.

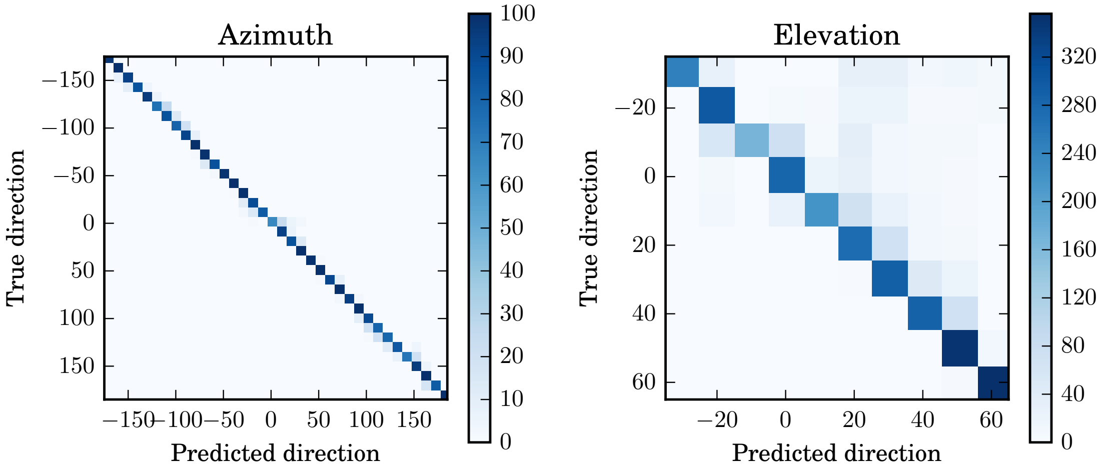

Figure 5: Confusion matrices for the vertical-polar classifiers, showing the horizontal and vertical classifier accuracy.

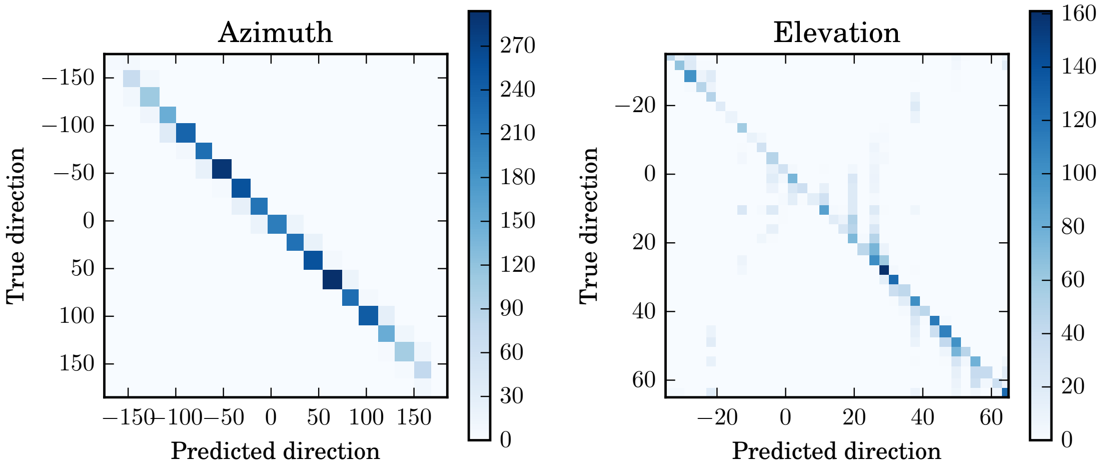

Figure 6: Confusion matrices for the interaural-polar classifiers, showing the horizontal and vertical classifier accuracy.