Audio Visual System with Cascade-Correlation Neural Network for Moving Audio Visual Robot

ALEXANDER BEKIARSKI
Department of Telecommunications
Technical University
Kliment Ohridski, 8 Sofia
aabbv@tu-sofia.bg

Abstract: - The development of a moving robot requires many researches to gain the desired performance. An obligatory and necessary stage of these researches is to choose the appropriate methods for each robot function. One of these functions is speaker localization, implemented with an existing in the moving robot audio and visual system. There are many possibilities to choose the methods for processing of audio and video information from the corresponding audio and video robot sensors. The goal of this paper is to propose and test the audio visual robot system using cascade-correlation neural network for calculating the speaker position and sent this information to moving robot system for correct robot movements control.

Key-Words: - Cascade-Correlation Neural Network, Audio and Visual Robots, Speaker localization

1 Introduction

The audio and video robot system uses the information from audio and video sensors to determine the position of a person speaking to the robot. The speech from the speakers is usually received with microphone array, which is then processing to calculate the direction of arrival (DOA) of sound [1]. The speaker image is transformed with a video camera to a video signal, which also is then processing to calculate the speaker co-ordinates [2].

The audio visual moving robot system can be achieved with multiple experiments and tests usingsimulations and real robot systems. In the stage of initial and preliminary tests it is more suitable to use methods of simulations of the proposed ideas, concepts and algorithms. For this purpose in this paper, the use in the audio visual robot system a cascade-correlation neural network for calculating the speaker position. The information for the speaker position can be:

- direction of arrival (DOA), determined from the audio robot system;
- or 2D co-ordinates, calculated from visual robot system.

This information is sent to the moving robot system to control the robot movements in direction to the speaker.

The reason for this proposition is based on the knowledge, that the most methods for calculating the speaker position use cross correlation in sound or visual information processing to derive the speaker position. Therefore, the cascade-correlation neural network has the needed capabilities.

2 Moving Robot Audio Visual System

The environments or the area of moving of the robot in the sense of audio visual robot system are presented in Fig.1.

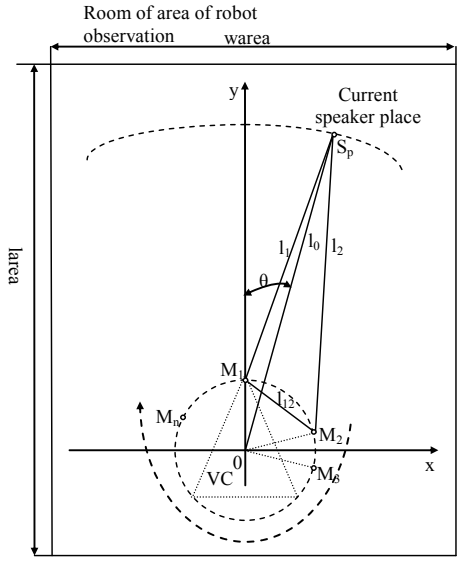

Fig.1.

In Fig.1 it is shown a particular situation of 2D area of robot observation in a open space or room. The space is presented in 2D (x,y) co-ordinate system with point of origin "0" in the center of an 2D microphone array M_1, M_2, M_3,…, M_n, placed on the moving device of the robot. In Fig.1 the place oo the robot video camera (VC) is shown. The microphone array and video camera are the sensors of the robot audio visual systems, respectively. The current speaker place is marked as S_p.

2.1 Sound Direction of Arrival from Audio System

A circular placement of the microphones with uniform spacing or distance between each of them is chosen:

$$l_{1,2} = l_{2,3} = ... = l_{n-1,n} = l_{n,1} = d_m \ . \qquad (1)$$

The microphone distances from equation (1) can be presented as number or samples:

$$n_{1,2} = \frac{l_{1,2}}{c}.f_s; \quad n_{2,3} = \frac{l_{2,3}}{c}.f_s; \ ... \ ; \quad n_{n-1,n} = \frac{l_{n-1,n}}{c}.f_s; \qquad (2)$$
$$n_{n,1} = \frac{l_{n,1}}{c}.f_s; \quad n_{d_m} = \frac{d_m}{c}.f_s \ ,$$

if the speech signal sampling frequency is f_s and the velocity of sound in the air is "c".

Also, the distance between each microphones M_1, M_2, M_3, …, M_n and the sound source or speaker source place S_p are defined as l_1, l_2, l_3, …, l_n . The distance between speaker source place S_p and origin point "0" of the co-ordinate system is defined too as l_0. These distances can be presented too as number of samples like $l_{1,2}$, $l_{2,3}$, …,$l_{n-1,n}$, $l_{n,1}$:

$$n_1 = \frac{l_1}{c}.f_s; \ n_2 = \frac{l_2}{c}.f_s; \ ... \ ; \ n_n = \frac{l_n}{c}.f_s; \ n_0 = \frac{l_0}{c}.f_s \ . \qquad (3)$$

One of the most important goal of an audio robot system is to find and estimate the sound source direction and localization. In most practical cases it is enough to estimate the sound source direction of arrival of speech from speaker or talker.

Here, the application of audio system for implementing and testing of a method, using the received speech signal from the microphone M_1, M_2, M_3, …, M_n, arranged as an 2D circular microphone array is presented.

For each pair of microphones in the microphone array it is possible to define and calculate the appropriate relation coefficient.

Using Generalized Cross-Correlation (GCC) method it is possible to estimate time delay:

$$R_{i,j}(k) = \frac{\sum_{m=0}^{N} S_i(m-k).S_j(m)}{\sqrt{\sum_{m=0}^{N} S_i(m-k)^2}.\sqrt{\sum_{m=0}^{N} S_m(m)^2}}, \qquad (4)$$

$$\text{for } i, j = 1,2,...,n \text{ and } i \neq j \ ,$$

where

k is the number of actual delay samples between the speech signals received from microphones i and j respectively;

N – number of samples in a chosen speech frame, for example 240 samples;

S_i and S_j - the speech signals received from microphones i and j respectively.

One of the microphones (for exmple M_1) in the microphone array can be chosen as front microphone and can be used as base point to calculate a set of relation coefficients between them (microphone M_1) and each other, that mean the microphones M_2, M_3, …, M_n, using equation (4):

$$R_{1,i}(k) = \frac{\sum_{m=0}^{N} S_i(m-k).S_1(m)}{\sqrt{\sum_{m=0}^{N} S_i(m-k)^2}.\sqrt{\sum_{m=0}^{N} S_1(m)^2}} \ , \qquad (5)$$

for $i, j = 2,3,..., n$.

The values of k represent the number of actual delay samples between the speech signals S_1 and S_i , received from microphone M_1 and each microphone M_i (for i=2, 3, …, n), respectively.

From equation (5) it is possible to calculate the value of of actual delay $\tau = k$, when $R_{1,i}(k)$ has a maximum value. The appropriate angle θ of the direction of arrival (DOA) can be calculated from τ .

2.2 Speaker Co-ordinates from Visual System

The algorithm of automatic speaker position finding and tracking uses a method of speaker face or body separation from the current image received with the video camera (VC). The speaker position is presented as co-ordinates of the centre of gravity x_{sp}, y_{sp} of speaker face or body:

$$x_{sp} = \frac{\sum_{i=1}^{n} a_i * x_i}{\sum_{i=1}^{n} a_i} \ , \ y_{sp} = \frac{\sum_{i=1}^{n} a_i * y_i}{\sum_{i=1}^{n} a_i}, \qquad (6)$$

where:

x_i, y_i are the current co-ordinates of image points;

a_i - the values of current image point brightness.

The audio and visual system processing give the reason to conclude, that the cross correlation is used directly in audio system (equations 4 and 5) and indirectly in visual system, when tracking and predict co-ordinates x_{sp}, y_{sp} of the speaker position. Therefore, in this paper, the use a cascade-correlation neural network for evaluating this cross correlation in audio visual robot system is proposed.

3 Cascade-Correlation Neural Network in Audio Visual Robot System

The structure of the cascade-correlation neural network proposed to implement in audio visual robot system as shown in Fig. 2. The inputs of this neural network receive the information in form of speech signals from

each microphone in microphone array. It also calculates the sequence of co-ordinates x_{sp}, y_{sp}, after processing, made in visual robot system.

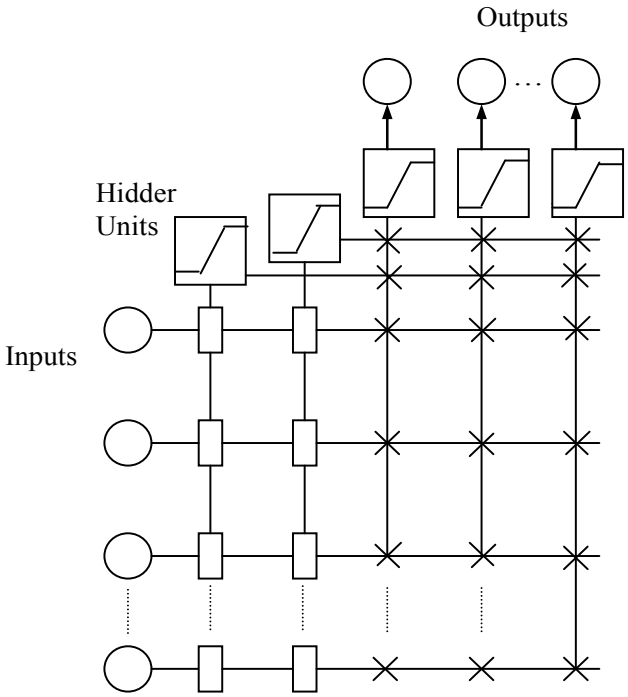

Fig.2.

The results of cascade-correlation neural network operation are presented in the outputs of the network. The structure of the cascade-correlation neural network is presented in the hidden units. They represent a sequence, which is combines the values of angle θ of sound direction of arrival (DOA) and also the predicted co-ordinates of the centre of gravity x_{sp}, y_{sp} of speaker face or body.

The training process of the cascade-correlation neural network can be explained with the following equation:

$$C = \sum_{o} \left| \sum_{p} (y_p - \bar{y})(e_{op} - \bar{e}_o) \right|, \qquad (7)$$

where

C is the value of correlation between the output of the candidate node and network output error, which can be maximized in order to find the candidate node weights update;

o – the symbol of output;

p – pattern of training input vector;

$\bar{y}$ and $\bar{e}_o$ - the mean values of the outputs y and output errors e over the all patterns of the training sample.

4 Testing Cascade-Correlation Neural Network in Audio Visual Robot System

The proposed structure of a moving robot audio system from Fig.1 is simulated and tested, using the cascade-correlation neural network structure from Fig.2. For the simulations, speech source signals as words and sentences pronounced from the speaker man or woman are used. In Fig.3 an example of speech source signal of the pronounced word “Five“is presented.

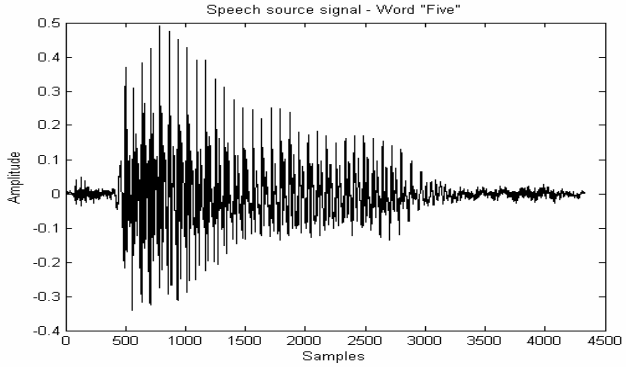

Fig.3.

In the first step of the simulation the space relations between the positions of speaker S_p, the center of robot “0” are defined and also the number and places of microphone array placed on the moving robot are chosen.

Each simulation is then executed, after the definitions of all of these necessary conditions. Some results are shown here and they represent the steps of testing the chosen operations for speaker direction finding, using the proposed moving robot audio visual system with implementation of a cascade-correlation neural network. On the Fig.4 are presented the time relations between speech source signal and the simulated signals S_1, S_2 and S_3 from the microphones M_1, M_2 and M_3. Comparing the first frames of each of four signals, it can be seen, that in the simulation a suitable time delay between speech source signal and the simulated received microphone signal are obtained.

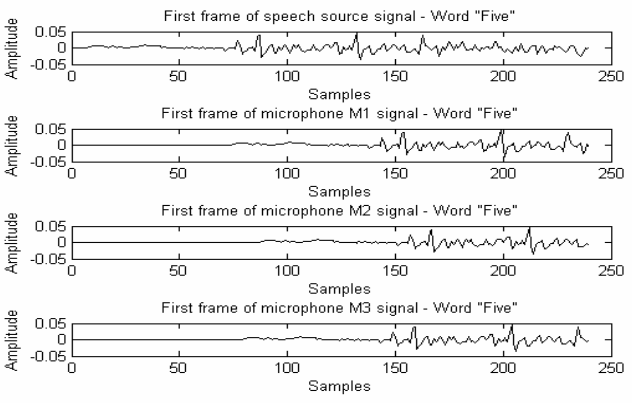

Fig.4.

The signal from microphone M_1 have some time delay, measured toward source speech signal, the signal from microphone M_3 some additional time delay, toward signal from microphone M_1 and the signal from the microphone M_2 have the time delay, toward the signal from microphone M_3. This means, that the speaker place in this simulation is chosen in such a way, that the distance of microphone M_3 to speaker place is smaller, than the distance of microphone M_2. In the next step of simulation the cascade-correlation neural network, two relation coefficients $R_{1,2}$ and $R_{1,3}$ between speech signal from microphones M_1, M_2 and between speech signal from microphones M_1, M_3, respectively are calculated, using the equations (5).

These relation coefficients are presented on Fig.5.

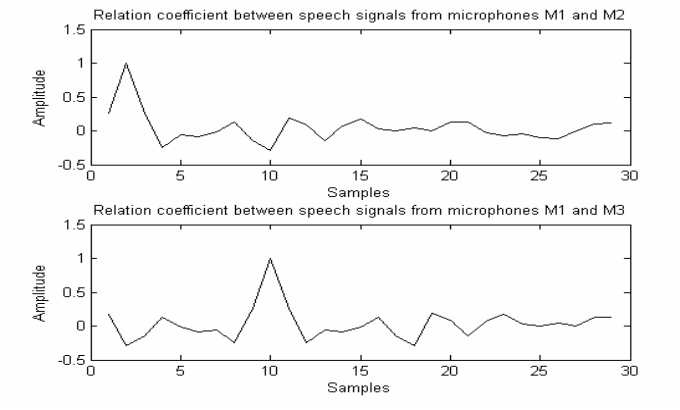

Fig.5.

On the Fig. 5 are seen the maximum values in each of the relation coefficients. These maximums confirm the existing of peaks in the correlation coefficients, which depend from the relative time delay between signal from microphones M_2 and M_3, respectively. Thus, we calculate the matrix of cross correlation $R_{i,j}$ using also cascade-correlation neural network. The results from these operations are shown in Fig.5.

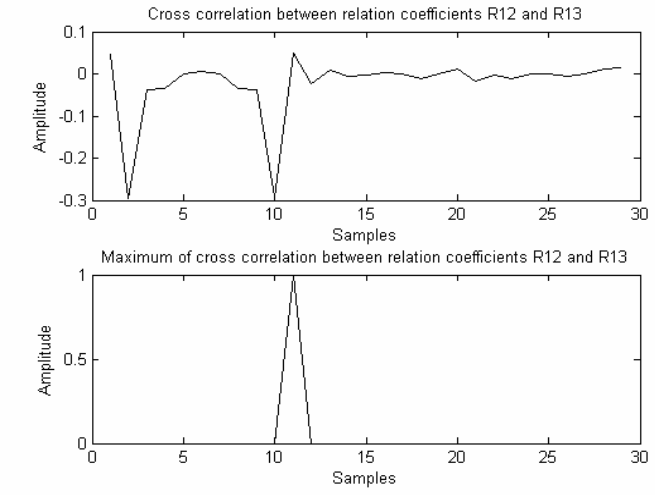

Fig.5.

The maximum value, which can be determined from the Fig.5, is equal to 11 and it is expressed as number of samples. It is possible to transform this value from the number of samples to value in degrees, i.e. to determine the value of angle θ, which gives the direction of sound arrival. Using value 11, calculated as number of samples, after transformation, an angle of direction of arrival $\theta = 23^o$ is determined.

4 Conclusion

The brief description, presented here of a case of the simulation of cascade-correlation neural network in audio visual robot system, demonstrate the correct work of the operations and calculations of each step of the simulation to determine the angle θ_d. The similar results can be reported for the speaker co-ordinates x_{sp}, y_{sp} prediction with cascade-correlation neural network. The performances of cascade-correlation neural network in the audio visual robot system are compared with the same test examples and robot movement situation, but without cascade-correlation neural network. It is possible to declare, that the advantages of using cascade-correlation neural network are connected with the possibility to adapt the cascade-correlation neural network structure to the current more simple or more complex case of moving robot situation in learning process. This is important mainly for the hardware or software realization of the algorithms of moving robot, tracking the direction to the speaker position and second for the speed of calculation important for real time robot moving, when processing audio and video sensor information in conjunction with cascade-correlation neural network.

Acknowledgements

This work was supported by National Ministry of Science and Education of Bulgaria under Contract BY-I-302/2007: "Audio-video information and communication system for active surveillance cooperating with a Mobile Security Robot".

References:

[1] Huang J., Supaongprapa T., Terakura I., Ohinishi N. and Sugie N. Mobile Robot and Sound Localization. Proc. IEEE Int. Conference on Intelligent Robots and Systems (IROS'97), September, 1997.

[2] Xiong Y. and Matthies L., Vision-Guided Autonomous Stair Clumbing. IEEE International Conference on Robotics and Automation. San Francisco, California, April, 2000.

[3] Zerbe V., M. Milushev Modular Control for a Joint Based on Antagonistic Pneumatic Motioning 53rd Internationales Wissenschaftliches Kolloquium -Technische Universität Ilmenau (53. IWK), 08 – 12 September, 2008, pp. 79-81.