

Intelligent Acquisition of Audio Signal Employing Neural Network and Rough Set Algorithms

Andrzej Czyzewski

Technical University of Gdansk, Sound & Vision Engineering Department
Narutowicza 11/12, Gdansk, Poland
e-mail: andcz@ieee.org

Abstract. The algorithms stemming from the neuro-rough computing approach were applied to digital acquisition of audio signals with regard to automatic localization of sound sources with the presence of noise and parasite echo. The application of neural networks to the automatic detection of sound arrival direction was tested first, then it was followed by some experiments employing rough sets and finally the neuro-rough approach to this problem solving was examined. The output of each tested algorithm was supposed to provide information about the direction of arriving sound. In the case of the neuro-rough algorithm the result of its action can be also available in the form of words defining the direction of arriving sound. Some details of the engineered systems and results of their experimental verification are compared and discussed.

1 Introduction

Sound source localization plays an important role in some spatial-filtering techniques applied to many telecommunication systems. Speech signals arriving from various directions may interfere with the target signal but also can mask it. Consequently, the main purpose of spatial filtering technique applied to telecommunication systems or to hearing aids is to attenuate the unwanted signal coming from other directions than the desired one. Another issue addressed by this kind of sound processing algorithms is automatic sound source tracking that is applicable to some advanced video teleconferencing systems employing automatically turned video camera.

Numerous source localization methods were investigated by various researchers [1][3][4][5]. These include frequency-domain processing introduced at the Sound Engineering Department [6][7][17][18][19][20][28]. Most of such systems are based on digital signal processing technology and are computationally intensive.

This paper presents some alternative method applications employing neural networks and rough set-based decision modules. The computer simulators (*Matlab* and *Rosetta*) were used in experiments. The new methods were investigated using some prerecorded sound excerpts providing experimental data stream.

The description of the proposed algorithms and some results of experiments are included. Some conclusions were drawn-out on the basis of carried-out experiments concerning learning algorithms applications to spatial filtering of sound.

2 Experimental Setup

The human hearing is known as most effective detector of the direction of arriving sounds. Therefore, in order to make the signal processing algorithms more efficient, their operation principle should follow some properties of human hearing. Despite the great development of science in the field of human perception, issues related to sound localization are not finally recognized, hence phenomena underlying thereof are still the subject of intense research [2][7][13]. According to the present state of knowledge, perception of sound directivity by the human binaural system is based on the following two principal entities [13]:

- Interaural Level Difference (ILD): difference of intensities of waveforms in the left and in right ears;
- Interaural Time Difference (ITD): difference of arrival times of relevant waveforms in the both ears, which is equivalent to a phase difference of the waveforms.

In the field of digital signal processing, the identifying of sound source localization can be performed by means of a microphone array which can be either linear or non-linear [16] [19]. The lay-out of the sound acquisition system is presented in Fig. 1. There were used 8 microphones placed symmetrically on a 30th cm in diameter rim.

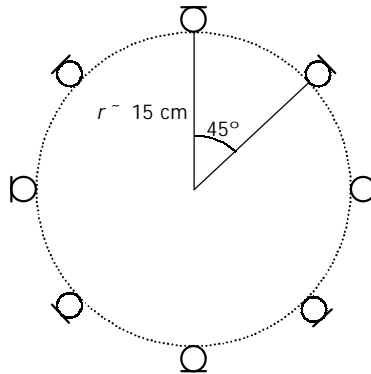


Fig. 1. The diagram of the microphone array shown in horizontal plane

Under the ideal conditions, the signal $x_i(t)$ received from i -th microphone of the circular array of microphones and in t -th moment of time can be described as follows:

$$x_i(t) = \alpha_i \cdot s[t - (i - 1) \cdot \tau] \quad (1)$$

- where: α_i - attenuation coefficient for i -th microphone,
 $s(t)$ - source signal,
 τ - time delay of acoustic wave between adjoining microphones,

thus the estimation of the source location based on the processing of acoustic signals with a microphone array provides a deterministic problem.

However, under real conditions there various distortions occur, and interference signals such as: background noise and reverberated sounds and others. Hence, the signals received by a circular microphone array are expressed by the following relationships:

$$\begin{cases} x_1(t) = \alpha_1 \cdot h_1(t) * s(t) + n_1(t) \\ x_2(t) = \alpha_2 \cdot h_2(t) * s(t - \tau) + n_2(t) \\ \vdots \\ x_i(t) = \alpha_i \cdot h_i(t) * s(t - (i-1) \cdot \tau) + n_i(t) \\ \vdots \end{cases} \quad (2)$$

where: $h_i(t)$ - impulse response of the reverberant channel associated with i -th microphone,
 $n_i(t)$ - ambient noise received by i -th microphone.

These conditions make the task of sound source localization more complex, and therefore a number of various methods have been proposed to solve the problem. Most of them are based on estimation of the sound source position on the basis of signals received by microphones in the matrix, including cross-correlation techniques [3], adaptive filtration [5] or computation of relevant eigenvalue vectors and matrices [1]. In turn, in the case of tracking or localizing a number of sources, the Maximum Likelihood-based methods are exploited [31]. More details can be found in the abundant literature on the localization of acoustic sources for multimedia applications [15] [16] [23] [29] [31].

2.1 Sound Acquisition Setup

As was said the array consisted of 8 electret microphones set on the circumference of a 15 cm radius (30th cm in diameter) rim. The set of microphones showed in Fig. 2 was fixed 1.58 m from the floor. The recording parameters were as follows: 16 bit/sample, sampling frequency equal to 48 kHz. There was one male speaker, distanced 1.5 m from the array. The speaker read a logatom list from the consecutive spots differing in 5° . In result 72 eight-track recordings were made, and every recording lasted approx. 55 s. For purposes of the experiments, eight additional excerpts were also prepared representing the sound directivity from -270° to $+270^\circ$ recorded every 15° . Another set of signals recorded for experimental purposes was derived from angles 0° - 90° with resolution of 5° .



Fig. 2. The microphone matrix used for sound acquisition

2.2 Extracting Feature Vectors

The direct processing of sound samples stream by learning algorithms might be impractical, because the volume of data can be very high in this case and there are too many indirect dependencies between consecutive sample packets which may not be interpreted easily. Therefore, during the feature extraction process the signal was divided into frames of the length N equal to 512, 1024 and 2048 samples and then processed by some feature extraction algorithms. As was verified earlier, in the practical application of spatial filtration (beamforming) in hearing aids, the following parameters can be efficiently exploited [18]:

$$M_i = \frac{\min(|L_i|, |R_i|)}{\max(|L_i|, |R_i|)} \quad D_i = \frac{|L_i - R_i|}{|L_i| + |R_i|} \quad (3)$$

$$A_i = |\angle L_i - \angle R_i|$$

where: L_i and R_i are magnitudes of the i -th spectral bin for the left and right channel, respectively.

Considering that the above parameters concern pairs of channels Ch_i^k and Ch_j^k , these parameters for the k -th spectral bin can be rewritten as below:

$$M_{ij}^k = \frac{\min(|Ch_i^k|, |Ch_j^k|)}{\max(|Ch_i^k|, |Ch_j^k|)} \quad D_{ij}^k = \frac{Ch_i^k - Ch_j^k}{|Ch_i^k| + |Ch_j^k|} \quad A_{ij}^k = \angle Ch_i^k - \angle Ch_j^k \quad (4)$$

It can be shown, that the parameters M_{ij}^k and D_{ij}^k are in a simple functional relationship and therefore one of them is superfluous and can be dropped. In such a case, parameters representing a single spectral bin are as follows:

$$M_{ij}^k = \frac{\min(|Ch_i^k|, |Ch_j^k|)}{\max(|Ch_i^k|, |Ch_j^k|)} \quad A_{ij}^k = \angle Ch_i^k - \angle Ch_j^k \quad (5)$$

As was said, in the experiments, 8-channel signals were examined, thus the following sets of parameters can be considered:

- *type A*: all mutual combinations of channels, which yields 56 parameters per spectral bin;
- *type B*: combination of opposite channels, which yields 8 parameters per spectral bin.

On account of the fact that the above parameters are to be fed to a learning algorithm, they are grouped into input vectors. The following three types of such vectors can be considered:

- *type V1*: all spectral bins are included in a vector;
- *type V2*: an input vector consists of parameters for a single bin and the additional information on the bin's frequency;
- *type V3*: an input vector consists only of parameters for a single spectral bin. In this case, the learning algorithm assumes a modular structure where a separate subsystem is dedicated for each spectral bin. The final decision is made on the basis of the interpreting of output values of all sub-algorithms.

The particularly interesting is the modular concept related to the feature vectors of the type **V3**. The description of data to be fed to the modular decision algorithm in this case is gathered in the Table 1.

Table. 1 Analysis of training conditions for the input vector **V3**

$N = 512; N/2 = 256$	$N = 1024; N/2 = 512$	$N = 2048; N/2=1024$
A : <i>vectorSize</i> = 56	A : <i>vectorSize</i> = 56	A : <i>vectorSize</i> = 56
B : <i>vectorSize</i> = 8	B : <i>vectorSize</i> = 8	B : <i>vectorSize</i> = 8
256 decision modules; material for training: 186 vectors /s	512 decision modules; material for training: 92 vectors /s	1024 decision modules; material for training: 45 vectors /s

3 Application of Neural Networks

3.1 Training Algorithms

Some heuristic algorithms for the neural network training were chosen, namely the general and simplified Fahlman's algorithm (QuickPROP) [24] and the **Resilient PROP**agation (RPROP) [25].

In the general Fahlman's algorithm (denoted further as Fahlman I) the weight update rule for a single weight w_{ij} in the k -th cycle is for the Fahlman's algorithm computed as below:

$$\Delta w_{ij}^k = -\eta^k \cdot S_{ij}^k + \alpha_{ij}^k \cdot \Delta w_{ij}^{k-1} \quad (6)$$

where the error gradient term S_{ij}^k assumes:

$$S_{ij}^k = \nabla E(\Delta w_{ij}^k) + \gamma \cdot \Delta w_{ij}^k \quad \gamma = 10^{-4} \quad (7)$$

and the learning rate η^k and the momentum ratio α_{ij}^k vary according to the following formulae:

$$\eta^k = \begin{cases} \eta_0 & ; \text{for } k=1 \vee S_{ij}^k \cdot \Delta w_{ij}^{k-1} > 0 \\ 0 & ; \text{otherwise, i.e.: } k \neq 1 \wedge S_{ij}^k \cdot \Delta w_{ij}^{k-1} \leq 0 \end{cases} \quad (8)$$

$$\alpha_{ij}^k = \begin{cases} \alpha_{\max} & ; \text{for } \beta_{ij}^k > \alpha_{\max} \vee S_{ij}^k \cdot \Delta w_{ij}^{k-1} \cdot \beta_{ij}^k < 0 \\ \beta_{ij}^k & ; \text{otherwise, i.e.: } \beta_{ij}^k \leq \alpha_{\max} \wedge S_{ij}^k \cdot \Delta w_{ij}^{k-1} \cdot \beta_{ij}^k \geq 0 \end{cases} \quad (9)$$

In the formulae above, the constant values of the training parameters assume: $0.01 \leq \eta_0 \leq 0.6$, $\alpha_{\max} = 1.75$, and the denotation β_{ij}^k stands for:

$$\beta_{ij}^k = \frac{S_{ij}^k}{S_{ij}^{k-1} - S_{ij}^k} \quad (10)$$

In the **simplified Fahlman's algorithm** denoted further as **Fahlman II** the weight update rule is expressed by the following relationship:

$$\Delta w_{ij}^k = \begin{cases} \alpha_{ij}^k \cdot \Delta w_{ij}^{k-1} & ; \text{for } \Delta w_{ij}^{k-1} \neq 0 \\ -\eta_0 \cdot \nabla E(\Delta w_{ij}^k) & ; \text{otherwise, i.e.: } \Delta w_{ij}^{k-1} = 0 \end{cases} \quad (11)$$

where the momentum ratio α_{ij}^k changes according to the expression below:

$$\alpha_{ij}^k = \min \left\{ \frac{\nabla E(\Delta w_{ij}^k)}{\nabla E(\Delta w_{ij}^{k-1}) - \nabla E(\Delta w_{ij}^k)}, \alpha_{\max} \right\} \quad (12)$$

and the constant values of the training parameters are the same as in the general QuickPROP, i.e.: $0.01 \leq \eta_0 \leq 0.6$, $\alpha_{\max} = 1.75$.

In the case of the **RPROP algorithm**, the weight update rule is given by the following formula based on the *sgn* function:

$$\Delta w_{ij}^k = -\eta_{ij}^k \cdot \text{sgn}(\nabla E(\Delta w_{ij}^k)) \quad (13)$$

where the learning rate η^k assumes values according to the rules below:

$$\eta_{ij}^k = \begin{cases} \min\{\mu^+ \cdot \eta_{ij}^{k-1}, \eta_{\max}\} & ; \text{for } \nabla E(\Delta w_{ij}^k) \cdot \nabla E(\Delta w_{ij}^{k-1}) > 0 \\ \max\{\mu^- \cdot \eta_{ij}^{k-1}, \eta_{\min}\} & ; \text{for } \nabla E(\Delta w_{ij}^k) \cdot \nabla E(\Delta w_{ij}^{k-1}) < 0 \\ \eta_{ij}^{k-1} & ; \text{otherwise, i.e.: } \nabla E(\Delta w_{ij}^k) \cdot \nabla E(\Delta w_{ij}^{k-1}) = 0 \end{cases} \quad (14)$$

and the constant values are set as follows:

$$\eta_{\min} = 10^{-6} \quad \eta_{\max} = 50 \quad \mu^- = 0,5 \quad \mu^+ = 1,2 \quad 0 < \mu^- < 1 < \mu^+ \quad (15)$$

3.2 Results of Modular Neural Network Application

Tables 2 - 4 present the results of the experiments with neural networks applied to the detection of direction of the incoming sound. These results are presented with regard to training algorithm and sound arrival direction. Additionally, the length of the sample packet and the number of training and testing vectors were shown in these tables. It was assumed that the maximum value found on neural networks outputs identifies the particular network which detected the appropriate direction of the arriving sound. The percentage of accurate scores obtained with this method is shown in the tables.

Table 2. Results of direction detection for the vector type - $V3$, $N = 512$; parameter type - C , training / testing vectors: 1042 / 446

<i>Direction</i>	Fahlman I		Fahlman II		RPROP	
	<i>cycles</i>	<i>scores</i>	<i>cycles</i>	<i>scores</i>	<i>cycles</i>	<i>scores</i>
-45°	16.335	82 %	15.893	83 %	23.119	85 %
-30°	14.239	84 %	18.991	83 %	21.092	80 %
-15°	15.268	78 %	16.453	80 %	19.672	82 %
0°	17.218	79 %	18.002	81 %	20.999	81 %
15°	16.001	81 %	19.979	83 %	21.017	82 %
30°	19.965	80 %	21.310	82 %	20.775	82 %
45°	18.342	79 %	21.367	78 %	25.901	80 %

Table 3. Results of direction detection for the vector type - $V3$, $N = 1024$; parameter type - A ; training / testing vectors: 515 / 221

<i>Direction</i>	Fahlman I		Fahlman II		RPROP	
	<i>cycles</i>	<i>scores</i>	<i>cycles</i>	<i>scores</i>	<i>cycles</i>	<i>scores</i>
-45°	27.890	90 %	37.199	92 %	41.092	89 %
-30°	32.893	89 %	34.269	87 %	39.501	88 %
-15°	32.672	88 %	31.474	89 %	42.277	90 %
0°	29.994	90 %	35.892	90 %	37.512	88 %
15°	30.173	86 %	40.866	82 %	50.899	85 %
30°	29.980	85 %	30.101	85 %	35.924	84 %
45°	27.559	87 %	38.943	88 %	39.994	88 %

Table 4. Results of direction detection for the vector type - $V3$, $N = 2048$; parameter type - A ; training / testing vectors: 252 / 108

<i>Direction</i>	Fahlman I		Fahlman II		RPROP	
	<i>cycles</i>	<i>scores</i>	<i>cycles</i>	<i>scores</i>	<i>cycles</i>	<i>scores</i>
-45°	21.218	86 %	22.190	85 %	30.168	86 %
-30°	19.900	87 %	26.886	87 %	27.110	86 %
-15°	20.457	87 %	21.106	88 %	28.249	88 %
0°	28.189	88 %	27.000	89 %	39.271	89 %
15°	25.190	85 %	32.981	86 %	31.992	87 %
30°	23.267	86 %	24.119	85 %	28.428	86 %
45°	24.219	87 %	31.148	87 %	28.550	87 %

As results from the presented data, the direct application of modular network structure allowed detection of the direction of arriving sound with quite a good accuracy, however not better than 92%.

4 Application of Rough Sets

The rough set based decision-making system was proposed to allow a better sound source localization performance in case of noisy and distorted signals, introducing much uncertainty related to the decision-making process. A block diagram of the introduced sound source localization method is shown in Fig. 3.

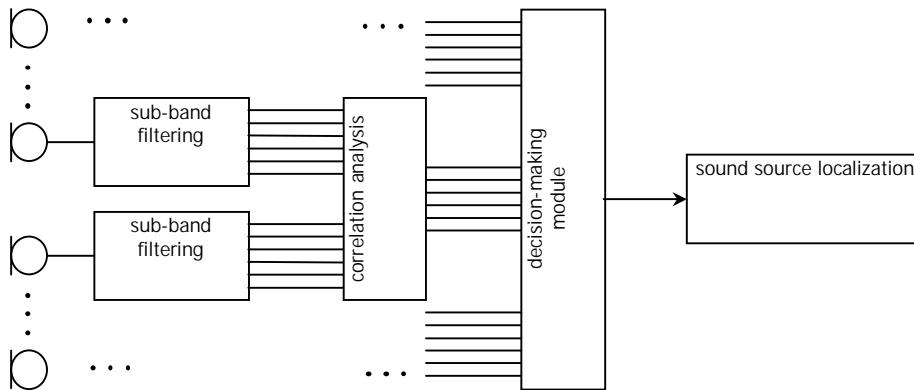


Fig. 3. Rough-set supported sound source localization lay-out

The input signal of each microphone is first passed through a set of band-pass filters. Subsequently, correlation analysis is performed for each pair of microphone signals and each sub-band. As an output, a set of correlation parameters is calculated. Sound source is then localized using decision-making unit, upon a set of correlation parameters for subsequent pairs of microphones and following frequency sub-bands. Since the input signal may contain noise and distortions a rule-based rough-set algorithm was employed to this task.

Concerning all microphones within the array the processing might be performed for all combinations of microphones. However, such an approach demands significant computing power to perform correlation analysis and tends to load large amount of data to the decision-making module. Therefore, in the performed experiments only pairs of counter-positioned microphones were considered. It gained significant reduction of computing power requirements without losing performance capabilities.

4.1 Correlation Parameters

Correlation parameters are calculated for each pair of counter-positioned microphones within subsequent frequency sub-bands. Correlation analysis is performed within the octave sub-bands. Boundary frequencies of sub-bands are presented in Table 5.

Table 5. Boundary frequencies of sub-bands

Band No.	lower boundary	higher boundary
1	20 Hz	100 Hz
2	100 Hz	200 Hz
3	200 Hz	400 Hz
4	400 Hz	800 Hz
5	800 Hz	1600 Hz
6	1600 Hz	3200 Hz
7	3200 Hz	6400 Hz
8	6400 Hz	20000 Hz

The sub-band filtering was performed using spectral filtering method. The *Mathematica* notebook performing spectral filtration was developed. The following signal processing constraints were used:

- sampling frequency: 48 kHz,
- window size: 2048 samples,
- overlap: 1024 samples,
- windowing function: Hamming.

Initially, standard autocorrelation function was applied accordingly to the Pearson's formula:

$$\rho(n) = \sum_i \frac{[x(t)_i - \bar{x}(t)][y(t+n)_i - \bar{y}(t+n)]}{\sqrt{\sum_i [x(t)_i - \bar{x}(t)]^2} \sqrt{\sum_i [y(t+n)_i - \bar{y}(t+n)]^2}} \quad (16)$$

and its simplified form is as follows:

$$\rho(n) = \frac{\sum_i x(t)_i y(t+n)_i}{\sum_i x(t)_i y(t)_i} \quad (17)$$

The correlation maximum should correspond to time-alteration of signals between microphones of concern at the given moment of time. However, since speech signal may include significant energy alterations, correlation function maxima may correspond to energy peaks. Therefore, as an alternate solution, the AMDF function was introduced to allow correlation analysis. The AMDF function is given as follows:

$$AMDF(n) = \sum_i |x(t)_i - y(t+n)_i| \quad (18)$$

Based on the AMDF function, a signal time-lag between microphones can be estimated upon location of global minimum. An example of AMDF function plot for a speech signal within 5th sub-band (see Table 5) is presented in Fig. 4. To allow better illustration the reverse signed AMDF denoted as (-AMDF) was shown. It should be noted, that the zero-lag location corresponds to 100 on the n axis.

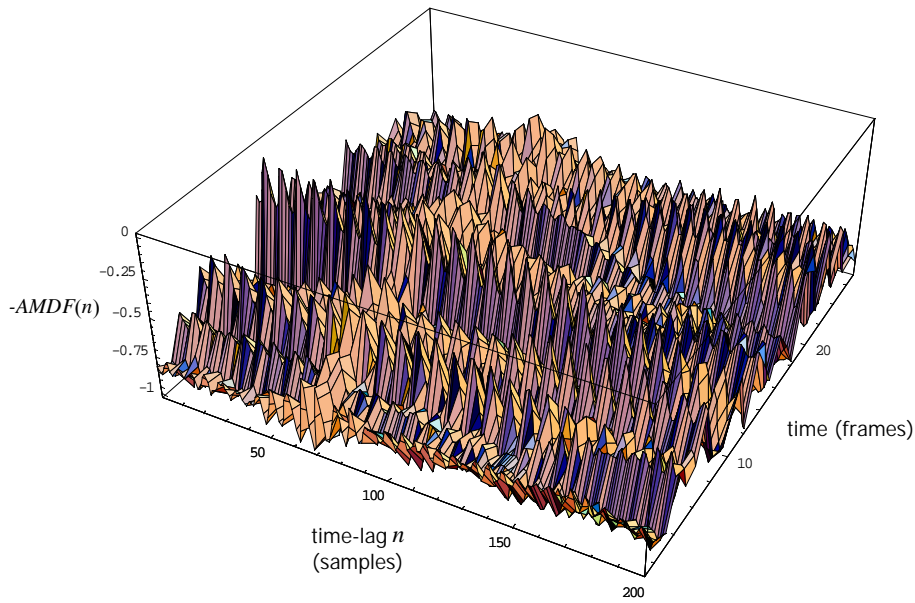


Fig. 4. (-AMDF) plot for a speech signal within 5th sub-band (800-1600 Hz)

The peak corresponding to the time-lag between microphones can be seen clearly in the plot in Fig. 4. It can be also observed, that in some frames actual peaks do not correspond to the time-lag. Consequently, the AMDF function is accumulated through frames of processing according to the formula:

$$AMDF_{ac}(n) = \frac{1}{M} \sum_{m=1}^M AMDF_m(n) \quad (19)$$

where: M – number of frames of analysis

A time-lag between microphones can be estimated upon location of the minimum within accumulated AMDF function. However, according to speech signal character-

istics as well as potential noise and distortions presence, such an estimation may lead to erroneous source localization. Therefore, for each pair of microphones and subsequent sub-bands the accumulated AMDF function is represented using two parameters: location of minimum of accumulated AMDF within a sub-band and an average signal energy within the sub-band defined as:

$$E_b = \frac{1}{N} \sum_{n=1}^N |AMDF_b(n)| \quad (20)$$

where: N is a number of samples and b is current sub-band

These parameters are then provided for further processing using the decision system.

5.2 Rough Decision System

The presented method is based on the rule-based rough set decision system. For the purpose of the presented experiments the rough-set software toolbox – *Rosetta* – was used [34].

Learning data are processed by the following way. First, the data set – knowledge base is acquired. Knowledge base consists of objects, which are represented using conditional attributes and decision parameter. As an input for the decision-making system a set of correlation parameters: location of AMDF minima (denoted Dx) along with signal energy (denoted Ax) within sub-bands are used. For each pair of microphones each sub-band is represented using two parameters, giving the total amount of 64 parameters representing the input pattern. The input file for the rough-set processing consists of a header and a dataset as shown in Table 6.

Table 6. Data layout for rough-set based sound source localization

Parameter type	D1 integer	A1 float(4)	D2 integer	A2 float(4)	D3 integer	A3 float(4)	...	angle string(3)
#1	-30	32.029	-52	112.24	4	180.875	...	000
		2						
#2	-37	32.116	-87	96.1503	-37	181.063	...	090
		3						
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
#N	-60	35.595	-48	151.269	-40	395.7	...	270
		2						

Consequently, acquired data are quantized to convert real attribute values into discretized form allowing further rule-based processing. Based on the discrete values, attributes are analyzed in terms of discernibility investigation. Sets of attributes al-

lowing partition of object classes are then revealed. These sets provide reducts. Consequently, rules are generated upon reducts.

The *Rosetta* system supports a variety of quantization as well as reduct and rule generation procedures, however details on these procedures lay beyond the scope of this paper [34]. For the purpose of the presented experiments the following processing parameters were used:

- discretization – equal frequency binding using three intervals,
- reduct and rule generation – object related genetic algorithm producing a set of rules via minimal attribute subsets that discern object classes; reducts and rules are generated upon analysis of all learning patterns.

These processing parameters were chosen during a preliminary research aimed on system efficiency and generalization ability optimizing.

4.3 Experiments on Sound Source Localization

For the purpose of the presented experiments speech recordings performed within an anechoic chamber were used. The experiments were divided into three subsequent parts: “low resolution source localization”, “low resolution source localization in the presence of wide-band noise” and “high resolution source localization”.

Low Resolution Source Localization

Initially, experiments on low-resolution source localization were performed. In this phase sound source was located at the angle of 0° , 90° , 180° and 270° accordingly. Five sound samples for each angle were used in the experiments. Two series of experiments were performed. In the first phase, one instance representing each angle was used for training, whereas the trained system was tested using the other patterns. In the second phase, the patterns used previously for testing were used for training and vice versa. Source localization scores for both phases of the experiments are shown in Table 7. The numbers given in brackets represent the number of properly classified examples versus the number of all tested examples.

Table 7. Sound localization accuracy for low-resolution analysis

accuracy	phase	
	first	second
minimum	68.75% (11/16)	100% (4/4)
maximum	100% (16/16)	100% (4/4)
average	80% (64/80)	100% (20/20)

An example of sound localization rule-based decisions is shown in Fig. 5 on the basis of the *Rosetta* program window display.

The screenshot shows a window titled "Rosetta - [Results]" with a menu bar (File, Edit, View, Window, Help) and a toolbar. The main content area displays a confusion matrix and ROC curve information.

		Predicted				
		000	090	180	270	
Actual	000	4	0	0	0	1.0
	090	1	3	0	0	0.75
	180	2	0	2	0	0.5
	270	0	0	0	4	1.0
		0.571429	1.0	1.0	1.0	0.8125
ROC	Class	Undefined				
	Area	3.402820e+038				
	Std. error	3.402820e+038				
	Thr. (0, 1)	3.402820e+038				
	Thr. acc.	3.402820e+038				

At the bottom of the window, it says "Ready" and "NUM Tuesday, June 12, //".

Fig. 5. Example of sound localization decisions for low-resolution analysis. In the region of the window denoted as “Actual” the predicted angle values versus really existing ones is shown and the number of relevant cases is displayed

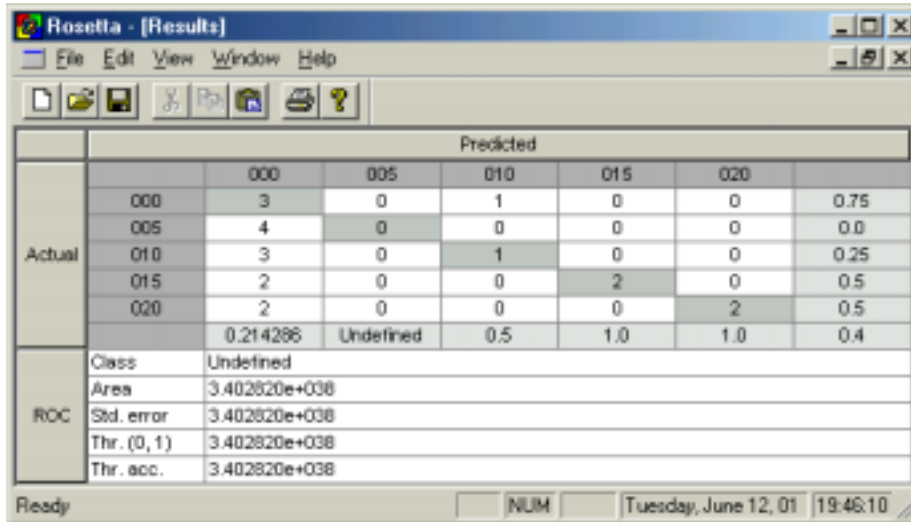
Low Resolution Source Localization of Noisy Signal

For the purpose of the experiments concerning source localization of noisy signal white noise was mixed with the sound samples. Experiments were performed for the noise levels relative to the maximum RMS level of the speech signal as follows: 0 dB, -20 dB and -40 dB. Experiments were performed in two phases, according to the procedure illustrated above for low-resolution source localization without noise. Results of the experiments are presented in Table 8 (again the numbers given in brackets represent properly classified examples versus all tested examples).

Table 8. Localization accuracy of noisy signal for low-resolution analysis

accuracy	phase	
	first	second
minimum	40% (8/20)	60% (3/5)
maximum	100% (20/20)	100% (5/5)
average	63% (63/100)	88% (22/25)

An illustration of decisions taken for the case noisy sound source localization is presented in Fig. 6.



		Predicted					
		000	005	010	015	020	
Actual	000	3	0	1	0	0	0.75
	005	4	0	0	0	0	0.0
	010	3	0	1	0	0	0.25
	015	2	0	0	2	0	0.5
	020	2	0	0	0	2	0.5
		0.214286	Undefined	0.5	1.0	1.0	0.4
ROC	Class	Undefined					
	Area	3.402820e+038					
	Std. error	3.402820e+038					
	Thr. (0, 1)	3.402820e+038					
	Thr. acc.	3.402820e+038					

Fig. 6. Decisions in the case of noisy sound source localization. As previously, in the region of the window denoted as “Actual” the predicted angle values versus really existing ones is shown and the number of relevant cases is displayed

4.4 High Resolution Source Localization

The experiments on high-resolution source localization were also performed. Sound source was located at the angle of 0°, 5°, 10°, 15° and 20°. According to the preliminary experiments the number of discretization intervals was increased up to five. Experiments were performed in two phases according to testing results completed for low resolution source localization. The results are presented in Table 9.

Table 9. High resolution sound localization accuracy

	phase	
	first	second
accuracy	first	second
minimum	40%	66,6%
maximum	100%	100%
average	63%	89,2%

An exemplary decision set for high-resolution source localization is presented in Fig. 7.

		Predicted					
		000	005	010	015	020	
Actual	000	3	0	0	1	0	0.75
	005	1	2	0	1	0	0.5
	010	1	0	2	1	0	0.5
	015	0	0	0	4	0	1.0
	020	0	0	0	1	3	0.75
		0.6	1.0	1.0	0.5	1.0	0.7
ROC	Class	Undefined					
	Area	3.402820e+038					
	Std. error	3.402820e+038					
	Thr. (0, 1)	3.402820e+038					
	Thr. acc.	3.402820e+038					

Fig. 7. Sound source localization decisions for high-resolution analysis. In the region of the window denoted as “Actual” the predicted angle values versus really existing ones is shown and the number of relevant cases is displayed as previously

All presented results were obtained using a randomly selected syllable, meanwhile system learned employing all other prerecorded syllables. As it results from the presented data, the direct application of modular network structure allowed detection of the direction of arriving sound with good accuracy, in some cases equal even to 100%.

5 Application of the Neuro-Rough Algorithm

The last group of experiments employed hybridized neuro-rough system as is presented in the block diagram in the Fig. 8. The system consists of some consecutive blocks described below.

Multilayer Neural Network

The system uses 256 separated multilayer neural networks (MLN). Each of MLN is related to a separate spectral component. Also, each MLN has an input layer, a hidden layer, and an output layer. The input layer consists of 56 neurons. The output layer includes 5 neurons (providing binary representation of DSA). Number of neurons in the hidden layer was altered during experiments. Each neuron adopted continuous nonlinear function varying in the range from 0 to 1. The MLNs were trained employing the Resilient Back Propagation method (see eqs. 6-9).

Classification Module

The direction of sound arrival (DSA) was encoded in 5-bit vectors, and such a representation was used to train MLNs. As a result of DSA estimation by 256 MLNs, 256

vectors were obtained consisting of 5 elements having values between 0 and 1. The discretisation module was connected to each neuron in the output layer of each of 256 MLN, and realized a division of real numbers to some subintervals. The division \prod_A on $[a, b]$ is defined as the set of k subintervals:

$$\prod_A = \{[a_0, a_1), [a_1, a_2), \dots, [a_{k-1}, a_k]\} \quad (21)$$

where: $a_0 = a, a_{i-1} < a_i, i = 1, \dots, k, a_k = b$

This approach to quantization is based on calculating division points a_i . After quantization, the parameter value is transformed into the number of the subinterval to which this value belongs. In most experiments the simplest division called binary quantization was used, where: $|\prod_A| = 2$.

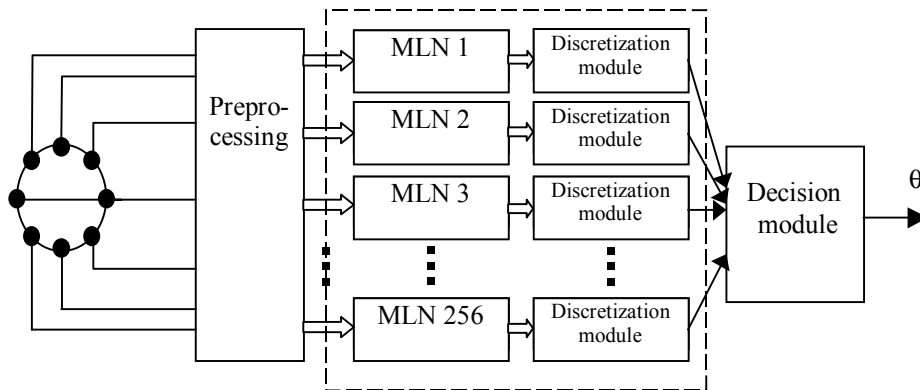


Fig. 8. Block diagram of speaker localization system based on neuro-rough approach

Voting module

The voting module is the last part of the presented system. Its main role is interpretation of binary vectors as a concrete DSA, and storing each of 256 answers in the memory. Finally, estimation of DSA is made by sorting the table content of DSA according to the rules acquired during the training process. Again, the *Rosetta* system was used as a decision tool based on the rough set theory.

Tests and results

The set of 18 directions (DSAs) defined as: *rear_left, left, ..., front, front_right, right, ..., rear_left* was defined. The results of first tests are presented in Tab.10. The 15

learning vectors for each of DSA were used (15 vectors x 18 DSA = 270 learning vectors). A continuous linear unipolar, neuron activation function was used. The three different numbers of neurons in the hidden layer were tested. The percentage value in this table indicates the range of wrong answers of 256 classification modules estimating a single DSA.

Table 10. Percentage ranges of wrong decisions among individual networks contained in the modular structure – employment of 15 learning vectors for each DSA – various neurons number in hidden layer

Number of neurons in hidden layer		21	35	45
Crisp analysis	Loud sounds	63.5%- 67.66%	56.58%- 65.84%	61.78% - 66.28%
	Quiet sounds	71.72% - 77.19%	70.29% - 74.98%	69.49% - 75.48%
	Noise	89.08% - 91.78%	88.72% - 91.45%	88.98% - 91.75%
Analysis allowing uncertainty margin 5 ⁰	Loud sounds	43.45% - 48.78%	41.82% - 46.07%	40.65%-46.27%
	Quiet sounds	53.08% - 56.21%	49.85% - 54.06%	48.94%- 54.64%
	Noise	75.5% - 79.9%	76.22% - 78.58%	75.35% - 79.82%

The next experiments were made with increased number of learning vectors. The results derived from the system with 36 learning vectors for each DSA is presented in Tab. 11.

Table 11. Percentage ranges of wrong decisions among individual networks contained in the modular structure – employment of 36 learning vectors for each DSA

Number of neurons in hidden layer		21
Crisp analysis	Loud sounds	52.76% - 54.56%
	Quiet sounds	61.81% - 68.29%
	Noise	87.35% - 89.54%
Analysis allowing uncertainty margin 5 ⁰	Loud sounds	34.64% - 37.70%
	Quiet sounds	43.10% - 47.48%
	Noise	74.33% - 76.56%

Subsequent experiments were organized employing 54 learning vectors for each DSA. The MLNs in the decision module included 45 neurons in the hidden layer. In this test three boundary values in the classification module were used. The obtained results are shown in Tab. 12.

Table 12. Percentage ranges of wrong decisions among individual networks contained in the modular structure – employment of 54 learning vectors for each DSA

Boundary value			0.4	0.5	0.6
Crisp analysis	Loud sounds		50.24% - 53.26%	49.63% - 52.97%	49.44% - 53.54%
	Quiet sounds		58.29% - 65.65%	58.12% - 65.71%	58.01% - 65.80%
	Noise		87.52% - 89.78%	87.50% - 89.50%	87.65% - 89.56%
Analysis allowing uncertainty margin 5 ⁰	Loud sounds		32.86% - 37.13%	32.51% - 37.28%	32.51% - 37.74%
	Quiet sounds		40.71% - 46.05%	40.47% - 45.70%	40.71% - 45.96%
	Noise		73.35% - 77.34%	73.44% - 77.24%	73.46% - 77.06%

The obtained results indicate high correlation between number of learning vectors and decision quality. It was required to increase the number of neurons in the hidden layer simultaneously with increasing number of learning vectors employed during the learning phase. In the test with 54 learning vectors about 50% of 256 outputs of the decision module provided correct answers. With this value it was possible to create properly working DSA recognition system. As was observed, the number of correct answers was higher than the number of answers indicating other (wrong) DSAs. Taking advantage of this dependence the voting module was created. The rough rules were employed to the interpretation of current state of modular neural network outputs. Despite than considerable portion of individual neural network outputs might provide inaccurate decisions, by using this simple method it was possible to get 100% efficiency of correct DSA recognition in each test run. The best solution was to use 54 learning vectors to train MLNs and 45 neurons in hidden layer in each MLN. Such system is characterized by best results in DSA estimation. In the other hand the whole decision system is quite computationally intensive.

The presented work show, that localization of DSA with modular neural network structure hybridized with rough sets was possible and provided perfect results.

6 Conclusions

The performed experiments proved that sound source position can be localized successfully using neural networks or employing a combination of correlation analysis and rough-set rule-based processing or modular neural networks plus rough set decision module (most effective solution). The rough-set approach provides generalization ability to the system allowing proper source localization even with noisy and reverberated sound patterns.

The obtained results demonstrate also that non-linear filters based on learning decision algorithms may provide quite an effective tool for the detection of sound source position in case of non-correlated noise presence. The analyzed problem is highly non-deterministic one in this case. Consequently, intelligent filters used in a sound acquisition system can cause a significant improvement in speech intelligibility and an increase in the signal-to-noise ratio. The results open also a possibility to employ the intelligent sound localization algorithms to some experimental teleconference systems. Additionally, they provide a practical example how through the use of

neuro-rough hybridization, real numbers representing angles of direction of sound arrival could be associated with words describing directions in natural language (such as for example: *rear left; left, front left, front, front right,..., rear right*). Therefore, the discussed problem and its solution demonstrate another way to computing with words.

7. References

1. Berdugo, B., Doron, M.A., Rosenhouse, J., Azhari, H.: On Direction Finding of an Emitting Source from Time Delays. *J. Acoustical Society of America* 106 (1999) 3355–3363
2. Bodden, M.: Modeling Human Sound-Source Localization and the Cocktail-Party-Effect. *Acta Acustica* 1 (1993) 43–55
3. Brandstein, M.S.: A Pitch-Based Approach to Time-Delay Estimation of Reverberant Speech. *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, Mohonk, New Paltz, NY, USA* (1997)
4. Chang, W.-F., Mak, M.W.: A Conjugate Gradient Learning Algorithm for Recurrent Neural Networks. *Neurocomputing* 24 (1999) 173–189
5. Chern, S.-J., Lin, S.-H.: An Adaptive Time Delay Estimation with Direct Computation Formula. *J. of Acoustical Society of America* 96 (1994) 811–820
6. Czerniawski, J.: The verification of new algorithms for identifying sound source position and sound acquisition methods based on spatial filtering. M.Sc. Diploma Thesis. Technical University of Gdansk, Gdansk, Poland (2001) (*in Polish*)
7. Czyzewski, A., Krolikowski, R.: Neuro-Rough Control of Masking Thresholds for Audio Signal Enhancement. *Neurocomputing* 36 (2001) 5-27
8. Datum, M.S., Palmieri, F., Moiseff, A.: An Artificial Neural Network for Sound Localization Using Binaural Cues. *J. of Acoustical Society of America* 100 (1996) 3372–3383
9. Day, S.P., Davenport, M.R.: Continuous-Time Temporal Back-Propagation with Adaptable Time Delays. *IEEE Trans. on Neural Networks* (1993)
10. Elman, J.L.: Finding Structure in Time. *Cognitive Science* 14 (1990) 179–211
11. Goudreau, M.W., Giles, C.L., Chakradhar, S.T., Chen, D.: First-Order vs. Second-Order Single Layer Recurrent Neural Networks. *IEEE Trans. on Neural Networks* 5 (1994) 511–518
12. Goudreau, M.W., Giles, C.L.: Using Recurrent Neural Networks to Learn Structure of Interconnection Networks. *Neural Networks* 8 (1995) 793–820
13. Hartmann, W.M.: How We Localize Sound. *Physics Today* 11 (1999) 24–29
14. Horne, B.G., Giles, C.L.: An Experimental Comparison of Recurrent Neural Networks. In: Tesauro, G., Touretzky, D., Leen, T. (eds.): *Neural Information Processing Systems, Vol. 7*. MIT Press (1995) 697–705
15. Jacovitti, G., Scarano, G.: Discrete Time Techniques for Time Delay Estimation. *IEEE Trans. on Signal Processing* 41 (1993) 525–533
16. Khalil, F., Lullien, J.P., Gilloire, A.: Microphone Array for Sound Pickup in Teleconference Systems. *J. of Audio Engineering Society* 42 (1994) 691–700

17. Czyżewski, A., Kostek, B., Lasecki, J.: Microphone Array for Improving Speech Intelligibility. Proc.: 20. Tonmeistertagung, International Convention on Sound Design, 20-23.11.1998, Stadthalle Karlsruhe, Germany, pp. 428-434
18. Kostek, B., Czyżewski, A., Lasecki, J.: Spatial Filtration of Sound for Multimedia Systems, IEEE Signal Processing Society 1999 Workshop on Multimedia Signal Processing, Vol. CD-ROM Proceedings, Copenhagen, Denmark (1999) 209-213
19. Kostek, B., Czyżewski, A., Lasecki, J.: Computational Approach to Spatial Filtering. 7th European Congress on Intelligent Techniques and Soft Computing, (EUFIT'99), Vol. CD-ROM Proceedings, Aachen, Germany (1999) 242
20. Lasecki, J., Kostek, B., Czyżewski, A.: Neural Network-based Spatial Filtration of Sound. 106th Audio Eng. Soc. Convention, Preprint No. 4918, Munich, Germany (1999)
21. Lin, T., Giles, C.L., Horne, B.G., Kung S.Y.: A Delay Damage Model Selection Algorithm for NARX Neural Networks. IEEE Trans. on Signal Processing, "Special Issue on Neural Networks", 11 (1997) 2719-2730
22. Lin, T., Horne, B.G., Tino P., Giles, C.L.: Learning Long-Term Dependencies in NARX Recurrent Neural Networks. IEEE Trans. on Neural Networks 7 (1996) 1329-1351
23. Mahieux, Y., le Tourneur, G., Saliou, A.: A Microphone Array for Multimedia Workstations. J. of Audio Engineering Society 44 (1996) 365-372
24. Fahlman, S.: An Empirical Study of Learning Speed in Back-Propagation Networks. Technical Report CMU-CS-88-162 of Carnegie Mellon University in Pittsburgh, USA, September 1988.
25. Riedmiller, M., Braun H.: A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm, Proc. of IEEE International Conference on Neural Networks, San Francisco (1993) 586-591.
26. Siegelmann, H.T., Horne, B.G., Giles, C.L.: Computational Capabilities of Recurrent NARX Neural Networks. IEEE Trans. on Systems, Man and Cybernetics - Part B: Cybernetics 2 (1997) 208-228
27. Sum, J.P.F., Kan, W.-K., Young, G.H.: A Note on the Equivalence of NARX and RNN. Neural Computing & Applications 8 (1999) 33-39
28. Szczerba, M. Sound Source Localization Based on Rough-Set Approach. Technical Report. Sound & Vision Engineering Department, TU Gdansk, Poland, (2001)
29. Wang, H., Chu, P.: Voice Source Localization for Automatic Camera Pointing System in Videoconferencing. Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, Mohonk, New Paltz, NY, USA (1997)
30. Williams, R.J., Zipser, D.: A Learning Algorithm for Continually Running Fully Recurrent Neural Networks. Neural Computation 1 (1989) 270-280
31. Zhang, M., Er, M.H.: An Alternative Algorithm for Estimating and Tracking Talker Location by Microphone Arrays. J. of Audio Engineering Society 44 (1996) 729-736
32. Ziskind, I., Wax, M.: Maximum Likelihood Localization of Multiple Sources by Alternating Projection. IEEE Trans. on Acoustics, Speech and Signal Processing 36 (1988) 1553-1560
33. Zurada, J.M.: Introduction to Artificial Neural Networks. West Publishing Company, St. Paul New York Los Angeles San Francisco (1992)

34. Øhrn, A.: Discernibility and Rough Sets in Medicine: Tools and Applications, Ph.D. Thesis, Department of Computer and Information Science, Norwegian University of Science and Technology, Trondheim, NTNU Report 1999:133, IDI Report (1999)

Acknowledgments

The research was sponsored by the Committee for Scientific Research, Warsaw, Poland. Grant No. 8 T11D 00218