Available at www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT®

Pattern Recognition Letters 24 (2003) 921–933

Pattern Recognition Letters

www.elsevier.com/locate/patrec

Automatic identification of sound source position employing neural networks and rough sets

Andrzej Czyzewski

Sound and Vision Engineering Department, Technical University of Gdansk, ul. Narutowicza, 80-952 Gdansk, Poland

Abstract

Methods for the identification of direction of the incoming acoustical signal in the presence of noise and reverberation are investigated. Since the problem is a non-deterministic one, thus applications of two learning algorithms, namely neural networks and rough sets are developed to solve it. Consequently, two sets of parameters have been formulated in order to discern target source from unwanted sound source position and then processed by learning algorithms. The applied feature extraction methods are discussed, training processes are described and obtained sound source localizing results are demonstrated and compared.

Keywords: Neural networks; Rough sets; Signal processing; Beamforming

1. Introduction

The problem of separating a desired signal in noisy conditions is vital in many domains related to communications, especially in teleconferencing applications. Speech signals incoming from various directions not only interfere with the target signal but also can obscure it. Consequently, the main purpose of spatial filtering technique applied to advanced teleconferencing is to attenuate the unwanted signal through the recognition of direction of arrival of target sound. Another issue addressed by this kind of applications is automatic sound source tracking that is applicable to advanced video teleconferencing.

Often, spatial filtering algorithms introduce signal distortions disturbing the sound perception while providing low increase in the signal-to-noise ratio. The problem of separating sound from a particular direction is especially difficult in the presence of reverberation and background noise, because these parasite sounds are usually omnidirectional. Since the reverberation and noise are stochastic, the problem of recognizing the arrival direction of sound provides practically a non-deterministic problem. Therefore, statistical or learning algorithms should be employed to solve it.

Some previous investigations performed by authors showed that a neural network may provide an effective non-linear filtering of an acoustic signal transformed into the frequency domain (Czyzewski et al., 1998; Kostek et al., 1999; Czyzewski et al., 1999; Lasecki et al., 1999; Czyzewski and

E-mail address: andcz@sound.eti.pg.gda.pl (A. Czyzewski).

0167-8655/03/$ - see front matter
PII: S0167-8655(02)00204-0

Krolikowski, 1999; Czyzewski and Krolikowski, 2001). However, the previously used set of parameters based on psychoacoustical principles was reviewed and found to be sub-optimal, mainly because some of these parameters were not orthogonal; thus they might be eliminated from the feature vector and other orthogonal parameters could be added instead of them. Consequently, a new set of parameters was formulated to identify signal localization. Moreover, a decision system, based on rough sets, was applied to the task of processing parameters representing acoustic signals.

The results of investigations in the domain of automatic recognition of sound arrival direction are presented in the paper. Some conclusions are drawn on the basis of experiments concerning the application of learning algorithms to spatial filtering of sound.

2. Experimental background

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 (Bodden, 1993; Hartmann, 1999). According to the present state of knowledge, perception of sound directivity by the human binaural system is based on the following two principal entities (Hartmann, 1999):

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 identification of sound source localization can be performed by means of a microphone array which can be either linear or non-linear (Khalil et al., 1994; Czyzewski et al., 1999). The lay-out of the sound acquisition system is presented in Fig. 1. Theoretically, under the ideal conditions, a signal $x_i(t)$ received from ith microphone of the linear array of microphones and in tth moment of time can be described as follows:

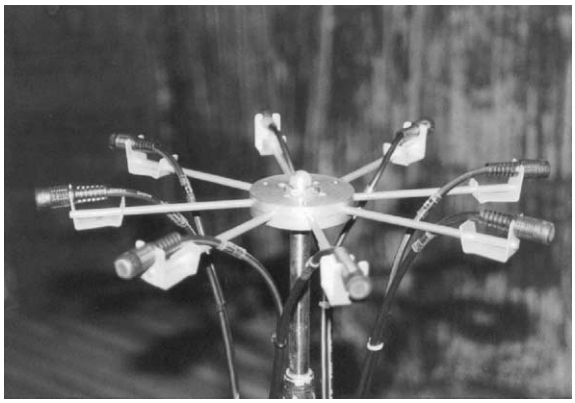

Fig. 1. The microphone matrix used for sound acquisition.

$$x_i(t) = \alpha_i s[t - (i-1)\tau], \tag{1}$$

where α_i, the attenuation coefficient for ith microphone; $s(t)$, the source signal; and τ, the 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 various distortions occur, and interference signals such as background noise and reverberated sounds, and others may be present. Hence, the signals received by a linear microphone array are expressed by the following relationships:

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

where $h_i(t)$, the impulse response of the reverberant channel associated with ith microphone; $n_i(t)$, the ambient noise received by ith microphone; and $*$, the convolution operator.

The above set of equations reflects the basic principle of linear systems, namely that the response signal can be calculated as convolution of

excitation and impulse response of the system. Considering the reverberation, one can notice that it could be represented by repeated excitations, because the sound is reflected many times in a room and it returns to the microphone with some progressive delays. Moreover, the system introduces some noise which normally has additive character.

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 (Brandstein, 1997), adaptive filtration (Chern and Lin, 1994) or computation of relevant eigenvalue vectors and matrices (Berdugo et al., 1999). In turn, in the case of tracking or localizing a number of sources, the maximum likelihood-based methods are exploited (Ziskind and Wax, 1988) representing the statistical approach. More details can be found in the abundant literature on the localization of acoustic sources for multimedia applications (Jacovitti and Scarano, 1993; Khalil et al., 1994; Mahieux et al., 1996; Wang and Chu, 1997; Zhang and Er, 1996).

2.1. Acquisition of sound material

The experiments carried-out at the Sound and Vision Engineering Department of the TU Gdansk consisted of several stages. First, phonemes were recorded in an anechoic chamber, then in a standard room. All signals coming from other than front direction were treated as unwanted ones. The recordings were made with the use of a circular array of microphones. The array consisted of eight electret microphones set on the circumference of a 15 cm radius rim, and was fixed 1.58 m from the floor (see Fig. 1). Eight mono tracks were recorded simultaneously. The recording parameters were as follows: 16 bit/sample and the sampling frequency was set 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 $-45°$ to $+45°$ recorded at every 15° intervals.

The feature extraction process was carried out as is described in the next paragraph. The learning algorithms used there are also described. The desired attenuation level was defined for each pattern.

3. Feature extraction process

During the feature extraction process the signal was divided into frames of the length of 256, 512 or 1024 samples. Multiplying it by the sample bit resolution constant and the size of the feature vector gives the size of memory buffer needed for further processing. In some cases it may result approximately in 20 GB of the engaged mass memory volume.

In the practical application of spatial filtration (beam forming) in hearing aids, the following parameters coming from principles of psychoacoustics can be efficiently exploited (Kostek et al., 1999):

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

$$A_i = |\angle L_i - \angle R_i|, \tag{3}$$

where L_i and R_i are the magnitudes of the ith spectral bin for the left and right channel, respectively and $\angle L_i$, $\angle R_i$ represent the phases of individual spectral components.

The above definition of sound features is justified by psychoacoustic principles causing the spatial hearing in humans. According to the present state of knowledge, perception of sound directivity by the human binaural system is based on the following two principal entities (Hartmann, 1999):

- difference of intensities of waveforms in the left and in right ears;
- difference of arrival times of relevant waveforms in the both ears, which is equivalent to a phase difference of the waveforms.

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

$$M_{ij}^k = \frac{\min(|\mathrm{Ch}_i^k|, |\mathrm{Ch}_j^k|)}{\max(|\mathrm{Ch}_i^k|, |\mathrm{Ch}_j^k|)};$$
$$D_{ij}^k = \frac{\mathrm{Ch}_i^k - \mathrm{Ch}_j^k}{|\mathrm{Ch}_i^k| + |\mathrm{Ch}_j^k|}; \quad A_{ij}^k = \angle\mathrm{Ch}_i^k - \angle\mathrm{Ch}_j^k. \tag{4}$$

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:

$$M_{ij}^k = \frac{\min(|\mathrm{Ch}_i^k|, |\mathrm{Ch}_j^k|)}{\max(|\mathrm{Ch}_i^k|, |\mathrm{Ch}_j^k|)};$$
$$A_{ij}^k = \angle\mathrm{Ch}_i^k - \angle\mathrm{Ch}_j^k. \tag{5}$$

In the experiments, eight-channel signals were examined, and hence the following sets of parameters can be considered:

Type ***A*** all mutual combinations of channels yielding 56 parameters per spectral bin;
Type ***B*** combination of opposite channels yielding 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, a learning algorithm assumes a structure of a modular network where a separate sub-system is dedicated for each spectral bin. The final decision is made on the basis of the maximum outputs of all sub-algorithms.

Table 1 shows the size of input vector (*vectorSize*) and storage consumption (*storage*) in relation to parameter types (***A*** or ***B***) and various lengths of analysis frame (512, 1024 and 2048). The data refer to single direction of incoming sound. Tables 2–4 assemble information on conditions of learning algorithm with regard to various types of input vector (***V1***, ***V2*** and ***V3***).

3.1. Selection of parameters

The selection of the parameters was made considering the following issues:

- the size of an input vector;

Table 1
Analysis of storage load

$N = 512$	$N = 1024$	$N = 2048$
A: vectorSize = 14,336 ***B***: vectorSize = 2,048	***A***: vectorSize = 28,672 ***B***: vectorSize = 4,096	***A***: vectorSize = 57,344 ***B***: vectorSize = 8,192
A: storage ≈ 20.34 MB ***B***: storage ≈ 2.91 MB	***A***: storage ≈ 20.13 MB ***B***: storage ≈ 2.88 MB	***A***: storage ≈ 19.69 MB ***B***: storage ≈ 2.81 MB

Denotation: N—size of a sample frame taken for spectral analysis.

Table 2
Analysis of training conditions for the input vector ***V1***

$N = 512$	$N = 1024$	$N = 2048$
A: vectorSize = 14,336 ***B***: vectorSize = 2,048	***A***: vectorSize = 28,672 ***B***: vectorSize = 4,096	***A***: vectorSize = 57,344 ***B***: vectorSize = 8,192
Material for training: 186 vectors/s	Material for training: 92 vectors/s	Material for training: 45 vectors/s

Table 3
Analysis of training conditions for the input vector ***V2***

$N = 512$	$N = 1024$	$N = 2048$
A: vectorSize = 57 ***B***: vectorSize = 9	***A***: vectorSize = 57 ***B***: vectorSize = 9	***A***: vectorSize = 57 ***B***: vectorSize = 9
Material for training: 256 vectors/frame; 47,616 vectors/s	Material for training: 512 vectors/frame, 47,104 vectors/s	Material for training: 1024 vectors/frame, 46,080 vectors/s

Table 4
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: vectorSize = 56 ***B***: vectorSize = 8	***A***: vectorSize = 56 ***B***: vectorSize = 8	***A***: vectorSize = 56 ***B***: vectorSize = 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: 1024 decision/s

- number of training vectors (the cardinality of a training set);
- storage complexity.

In the first case, the large size of an input vector results in the heavy load of decision algorithms. In turn, in the second case, assuming also that the neural network is used as a decision algorithm, the large number of training vectors evokes the problem of capacity of a neural net and selection of its architecture. However, the capacity can be increased by incrementing neural connections, which is strictly related to the increment of the size of the weight matrices. As regards the large storage complexity, in such a case there arises a problem of processing of such large structures in the memory of personal computers or specialized digital signal processors. Taking all the above into account the following sets of parameters were chosen for experiments:

1. vector type ***V1***, parameters type ***A***, the size of an analysis frame $N = 512$.
2. vector type ***V3***, parameters type ***A***, all sizes of an analysis frame ($N = 512$, $N = 1024$, $N = 2048$).
3. vector type ***V3***, parameters type ***B***, all sizes of an analysis frame ($N = 512$, $N = 1024$, $N = 2048$).

4. Neural network as a spatial filter (beam former)

Artificial neural networks were applied in many areas of engineering and audio signal processing (Czyzewski and Krolikowski, 1999; Czyzewski and Krolikowski, 2001), since they are capable to process uncertain information. The rationale of neural nets application to the present problem solving results from the non-deterministic nature of the features of sound accompanied by reverberant reflections and noise. Neural nets have already been applied for sound localization (Datum et al., 1996), however those attempts were based on some feed-forward structures (Zurada, 1992) and not on the set of parameters as above. The feed-forward networks do not offer such feasibility as recurrent ones do, especially in the field of time series modeling (Day and Davenport, 1993) or mapping of a complex process dynamics (Chang and Mak, 1999; Elman, 1990).

4.1. Neural network training algorithms

The training algorithms considered here are: the general and simplified Fahlman's algorithm (QuickPROP) (Fahlman, 1988) and the resilient propagation (RPROP) (Reidmiller and Braun, 1993).

4.2. The general Fahlman's algorithm (Fahlman I)

The weight update rule for a single weight w_{ij} in the kth cycle is:

$$\Delta w_{ij}^k = -\eta^k S_{ij}^k + \alpha_{ij}^k \Delta w_{ij}^{k-1}, \tag{6}$$

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

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

and the learning rate η^k and the momentum ratio α_{ij}^k vary according to formulae that can be found in the literature (Fahlman, 1988).

4.3. The simplified Fahlman's algorithm (Fahlman II)

In the simplified version of the QuickPROP algorithm implemented for the purpose of experiments, the weight update rule is expressed by the following relationship:

$$\Delta w_{ij}^k = \begin{cases} \alpha_{ij}^k \Delta w_{ij}^{k-1}, & \text{for } \Delta w_{ij}^{k-1} \neq 0, \\ -\eta_0 \cdot \nabla E(\Delta w_{ij}^k), & \text{otherwise, i.e. } \Delta w_{ij}^{k-1} = 0, \end{cases} \tag{8}$$

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

$$\alpha_{ij}^k = \min\left\{ \frac{\nabla E(\Delta w_{ij}^k)}{\nabla E(\Delta w_{ij}^{k-1}) - \nabla E(\Delta w_{ij}^k)}, \alpha_{\max} \right\} \tag{9}$$

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

4.4. The RPROP algorithm

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

$$\Delta w_{ij}^k = -\eta_{ij}^k \operatorname{sgn}(\nabla E(\Delta w_{ij}^k)), \tag{10}$$

where the learning rate η^k assumes values according to the rules defined in (Reidmiller and Braun, 1993).

5. Results of neural nets application

Tables 5–7 show results of experiments with neural networks. These are presented with regard to training algorithm and signal direction. The number of training and testing vectors are also shown in these tables.

Table 6 represents best results obtained from all the tested RNN-based beam formers. Here, the number of training vectors was 515 and the testing ones 221. The results were obtained with the modular neural network employed to the processing of ***V*3**-type vectors. As is seen from the table, these results depend on both: the training algorithm and the signal direction. The best scores were obtained for azimuths 0° and 45°. The feature vectors of the type ***A*** representing all mutual combinations of channels yielding 56 parameters per spectral bin were employed in the relevant experiment. On the other hand, slightly better results were obtained with the simplified Fahlman's

Table 5
Results of direction detection for the vector type ***V*1**

Direction	Fahlman I		Fahlman II		RPROP	
	Epochs	Scores (%)	Epochs	Scores (%)	Epochs	Scores (%)
−45°	44,168	**83**	59,136	**86**	75,246	**85**
−30°	37,290	**78**	40,206	**82**	71,329	**83**
−15°	50,891	**84**	48,282	**81**	73,590	**82**
0°	29,190	**80**	48,257	**86**	68,124	**84**
15°	34,506	**82**	55,902	**79**	81,299	**78**
30°	39,251	**84**	53,617	**78**	59,783	**80**
45°	41,889	**80**	42,689	**83**	63,775	**82**

$N = 512$; parameter type ***A***; training/testing vectors: 1042/446.

Table 6
Results of direction detection for the vector type ***V*3**

Direction	Fahlman I		Fahlman II		RPROP	
	Epochs	Scores (%)	Epochs	Scores (%)	Epochs	Scores (%)
−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**

$N = 1024$; parameter type ***A***; training/testing vectors: 515/221.

Table 7
Results of direction detection for the vector type ***V3***

Direction	Fahlman I		Fahlman II		RPROP	
	Epochs	Scores (%)	Epochs	Scores (%)	Epochs	Scores (%)
−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**

$N = 2048$; parameter type ***A***; training/testing vectors: 252/108.

algorithm (Fahlman II) than for other training algorithms. Though the rate of convergence is higher and the detection of the direction of sound arrival is slightly better, no ultimate conclusion as to supremacy of this training algorithm could be derived on the basis of obtained results.

6. Rough set theoretic approach

As stated before, the rationale for extraction of sound features for the processing by the neural nets is stemming from some psychoacoustic principles. The features basing on spectral components magnitude or phase differences plus the neural network acting as decision system emulate some means for the recognition of sound arrival direction possessed by humans. Another approach could be based on mathematical modeling of dependencies among data. Therefore, the correlation analysis of signals arriving from various directions is employed and the rough set algorithm capable of finding hidden relations among data (Pal and Skowron, 1999; Polkowski and Skowron, 1998) was used in subsequent experiments. The results of experiments with this "rough-correlation" method is presented below (Szczerba, 2001).

The block diagram of the sound source localization method, used here, is shown in Fig. 2.

The input signal of each microphone is first passed through a set of comprehensive band-pass filters. Subsequently, correlation analysis is performed for each pair of microphones and each sub-band. As an output, a set of correlation parameters is calculated. The sound source is then

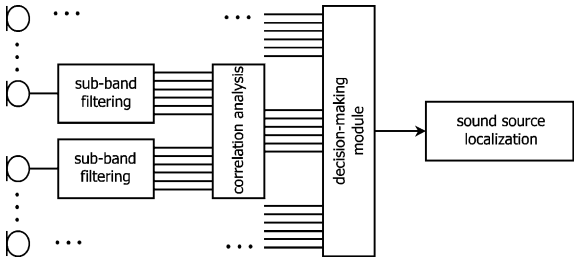

Fig. 2. Rough-set supported sound source localization lay-out.

localized using the 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 experiments only pairs of counter-positioned microphones were considered. It gained significant reduction of computing power requirements without losing performance capabilities.

6.1. 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 8.

Sub-band filtering was performed using spectral filtering. A *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:

Table 8
Boundary frequencies of sub-bands

Band no.	Lower bound (Hz)	Higher bound (Hz)
1	20	100
2	100	200
3	200	400
4	400	800
5	800	1600
6	1600	3200
7	3200	6400
8	6400	20,000

$$\rho(n) = \sum_i \frac{[x(t)_i - \bar{x}(t)][y(t+n)_i - \bar{y}(t+n)]}{\sqrt{\sum_i [x(t)_i - \bar{x}(t)]^2}\sqrt{\sum_i [y(t+n)_i - \bar{y}(t+n)]^2}} \quad (11)$$

and it is simplified form as follows:

$$\rho(n) = \frac{\sum_i x(t)_i y(t+n)_i}{\sum_i x(t)_i y(t)_i}. \quad (12)$$

The correlation maximum should correspond to time-alteration between microphones of concern. 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 defined as (Cook et al., 1992):

$$\mathrm{AMDF}(n) = \sum_i |x(t)_i - y(t+n)_i|. \quad (13)$$

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 fifth sub-band (see Table 8) is presented in Fig. 3. 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.

The peak corresponding to the time-lag between microphones can be seen clearly on the plot in Fig. 3. It can be also observed, that in some frames actual peaks do not correspond to the time-lag. Consequently AMDF function is accumulated through frames of processing according to the formula:

$$\mathrm{AMDF}_{\mathrm{ac}}(n) = \frac{1}{M}\sum_{m=1}^{M} \mathrm{AMDF}_m(n), \quad (14)$$

where M is the 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 characteristics as well as the presence of potential noise and distortions, 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 given as:

$$E_b = \frac{1}{N}\sum_{n=1}^{N} |\mathrm{AMDF}_b(n)|, \quad (15)$$

where N is the number of samples and b is an actual sub-band. These parameters are then provided for further processing using the decision system.

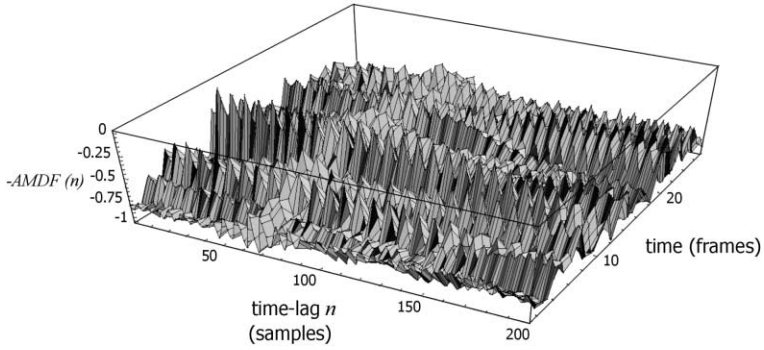

Fig. 3. (–AMDF) plot for a speech signal within fifth sub-band (range of 800–1600 Hz).

7. Decision system

The method used here is a rule-based rough-set decision system. For the purpose of the experiments the rough-set software toolbox *Rosetta* was used (Polkowski and Skowron, 1998; Øhrm, 1999).

Learning samples are processed in the following way. First, the data set—knowledge base is acquired. Knowledge base consists of objects, which are represented using conditional attributes and decision parameters. As input to 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 a total 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 9.

Consequently, the acquired data are quantized to convert real attribute values into discretized form allowing further rule processing. Based on the discrete values, attributes are analyzed in terms of discernibility investigation. Sets of attributes allowing partition of object classes are then revealed. These sets are called reducts. Consequently, rules are generated based upon reducts.

The *Rosetta* system supports a variety of quantization as well as reduct and rule generation procedures, however details on these procedures lie beyond the scope of this report (Øhrm, 1999). For the purpose of our 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 at optimizing the system efficiency and generalization ability.

8. Experiments on sound source localization

For the purpose of the 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.

8.1. 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° respectively. 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 accuracy for both phases of the experiments are shown in Table 10. An example of rule-based decisions is shown in Fig. 4.

8.2. Low resolution source localization of noisy signal

For the purpose of the experiments concerning source localization of noisy signal a white noise

Table 9
Data layout for rough-set based sound source localization

Parameter	*D1*	*A1*	*D2*	*A2*	*D3*	*A3*	...	Angle
Type	Integer	Float(4)	Integer	Float(4)	Integer	Float(4)	...	String(3)
#1	−30	32.0292	−52	112.24	4	180.875	...	000
#2	−37	32.1163	−87	96.1503	−37	181.063	⋮	090
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
#*N*	−60	35.5952	−48	151.269	−40	395.7	...	270

Table 10
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)

The numbers given in brackets represent the number of properly classified examples versus the number of all tested examples.

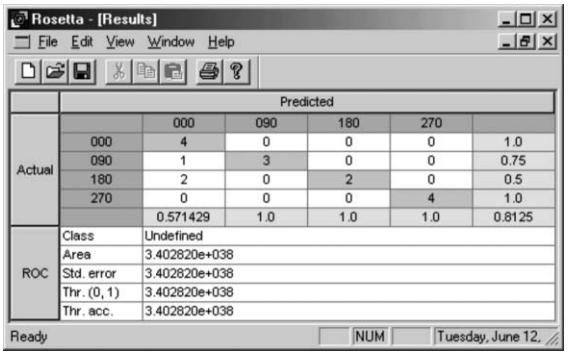

Fig. 4. Example of sound localization decisions for low-resolution analysis. In the region of the window denoted as "Actual" the predicted angle values versus really existing ones are shown and the number of relevant cases is displayed.

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 11.

An example of decisions for noisy sound source localization is presented in Fig. 5.

8.3. High resolution source localization

The experiments on high-resolution source localization were 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 the tests completed for low resolution source localization. The results are presented in Table 12.

An example of a decision set for high-resolution source localization is presented in Fig. 6.

As seen from the results of experiments, the method employing correlation analysis-based sound features and the rough set-based decision

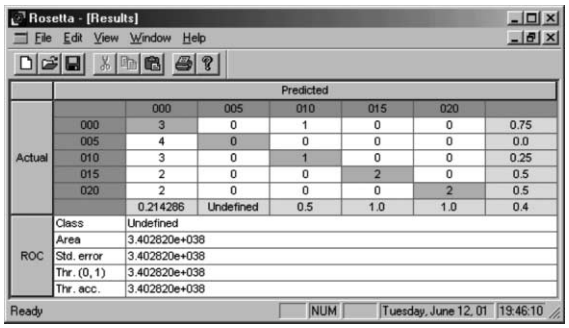

Fig. 5. Example of sound localization decisions for low-resolution analysis and noisy signal case. In the region of the window denoted as "Actual" the predicted angle values versus really existing ones are shown and the number of relevant cases is displayed.

Table 11
Localization accuracy of noisy signal for low-resolution analysis

Accuracy	Noise level					
	−0 dB		−20 dB		−40 dB	
	Phase		Phase		Phase	
	First (%)	Second (%)	First (%)	Second (%)	First (%)	Second (%)
Minimum	6.25	18.75	25	43.75	68.75	100
Maximum	31.25	68.75	56.25	93.75	100	100
Average	21.88	52.23	66.6	72.5	80	100

Table 12
High resolution sound localization accuracy

Accuracy	Phase	
	First (%)	Second (%)
Minimum	40	66.6
Maximum	100	100
Average	63	89.2

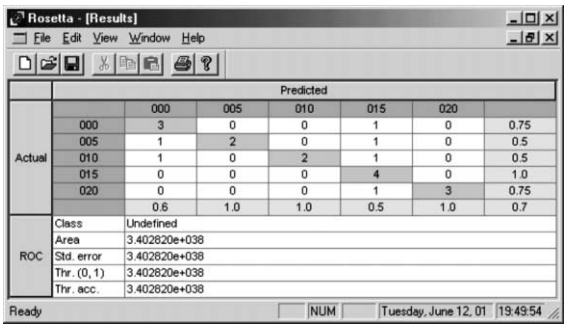

Fig. 6. Example of sound localization decisions for high-resolution analysis.

algorithm allows finding relations among audio data which makes possible to recognize efficiently the direction of sound arrival. This approach is purely mathematical, thus it differs from the previous application based on the psychoacoustically justified sound features and neural network emulating the process of acquiring information by human brains. Nevertheless, the results of both kinds of approaches are generally comparable as to their efficiency. The ''soft-correlation'' or ''rough-correlation'' algorithm brought more diversified scores from the range 40–100%, whereas the results of the previous method are more stable (78–92%). However, the neural network algorithm never produced 100% average scores, while the ''rough-correlation'' algorithm showed this possibility in many experiments.

9. Conclusions

The results demonstrate that non-linear filters based on learning decision systems may provide an effective tool for the detection of sound arrival direction, even in the presence of reverberation and noise for which the analyzed problem is highly non-deterministic one. Thus, the rationale of soft computing algorithms applications to the problem of recognition of sound direction arrival results from the non-deterministic nature of features of sound accompanied by reverberant reflections and noise. The experiments proved that the sound source can be localized successfully using psychoacoustically oriented sound features and neural networks or employing a combination of correlation analysis and rough-set rule-based processing.

The earlier application of neural networks to this task made by others was based on some feed-forward structures and different sets of parameters. Moreover, only artificially synthesized multitone test signals were employed without natural reverberation and noise. Additionally, the way earlier results were presented do not allow for assessment of the applicability of neural networks to the practical sound arrival detection systems. That is because only the minimum spatial resolution obtained in those experiments employing simulated acoustical setup was discussed by authors. In the present application the way of recognition of sound spatial properties by humans was considered and in some series of extensive experiments various feature vectors and different neural network structures and training algorithms were studied, and their effectiveness was compared employing real speech signals recorded with a microphone array.

Another approach was based on the mathematical modeling of some dependencies among data. The correlation analysis of sounds arriving from various directions was employed to that end and the rough set algorithm was tried as a decision tool in this system. The ''rough-correlation'' algorithm created in this way reveals strong knowledge generalization ability; thus the investigated system allowed proper source localization even with training employing only a small number of patterns. The method also proves its ability for sound source localization in the presence of non-correlated wide-band noise.

Consequently, intelligent filters used in sound acquisition systems may cause an increase in the

signal-to-noise ratio, thus an improvement of speech intelligibility can be expected. The results also open a possibility to employ intelligent sound localization algorithms to some experimental teleconference systems.

Acknowledgements

The research is sponsored by the Committee for Scientific Research, Warsaw, Poland, grant no. 8 T11D 00218.

References

Berdugo, B., Doron, M.A., Rosenhouse, J., Azhari, H., 1999. On direction finding of an emitting source from time delays. J. Acoust. Soc. Amer. 106, 3355–3363.

Bodden, M., 1993. Modeling human sound-source localization and the cocktail-party-effect. Acta Acust. 1, 43–55.

Brandstein, M.S., 1997. A pitch-based approach to time-delay estimation of reverberant speech. In: Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, Mohonk, New Paltz, NY, USA.

Chang, W.-F., Mak, M.W., 1999. A conjugate gradient learning algorithm for recurrent neural networks. Neurocomputing 24, 173–189.

Chern, S.-J., Lin, S.-H., 1994. An adaptive time delay estimation with direct computation formula. J. Acoust. Soc. Amer. 96, 811–820.

Cook, P.R., Morill, D., Smith, J.O., 1992. An automatic pitch detection and MIDI control system for brass instruments. In: Proceedings of Special Session on Automatic Pitch Detection ASA, New Orleans.

Czyzewski, A., Krolikowski, R., 1999. Application of fuzzy logic and rough sets to audio signal enhancement. In: Pal, S.K., Skowron, A. (Eds.), Rough Fuzzy Hybridization. A New Trend in Decision-Making. Springer-Verlag, Berlin, Heidelberg, New York, pp. 397–409.

Czyzewski, A., Krolikowski, R., 2001. Neuro-rough control of masking thresholds for audio signal enhancement. Neurocomputing 36, 5–27.

Czyzewski, A., Kostek, B., Lasecki, J., 1998. Microphone array for improving speech intelligibility. In: Proceedings 20 Tonmeistertagung, International Convention on Sound Design, Stadthalle, Karlsruhe, Germany, pp. 428–434.

Czyzewski, A., Lasecki, J., Kostek, B., 1999. Computational approach to spatial filtering. In: 7th European Congress on Intelligent Techniques and Soft Computing, (EUFIT'99), vol. CD-ROM Proceedings, Aachen, Germany, p. 242 (abstract).

Datum, M.S., Palmieri, F., Moiseff, A., 1996. An artificial neural network for sound localization using binaural cues. J. Acoust. Soc. Amer. 100, 372–3383.

Day, S.P., Davenport, M.R., 1993. Continuous-time temporal back-propagation with adaptable time delays. IEEE Trans. Neural Networks.

Elman, J.L., 1990. Finding structure in time. Cognitive Sci. 14, 179–211.

Fahlman, S., 1998. An empirical study of learning speed in back-propagation networks. Technical Report CMU-CS-88-162 of Carnegie, Mellon University in Pittsburgh, USA.

Hartmann, W.M., 1999. How we localize sound. Phys. Today 11, 24–29.

Jacovitti, G., Scarano, G., 1993. Discrete time techniques for time delay estimation. IEEE Trans. Signal Process. 41, 525–533.

Khalil, F., Lullien, J.P., Gilloire, A., 1994. Microphone array for sound pickup in teleconference systems. J. Audio Engng. Soc. 42, 691–700.

Kostek, B., Czyzewski, A., Lasecki, J., 1999. Spatial filtration of sound for multimedia systems. In: IEEE Signal Processing Society 1999 Workshop on Multimedia Signal Processing, vol. CD-ROM Proceedings, Copenhagen, Denmark, pp. 209–213.

Lasecki, J., Kostek, B., Czyzewski A., 1999. Neural network-based spatial filtration of sound. In: 106th Audio Eng. Soc. Convention, preprint no. 4918, Munich, Germany.

Mahieux, Y., le Tourneur, G., Saliou, A., 1996. A microphone array for multimedia workstations. J. Audio Engng. Soc. 44, 365–372.

Øhrm, A., 1999. 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.

Pal, S.K., Skowron, A. (Eds.), 1999. Rough-Fuzzy Hybridization: A New Trend in Decision-making. Springer-Verlag, Singapore.

Polkowski, L., Skowron, A., 1998. Rough Sets in Knowledge Discovery. Physica-Verlag, Wurzburg, Wien.

Riedmiller, M., Braun, H., 1993. A direct adaptive method for faster backpropagation learning the RPROP algorithm. In: Proceedings of the IEEE International Conference on Neural Networks, San Francisco, pp. 586–591.

Szczerba, M., 2001. Sound source localization based on rough-set approach. Technical Report. Sound and Vision Engineering Department, TU Gdansk, Poland.

Wang, H., Chu, P., 1997. Voice source localization for automatic camera pointing system in videoconferencing. In: Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, Mohonk, New Paltz, NY, USA.

Zhang, M., Er, M.H., 1996. An alternative algorithm for estimating and tracking talker location by microphone arrays. J. Audio Engng. Soc. 44, 729–736.

Ziskind, I., Wax, M., 1988. Maximum likelihood localization of multiple sources by alternating projection. IEEE Trans. Acoustics, Speech Signal Process. 36, 1553–1560.

Zurada, J.M., 1992. Introduction to Artificial Neural Networks. West Publishing Company, St. Paul, New York, Los Angeles, San Francisco.