16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

BINAURAL SOUND LOCALIZATION FOR UNTRAINED DIRECTIONS BASED ON A GAUSSIAN MIXTURE MODEL

Takanori Nishino[†] *and Kazuya Takeda*[‡]

[†] Center for Information Media Studies, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8603, Japan
phone: + (81) 52-789-4210, fax: + (81) 52-789-3172,
email: nishino@media.nagoya-u.ac.jp
web: http://www.sp.m.is.nagoya-u.ac.jp/%7Enishino/

[‡] Graduate School of Information Science, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8603, Japan
phone: + (81) 52-789-3629, fax: + (81) 52-789-3172,
email: kazuya.takeda@nagoya-u.jp

ABSTRACT

We propose and evaluate estimation methods for all sound source directions on the horizontal plane based on a Gaussian mixture model (GMM) using binaural signals. An estimation method based on GMM can estimate a sound source direction on which GMM has already been trained; however, it cannot estimate without a model that corresponds to a sound source direction. Three methods with interpolation techniques are investigated. Two generate GMMs for all directions by interpolating an acoustic transfer function or statistical values of GMM, and the other calculates the posterior probability for all directions with a limited number of GMMs. In our experiments, we investigated six interval conditions. From the results, the interpolation methods of an acoustic transfer function and the statistical values of GMM achieve better performance. Although there were 12 trained GMMs for the 30° intervals, the interpolation method of the statistical values of GMM estimated 62.5 % accuracy (45/72) with 2.8° estimation error. These results indicate that the proposed method can estimate all sound source directions with a small amount of known information.

1. INTRODUCTION

The detection of sound source direction is a crucial technique widely used in such fields as speech enhancement, sound recording, security systems, and so on. Studies based on microphone arrays are abundant and employ many microphones to obtain high detection performance. Reducing the number of microphones would lower costs and facilitate maintenance.

We can perform a sound localization using both ears. A binaural signal can be represented as a convolution of a sound source signal and a binaural room impulse response (BRIR). A BRIR is composed of a head-related transfer function (HRTF) and a room impulse response that represents an acoustic environment. We can perceive the sound source direction with an interaural time difference and an interaural level difference (ILD) that are included in binaural signals. Previous research examined the probability distributions of these interaural differences[1], and a sound localization model using its distribution has also been proposed[2]. Estimation methods using HRTF were also examined for robot hearing [3, 4, 5, 6].

The detection of sound source direction is performed by comparing the acquired memory of the sound localization and the information obtained from the current sound [7]. A method based on a Gaussian mixture model (GMM) was investigated as one training-based scheme [8]. In that study, experiments for estimating sound source direction were conducted in eight reverberation conditions, and the results showed that this method could estimate any sound source direction in a room that has several reverberation times. Moreover, we conducted experiments for an in-car environment[9], and the results showed that our evaluated method is robust to sound source situations and the influence of noise.

These training-based methods can estimate the trained direction; however, basically, the sound source of the untrained direction cannot be estimated. GMMs must be prepared for all directions, which is difficult because measuring the binaural signals for all sound source directions and computer resources is expensive.

In this paper, we propose a novel estimation method for all sound source directions using a binaural signal based on GMM. Three estimation methods for all sound source directions were investigated using interpolation techniques. Our method applied interpolation methods to a BRIR, the statistical value of GMM, and posterior probability. Estimation performances were evaluated by correct rate and estimation error.

2. SOUND LOCALIZATION BASED ON GMM

2.1 Feature parameter

In our method, ILD is represented by a cepstrum-like parameter called the ILD cepstrum that represents its rough tendencies. It also divides components based on HRTFs and room environments. The ILD cepstrum was obtained with the following procedure:

1. The signals that arrive at the left and right ears are denoted as $s_{\mathrm{L}}[t]$ and $s_{\mathrm{R}}[t]$, respectively. Both are truncated with a Hamming window whose frame length and frame shift is $l = 128$ and $l_s = 32$, respectively, based on the results of our preliminary experiment.
2. ILD is calculated using Eq. (1). However, when one of the signals has a lower absolute value of amplitude than the threshold, ILD calculation is not conducted:

$$|S_{\mathrm{LR}}(f_k)| = \frac{|S_{\mathrm{L}}(f_k)|}{|S_{\mathrm{R}}(f_k)|}, \tag{1}$$

where $S_{\mathrm{L}}(f_k)$ is the magnitude response of the left ear's signal, $S_{\mathrm{R}}(f_k)$ is the right ear, and f_k denotes the frequency. We used a threshold of 0.005 based on the results of our preliminary experiment.

3. The Fourier transform is applied to the logarithm of ILD, and feature parameter $c[n]$ is obtained using Eq. (2). In our study, the ILD cepstrum is denoted by $c[n]$. Thus, the ILD envelope is obtained by lower-order ILD cepstrums:

$$c[n] = \frac{1}{N}\sum_{l=0}^{N-1} 10\log_{10}|S_{\mathrm{LR}}(f_k)|e^{j2\pi ln/N} \quad (n = 0,1,\cdots,N). \tag{2}$$

The condition of $N = 15$ was also investigated. If a BRIR consists of the convolution of the room impulse response and HRTF, the ILD cepstrum can be represented as the sum of components resulting from the room impulse response and HRTF.

2.2 GMM training and sound source direction estimation with GMM

GMM, a statistical model that represents the linear combination of Gaussian distributions, is often used for speech recognition, speaker identification, and so on. In our method, a Gaussian model for every direction was prepared and trained using binaural signals. The training procedure is described as follows:

1. The distribution of the ILD cepstrum $c[n]$ given by Eq. (2) is approximated with Gaussian distribution, which is considered the statistical model for estimating sound source direction. Statistical models for every sound source direction can be represented by:

$$\lambda_\theta = \left\{w_{\theta,m}, \mu_{\theta,m}, \Sigma_{\theta,m} | m = 1,2,\cdots,M\right\}.$$

2. The expectation maximization (EM) algorithm gives the weight for each distribution w_m, mean μ_m, and covariance matrix Σ_m. Then estimation model λ_θ is trained for every sound source direction. Our method uses the diagonal covariance matrix. In our experiments, a single mixture model ($M = 1$) was used because distributions of the ILD cepstrum were unimodal.

Below is the procedure for estimating sound source direction with the Gaussian model:

1. The ILD cepstrum $c[n]$ of the input signals is calculated with the same procedure as in Section 2.1.
2. The posterior probability between the ILD cepstrums of the input signals and every trained Gaussian model is calculated. The direction of the model that gives maximum posterior probability is considered the sound source direction.

Figure 1 shows a posterior probability map when GMMs for all directions were prepared. The input signal (test signal) is identical as a training signal for every azimuth. In this figure, the horizontal axis is a target azimuth that corresponds to an input signal, and the vertical is an evaluated azimuth that corresponds to the trained GMMs. Colors represent posterior probability. Posterior probabilities were normalized at every target azimuth because the calculations of posterior probability are independently performed every target azimuth. If sound localization is performed correctly, as shown in Fig. 1, a dark red appears on a diagonal. In addition, note that higher probabilities appear at directions that cause front-back confusion.

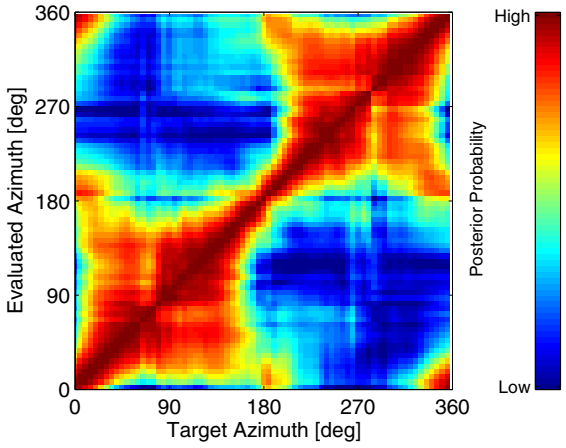

Figure 1: Posterior probability map when training and test signals are identical.

3. METHOD FOR UNTRAINED SOUND SOURCE DIRECTION ESTIMATION

Three methods to estimate every sound source direction were investigated using interpolation techniques. The first and second methods prepare GMMs for all directions with an interpolation method. In our study, since binaural signals were obtained by convolving a BRIR and an evaluation signal, an interpolation method was applied to the BRIRs. If interpolation is performed precisely, good estimation can be achieved. However, neither method decreases the number of calculations of the posterior probability. The third method, which interpolates posterior probability, can decrease the number of calculations of the posterior probability.

- Method 1:
  BRIR was interpolated by a simple linear interpolation method[10]. Interpolated BRIR $\hat{h}[t]$ is given by

$$\hat{h}[t] = rh_{\theta_1}[t] + (1-r)h_{\theta_2}[t-\tau], \quad 0 \le r \le 1, \tag{3}$$

  where $h_{\theta_1}[t]$ and $h_{\theta_2}[t]$ are measured BRIRs, r is a dividing ratio, and τ is decided by the cross correlation coefficient between $h_{\theta_1}[t]$ and $h_{\theta_2}[t]$. To obtain higher interpolation performance, first, the BRIRs were applied 10 times upsampling [11]. After interpolation, a downsample method was applied. Except for the measured directions, the GMM of every direction was trained with the interpolated BRIR.
- Method 2:
  A GMM has means, variances, and weight coefficients. GMMs were trained using the measured BRIR to decide statistical values. A statistical vector was defined with calculated values, for example:

$$\mu = \{\mu_{0^\circ},\cdots,\mu_\theta,\cdots,\mu_{360^\circ}\}.$$

  The statistical values of an arbitrary direction were obtained by interpolating the vectors with the cubic spline method.
- Method 3:
  GMMs were trained using the measured BRIR and posterior probabilities P_θ between the evaluated signals and

then calculated. The posterior probability vector was defined as follows for interpolation:

$$\mathbf{P} = \{P_{0^\circ}, \cdots, P_{\theta}, \cdots, P_{360^\circ}\}.$$

A posterior probability of a desired direction was obtained by interpolating $\mathbf{P}$ with the cubic spline method.

4. EXPERIMENTS

4.1 BRIR measurement

BRIRs were measured with a head-and-torso simulator (HATS, B&K 4128) in an echoic measurement room with a swept sine signal [12] at durations of 1.365 s transduced by a loudspeaker (BOSE Acoustimass). The HATS was positioned on a turntable, and the loudspeaker was placed on an arched traverse. Both the turntable and the arched traverse could be moved at intervals of 1° with 0.3° accuracy. The distance between the loudspeaker and the center of the bitragion was 1.2 m. BRIRs were measured for 72 azimuths on the horizontal plane at a sampling frequency of 48 kHz. The reverberation time of the measurement room was 151 ms, and the background noise level of the measurement room was 19.1 dB(A). The azimuth angles of the sound source (loudspeaker) corresponded to the following: the front was 0°, the left was 90°, the right was 270°, and the angle directly behind the HATS was 180°.

4.2 Experimental conditions

In the experiment, a binaural signal was obtained by convolution of the BRIR, and a kind of bubble noise was generated by superposing many speech signals. We used the bubble noise with 24 superpositions, whose signal durations were 2 s. Signal durations for training and evaluation were identical.

The evaluated intervals were 10, 15, 30, 45, and 60°. For example, in the case of 10° intervals, GMMs were trained for 0°, 10°, ···, 360° (same as 0°), and 72 azimuths (0°, 5°, ···, 355°) were estimated.

Method 1, the BRIRs must be interpolated for an arbitrary azimuth from the BRIRs that were measured at the evaluated intervals. Interpolation performances were evaluated by the signal-to-deviation ratio (SDR) and the spectral distortion (SD):

$$\mathrm{SDR} = 10\log_{10} \frac{\sum_{n=1}^{N} h^2[n]}{\sum_{n=1}^{N} \{h[n] - \hat{h}[n]\}^2} \quad [\mathrm{dB}], \tag{4}$$

where $h[n]$ is the measured BRIR and $\hat{h}[n]$ is the interpolated BRIR. N is the duration of the BRIR, and $N = 65,536$ was used. The interpolated BRIR is more similar to the measured BRIR when a large SDR is obtained.

$$\mathrm{SD} = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(20\log_{10}\frac{|H(f_k)|}{|\hat{H}(f_k)|}\right)^2} \quad [\mathrm{dB}], \tag{5}$$

where $|H(f_k)|$ is the frequency magnitude response of the measured HRTF, $|\hat{H}(f_k)|$ is that of the interpolated HRTF, and f_k is the frequency. K is the number of frequency bins, and $K = 513$ was used. The interpolated HRTF is more similar to the measured HRTF when a small SD score is obtained.

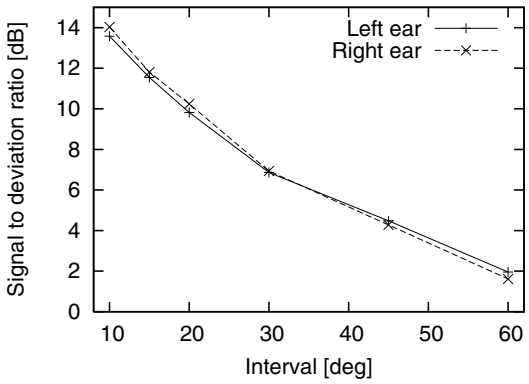

Figure 2: Average signal-to-deviation ratio of interpolated BRIRs. Solid and dashed lines represent performance of left- and right-ear BRIRs, respectively.

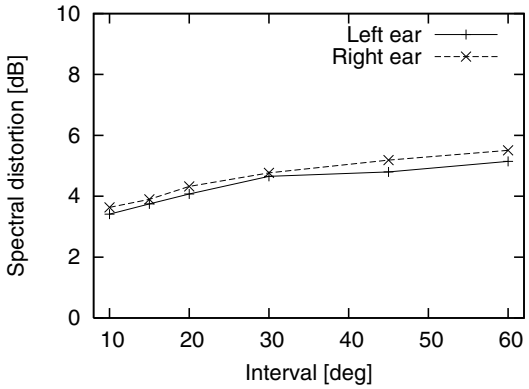

Figure 3: Average spectral distortion of interpolated BRIRs. Solid and dashed lines represent performance of left- and right-ear BRIRs, respectively.

Figures 2 and 3 show the performances. Since the interval was large, the interpolation performances were worse. The interpolation performances of BRIR were worse than the HRTF interpolation [10] because BRIR's duration was long and it includes a component concerned with the reverberation.

4.3 Results

Figures 4 and 5 show the correct rate and the estimation error, respectively. The correct answer was that the target and the estimated azimuth were identical. Estimation error was calculated by

$$e = \frac{1}{72}\sum_{i=1}^{72} |\theta_i - \hat{\theta}_i| \quad [\mathrm{deg}], \tag{6}$$

where θ_i is the i-th target azimuth and $\hat{\theta}_i$ is the i-th estimated azimuth. Trained directions were included in the calculation of the correct rate and error due to reflect the possibility of mistakes at the trained directions.

From the results, since the interval was large, the performances were worse. The difference between Methods 1

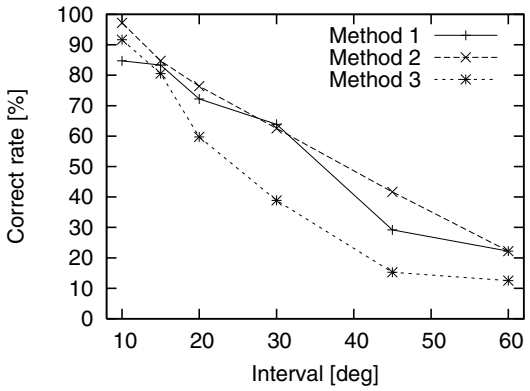

Figure 4: Correct rate.

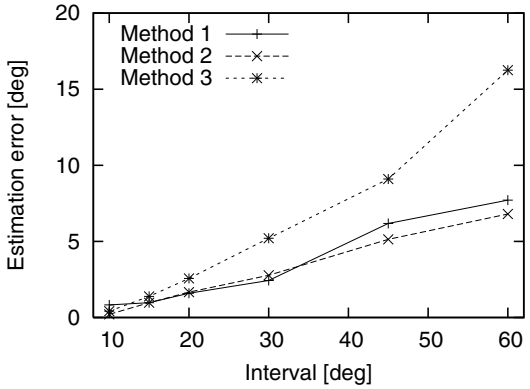

Figure 5: Estimation error.

and 2 is small, and both methods could estimate greater than 60 % accuracy within 3° estimation error under 30° intervals. Method 1, the estimation performances have a strong correlation with the SDR instead of the SD score. The correlation coefficient between the results of Method 1 and the SDR was 0.96.

Therefore, the interpolation method must be improved to obtain good estimation performances.

Figures 6 to 8 show the posterior probability maps and estimated directions for 30° intervals, and Figures 9 to 11 are for the 60° intervals. In these figures, the horizontal axis is a target azimuth that corresponds to an input signal, and the vertical is an evaluated azimuth that corresponds to the trained GMMs. Colors represent the posterior probability, and dots are the estimated direction. Except for Figure 11, these figures all show that the posterior probability map resembles the original map (Fig. 1). Therefore, the sound source direction can be estimated using a small amount of known information. However, since estimation performances were worse at directions around both ears, improvement is still necessary. Method 3 for large intervals, variable intervals must be examined instead of equivalent intervals.

5. CONCLUSION

Three methods based on GMM for estimating all sound source directions on the horizontal plane were investigated. The results showed that good performances were obtained by preparing the GMM for all directions. Both methods estimated with greater than 60 % accuracy within 3° estimation error under intervals of 30°. Even though the method of interpolating the posterior probability did not obtain good performance, it is still considered useful because the posterior probability maps of the three methods are similar and the number of calculations was decreased significantly.

Future work includes conducting experiments under several reverberation conditions, using an interaural time difference, and applying it to robot hearing.

Acknowledgment This work was supported by a Grant-in-Aid for Scientific Research (No. 17700183 and No. 18300064).

REFERENCES

[1] P. M. Zurek, "Probability distributions of interaural phase and level differences in binaural detection stimuli," *J. Acoust. Soc. Am.*, vol.90, pp. 1927–1932, 1991.

[2] D. Heavens and T. Pearce, "Binaural sound localization by level distributions," *Proc. ICA2007*, PPA-05-012, 2007.

[3] F. Keyrouz, Y. Naous, and K. Diepold, "A new method for binaural 3-D localization based on HRTFs," *Proc. ICASSP2006*, vol.5, pp. 341–344, 2006.

[4] J. Hörnstein, M. Lopes, J. Santos-Victor, and F. Lacerda, "Sound localization for humanoid robots - building audio-motor maps based on the HRTF," *Proc. IROS2006*, 2006.

[5] Z. Zhang, K. Itoh, S. Horihata, T. Miyake, and T. Imamura, "Estimation of sound source direction using a binaural model," *Proc. SICE-ICASE 2006 International Joint Conference*, pp. 4051–4056, 2006.

[6] L. Calmes, G. Lakemeyer, and H. Wagner, "Azimuthal sound localization using coincidence of timing across frequency on a robotic platform," *J. Acoust. Soc. Am.*, vol. 121, pp. 2034-2048, 2007.

[7] G. Plenge, "Über das problem der im-kopf-lokalization [On the problem of in head localization]," *Acustica*, vol. 26, pp. 241–252, 1972.

[8] T. Nishino, N. Inoue, K. Itou, and K. Takeda, "Estimation of sound source direction based on Gaussian model using envelope of interaural level difference," *J. Acoust. Soc. Jpn.*, vol. 63, pp. 3–12, 2007.

[9] M. Takimoto, T. Nishino, H. Hoshino and K. Takeda, "Estimation of speaker and listener positions in a car using binaural signals," *Acoust. Sci. & Tech.*, vol. 29, pp.110–112, 2008.

[10] T. Nishino, S. Mase, S. Kajita, K. Takeda, and F. Itakura, "Interpolating HRTF for auditory virtual reality," *J. Acoust. Soc. Am*, vol. 100, p. 2602, 1996.

[11] K. Watanabe, S. Takane, and Y. Suzuki, "Interpolation of head-related transfer functions based on the common-acoustical-pole and residue model," *Acoust. Sci. & Tech.*, vol. 24, pp. 335–337, 2003.

[12] N. Aoshima, "Computer-generated pulse signal applied for sound measurement," *J. Acoust. Soc. Am.*, vol. 69, pp. 1484–1488, 1981.

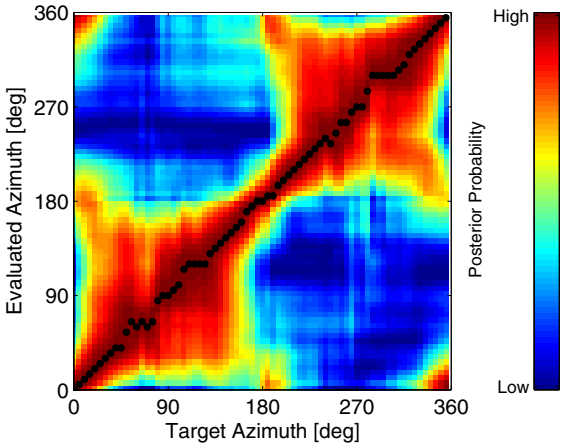

Figure 6: Posterior probability map for Method 1 for 30° intervals. Correct rate is 63.9 %, and estimation error is 2.4°.

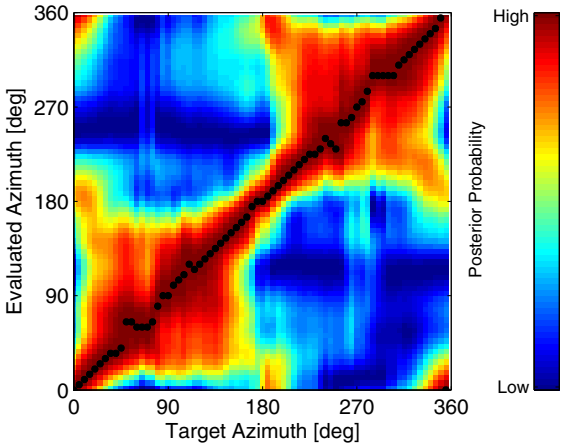

Figure 7: Posterior probability map for Method 2 for 30° intervals. Correct rate is 62.5 %, and estimation error is 2.8°.

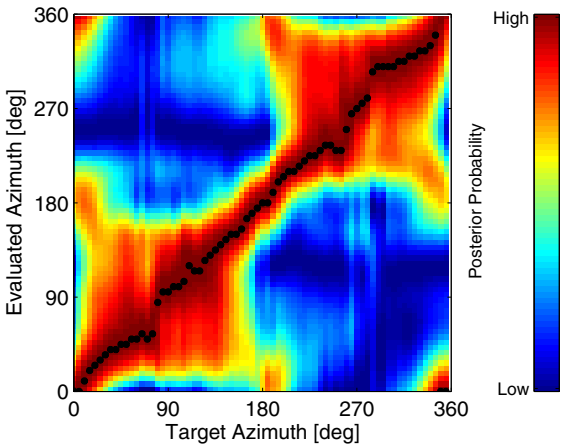

Figure 8: Posterior probability map for Method 3 for 30° intervals. Correct rate is 38.9 %, and estimation error is 5.2°.

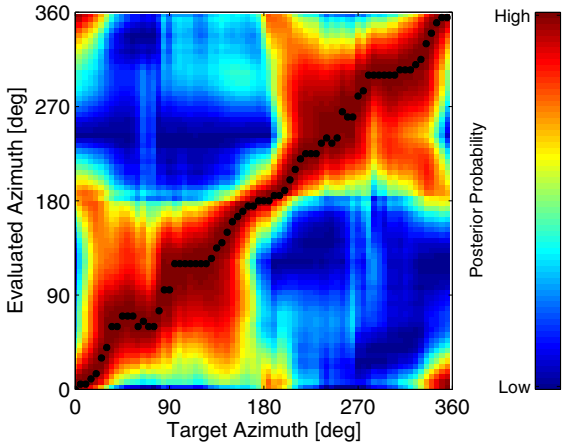

Figure 9: Posterior probability map for Method 1 for 60° intervals. Correct rate is 22.2 %, and estimation error is 7.7°.

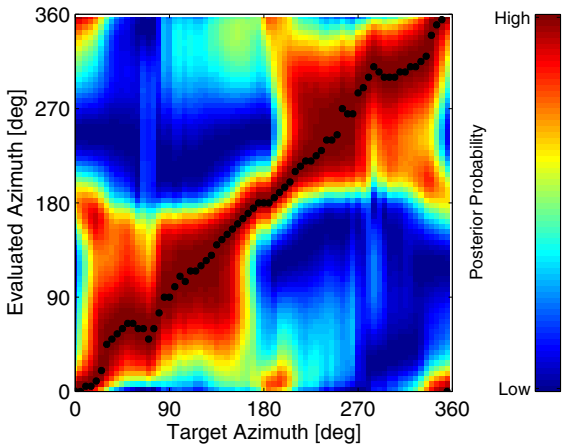

Figure 10: Posterior probability map for Method 2 for 60° intervals. Correct rate is 22.2 %, and estimation error is 6.8°.

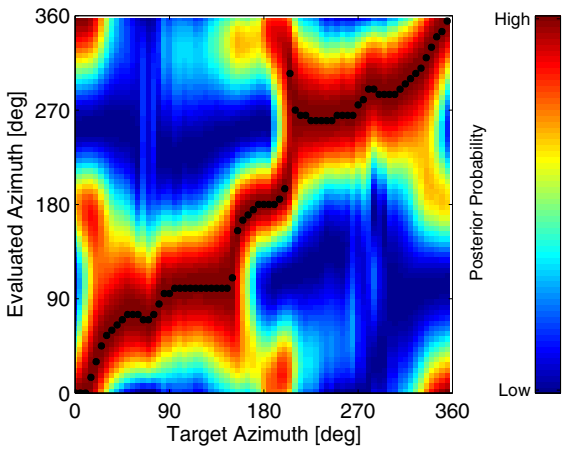

Figure 11: Posterior probability map for Method 3 for 60° intervals. Correct rate is 12.5 %, and estimation error is 16.3°.