Sound Localization for Humanoid Robots - Building Audio-Motor Maps based on the HRTF

Jonas Hörnstein, Manuel Lopes, José Santos-Victor
Instituto de Sistemas e Robótica
Instituto Superior Técnico
Lisboa, Portugal
Email: {jhornstein,macl,jasv}@isr.ist.utl.pt

Francisco Lacerda
Department of Linguistics
Stockholm University
Stockholm, Sweden
Email: frasse@ling.su.se

***Abstract*— Being able to locate the origin of a sound is important for our capability to interact with the environment. Humans can locate a sound source in both the horizontal and vertical plane with only two ears, using the head related transfer function HRTF, or more specifically features like interaural time difference ITD, interaural level difference ILD, and notches in the frequency spectra. In robotics notches have been left out since they are considered complex and difficult to use. As they are the main cue for humans' ability to estimate the elevation of the sound source this have to be compensated by adding more microphones or very large and asymmetric ears. In this paper, we present a novel method to extract the notches that makes it possible to accurately estimate the location of a sound source in both the horizontal and vertical plane using only two microphones and human-like ears. We suggest the use of simple spiral-shaped ears that has similar properties to the human ears and make it easy to calculate the position of the notches. Finally we show how the robot can learn its HRTF and build audio-motor maps using supervised learning and how it automatically can update its map using vision and compensate for changes in the HRTF due to changes to the ears or the environment.**

I. Introduction

Sound plays an important role in directing humans' attention to events in their ecological setting. The human ability to locate sound sources in potentially dangerous situations, like an approaching car, or locating and paying attention to a speaker in social interaction settings, is a very important component of human behaviour. In designing a humanoid robot that is expected to mimic human behaviour, the implementation of a human-like sound localization capability as a source of integrated information is therefore an important goal.

Humans are able of locating the sound sources in both the horizontal and vertical plane from exploring acoustic information conveyed by the auditory system, but in a robot that uses two simple microphones as ears there is not enough information to do the same. Typically the robot would be able to calculate or learn the positions of the sound source in the plane of the microphones, i.e. the azimuth which usually corresponds to the horizontal plane. This can be done by calculating the difference in time between the signal reaching the left and the right microphone respectively. This is called the interaural time difference (ITD) or the interaural phase difference (IPD) if we have a continuous sound signal and calculate the phase difference of the signal from the two microphones by cross-correlation. However, the ITD/IPD does not give any information about the elevation of the sound source. Furthermore it cannot tell whether a sound comes from the front or the back of the head. In robotics this is usually solved by adding more microphones. The SIG robot [1] [2] has four microphones even though two are mainly used to filter the sound caused by the motors and the tracking is mainly done in the horizontal plane. In [3] eight microphones are used, and in [4][5] a whole array of microphones is used to estimate the location of the sound.

While adding more microphones simplifies the task of sound localization, humans and other animals manage to locate the sound with only two ears. This comes from the fact that the form of our head and ears change the sound as a function of the location of the sound source, a phenomenon known as the head related transfer function (HRTF). The HRTF describes how the free field sound is changed before it hits the eardrum, and is a function $H(f, \theta, \phi)$ of the frequency, f, the horizontal angle, θ, and the vertical angle, ϕ, between the ears and sound source. The IPD is one important part of the HRTF. Another important part is that the level of the sound is higher when the sound is directed straight into the ear compared to sound coming from the sides or behind. Many animals, like cats, have the possibility to turn their ears around in order to get a better estimate of the localization of the sound source. Even without turning the ears, it is possible to estimate the location of the sound by calculating the difference in level intensity between the two ears. This is referred to as the interaural level difference (ILD). However, if the ears are positioned on each side of the head as for humans, ILD will mainly give us information about on which side of the head that the sound source is located, i.e. information about the azimuth which we already have from the ITD/IPD. In order to get new information from the ILD we have to create an asymmetry in the vertical plane rather than in the horizontal. This can be done by putting the ears on top of the head and letting one ear be pointing up while the other is pointing forwards as done in [6]. The problem with this approach is that a big asymmetry is needed to get an acceptable precision and ILD of human-like ears does not give sufficient information about the elevation of the sound source.

For humans it has been found that the main cue for

1-4244-0259-X/06/$20.00 ©2006 IEEE

estimating the elevation of the sound source comes from resonances and cancellation (notches) of certain frequencies due to the pinna and concha of the ear. This phenomenon has been quite well studied in humans both in neuroscience and in the field of audio reproduction for creating 3D-stereo sound [7][8][9][10][11][12][13], but has often been left out in robotics due to the complex nature of the frequency response and the difficulty to extract the notches. In this paper we suggest a simple and effective way of extracting the notches from the frequency response and we show how a robot can use information about ITD/IPD, ILD, and notches in order to accurately estimate the location of the sound source in both vertical and horizontal space.

Knowing the form of the head and ears it is possible to calculate the relationship between the features (ITD, ILD, and the frequencies for the notches) and the position the sound source, or even estimate the complete HRTF. However, here we are only interested in the relationship between the features and the position. Alternatively we can get the relationship by measuring the value of the features for some known positions of the sound source and let the robot learn the maps. Since the HRTF changes if there is some changes to the ears or microphones or if some object like for example a hat is put close to the ears, it is important to be able to update the maps. Indeed, although human ears undergo big changes from birth to adulthood, humans are capable of adapting their auditory maps to compensate for acoustic consequences of the anatomical changes. It has been shown that vision is an important cue for updating the maps [14], and it can also be used as a mean for the robot to update its maps [6].

In the rest of this paper, Section 2 will show how we can design a simple head and ears that give a HRTF similiar to the human head and ears' HRTF. In Section 3 we discuss how to extract features such as ITD/IPD, ILD, and notches from the signals provided from the ears. In Section 4 we show how the robot can use the features to learn its audio-motor-map. In Section 5 we show some experimental results. Conclusions and directions for future work are given in Section 6.

II. Design of Head and Ears

In this section we want to describe the design of a robot's head and ears with a human-like HRTF, and hence with acoustic properties for the ITD, ILD, and frequency notches similar to those observed in humans. The HRTF depends on the form of both the head and the ears. The ITD/IPD depends on distance between the ears and the ILD is primarily dependent on the form of the head and to less extent also the form of the ears, while the peaks and notches in the frequency response mainly are related to the form of the ears.

For the sake of calculating the HRTF, a human head can be modeled by a spheroid [15] [16]. The head used in this work is the iCUB head which is close enough to a human head to expect the same acoustic properties. The detailed design of the head is described in [17] but here we can simply consider it a sphere with a diameter of 14 cm.

The ears are more complex. Each of the robot's ears was built by a microphone placed on the surface of the head and a reflector simulating the pinna/concha, as will be described in detail below. The shape of human ears differs substantially between individuals, but a database with HRTF for different individuals [18] provides some general information on the frequency spectra created for various positions of the sound source by human ears. Obviously one way to create ears for a humanoid robot would be simply to copy the shape of a pair of human ears. That way we can assure that they will have similar properties. However we want to find a shape that is easier to model and produce while preserving the main acoustic characteristics of the human ear. The most important property of the pinna/concha, for the purpose of locating the sound source, is to give different frequency responses for different elevation angels. We will be looking for notches in the frequency spectra, created by interferences between the incident waves, reaching directly the microphone, and their reflections by the artificial concha, and want the notches to be produced at different frequencies for different elevations. A notch is created when a quarter of the wavelength of the sound, λ, (plus any multiple of $\lambda/2$) is equal to the distance, d, between the concha and the microphone:

$$n * \frac{\lambda}{2} + \frac{\lambda}{4} = d \, (n = 0, 1, 2, ...)$$

For these wavelengths, the sound waves that reach the microphone directly are cancelled by the waves reflected by the concha. Hence the frequency spectra will have notches for the corresponding frequencies:

$$f = \frac{c}{\lambda} = \frac{(2 * n + 1) * v}{4 * d} \quad (1)$$

$$c = \{speed\, of\, sound\} \approx 340\, m/s \quad (2)$$

To get the notches at different frequencies for all elevations we want an ear-shape that has different distance between the microphone and the ear for all elevations. Lopez-Poveda and Meddis suggest the use of a spiral shape to model human ears and simulate the HRTF [19]. In a spiral the distance between the microphone, placed in the center of the spiral, and the ear increases linearly with the angle. We can therefore expect the position of the notches in the frequency response to also change linearly with the elevation of the sound source.

We used a spiral with the distance to the center varying from 2 cm below to 4 cm in the top, Figure 1. That should give us the first notch at around 2800 Hz for sound coming straight from the front and with the frequency increasing linearly as the elevation angle increases, Figure 2. When the free field sound is white noise as in the figure it is easy to find the notches directly in the frequency spectra of either ear. However, sound like spoken language will have its own maxima and minima in the frequency spectra depending on what is said. It is not clear how humans separate what is said from where it is said [20]. One hypothesis is that we perform a binaural comparison of the spectral patterns, as have also been suggested for owls [21]. Both humans and owls have small asymmetries between

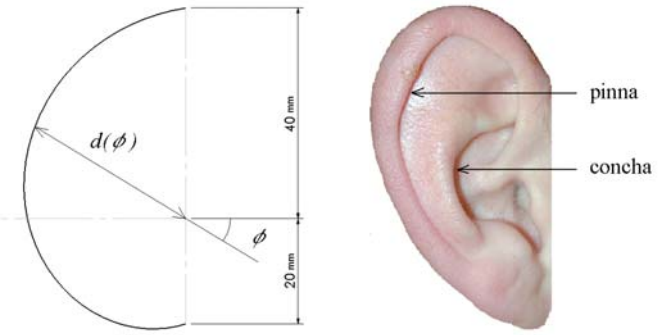

Fig. 1. Pinna and concha of a human ear (right), and the artificial pinna/concha (left)

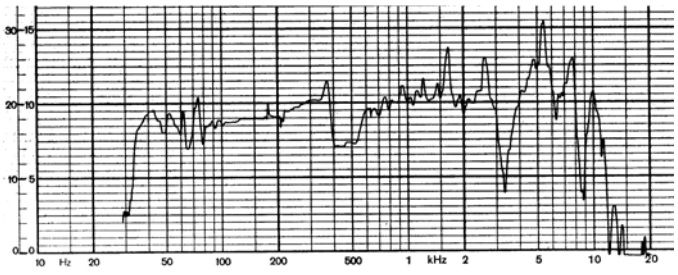

Fig. 2. Example of the HRTF for a sound source at around 20 degrees above

the left and right ear that can give excellent cues to vertical localization in the higher frequencies. These relatively small asymmetries that provide different spectral behaviour between the ears should not be confused with the large asymmetries needed to give any substantial difference for the ILD. Here we only need the difference in distance between the microphone and the pinna for the right and left ear to be enought to separate the spectral notches. In the optimal case we would like the right ear to give a maximum for the same frequency that the left ear has a notch and hence amplify that notch. This can be done by choosing the distance for the right ear, d_r, as:

$$d_r(\phi) = 2\frac{m_r + 1}{2 * n_l + 1} * d_l(\phi)$$

where m_r=maxima number for right ear, n_l=notch number for left ear, and d_l=distance between the microphone and ear for left ear.

If we for example want to detect the third notch of the left ear, and want the right ear to have its second maxima for the same frequency, we should choose the distance between the microphone and ear for the right ear as:

$$d_r(\phi) = 2\frac{2 + 1}{2 * 3 + 1} * d_l(\phi) = 6/7 * d_l(\phi)$$

In the case of two identical ears we can not have a maxima of the right ear at the same place as the left ear has a notch for all elevations. The best we can do is to choose the angle between the ear so that the right ear has a maxima for the wanted notch when the sound comes from the front. In the specific case of the ears in Figure 1 we choose an angle of 18 degrees. Prototypes for the ears were constructed from a thin metal band which was easy to bend into a spiral shape. The final ears were made from plastic and designed to esthetically fit the head of the iCub, Figure 3.

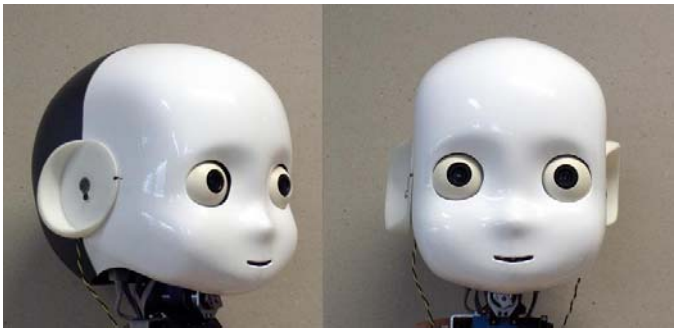

Fig. 3. The iCub with spiral shaped ears

III. Features

As discussed in the introduction, humans mainly depend on three different features for locating the sound source: the interaural time difference ITD, the interaural level difference ILD, and the notches in the frequency response of each ear. In this section we show how to extract the features given the signals $s_l(t)$ and $s_r(t)$ from the left and right ear respectively.

The first step is to sample the sound. We sample the sound for 1/10 s using the sample frequency F_s=44100 Hz. This gives us k=4410 samples. We then calculate the mean energy as a sum of square of the given samples divided by the number of samples: $\sum_k \left(s_l^2(k) + s_r^2(k)\right)/k$. A simple threshold value is used to decide if the sound has enough energy to make it meaningful to try to extract the features and try to locate the sound source. The calculation of the individual features is explained below.

A. Interaural time difference, ITD

The interaural phase difference is calculated by doing a cross-correlation between the signals arriving to the left and right ear/microphone. If the signals have the same shape we can expect to find a peak in the cross-correlation for the number of samples that corresponds to the interaural time difference, i.e. the difference in time at which the signal arrives at the microphones. We can easily find this by searching for the maximum in the cross correlation function. Knowing the sampling frequency F_s and the number of samples n that corresponds to the maximum in the cross-correlation function we can calculate the interaural time difference as:

$$ITD = \frac{n}{F_s}$$

If the distance to the sound source is big enough in comparison to the distance between the ears, l, we can approximate the incoming wave front with a straight line and the difference in distance Δl traveled by the wave for the left and right ear can easily be calculated as:

$$\Delta l = l \sin(\theta)$$

where θ is the horizontal angle between the head's mid sagittal plane and the sound source. Knowing that the distance traveled is equal to the time multiplied with the velocity of the sound we can now express the angle directly as a function of the ITD:

$$\theta = \arcsin\left(ITD * \frac{c}{l}\right)$$

However for the sake of controlling the robot we are not interested in the exact formula since we want the robot to be able to learn the relationship between the ITD and the angle rather than hard coding this into the robot. The important thing is that there exists a relationship that we will be able to learn. We therefore measured the ITD for a number of different angles in an anechoic room, figure 4.

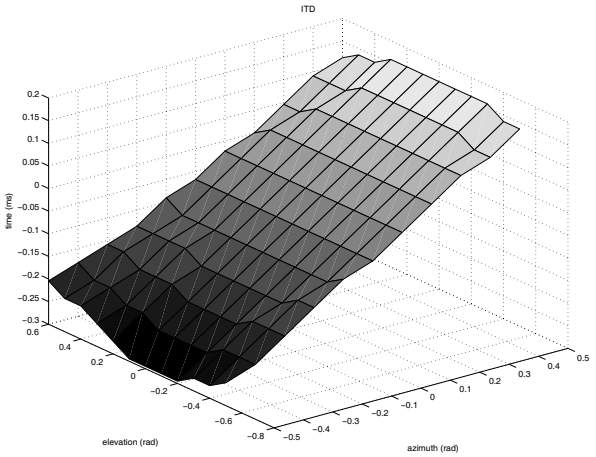

Fig. 4. ITD for different positions of the sound source

B. Interaural level difference, ILD

The interaural level difference ILD, is calculated as a function of the avarage power of the sound signals reaching the left and right ear.

$$ILD = 10 * \log_{10}\left(\frac{\sum_k s_l^2(k)}{\sum_k s_r^2(k)}\right)$$

Sometimes the ILD is calculated from the frequency response rather than directly from the temporal signal. It is easy to go from the temporal signal to the frequency response by applying a fast Fourier transform FFT. The reason for working with the frequency response instead of the temporal signal is that it makes it easy to apply a high-pass, low-pass, or band-filter on the signal before calculating its average power. Different frequencies have different properties. Low frequencies typically pass more easily through the head and ears while higher frequencies tend to be reflected and their intensity more reduced. One type of filtering that is often used is dBA which corresponds to the type of filtering that goes on in human ears and which mainly takes into account the frequencies between 1000 Hz and 5000 Hz. In [6] a band-pass filter between 3-10 kHz have been used which gives them a better calculation of ILD. Different types of head and ears may benefit from enhancing different frequencies. Here we calculate the ILD directly from the temporal signal which is equivalent to considering all frequencies. The response for a sound source placed at different angles from the head is shown in figure 5.

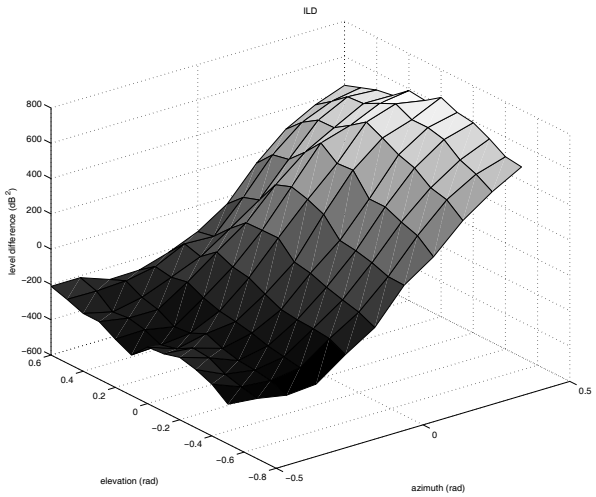

Fig. 5. ILD for different positions of the sound source

C. Spectral notches

Existing methods for extracting spectral notches such as [22][23][24] focus on finding the notches in spectral diagrams obtained in anechoic chambers using white noise, such as the diagrams presented in Figure 2. For a humanoid robot that has to be able to turn towards any type of sound these methods are not suitable. In this section we suggest a novel method to extract the frequency notches that is reasonable fast and simple to implement while giving better accuracy for calculating the elevation of the sound source than methods based on ILD. The method makes use of the fact that we have a slight asymmetry between the ears and has the following steps:

1) Calculate the power spectra density for each ear
2) Calculate the interaural spectral differences
3) Fit a curve to the resulting differences
4) Find minima for the fitted curve

To calculate the power spectra we use the Welch spectra [25]. Typical results for the power spectra density, $H_l(f)$ and $H_r(f)$, for the left and right ear respectively are shown in figure 6. As seen the notches disappears in the complex spectra of the sound, which makes it very hard to extract them directly from the power spectra. To get rid of the maxima and minima caused by the form of the free field sound, i.e. what is said, we calculate the interaural spectral difference as:

$$\Delta H(f) = 10 * \log_{10} H_l(f) - 10 * \log_{10} H_r(f) =$$
$$= 10 * \log_{10}\left(\frac{H_l(f)}{H_r(f)}\right)$$

Finally we fit a 12 degree polynomial to the interaural spectral difference. The interaural spectral difference and the fitted polynomial for zero azimuth and elevation, are shown in Figure 6. From the fitted function it is easy to extract the minima. In this work we have chosen to use the first minima over 5000 Hz as a feature. In this specific setup that corresponds to the second notch. However the position of the notches depends on the design of the ears and for other types of ears other frequencies will have to be used.

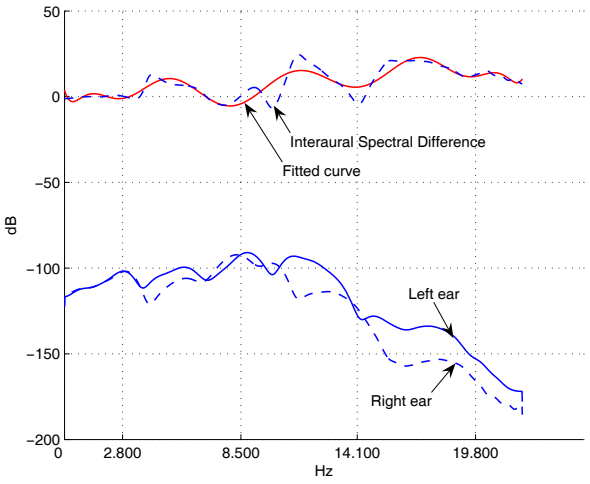

Fig. 6. The interaural spectral difference (dotted line above) and the fitted curve (solid line above), and the original power spectra for left ear (solid line below) and right ear (dotted line below). Note that the spectras are shown in dB, i.e. $10*log(H_x(f))$. The vertical lines represent the expected frequencies for the notches.

As seen in Figure 6 the minima more or less corresponds to the expected frequencies of the notches for the left ear. This is because we carefully designed the ears so that the notches from the two ears would not interfere with each other. In this case we could actually calculate the relationship between the frequency of the notch and the position of the sound source. However, in the general case it is better to let the robot learn the HRTF than to actually calculate it since the position of the notches is critically affected by small changes in the shape of the conchas or the acoustic environment. Also, if we learn the relationship rather than calculating it we do not have to worry about the fact that the minimas that we find do not directly correspond to the notches as long as they change with the elevation of the sound source. In figure 7 we show the extracted feature with the sound source placed at a number of different positions in relation to the head.

IV. Audio-motor maps

In this section we are going to map sound features to the corresponding localization of sound sources. We present solutions to: i) learn this map, ii) use it to control a robot toward a target and iii) to improve its quality online.

As seen before, differences in the head size, ear shapes or even a hat in the robot can dramatically change the frequency

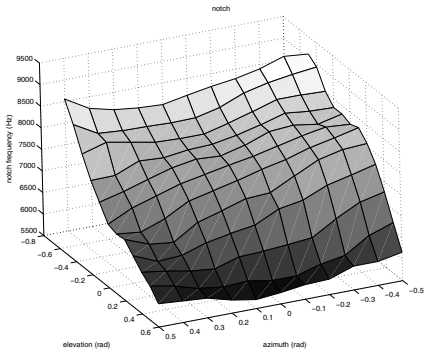

Fig. 7. Notch frequencies for different position of the sound source

response of the microphones. The coordinate transformation from the ears to the vision and the control motors is difficult to calibrate. Because of this, the system should be able to adapt to these variations. The exact solution we have for the pan localization is only valid when the sound source is in the same plane as the receptors, by having a learning mechanism we can combine the vertical and horizontal information to have a better localization.

A map m from the sound features $\mathcal{C}$ to its location written in head spherical coordinates S_H can be used to direct the head toward a sound source. This map can be represented by $S_H = m(\mathcal{C})$. A simple way to move the head toward the target is to move the head pan and tilt by an increment $\Delta\theta$ equal to the position of the sound source, i.e. $\Delta\theta = S_H$.

Although the function is not linear, if we restrict to the space of motion of the head, we have an almost linear relationship between the position of the head and the ITD and notch frequency. This can be seen in the Figures 4 and 7. There is a more complex relationship between the position of the head and the ILD, Figure 5. However, since the ITD and notches already contain enought information to calculate both the pan and the tilt, we ILD will only give us redundant information. To get a simpler model that allows a faster and more robust learing, we have left out the ILD and approximate the nonlinear function by a linear function:

$$S_H = M\mathcal{C}$$

To estimate the value of M a linear regression with the standard error criteria was selected:

$$\hat{M} = \arg\min_{M} \sum_{i=1} \|\Delta\theta - M\mathcal{C}_i\|^2 \tag{3}$$

This solution gives an offline batch estimate, for online estimation, a Broyden update rule [26] was used. This rule is very robust, fast, has just one parameter and only keeps in memory the actual estimative of M. Its structure is as follows:

$$\hat{M}(t+1) = \hat{M}(t) + \alpha \frac{\left(\Delta\theta - \hat{M}(t)\mathcal{C}\right)\mathcal{C}^\top}{\mathcal{C}^T\mathcal{C}} \tag{4}$$

where α is the learning rate. This method is useful to be used in a supervised learning method, where the positions corresponding to a certain sound are given. This is not the case for an autonomous system. A robot needs an automatic feedback mechanism to learn the audio-motor relation autonomously. Vision can provide information in a robust and non-intrusive way. As the goal of this robot is to interact with humans the test sound will be produced by humans and so the robot knows the visual appearance of the sound source. A face detection algorithm based on [27] and [28] was used. After hearing a sound the robot moves to the position given by the map. If the human head is not centered in the image then a visual servoing behavior is activated in order to bring the observed face to the image center, for this the robot is controlled with the following rule:

$$\dot{\theta} = J_H^+ \dot{F}_p$$

where $\dot{\theta}$ is the velocity for the head motors, J_H^+ is the pseudo-inverse of the head jacobian and $\dot{F}_p$ is the desired motion of the face in the image. When the face is in the center of the image the audio-motor map can be updated using the learning rule of Eq. 4. Table I presents the final algorithm.

TABLE I

ALGORITHM FOR AUTONOMOUSLY LEARNING A AUDIO-MOTOR MAP BY INTERACTING WITH A HUMAN

1) listen to sound
2) move head toward the sound using the map
3) locate human face in the image
4) if face not close enough to the center
   a) do a visual servoing loop to center the face
   b) update the map

V. EXPERIMENTAL RESULTS

We acquired a dataset in a silent room with a white-noise sound-source located $1.5m$ from the robot. We recorded $1\ second$ of sound in 132 different head positions θ by moving the head with its own motors. A set of features was evaluated from this data consisting of ITD and the notch frequency evaluate on $0.1\ second$. Figure 7 shows the resulting feature surfaces after averaging them on 11 samples. This is used as the training dataset.

A second dataset was created to test the learning method. The procedure was similar to the previous one but the sound-source was replaced by a human voice sound. This was done because the system should operate in an human-robot interaction and also to evaluate the generalization properties of the method.

The map was then estimated with the optimization problem in Eq. 3, Figure 8 presents the reconstruction error by showing for each head position the corresponding error in reconstruction. We can see that the worst case corresponds to the joint limits of the head but it is always less than $0.1\ rad$, which is very small. As a comparison we can note that $0.1\ rad$ is the size of a adult human face when seen from $1.5\ m$ of distance. The error increase near the limits is due to the non-linearity of the features being approximated by a linear model, however with this small error the computational efficiency and robustness makes us choose this model.

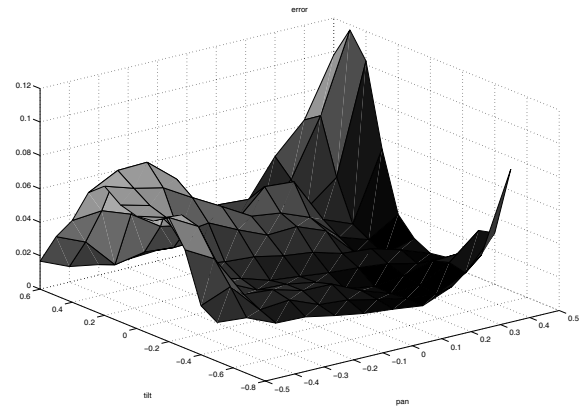

Fig. 8. Audio-motor map reconstruction error for each head position

Finally we have done a test in a real world environment and interaction. The experiment was done in our offices having a trainer calling the robot at random positions within the full movement of the head. The previously learned map was used as a bootstrap, but to make the learning task more interesting we manually introduced an offset of 0.4 radians in the map for both the pan and tilt estimates. Even with the introduced offset, the estimated position was good enough to guaranty that the trainer's face would be within the visual field of the robot, allowing us to use the algorithm presented in Table I. Figure 9 presents the evolution of the error during the experiment. The error represents the difference, in radians, from the position mapped by the model and the real location of the person. We can see that learning algorithm manages to bring the error back to around 0.1 radians.

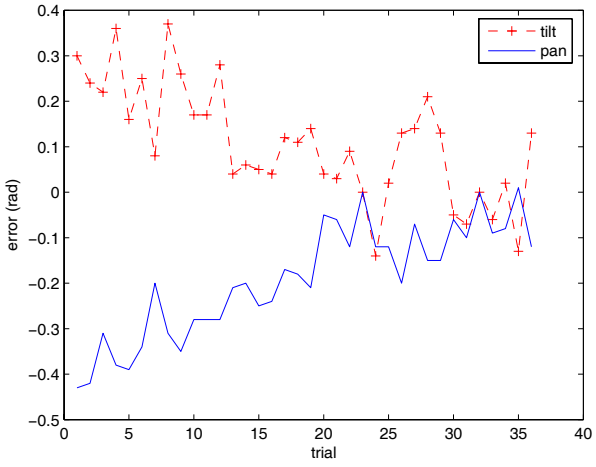

Fig. 9. Convergence rate of the audio-motor map learning algorithm when running online with feedback given by a face detector.

VI. Conclusion

In this paper we have presented a novel method for estimating the location of a sound source using only two ears. The method is inspired by the way humans estimate the location from features such as interaural time difference ITD, interaural level difference ILD, and spectral notches. We suggest the use of spiral formed ears of slightly different size or with different inclination to make it easy to extract the notches. The spiral form also makes it easy to mathematically derive the HRTF for the robot, thus making it possible to simulate the features and/or building controllers based on the HRTF.

We have also shown that the robot can learn the HRTF either by supervised learning or by using vision. Initial audio-motor maps either calculated or learnt combined with an online vision-based update of the maps are suggested for the control of the robot. This makes it possible for the robot to compensate for small changes in the HRTF caused by dislocations of the ears, exchange of microphones, or placement of objects like a hat close to the ear.

The suggested method has good accuracy within the possible movements of the head used in the experiments. The error in the estimated azimuth and elevation is less than 0.1 radians for all angles, and less than 0.02 radians for the center position.

The method is especially suitable for humanoid robots where we want ears that both look like human ears and perform like them. The precision in the estimate of the position is more than enough for typical applications for a humanoid like turning towards the sound.

Future work includes resolving front-back ambiguities and learning the HRTF for wider angles than the movements of the head which demands other approximations of the audio-motor map than linear.

Acknowledgment

Work partially supported by EU Project NEST-5010 CONTACT, by a FCT Ph.D. scholarship and by the FCT POS Conhecimento Program that includes FEDER funds. We would also like to thank Lisa Gustavsson, Eeva Klintfors, Ellen Marklund, Peter Branderud, and Hassan Djamshidpey at the Department of Linguistics, Stockholm University, and Rafael Serrenho and Onofre Moreira at Centro de Análise e processamento de sinais, IST, who helped us with the measurements in the anechoic chamber.

References

[1] H. Okuno, K. Nakadai, and H. Kitano, "Social interaction of humanoid robot based on audio-visual tracking," in *Proc. of 18th Intern. Conf. on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE-2002)*, June 2002.

[2] H. Okuno, "Human-robot interaction through real-time auditory and visual multipletalker tracking," in *IROS*, 2001.

[3] J. Valin, F. Michaud, J. Rouat, and D. Létourneau, "Robust sound source localization using a microphone array on a mobile robot," in *Proceedings International Conference on Intelligent Robots and Systems*, 2003.

[4] D. Sturim, H. Silverman, and M. Brandstein, "Tracking multiple talkers using microphone-array measurements," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'97) - Volume 1*, 1997, p. 371.

[5] K. Guentchev and J. Weng, "Learning based three dimensional sound localization using a compact non-coplanar array of microphones," in *AAAI Spring Symp. on Int. Env.*, Stanford CA, March 1998.

[6] S. G. Natale L., Metta G., "Development of auditory-evoked reflexes: Visuo-acoustic cues integration in a binocular head," *Robotica and Autonomous Systems*, vol. 39, pp. 87–106, 2002.

[7] V. Algazi, R. Duda, R. Morrison, and D. Thompson, "Structural composition and decomposition of hrtfs," in *Proc. IEEE WASPAA01*, New Paltz, NY, 2001, pp. 103–106.

[8] B. Gardner and K. Martin, "Hrtf measurements of a kemar dummy-head microphone," MIT Media Lab Perceptual Computing, Tech. Rep. 280, May 1994.

[9] S. Hwang, Y. Park, and Y. Park, "Sound source localization using hrtf database," in *ICCAS2005*, Gyenggi-Do, Korea, June 2005.

[10] L. Savioja, J. Houpaniemi, T. Huotilainen, and T. Takala, "Real-time virtual audio reality," in *Proceedings of ICMC*, 1986, pp. 107–110.

[11] J. Huopaniemi, L. Savioja, and T. Takala, "Diva virtual audio reality system," in *Proc. Int. Conf. Auditory Display (ICAD'96)*, Palo Alto, California, Nov 1996, pp. 111–116.

[12] C. Avendano, R. Duda, and V. Algazi, "Modeling the contralateral hrtf," in *Proc. AES 16th International Conference on Spatial Sound Reproduction*, Rovaniemi, Finland, 1999, pp. 313–318.

[13] B. C. J. Moore, "Interference effects and phase sensitivity in hearing," *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, vol. 360, no. 1794, pp. 833 – 858, May 2002.

[14] M. Zwiers, A. V. Opstal, and J. Cruysberg, "A spatial hearing deficit in early-blind humans," *JNeurosci*, vol. 21, 2001.

[15] V. Algazi, C. Avendano, and R. Duda, "Estimation of a spherical-head model from anthropometry," *J. Aud. Eng. Soc.*, 2001.

[16] R. Duda, C. Avendano, and V. Algazi, "An adaptable ellipsoidal head model for the interaural time difference," in *Proc. ICASSP*, 1999. [Online]. Available: citeseer.csail.mit.edu/duda99adaptable.html

[17] R. Beira, M. Lopes, M. Praça, J. Santos-Victor, A. Bernardino, G. Metta, F. Becchi, and R. Saltaren, "Design of the robot-cub (icub) head," in *IEEE ICRA*, 2006.

[18] V. Algazi, R. Duda, and D. Thompson, "The cipic hrtf database," in *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, NY, USA, 2001.

[19] E. A. Lopez-Poveda and R. Meddis, "Sound diffraction and reflections in the human concha," *J. Acoust. Soc. Amer.*, vol. 100, pp. 3248–3259, 1996.

[20] C. Jin, A. Corderoy, S. Carlile, and A. van Schaik, "Contrasting monaural and inteaural spectral cues for human sound localization," *J. Acoust. Soc. Am.*, vol. 6, no. 115, pp. 3124–3141, June 2004.

[21] R. A. Norberg, "Skull asymmetry, ear structure and function, and auditory localization in tengmalm's owl, aegolius funereus (linne)," *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, vol. 282, no. 991, pp. 325–410, Mar 1978.

[22] S. G. R. M. A.Ramírez, "Extracting and modeling approximated pinna-related transfer functions from hrtf data," in *Proceedings of ICAD 05-Eleventh Meeting of the International Conference on Auditory Display*, Limerick, Ireland, July 2005.

[23] V. C. Raykar, R. Duraiswami, L. Davis, and B. Yegnanarayana, "Extracting significant features from the hrtf," in *Proceedings of the 2003 International Conference on Auditory Display*, Boston, MA, USA, July 2003.

[24] V. C. Raykar and R. Duraiswami, "Extracting the frequencies of the pinna spectral notches in measured head related impulse responses," *J. Acoust. Soc. Am.*, vol. 118, no. 1, pp. 364–374, July 2005.

[25] P. Welch, "The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Trans. Audio Electroacoust.*, vol. AU-15, pp. 70–73, June 1967.

[26] R. Fletcher, *Practical Methods of Optimization*, 2nd ed. Chichester, 1987.

[27] P. Viola and M. J. Jones, "Rapid object detection using a boosted cascade of simple features," *IEEE CVPR*, 2001.

[28] R. Lienhart and J. Maydt, "An extended set of haar-like features for rapid object detection," in *IEEE ICIP*, Sep 2002, pp. 900–903.