Method of monaural localization of the acoustic source direction from the standpoint of the active perception theory

V E Gai[1], **I V Polyakov**[1], **M S Krasheninnikov**[1], **A A Koshurina**[1], **R A Dorofeev**[1]

[1]Nizhny Novgorod State Technical University n.a. R.E. Alekseev, Minin St., 24, Nizhny Novgorod, 603950

E-mail: iamuser@inbox.ru

Abstract. Currently, the scientific and educational center of the "Transport" of NNSTU performs work on the creation of the universal rescue vehicle. This vehicle is a robot, and intended to reduce the number of human victims in accidents on offshore oil platforms. An actual problem is the development of a method for determining the location of a person overboard in low visibility conditions, when a traditional vision is not efficient. One of the most important sensory robot systems is the acoustic sensor system, because it is omnidirectional and does not require finding of an acoustic source in visibility scope. Features of the acoustic sensor robot system can complement the capabilities of the video sensor in the solution of the problem of localization of a person or some event in the environment. This paper describes the method of determination of the direction of the acoustic source using just one microphone. The proposed method is based on the active perception theory.

1. Topicality

Hearing allows us to perceive sounds of various types (music, speech, etc.). The development of the robot, interacting with someone is impossible without an implemented acoustic sensor system. The sensor system of this type is required for the robot, first, to determine the direction of the acoustic source. Usually, in robotic systems, the problem of determination of the direction of the acoustic source is achieved by using two microphones. Taking into consideration the time difference between the arrival of the signal at the microphones, the direction of the acoustic source is estimated. The accuracy of determination of the direction of the acoustic source is limited by the size of the array of microphones. If the array size is small (microphones are located close to each other), it is difficult to determine the direction of the acoustic source accurately. But it is hard to use the arrays of large microphones for small robots. Consequently, the problem of determination of the direction of the acoustic source using just one microphone is quite important. The present work is devoted to solving the problem of monaural localization of the acoustic source direction.

There are several solutions to the problem of monaural localization of the acoustic source. The method, described in [1], uses the transfer functions of the head created by using a dummy and statistical speech models, based on the Gaussian mixture model for a particular speaker, to create a set of Gaussian mixture models for speech from various directions. Then, the models obtained at the testing stage are used to determine the direction of the source of speech.

In [2], the acoustic source localization is performed using a single microphone or an artificial ear. Acoustic signal $y(t)$, registered by the microphone is represented as a set of some acoustic sources $s(t)$, additive white Gaussian noise $w(t)$ and transfer function $h_\theta(t)$, depending on the direction of the acoustic source:

$$y(t) = h_\theta(t) * s(t) + w(t).$$

The calculation of signal decomposition *y(t)* into the most likely *s*(*t*) and $h_\theta(t)$ is performed in the frequency domain. For it, the hidden Markov models are generated for different audio signals *s*(*t*). Simulation $h_\theta(t)$ for different angles is performed using Gaussian noise recorded from different directions. When using the method, the assessment of angle θ is performed using the maximum likelihood method.

2. Localization in the direction of the acoustic source from the standpoint of the active perception theory

2.1. Introduction into the theory of active perception

The system of determination of the acoustic source direction can be represented as a pattern recognition system. It is known that the detection system comprises three processing stages: pre-treatment, calculation and decision making [3]. To implement the stages of pre-processing and calculation of the acoustic signal features, it is proposed to use the theory of active perception [4].

Pretreatment means the performance of the *Q*-transformation, which consists in the application of the addition operation of the original signal segments:

$$g(t) = \sum_{k=(t-1)\cdot L+1}^{t\cdot L} f(k), t = \overline{1, N},$$

where *L* – the number of readings, within the segment, *N* – the number of signal segments, *g* – the result of application of the *Q*-transformation to signal *f*, *f*(*k*) – *k* –reading of signal *f*, *g*(*t*) – *t*-reading of te signal *g*.

Formation of the feature description of the original signal is the application of the set of Walsh filters of the Hartmut system to signal *g*:

$$\mu(k, c(t)) = \sum_{i=0}^{M-1} F_k(i) \cdot g(((t-1)\cdot M + 1) : (t \cdot M)),$$

where $\mu(k, c(t))$ is the result of application of a sets of Walsh filters of the Hartmut system to signal *g*, $k = \overline{0, M-1}$, $t = \overline{0, |c|-1}$, $c = \{1, P, 2\cdot P, 3\cdot P, ..., N - T\cdot P\}$ is a set of the offset values for signal *g*, $|c|$ – is a cardinal number of set *c*, *P* - is an offset value for signal *g* ($1 \le P \le M$), *M* – is the number of the used filters. Thus, the feature description of the signal represents the size of matrix $M \times |c|$ and each line of the feature description is a result of the *U*-transformation the signal segment.

The consequent application of the *Q*-transformation and filter systems to the signal, implement *U*-transformation, which is basic in the theory of active perception.

U-transformation has the lowest possible computational complexity, since for its realization only simple operations are used – addition and subtraction. Standard transformations require the implementation of convolution, and the level of weighting coefficients – the operation of arithmetic multiplication.

The theory of active perception is not limited to the formation of the spectral representation of the signal. The structure of the theory includes the *Group algebra* section on the analysis of dependencies between the spectral expansion coefficients. Detected dependences allow us to use them at the stages of decision-making and understanding of the analyzed signal. Let each filter $F_i \in \{F_i\} \equiv \boldsymbol{F}$ correspond to coordinate-defined binary operator $V_i \in \{V_i\} \equiv \boldsymbol{V}$; Then components $\mu_i \neq 0$ of vector μ is permissible to put in line with operator *Vi* or $\overline{V_i}$ depending on the component sign. As a result, vector μ is associated with a subset of operators $\{V_i\}$, with a design similar to the filter, but having a different meaning of the vector elements ($+1 \leftrightarrow 1$; $-1 \leftrightarrow 0$). By specifying set-theoretic operations of multiplication and addition on set $\{V_i\}$, we have the algebra of signal description of one-dimensional

Boolean functions. Taking into account the reversals, all in all, there are 15 operators that can be used in the formation of feature descriptions, as the V_0 operator receives only the direct value.

On the set of operators, the algebra group is formed (the stage of synthesis) of the analyzed signal:

1) The family of algebraic structures (called the complete groups) $\{P_{ni}\}$ kind of $P_{ni} = \{V_i, V_j, V_k\}$ with the power of 35;

2) The family of algebraic structures (called the closed groups) $\{P_{si}\}$ kind of $P_{si} = \{V_i, V_j, V_k, V_r\}$ with the power of 105, where each group is formed of a pair of complete groups connected in a certain way.

Among the complete groups, we can allocate the complete groups for the operations of addition and the operation of multiplication, among the closed groups - closed groups and closed sets.

Two groups (complete or closed) are called incompatible if they include operators with the same numbers, but with different signs.

With closed and complete groups, we perform the spectral correlation analysis. Complete groups reveal the correlation between the operators. Closed group reveal the correlation between the complete groups. If the set of operators is an alphabet, then the set of groups is more complex grammatical descriptions of an observed signal: a complete group is a word, a closed group is a phrase.

Using the spectral representation of signal μ, formed a set of operators describing the signal, and then the set of complete and closed groups:

$V = GV[\mu]$, $P_{na} = GP_{na}[\mu, V]$, $P_{nm} = GP_{nm}[\mu, V]$, $P_s = GP_s[\mu, V, P_{na}, P_{nm}]$, $P_c = GP_c[\mu, V, P_{na}, P_{nm}]$,

where GV – is an operator for calculation of feature descriptions V using spectral representation of the signal on the basis of operators, GP_{na} (GP_{nm}) – on the basis of complete groups for the operations of addition P_{na} (multiplication, P_{nm}), GP_c (GP_s) – on the basis of closed groups P_s (closed sets P_c).

3.2. Proposed approach: the implementation of stages of the recognition system

Using feature descriptions described above, we can get integrated feature descriptions in the form of histograms of the frequency of the operator occurrence, complete and closed groups (two-dimensional, three-dimensional)

$$h_V = H[V, \Gamma], h_{na} = H[P_{na}, \Gamma], h_{nm} = H[P_{nm}, \Gamma], h_s = H[P_s, \Gamma], h_c = H[P_c, \Gamma],$$

where h_V – histogram of operators, h_{na} (h_{nm}) – histogram of complete groups for operations of addition (multiplication), h_s (h_c) – histogram of closed sets (closed groups), Γ – dimensions of histogram: $1d$ – one-dimensional histogram, $2d$ – two-dimensional histogram, $3d$ – three-dimensional histogram. In the two-dimensional histogram, we consider the possible occurrence of pairs of groups in the description of one of the signal segments, in the three-dimensional – triples are considered.

Two-dimensional histograms, which are the matrices, have the following properties:

1) Zero main diagonal;

2) Symmetry relating to the main diagonal.

Thus, the number of significant elements is a two-dimensional histogram $(H_{dim} \cdot H_{dim} - H_{dim}) / 2$, where H_{dim} – number of columns (rows) of the matrix.

Taking into consideration the incompatibility of some complete and closed groups, we may reduce the dimension of the proposed feature systems. The compression considering the incompatibility is only possible for the histograms with dimension greater than or equal to two. Thus, taking into account the given arguments, dimensions of the feature systems based on the histogram are shown in table 1. In brackets, we see the compression ratio, which is calculated as the ratio of the number of features before compression to the number of features after compression.

By increasing the dimension of the proposed systems of features, the number of features is growing as an exponential function. In this regard, it was decided to limit ourselves to two-dimensional histograms of groups and three-dimensional histogram of operators.

Table 1. The dimension of the feature systems before and after the compression

System of features	Before compression			After compression		
Dimension	$1d$	$2d$	$3d$	$1d$	$2d$	$3d$

h_V	1 × 30	30 × 30	30 × 30 × 30	1 × 435 (2,06)
h_{na} (h_{nm})	1 × 140	140 × 140	140 × 140 × 140	1 × 5050 (3,88)
h_s (h_c)	1 × 840	840 × 840	840 × 840 × 840	1 × 106030 (6,65)

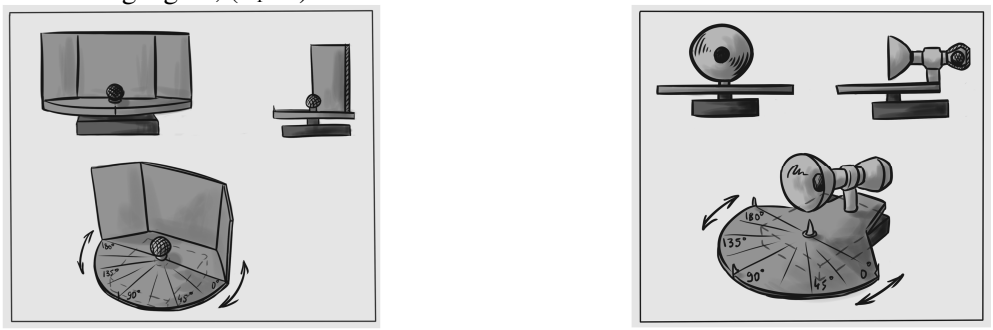
4. Simulation experiment

4.1. Installation for carrying out experiments

The experiment was performed in the laboratory. The installation for the experiment includes a recording device (a microphone connected to a sound card or a smart phone), a speaker and a device for turning the microphone and a reflector for a predetermined angle (see figure 1.a). As a sound recording device, we use an omnidirectional microphone. A speaker is a plate with a non-uniform surface. As a playback device, a broadband speaker is used. Figure 1.b shows a prototyping device consisting of a simplified model of the external human ear as the most successful natural catcher of speech sounds. The model includes a small elliptical screen and a channel with the length of about 3 cm up to the microphone.

4.2. Description of the experiment

The simulation experiment aims to study the accuracy of the direction localization of the acoustic source, depending on the layout, signal duration, as well as signs of the system. In the experiment, the analyzed range of angles is from 0 ... 180 degrees, at a pitch of 15 degrees. The process of teaching of the system is to register audio recording with duration of L_1 seconds from each direction (see figure 2). As a result, we get 13 records. Each record is divided into parts, each part is two seconds long. Then, for each of these parts a feature description is made, which is stored in the database. Thus, by L_1 seconds of a training signal, $(L_1 / 2) \cdot 13$ records are formed in the data base.



a) Installation model No. 1 b) Installation model No. 2

Figure 1. The installation used for audio recording

The process of determination of the acoustic source direction is as follows:

1) acoustic signal *h* with duration of L_2 seconds is registered;

2) the recorded signal is divided into segments; each segment is 2 seconds long;

3) for each segment a feature description is formed (see paragraph 3.2);

4) based on the classifier, an assumptive direction of the angle is determined for each segment;

5) the final decision about the direction to the acoustic source for signal, *h* is generally accepted as the most frequently occurring angle value obtained for the separate signal segments.

When performing a computational experiment, we will determine the accuracy of the direction of the acoustic source, depending on the features of the system, the installation model, the length of training signal (L_1), and the duration of test signal (L_2). The solution of the problem of classification is performed based on the method of support vector machine (SVM) and *k*-nearest neighbor (KNN).

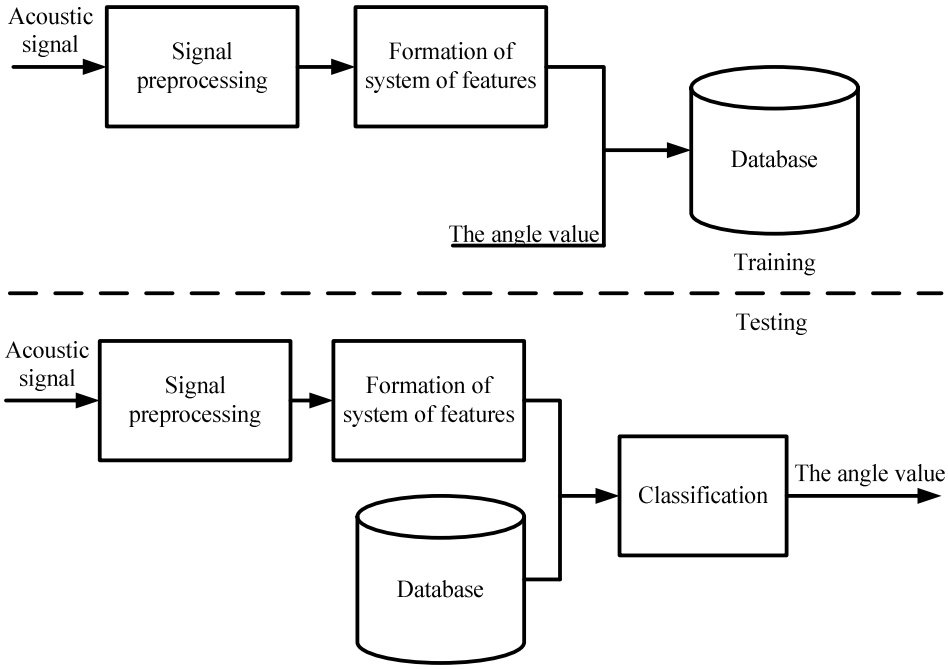


Figure 2. An information system model to determine the direction of the acoustic source

4.3. Notable results

Table 3 shows the error in determining the direction of the acoustic source under different conditions [2]. The experiments were conducted with different types of acoustics; the recording duration of the test was 5 minutes. The recording truck included four types of acoustic-reflecting walls (A, B, C, D) with varying degrees of complexity of the form (C – the most difficult).

Table 3. An average error in degrees of localization of the of the acoustic source direction

Model type of the recording system / Acoustic Type	A	B	C	D
Broadband noise	42.6	8.8	4.3	22.3
Speech	67.8	19.3	7.7	21.35
Dog's barking	55.7	14.2	18.3	60.28
The waterfall noise	42.1	11.8	9.3	42.7
Pure tone	88.7	89.1	86.4	82.6

4.4. Results

Tables 4 and 5 show the evaluation results of the average error in determining the direction of a sound source by using different signs systems classifiers and L_1 and L_2 values.

Table 4. The average error in determination of the direction (classifier KNN)

L_1	L_2	System of features						
		h_V, 2d	h_V, 3d	h_{na}, 1d	h_{na}, 2d	h_{nm}, 1d	h_{nm}, 2d	h_c, 1d
40	20	64.61	61.15	51.92	54.80	60.57	60.57	65.19
40	10	60.00	63.75	63.75	68.94	55.67	64.61	62.88
40	5	61.00	62.01	62.30	61.00	62.88	62.01	60.28
40	2	63.75	51.25	59.76	63.23	57.17	62.82	60.63

Table 5. The average error in determination of the direction (classifier SVM)

L_1	L_2	System of features

		h_V, 2d	h_V, 3d	h_{na}, 1d	h_{na}, 2d	h_{nm}, 1d	h_{nm}, 2d	h_c, 1d
40	20	2.30	2.57	2.88	2.15	2.57	2.19	5.30
40	10	12.98	4.59	10.67	3.01	6.63	0.20	8.07
40	5	22.50	10.52	16.87	10.52	12.98	6.40	8.94
40	2	34.78	27.98	25.90	24.92	26.19	22.26	25.73

Conclusions on the results of experiments:

1) the accuracy of the results obtained is compatible with known works [2], but in the proposed implementation the duration of the test signals is significantly reduced (20 seconds instead of 5 minutes);

2) using an SVM classifier provides better accuracy in determining the direction of the sound source, compared with the classifier KNN (tables 4 and 5);

3) using an SVM classifier provides the highly accurate determination of the sound source direction (according to the value of the average error – in the table 5);

4) reduction of the test signal duration leads to an increase of the average error in determination of the sound source direction (see table 5).

The developed method can be used in a universal rescue vehicle [5].

5. Conclusion

This paper describes the method of the monaural acoustic source localization. The solution is performed from the standpoint of the active perception theory. The investigations of the proposed method showed the possibility of reducing the time of decision-making at the expense of reducing the duration of the analyzed signal, as compared to known approaches. The software implementation of the method was used in the mobile robot control system. The performed work is a stepping-stone in the development of new methods and means of search and rescue of persons in distress at sea.

6. Acknowledgments

This work was carried out at the Nizhny Novgorod State Technical University named after R.E. Alekseev, with financial support from the government in the face of the Russian Ministry of Education (the unique identifier of the project: RFMEFI57714X0105.

References

[1] Fuchs A K, Feldbauer C, Stark M 2011 *Proc. Int. Conf. Speech Communication Association* (27-31 August 2011, Florence, Italy) p 2521

[2] Saxena A, Ng A Y 2009 *Proc. Int. Conf. on Robotics and Automation* p 1737

[3] Perez-Meana H 2007 *Advances in Audio and Speech Signal Processing: Technologies and Applications* (Igi Global) p 374

[4] Utrobin V A 2004 Physical interpretation of the elements of image algebra, J. *Advances in Physical Sciences* **47** 1017-1032

[5] Krasheninnikov M, Kulashov A, Shapkin V, Koshurina A 2013 The concept and methodology of creating the universal life-saver with rotary-screw mover *J. Lecture Notes in Electrical Engineering* **195** 477-490.