PARAMETRIZATION OF ACOUSTIC IMAGES FOR THE DETECTION OF HUMAN PRESENCE BY MOBILE PLATFORMS

M. Moebus, A. M. Zoubir

Technische Universität Darmstadt
Signal Processing Group
Merckstr. 25, D-64283 Darmstadt, Germany
{*moebus,zoubir*}*@ieee.org*

M. Viberg

Chalmers University of Technology
Department of Signals and Systems
SE-41296 Göteborg
viberg@chalmers.se

ABSTRACT

We address the problem of human detection with mobile platforms such as robots. Instead of using an optical system, we propose to employ an acoustic 2D array to reliably obtain an image of a human in a 3D spatial power spectrum which is independent of lighting conditions and uses cheap acoustic sensors. We show that humans have a distinct acoustic signature and propose to model the echoes from reflecting parts of objects in the scene by a Gaussian-Mixture-Model. When it is fitted to the acoustic image, we can extract geometric relations between the present echoes and represent the acoustic signatures in a low-dimensional parameter space. We present results based on real data measurements that demonstrate that different objects can be reconstructed from the data and discriminated. The obtained parameter space forms the basis for subsequent detection and classification of humans.

Index Terms— acoustic arrays, human detection, gaussian-mixture-model

1. INTRODUCTION

Mobile platforms that operate in human environments not only need to detect obstacles in their surroundings, but also become aware of the presence of humans. For example, detecting persons in the surroundings of a robot is crucial for control of its awareness and navigation. Based on the detection, the orientation and trajectory of the person can be estimated and the system can respond meaningfully, step out of the way of the human, addressing him or her etc. Existing human detection systems are mainly based on optical mono or stereo cameras, which have certain limitations, e.g. they depend on good lighting conditions or employ expensive infrared cameras. At the same time, optical systems can not always reliably detect range and often assume motion or a specific shape of the persons [1, 2]. Radar-based systems require expensive hardware and can be unreliable due to the very low reflection intensity from humans (e.g. [3]). Acoustic imaging provides a simple and cheap sensor alternative that allows for very precise range as well as angular information. Using an acoustic 2D array, objects can easily be detected in the environment and a 3D image of reflections in the surrounding scene can be created [4, 5]. Human detection based on such a cheap system can greatly enhance the overall system reliability. In this paper, we demonstrate that humans have a distinctive signature in acoustic images which can be exploited for detection. We present a method to model the acoustic signatures in the 3D spatial spectrum estimate of the objects by a Gaussian-Mixture-Model (GMM). Based on the parameters of this model, a detector can be designed to discriminate between person and non-person objects.

The paper is organized as follows: After a short problem formulation in Section 2, we present the approach in Section 3. In Section 4, we apply the approach to real data measurements recorded with an acoustic array from a set of indoor scenes and demonstrate that the important characteristics of the acoustic signatures are well-modeled such that they can be used in a subsequent detection and classification stage.

2. PROBLEM FORMULATION

We address the problem of human presence detection in the surroundings of a mobile platform such as a robot using acoustic imaging. The images are created using a 2D array which sends out a narrowband excitation signal in order to illuminate acoustically the scene of interest. The back-scattered echoes which are reflected from objects in the scene are recorded by the array. Based on this data, a three-dimensional power spectrum estimate is obtained via adaptive beamforming, resulting in $P(\theta, \phi, r)$ where θ is the elevation angle, ϕ the azimuth angle, and r denotes range (for more details, see [5]). The object classes that can be present in the scene vary greatly and are difficult to model by a closed set of prototypes, because the representation in $P(\theta, \phi, r)$ depends on both the relative position of objects to the array as well as the angle from which the excitation signal is sent into the scene. The strength of the reflected echo depends on the shape and texture of the reflecting surface and its orientation relative to the array. This leads to the effect that some solid reflective parts of the scene are weakly presented in the spatial

978-1-4244-4296-6/10/$25.00 ©2010 IEEE

power spectrum. However, there exist some reliable acoustic phenomena such as corner and edge reflections as well as reflections from rough surfaces which allow us to generically model object reflections. We are interested in features that discriminate between reflections from humans and those from other objects in the scene.

3. PROPOSED APPROACH

In Figure 1(a)-(d), we show respectively acoustic images from three persons and an office chair at a distance of 1.75m away from the array in broadside direction. The persons are facing the array (0°), looking to the side (90°) or looking directly away (180°), and were recorded either walking or standing. The office chair has a diameter of 0.4m, plastic arm rests and a textile seat cover and cushion. The images were processed on a large range bin around the maximal reflection in order to obtain a single image $P_r(\theta, \phi)$ (in dB) containing all spatial information about the person. However, the range information is not lost and can be used to determine the three-dimensional position of reflectors subsequently. As we can see, due to the complex texture of the human body, the excitation signal reliably reflects diffusely back to the array from both the torso and the head. As most objects only show specular reflections, this is quite unique to humans. Only when there are objects which possess a large complex surface texture, it is possible to obtain larger reflecting areas. Thus, already the occurrence of a torso-shaped reflection is rare and together with another reflector above, the likelihood of human presence is quite high. However, there are several problems that occur. Firstly, we can see in the figure that while all three images show the reflections from a human, the head echo is not always as strong as the torso echo. Additionally, the two echoes might not be well separated, but merge together, depending on the distance and location of the human or due to limited resolution of the beamformer in directions far off-broadside. Another aspect that can be seen is that the torso reflection often will not be uni-modal, as there are several parts of the torso that reflect well. To overcome these problems, we propose a procedure which is outlined below and illustrated in Figure 2.

3.1. Image Segmentation

Although the above mentioned problems lead to the fact that standard segmentation techniques will not reliably detect the head and torso echoes as two distinct segments in many cases, an initial application of a segmentation algorithm can serve as a first approximation to a foreground/background discrimination. However, in cases where echoes of head and torso are not distinctively separated, this will not allow to discriminate the two. We therefore employ a model-based parametrization of the image where the segmentation algorithm serves mainly as a tool to constrain the model and provide a good initialization of the formulated optimization problem (see Section 3.2). In order to reduce potential multi-modalities of the torso echoes, we smooth the image $P_r(\theta, \phi)$ in an initial step. We then use the EM algorithm to fit a GMM with two Gaussians $\mathcal{G}(\mu_1, \sigma_1), \mathcal{G}(\mu_2, \sigma_2)$ where $\mu_1 > \mu_2$ to the empirical pixel intensity distribution of the smoothed image. The foreground region $\mathcal{R}$ is formed by pixels in the image that have a higher probability to belong to the Gaussian $\mathcal{G}(\mu_1, \sigma_1)$. This approach is well-known as a simple segmentation technique (e.g., see [6]). In Figure 3, we show an example of the resulting foreground region where the head and torso echoes are not separated into distinct segments. Other, more complex segmentation techniques, e.g. involving active contours or min-cut/max-flow techniques do not work superior as they are gradient-based and assume sharp segment boundaries which are not present in the acoustic images (see e.g. [7, 8]).

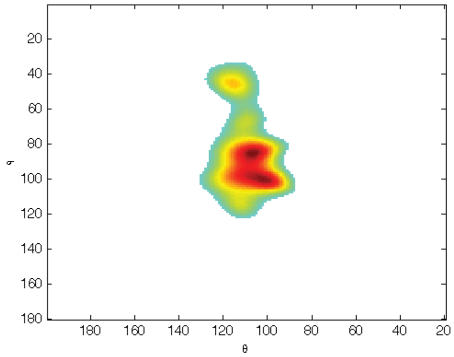
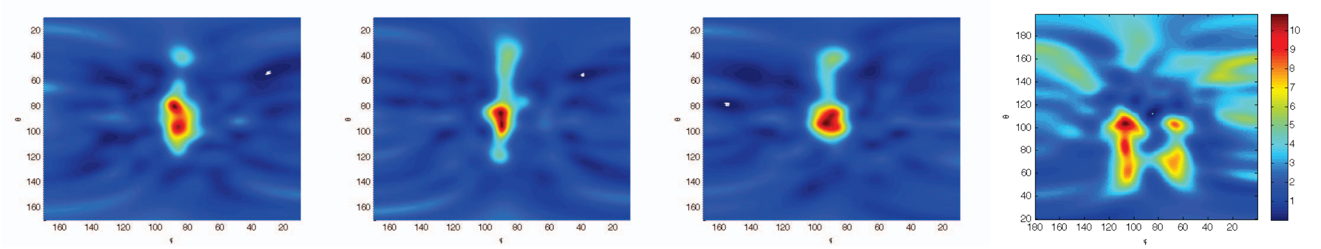


Fig. 3: Result of the segmentation of an image where the echoes of head and torso are in close distance.

3.2. Modeling the Acoustic Signature

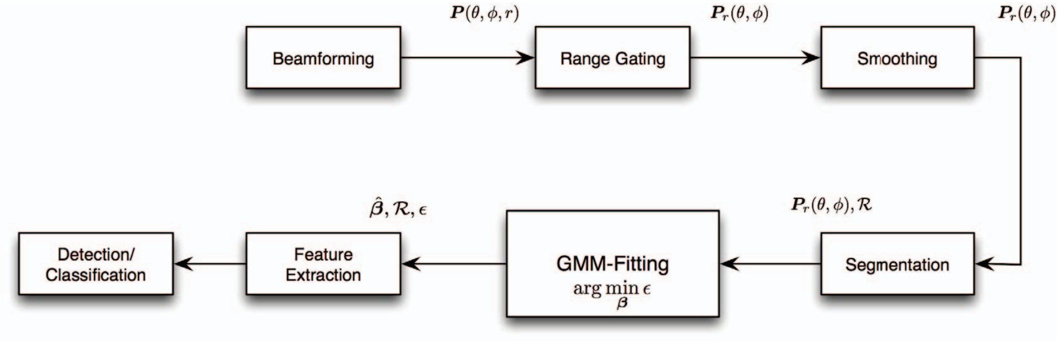
Based on the observations about the occurrence of head and torso echoes, we are interested in two aspects. First, we want to parametrize the image such that the acoustic signature is preserved and we can establish geometric properties in the image. Secondly, we aim to find clusters in the parameter space that allow us to discriminate between humans and other objects present in the scene, i.e., we want to obtain parameter sets that are unique to the presence of a large torso echo, a weaker head echo and possibly other echo sources in the scene. We therefore propose to model the spatial power spectra obtained from the acoustic array as a mixture of K two-dimensional Gaussians $\mathcal{G}(w_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), k = 1, \ldots, K$ in the (θ, ϕ)-domain where w_k is a weighting factor, $\boldsymbol{\mu}_k$ are the mean vectors, and, with $\rho, \sigma_{\theta,k}, \sigma_{\phi,k}$ being the correlation coefficient and the standard deviations in θ and ϕ dimensions, the covariance matrices

$$\boldsymbol{\Sigma}_k = \begin{pmatrix} \sigma_{\theta,k}^2 & \rho\sigma_{\theta,k}\sigma_{\phi,k} \\ \rho\sigma_{\theta,k}\sigma_{\phi,k} & \sigma_{\theta,k}^2 \end{pmatrix} .$$



(a) 0°, walking (b) 90°, standing (c) 180°, walking (d) Office chair

Fig. 1: Three spatial spectra of humans in azimuth and elevation from different orientations and one spatial spectrum of an office chair.



$\boldsymbol{P}(\theta,\phi,r)$
$\boldsymbol{P}_r(\theta,\phi)$
$\boldsymbol{P}_r(\theta,\phi)$
Beamforming
Range Gating
Smoothing
$\hat{\boldsymbol{\beta}}, \mathcal{R}, \epsilon$
$\boldsymbol{P}_r(\theta,\phi), \mathcal{R}$
Detection/ Classification
Feature Extraction
GMM-Fitting $\arg\min_{\boldsymbol{\beta}} \epsilon$
Segmentation

Fig. 2: Flowchart of the proposed human detection scheme.

We then fit the GMM to the image by solving the following optimization problem:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \epsilon = \arg\min_{\boldsymbol{\beta}} \int_{\mathcal{R}} ||\boldsymbol{P}_r(\theta,\phi) - f(\boldsymbol{\beta})||^2 \delta\theta\delta\phi \tag{1}$$

$$\text{s.t.} \quad \lambda_{1,k}, \lambda_{2,k} > 0 \quad \forall\, k = 1, \ldots, K \tag{1.1}$$

$$\mathcal{S}_i \not\cap \mathcal{S}_j \quad \forall\, i \neq j \tag{1.2}$$

$$\boldsymbol{\mu}_k \in \mathcal{R} \quad \forall\, k = 1, \ldots, K \tag{1.3}$$

where $\mathcal{R}$ is obtained by segmentation (see Section 3.1),

$$\begin{aligned} f(\boldsymbol{\beta}) &= f(w_1, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1^T, \ldots, w_K, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K^T) \qquad (2) \\ &= \sum_{k=1}^{K} w_k \mathcal{G}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \end{aligned}$$

$\boldsymbol{\beta}$ denotes the parameter vector to be estimated. $\lambda_{1,k}, \lambda_{2,k}$ are the eigenvalues of $\boldsymbol{\Sigma}_k$. Constraint (1.1) is required in order to guarantee positive definite covariance matrices for all $\boldsymbol{\Sigma}_k$. The second constraint (1.2) is introduced in order to ensure that each echo is modeled only from a single Gaussian. Here, $\mathcal{S}_i$ is an ellipsoid region in the (θ,ϕ)-domain that covers a certain fraction of the volume of the ith Gaussian such that all points (θ,ϕ) in $\mathcal{S}_i$ fulfill

$$((\theta,\phi) - \mu_i)\boldsymbol{\Sigma}_i^{-1}((\theta,\phi) - \mu_i)^T \leq C(1-\rho)^2 \quad . \tag{3}$$

Here, since we assume two-dimensional Gaussians, C is determined according to the inverse cumulative χ_2^2 distribution such that

$$\int_0^{C^2} \frac{e^{-t/2}}{2} \delta t = P \tag{4}$$

is satisfied.[1] Here, P denotes the fraction of the volume under the Gaussians to be covered by any $\mathcal{S}_i$. The last constraint (1.3) is based on the segmentation result. It is not strictly necessary to formulate the problem, but prevents divergence of the solving algorithm from reasonable solutions.[2] The problem formulated in equation (1) can also be initialized using knowledge of $\mathcal{R}$, e.g. the mean vectors $\boldsymbol{\mu}_k, k = 1, \ldots, K$ can be defined to be at the locations of the K largest extrema in $\mathcal{R}$. It can be solved numerically, e.g. using a Quasi-Newton algorithm, for different K, depending on the target object class, i.e.. for humans, knees, hands and feet are not always visible in the image. Therefore, we set $K = 2$ to model only the head and torso echoes. The solution provides

[1] Note that this holds exactly only if $\rho = 0$ and is an approximation otherwise. However, for $\rho > 0$, the ellipsoid is only rotated, thus the relation leads to the desired coverage of the Gaussian.

[2] Depending on the implementation of the employed optimization algorithm, it can be beneficial to reformulate some constraints into penalty terms of the cost function, e.g. a positive definite $\boldsymbol{\Sigma}_k$ can be favoured by a penalty term $\log(\det \boldsymbol{\Sigma}_k)$, (1.3) can be taken into account by penalizing the distance of $\boldsymbol{\mu}_k$ to $\mathcal{R}$.

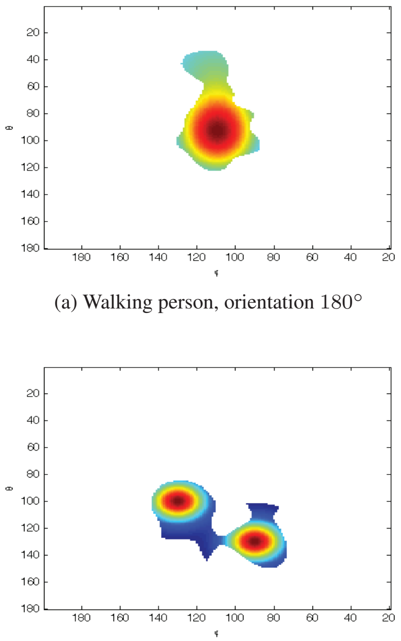
not only $\hat{\boldsymbol{\beta}}$, but also ϵ as a measure for the goodness-of-fit of the model. Based on $\hat{\boldsymbol{\beta}}$, we can then formulate other features that are meaningful to detect humans in the image, e.g. the relative position, distance and angle of the centroids. Additionally, we can expect the head echo to be much smaller than the torso echo, meaning that the ratio of variances is an additional feature. Together with the segmentation contours, these geometric features, $\hat{\boldsymbol{\beta}}$ and ϵ allow to represent humans in acoustic images uniquely and can be supplied to a subsequent classifier.

4. RESULTS AND DISCUSSION

We demonstrate the approach described above and apply it to data from two different scenes, each containing a single object. The data was recorded using a 30-element 2D array with omnidirectional acoustic receivers. In the first scene, a human was standing in front of the array at a distance of 1.75m, facing the array. In the second scene, the office chair from Figure 1 was placed at the same position. In Figure 4, we can see the reconstructed images based on the segmentation region $\mathcal{R}$ and the estimated parameters $\hat{\boldsymbol{\beta}}$ for a GMM with $K = 2$. The resulting image for the first scene (Fig. 4a) clearly shows the Gaussian that models the strong torso echo ($w_1 = 8.296$) while the head echo, modeled by a significantly weaker Gaussian above ($w_2 = 3.637$), is less strongly visible, but reliably located above the torso as expected. Thus, although the two echo sources were not well separated, the image was successfully parametrized. In the second scene, the two strongest echoes are reflected from both one of the arm rests and a corner of the seat cushion. This is also modeled in the reconstructed image (Fig. 4b), where only two smaller, almost equally strong reflections are modeled at both corresponding regions in the image. We see that modeling the spatial power spectra by a GMM allows to represent the spatial information about the position of reflecting surfaces in a parameter vector. The approach can be extended by application of the model with a varying order K if it is desired to model the image in more detail. Effectively, this would result in a model-order selection problem where the order has to be chosen based on different $\hat{\boldsymbol{\beta}}$s to obtain a reasonable trade-off between modeling objects and clutter or noise.

5. CONCLUSIONS

We have addressed the problem of human detection using a 2D array of acoustic receivers for mobile platforms such as robots. We have argued that humans have distinct acoustic signatures in the spatial power spectrum. We have proposed to generically model the echoes from reflecting parts of the object by a Gaussian-Mixture-Model. This allows a reliable representation of acoustic signatures in a low-dimensional parameter space, which can be used for a subsequent detection and classification of humans.



(a) Walking person, orientation 180°

(b) Office chair, orientation 0°

Fig. 4: Reconstructed images based on the estimated parameters for a person (a) and an office chair (b).

6. REFERENCES

[1] R. Cutler and L.S. Davis, “Robust real-time periodic motion detection, analysis, and applications,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, 2000.

[2] Z. Li, K. Wang, L. Li, and F.-Y. Wang, “A review on vision-based pedestrian detection for intelligent vehicles,” in *IEEE Int. Conf. Vehicular Electronics and Safety (ICVES)*, 2006.

[3] N. Yamada, Y. Tanaka, and K. Nishikawa, “Radar cross section for pedestrian in 76GHz band,” in *European Microwave Conference*, 2005, vol. 2.

[4] M. Moebus and A.M. Zoubir, *In-Vehicle Corpus and Signal Processing for Driver Behavior*, chapter 13, Springer, 2008.

[5] M. Moebus and A.M. Zoubir, “Three-dimensional ultrasound imaging in air using a 2D array on a fixed platform,” in *Proc. of the IEEE Int’l Conference on Acoustics, Speech and Signal Processing*, 2007, vol. 2.

[6] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 1st edition, October 2007.

[7] A. Blake and M. Isard, *Active Contours: The Application of Techniques from Graphics,Vision,Control Theory and Statistics to Visual Tracking of Shapes in Motion*, Springer-Verlag New York, Inc., Ney York USA, 1998.

[8] J. A. Sethian, *Level Set Methods and Fast Marching Methods: Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision, and Materials Science*, Cambridge University Press, June 1999.