Proceedings of the 2007 IEEE Symposium on Computational Intelligence in Image and Signal Processing (CIISP 2007)

ROBOTIC LOCALIZATION AND SEPARATION OF CONCURRENT SOUND SOURCES USING SELF-SPLITTING COMPETITIVE LEARNING

Fakheredine Keyrouz, Werner Maier, and Klaus Diepold

Technische Universität München
80290 Munich, Germany
{keyrouz,kldi}@tum.de

ABSTRACT

We combine binaural sound-source localization and separation techniques for an effective deployment in humanoid-like robotic hearing systems. Relying on the concept of binaural hearing, where the human auditory 3D percepts are predominantly formed on the basis of the sound-pressure signals at the two eardrums, our robotic 3D localization system uses only two microphones placed inside the ear canals of a robot head equipped with artificial ears and mounted on a torso. The proposed localization algorithm exploits all the binaural cues encapsulated within the so-called Head Related Transfer Functions (HRTFs). Taking advantage of the sparse representations of the ear input signals, the 3D positions of two concurrent sound sources is extracted. The location of the sources is extracted after identifying which HRTFs they have been filtered with using a well-known self-splitting competitive learning clustering algorithm. Once the location of the sources are identified, they are separated using a generic HRTF dataset. Simulation results demonstrated highly accurate 3D localization of the two concurrent sound sources, and a very high Signal-to-Interference Ratio (SIR) for the separated sound signals.

Index Terms— sound localization, source separation, HRTF, self-splitting competitive learning.

1. INTRODUCTION

Many biological organisms have evolved intelligent and efficient means for acoustic communication. Adaptation and optimization can be found in all components of the acoustic communication system: signal generation at the sender is optimized in such a way that the signal characteristics are tailored to the transmission channel. Receivers have developed sophisticated mechanisms for segregating the signals from different sound sources, and for analysing signal characteristics. The acoustics of the environment frequently imposes similar demands on the mechanisms for auditory analysis in different animal species. Thus, mechanisms of auditory analysis show a number of similarities in different animal species, ranging from insects to mammals. These similarities result either from convergent evolution of auditory systems that are selected to achieve a similar performance, given similar environmental conditions, or they are simply the consequence of the preservation of structures in evolutionary history [1].

In many everyday listening situations, humans benefit from having two ears, naturally evolved, to analyze concurrent sound sources in various listening environments. For more than a century, research has been conducted to understand which acoustic cues are resolved by the auditory system to localize sounds and to separate concurrent sounds. The term "binaural hearing" refers to the mode of functioning of the auditory system of humans or animals using two ears. These ear organs segregate acoustic cues to solve tasks related to auditory localization, detection, or recognition.

The process the auditory system undergoes in combining the single cues, of the impinging sound waves at the ear drums, to a single, or mutliple, auditory event is not trivial. This holds true, in particular since many psycho-acoustical details are still unknown, e.g., how the single cues have to be weighted in general. It also remains unclear whether the Interaural Time Difference (ITD) and Interaural Level Difference (ILD) cues are combined, in the central nervous system, before or after they are spatially mapped.

In humans, the term cocktail-party effect denotes the fact that listeners with healthy binaural hearing capabilities are able to concentrate on one talker in a crowd of concurrent talkers and discriminate the speech of this talker from the rest. Also, binaural hearing is able to suppress noise, reverberance and sound coloration to a certain extent.

In robotics, on the other hand, efficient and accurate binaural 3D localization of several sound sources is quite a challenging task. In the recent years, a good number of algorithms have been proposed to tackle this problem. Basically, most of the detection methods used rely on microphone arrays, where the number of microphones is more or equal to the number of sound sources to be localized concurrently in 3D [2]. Among them, some approaches deal with simultaneous localization and

separation of sound sources [3]. However, a more intriguing, and naturally more demanding scenario, is localization of sound sources that outnumber the available number of microphones. Very few approaches were able to estimate the position of the sound sources, while only providing azimuth angles [4]. Humans and most mammals, however, are capable of detecting multiple concurrent sound sources with two ears by assessing monaural cues and binaural cues like interaural level differences (ILD), and interaural time/phase differences (ITD/IPD), in several frequency bands.

Motivated by the decoding process which the human auditory system performs when transforming the two signals at the eardrums back into a 3D space representation, it has been suggested that robotics can benefit from the intelligence encapsulated within the Head Related Transfer Function (HRTF) to localize sound in 3D using only two microphones [5]. Furthermore, it has been shown that humanoids can separate and, at the same time, localize two concurrent sound signals in free space [6]. Exploiting some relationships between source separation and system identification we were able to find the right HRTFs the sound sources were filtered with, and hence the corresponding position of the source, in terms of azimuth and elevation angles. While sound sources situated on two different hemispheres around the head of the robot were correctly localized, the algorithm in [5] failed to properly locate sound sources situated nearby each other in the same hemisphere. Furthermore, the algorithm was unable to localize more than two concurrent sound sources.

Recent investigations on the auditory space map of the barn owl, a predator with an astonishing ability to localize sound, revealed that the ILD/ITD cues cluster around two positions in the auditory map when two uncorrelated sound sources are simultaneously present. These clusters stem from time-frequency instances when one source predominates the other, i.e. has a stronger intensity [7].

Using this fact, we present in this paper a new algorithm for the localization of two concurrent sound sources based on estimating appropriate HRTFs, and their corresponding 3D locations, using sparse representations of the ear input signals of the KEMAR humanoid head. If the concurrent signals are sparse, which is naturally the case with speech signals, there must exist many instances when one source predominates the other. In such cases, the ear signals cluster around the actual HRTF, corresponding to the correct source location, in the single frequency bands. Comparing the estimated HRTFs to all KEMAR HRTFs [8], the new algorithm proved to be able to find the azimuth and elevation positions of the concurrent sources, situated in any given 3D position, using only two microphones.

2. TIME-FREQUENCY DOMAIN HRTF RECOVERY

2.1. Blind System Identification Framework

Our approach to binaural sound localization is based on finding the HRTFs, the sound sources were filtered with, on their way to the robot's microphones, which, in our case, play the role of to the human eardrums. Applying Short-Time Fourier Transform (STFT), we can describe the ear input signals with the following equations:

$$X_1(f,\tau) = \sum_{j=1}^{M} H_{1j}(f) S_j(f,\tau) \quad (1)$$

$$X_2(f,\tau) = \sum_{j=1}^{M} H_{2j}(f) S_j(f,\tau) \quad (2)$$

where τ denotes the time frame. The term M is the number of sound sources, and $S_j(f,\tau)$ are windowed sound source signals in frequency-domain. It is known that speech signals are very sparse in time-frequency domain, more than in time-domain [9]. However, frequency domain Independent Component Analysis (ICA) introduces the inherent permutation problem in each frequency bin, to which we will later present a solution. Since a sparse signal is almost zero in most time-frequency instances, there are a many instances when only one source is active. Hence, the ear input signals can be rewritten as:

$$X_1(f,\tau) = H_{1J}(f) \cdot S_J(f,\tau) \quad (3)$$

$$X_2(f,\tau) = H_{2J}(f) \cdot S_J(f,\tau) \quad J \in \{1,....M\} \quad (4)$$

Assuming stationary source positions, the HRTFs $H_{1J}(f)$ and $H_{2J}(f)$ are constant for all time instances τ. Since they are related to the source positions they are different for each source. This means ideally, that the time-frequency samples of the ear input signals, $X_1(f,\tau)$ and $X_2(f,\tau)$, that originate from the J-th source, cluster at each frequency f around the corresponding complex HRTFs values. Additionally, the Fourier transforms of the ear input signals are phase and amplitude normalized:

$$X_1(f,\tau) \leftarrow \frac{X_1(f,\tau)}{\sqrt{X_1^2(f,\tau) + X_2^2(f,\tau)}} e^{-\varphi_{X_1}} \quad (5)$$

$$X_2(f,\tau) \leftarrow \frac{X_2(f,\tau)}{\sqrt{X_1^2(f,\tau) + X_2^2(f,\tau)}} e^{-\varphi_{X_1}} \quad (6)$$

where φ_{X_1} is the phase corresponding to the input signal of the left ear microphone, which is chosen as a reference sensor. Figure 1 illustrates the clustering of the normalized data in the feature space.

2. Self-splitting competitive learning

As pointed out earlier, the source separation problem needs to be solved in each frequency bin. This means that, our algorithm clusters the data in all frequency bins over several

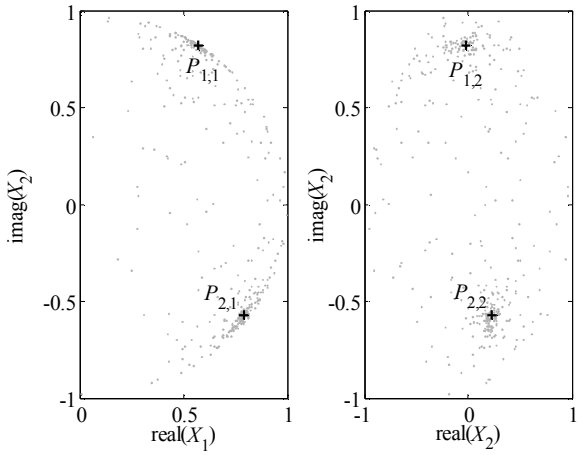

Fig. 1 Samples from the two ear microphones after STFT. The two subplots depict the real part of X_1 and the real part of X_2 versus the imaginary part of X_2, respectively. The data was gathered from 400 time-frames at a frequency of 538 Hz and normalized according to (5) and (6). The clusters show that there are two speakers present. Furthermore, the prototypes determined by Self-Splitting Competitive Learning are depicted in the cluster centers.

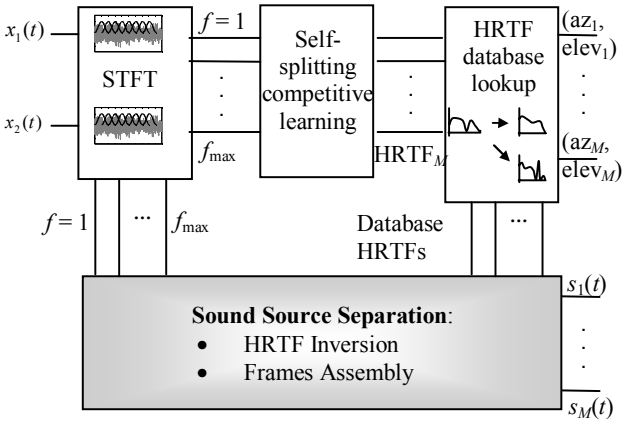

Fig. 2 The ear input signals are windowed and transformed to frequency domain. Afterwards, the self-splitting competitive learning algorithm finds the prototypes that represent the HRTFs in each frequency bin. By looking for the HRTFs that best match the estimated ones, the azimuth and elevation positions of the M sources are determined. With the aid of these HRTFs the sound sources are separated.

time frames separately. For further data analysis we use the clustering algorithm proposed in [10] which is based on self-splitting competitive learning (SSCL). In the following, we will briefly describe its principle.

The key issue in SSCL is the One-Prototype-Take-One-Cluster paradigm (OPTOC). this means that one prototype represents only one cluster. At first, a single prototype $\vec{P}_1 = [P_1\ P_2]^T, (P_1, P_2 \in C)$, is initialized randomly in the feature space. At the same time an asymptotic property vector (APV), $\vec{A}_1 = [A_1\ A_2]^T, (A_1, A_2 \in C)$, is created far away from the prototype. Its task is to guide the learning of the prototype making sure that, after some iterations, the prototype has settled in the center of a cluster (Fig. 3). The APV $\vec{A}_1$ defines a neighborhood around the prototype $\vec{P}_1$. If a randomly taken pattern, $\vec{X} = [X_1\ X_2]^T$, obtained using (5) and (6) lies within this neighborhood, it contributes to learning $\vec{A}_1$. It is observed that in the course of iterations, $\vec{A}_1$ moves towards $\vec{P}_1$. Learning stops when the Euclidean norm $\left\|\vec{P}_1 - \vec{A}_1\right\|$ falls below a constant ε_1. Now, in order to classify other clusters that may be present in the feature space, further prototypes have to be initialized. Hence, the following split validity criterion is introduced. If $\left\|\vec{P}_1 - \vec{C}_1\right\|$ is larger than a constant ε_2, a new prototype and a new APV are created in the feature space which are to lead to the center of another cluster (Fig. 4). The learning process starts anew. The term $\vec{C}_1$ is called the Center Property Vector (CPV) and determines the arithmetic mean of the input data points which have contributed to learning the prototype $\vec{P}_1$. In order to avoid unnecessary competition between the first and the new prototype, a distant property vector $\vec{R}_1$, adapted during the learning process, makes sure that the new prototype $\vec{P}_2$ is initialized far away from the first one. A detailed pseudo-code of the SSCL algorithm and update equations are given in [10]. A crucial key directly affecting the performance of the clustering algorithm is the choice of the two constants ε_1 and ε_2. As opposed to an adaptive choice, proposed in [9], that depends on the variances and number of elements of the clusters, we set ε_1 and ε_2 to a constant value, 0.01 in our case, and we confine the maximum number of prototypes to the number of present sound sources plus two. On the one hand, this has the disadvantage that the algorithm does not work completely blindly as the number of present sources has to be known but, on the other hand, it results in a robust classification of the clusters. The maximum number of clusters is chosen a little bit larger than the actual number of sources, since there are many data points, in the feature space, resulting from non-sparse time-frequency instances, see Fig. 1. These data points should not be represented by the prototypes in the center of the clusters, but by other prototypes. Of course, there has to be a criterion in order to choose the "right" prototypes that represent the HRTFs at a certain frequency.

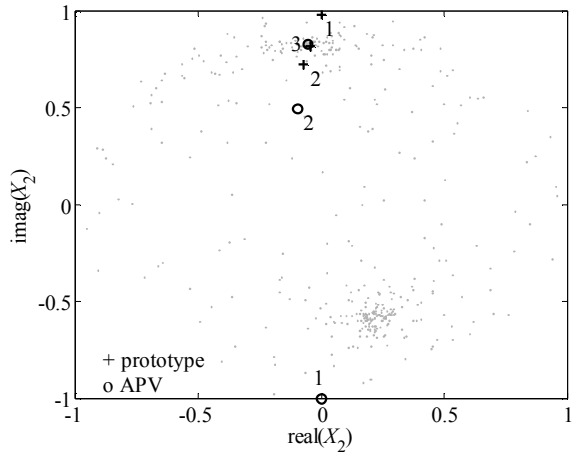

Fig. 3 Learning process of the first prototype. Step 1 shows the initialization of the second component of $\vec{P}_1$ and the APV $\vec{A}_1$. Step 2 shows their positions after 100 iterations. In step 3 the distance between $\vec{P}_1$ and $\vec{A}_1$ has fallen to 0.01 and learning stops.

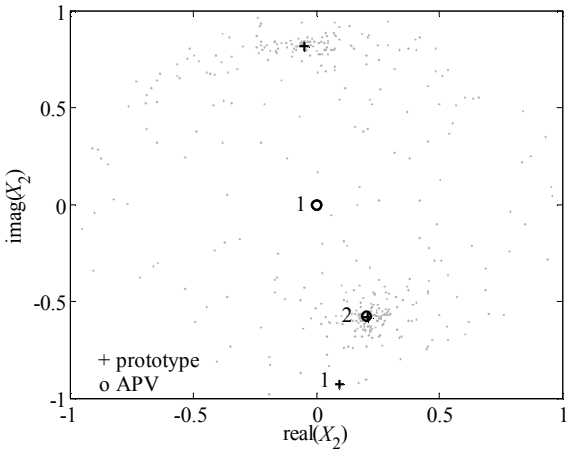

Fig. 4 Learning process of the second prototype $\vec{P}_2$ created when the first prototype has settled in the center of the topmost cluster. It is initialized together with an APV $\vec{A}_2$ at a certain distance from the first prototype (step 1), and is led to the center of the cluster at the bottom after some hundred iterations (step 2).

A criterion that proved to be appropriate goes as follows: If $\frac{N_i}{v_i} > \frac{1}{4}\max_i\left(\frac{N_i}{v_i}\right) \quad \forall\ i = 1,...,\textit{number of prototypes}$, then the *i*-th prototype represents a HRTF at a certain frequency. Hereby, N_i denotes the number of data points that have been assigned to the *i*-th prototype, and v_i is the variance of the cluster.

2.3. Solving the permutation problem

As mentioned earlier, we have to tackle the permutation problem introduced by the frequency-domain ICA. Once the clusters in the feature space in all frequency bins have been classified, one has to determine which of the clusters in the frequency bins belong to the same HRTF. Our approach is based on the assumption that the position of a cluster does not move a lot between adjacent frequency bins. The prototypes that remain, after applying the above-mentioned criterion in the frequency bin *f*, can be arranged in a matrix $\mathbf{H}(f) = \left[\vec{P}_1(f)\ \vec{P}_2(f)\ ....\ \vec{P}_{N_{\mathrm{Pr}}}(f)\right] \in C^{2xN_{\mathrm{Pr}}}$, Where N_{Pr} denotes the number of remaining prototypes that represent the HRTFs. Let $\wp = \{\mathbf{\Pi}_1, \mathbf{\Pi}_2, ..., \mathbf{\Pi}_{M!}\}$ be a group of permutation matrices of dimension *M x M*. Then, the correct permutation can be described by the following equations:

$$[\mathbf{d}_{i1}\ \mathbf{d}_{i2}\ ....\ \mathbf{d}_{iN_{\mathrm{Pr}}}] = \mathbf{H}(f)\cdot\mathbf{\Pi}_i^T - \mathbf{H}(f-1) \quad \forall\ \mathbf{\Pi}_i \in \wp \quad (7)$$

$$\hat{\mathbf{\Pi}} = \left\{ \mathbf{\Pi}_i \left| \min_i \sum_{j=1}^{N_{\mathrm{Pr}}} \left\| \mathbf{d}_{ij} \right\| \right. \right\} \quad (8)$$

starting with low frequencies and ending up at high frequencies. The term $\hat{\mathbf{\Pi}}$ assigns the prototypes of the current frequency bin to their corresponding HRTF values in the previous bin, such that the distance between them stays minimum.

3. HRTF DATABASE LOOKUP

In order to determine the correct position of the sound sources in azimuth and elevation angle, we have to find the HRTFs for the left and right ears, from the KEMAR CIPIC database, which correspond to our estimated HRTFs. Then, we can calculate, for each sound source, the interaural HRTF $A_{est}(f)$ by dividing our estimated HRTF of the right ear by the HRTF of the left ear. The ILD and IPD are calculated using the expressions $\Delta L_{est}(f) = 20\log\left|A_{est}(f)\right|$ and $b_{est}(f) = \angle A_{est}(f)$, respectively. The interaural HRTF, of the KEMAR database, denoted by $\hat{A}_{CIPIC}(f)$, that best matches $A_{est}(f)$, is determined as follows:

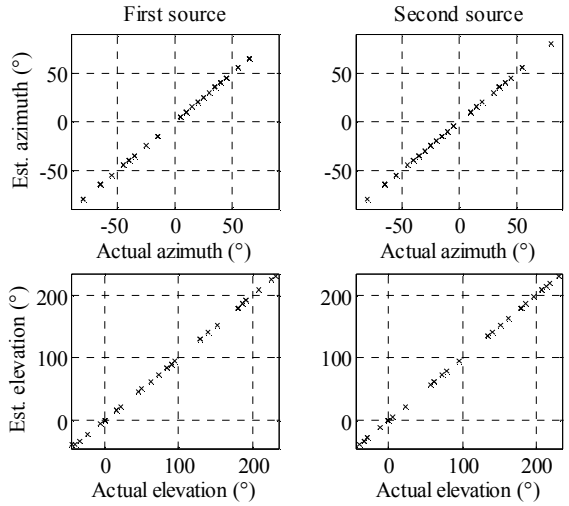

Fig. 5 Estimated azimuth (top) and elevation (bottom) angles obtained from 100 simulation runs with randomly chosen speaker positions in the whole 3D space.

$$\hat{A}_{CIPIC}(f) = \left\{ A_i(f) \middle| \min_i \left(\underset{f}{median}\left(\left| \Delta L_{est}(f) - \Delta L_i(f) \right| \right) \right) \wedge \min_i \left(\underset{f}{median}\left(\left| b_{est}(f) - b_i(f) \right| \right) \right) \right\} \quad \forall\, f, i \quad (9)$$

Equation (9) is evaluated for frequencies within the range 200 Hz to 11 kHz, since in this region binaural cues are very distinct. Having found the correct interaural HRTFs from the database, one can determine the azimuth and elevation angles of the sound sources since each HRTF is unique and corresponds to a unique position in the 3D space.

4. SOURCE SEPARATION

Using matrix-vector notation, we can express the windowed and Fourier-transformed ear signals in Eqs (3) and (4) in the case of two concurrent sound sources as follows:

$$\begin{bmatrix} X_1(f,\tau) \\ X_2(f,\tau) \end{bmatrix} = \begin{bmatrix} H_{11}(f) & H_{12}(f) \\ H_{21}(f) & H_{22}(f) \end{bmatrix} \cdot \begin{bmatrix} S_1(f,\tau) \\ S_2(f,\tau) \end{bmatrix} \quad (10)$$

Obviously, this system is mathematically determined, as the number of equations equals the number of unknowns, which are in this case $S_1(f,\tau)$ and $S_2(f,\tau)$. The HRTFs in the matrix are found using the lookup strategy described in the previous section. Consequently, in order to retrieve the source signals in the time-frequency domain, the matrix in (10) simply has to be inverted, whereby we assume that it has full has full rank due to distinct positions of the two sources within the 3D space. Thus, the sound sources are extracted:

$$\begin{bmatrix} S_1(f,\tau) \\ S_2(f,\tau) \end{bmatrix} = \begin{bmatrix} H_{11}(f) & H_{12}(f) \\ H_{21}(f) & H_{22}(f) \end{bmatrix}^{-1} \cdot \begin{bmatrix} X_1(f,\tau) \\ X_2(f,\tau) \end{bmatrix} \quad (11)$$

This inversion is done for each frequency *f*, this yields the Fourier spectrum of a complete time frame for the separated speech signals. Afterwards, applying inverse Fourier transform and assembling all time frames with the overlap-add method, we reconstruct the separated sources in time-domain. Figure 2 shows a block diagram that illustrates our localization and separation algorithm.

5. SIMULATION RESULTS

We tested our new sound localization algorithm by performing 100 simulation runs with two concurrent sound sources located in free space. In 40 simulations we positioned the sources in the horizontal plane (zero-elevation plane). In half of all the tests, both sounds were situated near each other in the same hemisphere around the KEMAR head. The concurrent sound sources were speech signals of two male speakers sampled at a rate of 44.1 kHz and 16 bit. For binaural synthesis these mono signals were convolved with the different HRTFs of the KEMAR database, simulating thus different locations in space. The ear input signals were windowed with a Hamming window of 1024 samples and an overlap of 50 % used for properly calculating the STFTs. For clustering, 400 time-frames of the ear input signals were acquired by the algorithm. This resulted in a signal length of approximately 4.7 seconds. Figure 5 shows the results of the 100 simulation runs. The subplots depict the estimated azimuth (above) and elevation (below) angles versus the actual ones for the first speaker (left) and the second speaker (right). Notably, the observed localization rate was 100 %. For all the simulation runs, both concurrent sound sources were located exactly at their target azimuth and elevation angles. The algorithm showed the same 100% localization performance in the case of both sound sources located close to each other in the same hemisphere around the KEMAR head.

In further 50 simulation runs a stationary noise source (computer fan) was introduced. This noise source was constantly located at 0° azimuth and 0° elevation. The mean *SNR*, i.e. the ratio of the mean power of the speech signals to the noise power was chosen to be 20 dB. Under these conditions, the localization percentage fell to 90 %. For 25 dB *SNR*, the algorithm is quite robust to the stationary noise since the localization accuracy rises to 97 %. Finally, we investigated the performance of source separation in the noise-free case. As proposed in [11], we computed the global signal-to-interference ratio (*SIR*) as follows:

$$SIR = 10 \log_{10} \frac{\|s_t\|^2}{\|e_i\|^2} \quad (12)$$

where

$$s_t = \left\langle y_j, s_j \right\rangle s_j \Big/ \left\| s_j \right\|^2 \quad (13)$$

$$e_i = \sum_{j' \neq j} \left\langle y_j, s_{j'} \right\rangle s_{j'} \Big/ \left\| s_{j'} \right\|^2 \quad (14)$$

Equation (12) represents the ratio between the power of the desired separated signal s_t, to the power of the interfering signal e_i, in the j-th output channel. The term $\langle \cdot, \cdot \rangle$ denotes the inner product of two signals. The median *SIR* value obtained after running 50 simulations, is 37.2 dB. In statistical terms, 50 % of the *SIR* values lay between 36.5 dB and 37.9 dB. Compared to other blind source separation methods, e.g. [7], trying to solve the same determined problem, our separation algorithm yields an average *SIR* that is more than 10 dB higher.

5. CONCLUSION AND FUTURE WORK

In this paper, we presented a new algorithm for binaural localization of two concurrent sound sources in both azimuth and elevation positions. By exploiting the ILD and IPD binaural cues that are encapsulated within the HRTFs, binaural 3D concurrent sound localization was made possible using only two microphones placed inside the artificial ears of the KEMAR head. Compared to existing techniques using microphone arrays for the same purpose, our algorithm is less complex and very accurate. It was shown, that two concurrent sound sources could be perfectly localized at their intended 3D locations even in the anti-causal case where both sources share the same hemisphere around the humanoid's head. This is, e.g. a remarkable improvement compared to the initially proposed algorithm in [6], where we attained a localization accuracy of 74 %.

The self-splitting competitive learning technique, mainly deployed in image processing, turned out to be very reliable for acoustical signal processing. It proved to be an intelligent tool to retrieve the exact cluster centers inside the feature space of the impinging sound signals, and consequently, to extract the 3D locations of the concurrent sound sources. After localization, the proposed sound source separation algorithm proved to be outperforming compared to other blind source separation methods solving the same determined problem under the same conditions.

Based on our new method, tow venues of future work are to be considered. One important venue is the localization of three or more concurrent sound sources, using only two microphones inserted in a humanoid ear canals. A second very challenging task is to determine the number of the concurrently active sound sources, in this manner, the proposed clustering algorithm is expected to work in a completely blind fashion.

ACKNOWLEDGMENT

This work is fully supported by the German Research Foundation (DFG) within the collaborative research center SFB453 "High-Fidelity Telepresence and Teleaction".

6. REFERENCES

[1] J. Blauert, "An introduction to binaural technology," in *Binaural and Spatial Hearing*, R. Gilkey, T. Anderson, Eds., Lawrence Erlbaum, USAHilldale NJ, 1997, pp. 593–609.

[2] J. Huang, N. Ohnishi, and N. Sugie, "Spatial Localization of Sound Sources: Azimuth and Elevation Estimation," *Proc. IEEE Instrumentation and Measurement Conference.*, USA, vol. 1, pp. 330-333, May 1998.

[3] Y. Tamai, Y. Sasaki, S. Kagami, and H. Mizoguchi, "Three ring microphone array for 3D sound localization and separation for mobile robot audition," *IEEE International Conference on Intelligent Robots and Systems*, pp. 4172-4177, Aug. 2005.

[4] C. Liu et al., "Localization of Multiple Sources with Two Microphones," *Journal of the Acoustical Society of America*, vol. 108, no. 4, pp. 1888-1905, Oct. 2000.

[5] F. Keyrouz, Y. Naous, and K. Diepold, "A New Method for Binaural 3D localization Based on HRTFs," *Proc. IEEE ICASSP*, France, pp. 341–344, May 2006.

[6] F. Keyrouz, W. Maier, and K. Diepold, "A Novel Humanoid Binaural 3D Sound Localization and Separation Algorithm," *Proc. IEEE-RAS Int. Conf. on Humanoid Robots*, Italy, December 2006.

[7] C.H. Keller, and T.T. Takahashi, "Localization and Identification of Concurrent Sounds in the Owl's Auditory Space Map," *Journal of Neuroscience.*, vol. 25, no. 45, pp. 10446-10461, Nov. 2005.

[8] V.R. Algazi, R.O. Duda, D.M. Thompson, and C. Avendano, "The CIPIC HRTF Database," *Proc. IEEE WASPAA*, USA, pp. 99-102, Oct. 2001.

[9] S. Winter, W. Kellermann, H. Sawada, and S. Makino, "MAP Based Underdetermined Blind Source Separation of Convolutive Mixtures by Hierarchical Clustering and L1-Norm Minimization," *EURASIP Journal on Advances in Signal Processing*, Article ID 24717, 2007.

[10] Y.-J. Zhang, and Z.-Q. Liu, "Self-Splitting Competitive Learning: A New On-Line Clustering Paradigm," *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 369-380, March 2002.

[11] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462-1469, July 2006.