SIMULATION OF AUDIO VISUAL ROBOT PERCEPTION OF SPEECH SIGNALS AND VISUAL INFORMATION

Alexander Bekiarski, Snejana Pleshkova-Bekiarska

Technical University – Sofia, Bulgaria
Kliment Ohridski, 8
Tel.: +359 965 3300; E-mail: snegpl@tu-sofia.bg

Abstract

Audio visual robots are intelligent robots that require many researches to gain the desired performance. An obligatory and necessary stage of these researches is simulation. In many articles for the moving robots there are simulations, but most of them investigate a small part of entire robot action or robot system. The goal in this article is to propose tests and simulation of the audio perception robot system with conjunction to the received visual information from the moving robot visual sensors. The achieved results from these simulations can be used in the future works to combine and improve the overall moving robot performances.

1. INTRODUCTION

Intelligent robots must involve audio and video sensors for perceiving the sounds in the area of observation [1] and also visual information [2]. The sound sources can be speech signals from the speakers. The robot perceives the speaker sounds, process them and determine the speaker localization or only determination of the direction of arrival (DOA) of sound [3]. Another more complicated task is the speech recognizing or speaker identification, performed from the robot to identified or recognized the talker and then to choose the moving direction [4] and [5].

The moving robot system must be designed correctly after multiple experiments and tests using real robot systems as designed tool, but in the beginning of tests it is suitable to use simulations of the proposed ideas, concepts and algorithms. In this article is proposed a method of simulation of moving robot audio perception system in conjunction with received video information. The steps of the proposed simulation are:

- speech source signal simulation as isolated words;
- speech sound waves propagation in area of robot observation taken as a room space;
- 2D room area of space of robot observation simulation;
- simulation of the types and the characteristics of robot audio and video sensors (linear, 2D microphone arrays, mono or stereo video cameras);
- simulation and test of speaker localization;
- calculation of speaker co-ordinates from video information;
- combining two information from speech signal localization and speaker co-ordinates etc.

2. SPEECH SOURCE SIGNAL SIMULATION

For the simulations of the received from the robot audio system speech signal are used speech source signals as words and sentences pronounced from the speakers man or woman. One of these chosen source signals is presented in Fig. 1 as an example of speech source signal of the pronounced word “Five”, part of the speech signals collected as a specific data base used these tests presented in this article for determine the best characteristics for audio system of a real moving robot with audio and video sensors.

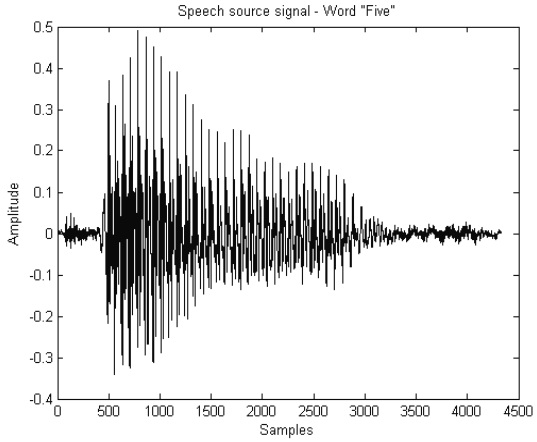

Fig. 1. Speech source signal – Word "Five"

For the simulation of speech signals, received from a set of microphones, which serves as robot audio sensors, it is necessary to define the space relations between speaker position S_p, the center "0" of the robot co-ordinate system and also the number and places of microphones situate on the moving robot.

3. SIMULATION OF 2D ROOM AREA OF SPACE OF ROBOT OBSERVATION

In the presented simulation it is chosen to use a circular 2D arrangement of microphone array. The number of the microphones is "n", chosen here as n=3, the simplest case (microphones M_1, M_2 and M_3), but can be more than n=3. Places of the microphones M_1, M_2 and M_3 are chosen with the appropriate 2D co-ordinates x_{M1}, y_{M1}; x_{M2}, y_{M2}; x_{M3}, y_{M3}. The co-ordinates and the positions of the microphones must be defined in a way that they are placed in the appropriate corners of an equilateral triangle, inside of a circle, and the place of the microphone M_1 is chosen to be in front of microphone array, i.e. in the direction of speaker source. The distance between speaker place S_p and the center of robot "0", defined as l_0, is placed in the center of gravity of the equilateral triangle. This distance is variable and depend from the current place and movements of the robot, but here it is and is kept constant, for example l_0=1 m. Only the angle between speaker position S_p and center of robot "0" is changed in the time of simulations.

4. SIMULATION AND CALCULATION OF SPEAKER LOCALIZATION

The angle θ_d between axis "y" and l_0 define the direction of speech sound source and can be calculated as:

$$\theta_d = \frac{\sum_{\theta=1^\circ}^{360^\circ} R_{i,j}[k(\theta)] \times \theta}{\sum_{\theta=1^\circ}^{360^\circ} R_{i,j}[k(\theta)]}, \quad (1)$$

where

$R_{i,j}[k(\theta)]$ - is the relation coefficient;

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

θ – the current value of the angle.

The relation coefficient $R_{i,j}[k(\theta)]$ can be calculated as correlation between appropriate pairs of speech signals S_i and S_j (i and j = 1,2,3), received from micro-phones M_1, M_2 and M_3 in the 2D circular microphone array placed in the moving robot:

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

where

N is the number of speech signal samples in a frame, defined usually as $N = 240$;

m – the current number of each speech signal sample.

5. SIMULATION OF THE RECEIVED MICROPHONE SIGNALS

Simulation of the speech signals S_i and S_j (i and j = 1,2,3) is made as a modification of the source speech signal from the talker. This modification means, that the speaker source speech signal is delayed with the different values for producing each microphone speech signal.

The values of the delays and displacements for each received from microphone speech signals are different and are calculated in discrete form or number of samples as:

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

$$n_{1,2}=\frac{l_{1,2}}{c}.f_s;\ \ n_{2,3}=\frac{l_{2,3}}{c}.f_s;...;n_{n-1,n}=\frac{l_{n-1,n}}{c}.f_s;\ \ n_{n,1}=\frac{l_{n,1}}{c}.f_s;\quad (4)$$
$$n_{d_m}=\frac{d_m}{c}.f_s,$$

where

$n_1, n_2, ... n_n, n_0$ are the delays in discrete form for each microphone;

$l_1 = l_2 = ... = l_n = l_0$ - the corresponding distances of microphones from speaker;

$l_{1,2} = l_{2,3} = ... = l_{n-1,n} = l_{n,1} = d_m$ are the distances between each pair of microphones;

c - velocity of sound in the air;

f_S - speech signal sampling frequency.

Equations (3) and (4) are presented in general form, i.e. multiple microphones, but in this simulation it is considered the case for n=3 as the number of microphones. Also, without breaking the defined space conditions, in the beginning of each simulation, it is possible to vary be distances between three microphones $l_{1,2}$, $l_{2,3}$ and $l_{3,1}$, defined in equation (4), but keep them equal, i.e. $l_{1,2}=l_{2,3}=l_{3,1}=d_m$. Here in Fig. 2. are shown the simulations for the received microphone signals with the distance d_m =0,2 m between three microphones, but it can be changed, for example between 0,1 and 0,6 m, or more.

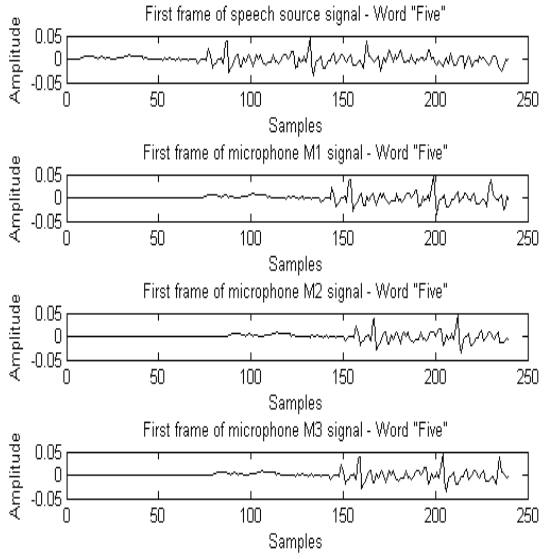

Fig. 2. First frames of speech signals of the pronounced word "Five" in appropriate places of the space of robot area of observation

From equation (4) it is clear, that increasing distance d_m leads to large value on $n_{1,2}=n_{2,3}=n_{3,1}=d_m$, representing the distance between the microphones as number of samples. Finally this gives more precision in simulation, when calculating the angle θ_d, which represent the sound direction of arrival.

6. RESULTS OF ROBOT AUDIO SYSTEM SIMULATIONS

The described parts of an audio system are first simulated separately for verify their correct works.

Then all of these parts are connected together in an simulation algorithm to test and simulate the whole work of the audio robot system. In this simulation are used the equations presented here. The main goal is to prove with the appropriate measurements and calculation in the time of simulation the proposed method for speech source localization.

Here are presented some conditions for these simulation.

As an area of robot observation is used a 2D model of a room with following dimensions: length of the room larea = 10 m and width of the room varea = 5 m. The speaker position S_p is defined and varied with choosing the coordinates x_{sp} and y_{sp}, according to the origin of the coordinate system placed at the center of robot "0". There are simulated many concrete situations of 2D area of robot observation space or room. In this space it is chosen an 2D (x,y) coordinate system with origin point "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.

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. In a chosen situation of simulation it is necessary to choose the range of variation of k (equation 1), for calculated the relation coefficients $R_{1,i}$ (k). Here it is chosen the changes of angle θ between axis y and the direction l_0 of sound source or speaker place from 0^0 to 360^0 and to keep the distance between robot center "0" and speaker place constant (l_0=const). In this

case the range of variation of actual delay k as number of samples depend on actual differences between distance l_1 and each distance l_2, l_3, …, l_n of the microphones M_1, M_2, M_3, …, M_n from speaker place S_p. For changes of angle $\theta=0^0\div360^0$ these distance range from:

$$l_{1d}^{\min} = \min\left(-l_{1,i}\right) \text{ to } l_{1d}^{\max} = \max\left(l_{1,i}\right), \quad (5)$$

for i=1,2,…, n, which correspond to earlier/later sounds enters for microphones M_2, M_3, …, M_n toward microphone M_1. Actually for this simulation n=3.

More precisely, from equations (3) and (4), it is possibly to represent the range of variations of actual delay k as number of samples:

$$k_{\min} = \frac{l_{1d}^{\min}}{c}.f_s = \frac{\min\left(-l_{1,i}\right)}{c}.f_s \quad (6)$$

$$k_{\max} = \frac{l_{1d}^{\max}}{c}.f_s = \frac{\max\left(l_{1,i}\right)}{c}.f_s \quad \text{for } i=1,2,...,n. \quad (7)$$

Under these circumstances, the range of variations of k from k_{min} to k_{max} can be described as a function of angle θ k(θ), for $\theta=0^0\div360^0$.

The last representation of k gives the reason to describe the relation coefficients $R_{1,i}(k)$ from equation (5) also as a function of angle θ in the range of $0^0\div360^0$, that means $R_{1,i}$ [k(θ)]. Therefore, the sound or speaker direction can be estimated from the matrix of cross correlation of $R_{1,i}$ [k(θ)] and $R_{1,j}$ [k(θ)]:

$$R_{i,j}\left[k(\theta)\right] = R_{1,i}\left[k(\theta)\right].R_{1,j}\left[k(\theta)\right] \quad (8)$$

$$\text{for } i=1,2,...,n;\ i\neq j \text{ and } \theta=1^\circ,2^\circ,...,360^\circ.$$

From the equation (8) it is possible to find the maximum of cross correlation:

$$R_{\max} = \max\left\{R_{i,j}\left[k(\theta)\right]\right\}. \quad (9)$$

Also it is proposed to define a threshold value R_t from the maximum of cross correlation:

$$R_t = 0{,}99.\max\left\{R_{i,j}\left[k(\theta)\right]\right\}. \quad (10)$$

The threshold R_t from equation (10) is used to perform normalization to the cross correlation $R_{i,j}$ [k(θ)]:

$$R_{i,j}\left[k(\theta)\right] = 0, \text{ if } R_{i,j}\left[k(\theta)\right] < R_t$$

$$R_{i,j}\left[k(\theta)\right] = \frac{R_{i,j}\left[k(\theta)\right] - R_t}{R_{\max} - R_t},$$

$$\text{If } R_{i,j}\left[k(\theta)\right] \geq R_t \quad (11)$$

for θ = 1°, 2°,..., 360°

These conditions and calculations are used in simulation and allow to perform a weighted average to the $R_{i,j}$ [k(θ)] from equation (11), which allow to find from equation (1) the correct value of angle θ_d as a sound or speaker direction. for the cases, where there are not reverberations, noise signals or words with consonants, that have weakly periodic signals. But these cases are not a real situation for the moving robot. Therefore, in order to find accurate determination of sounds direction of speech signals, it is possible to determine and calculate sound direction only at the speech signal frame that has the maximum energy within a period of speech signal.

The lots of simulations with the proposed robot audio system shown, that on the range of all angles ($1^0\div360^0$), the difference between magnitudes of the cross correlation is very informative to assist in finding reliable detection of the sounds direction.

After these definitions and presentation of some equations, from which it is possible to find, in the time of simulation, the direction of arrival of the sounds from the speaker it is possible to use them to make the tests and examinations to confirm the correctness of the proposed moving robot audio system simulation. If this correctness existed, there are the reasons to apply these methods and operations in a real moving robot system.

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 system. On the Fig.2 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 sig-

nals, it can be seen, that in the simulation it is realized a suitable time delay between speech source signal and the simulated received microphone signal.

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 are calculated, using the equations (2), 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.

These relation coefficients are presented on Fig. 3. On the Fig. 3 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.

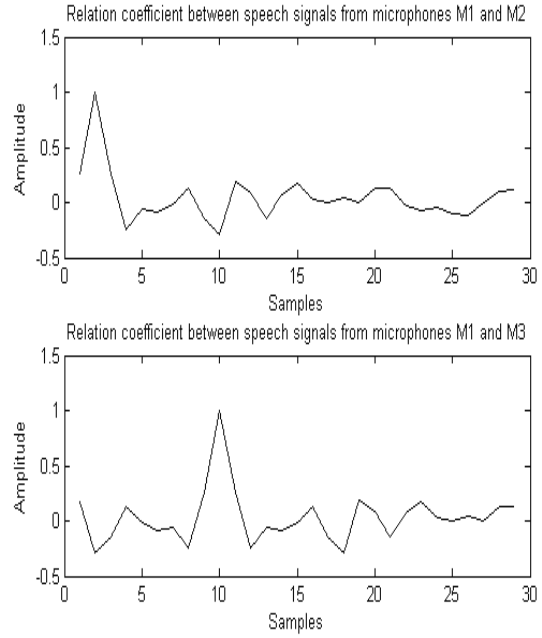

Fig. 3. Relation coefficients between speech signals from microphones

This gives the reason to calculate the matrix of cross correlation $R_{i,j}$ using equation (8) and then to make the normalization of the cross correlation using equations (9), (10) and (11). The results from these operations are shown in Fig. 4.

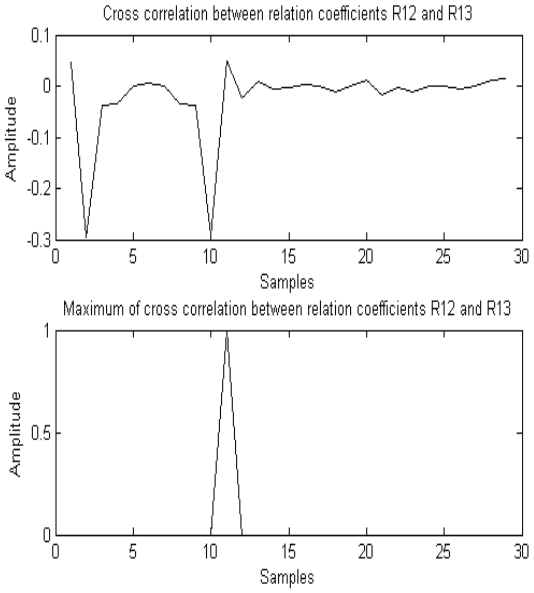

Fig. 4. Cross correlation between relation coefficients

The maximum value, which can be determined from the Fig.5, is equal to 11 and it is expressed as number of samples. Using equation (12) in the simulation 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_d, which gives the direction of sound arrival. Using value 11, calculated as number of samples, gives after transformation with equation (12) an angle $\theta_d = 23^0$.

7. CONCLUSION

The proposed simulation and the results, presented as the steps of simulation, using some proposed equations demonstrate the correct work of the operations and calculations of each step of the simulation to determine the angle θ_d, describing the direction of arrival of the sounds from the speaker to the robot. All of these steps are simulated and tested separately and then as an entire algorithm to test the relation between each step.

In the future works these results will be combined with the results from the simulation of the video sensor robot system, which give as the results the co-ordinates of speaker or talker

calculated after their visual identification from the visual robot perception system.

ACKNOWLEDGMENT

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] W.K. Ma, B. N. Vo, S. Singh, "Tracking in Onknown Time-Varying Number of Speakers using TDOA Measurements: A Random Finite Set Approach", IEEE Trans. on Signal Proc., vol.54, No9, 1993, pp. 3291–3296, Sept. 2006.

[2] J. Fritisch, M. Kleinehagenbrock, S. Lang "Audiovisual person tracking with a mobile robot", in Proc. Int. Conf. on Intelligent Autonomous System, pp. 898-906, IOS Press, 2004.

[3] H. Savada, R. Mucai, S. Araki,"Direction of arrival estimation for multiple source signals using independent component analysis", Proc. ISSPA, Paris, France, 2003.

[4] J. M. Valin, F. Michaud, B. Hadjou, "Localization of simultaneous moving sound sources for mobile robot using a frequency-domain steered beam former approach", IEEE Proc. Robotics and Automation, pp.236-242, 2002.

[5] A. Master, "Speech spectrum modeling from multiple sources", Master Thesis, Cambridge University, Engineering Department Cambridge, England, 2005.