Abstract: - A moving robot audio visual system consist from simple microphone or more precise microphone array as an audio sensor and also from a mono or stereo visual robot sensor or video cameras. One of the main functions of audio system in the mobile robots is to help them in speaker localization, in cases of robot interaction with a speaking person via human robot interface. The methods for processing of audio information received with the corresponding audio robot sensors or microphone arrays, as a part of the human robot interface are different, but here in this article is analyzed and tested the ability to integrate the audio information processing and result of sound localization with the results of video information processing in the video part of the human robot interface. It is expected, that this integration can improve both audio sound localization precision and visual finding and tracking accuracy, when the mobile robot use these important information together in his movement to the speaking person.

Key-Words: - Human Robot Interface, Microphone Array Beamforming , Speaker localization, Audio Visual Robot Systems

1 Introduction

The purpose of a human robot interface is to accomplish the audio and visual communications between the persons and robot [1]. Usually in these situations the robot is realized as a mobile robot, which must tracks and localizes the person position, using the existing for the intelligent mobile robots video and audio sensors like video cameras and microphone array mounted on the mobile robot platform, respectively. The audio and video information can be processed autonomous with existing on the robot resources or send via Web Services [2] to a stationary host computer with more resources for intensive processing. Each of these two audio and video systems can be treated as separated and independent and they use the specific methods for processing of incoming to the robot audio and video information. The results of these two different methods of processing are addressed to one goal, to help the robot in the interaction with speaking persons, using a human robot interface. Therefore, the integration of visual and audio robot systems in calculating and using the results from processing of visual and audio information can probably improve the possibilities of robot interaction with a speaking person and can approach them more closely to the human ability in voice communications between each other. To realize this proposed idea first it is necessary to analyze the performance of the methods used separately in each of these two robot audio and video subsystems and then to decide and propose some new and advanced methods for integration of audio and visual information processing as a goal of a more suitable human robot interface with an improved intelligence and precision in audio visual robot movement control.

The focus in this article is concentrated mainly to the processing of audio information, collected with audio sensors in the audio system of the mobile robot. In the same time the results from the processing of the visual information are taken in account in each current step of processing of audio information, to realize the mentioned joined processing, resulting to precision improvement of the human robot interface when the mobile robot interacting and moving to a speaking person.

2 The methods and strategies in sound localization for human robot interfaces

There are a lot of methods and strategies in the field of sound localization:
- steered beamformer based locators;
- time difference of arrival (TDOA) based locators;
- robust sound source localization algorithm using microphone arrays.

The purpose of each of them is to perceive and to process the incoming, from the speakers or other sound source, audio information and to calculate the space position or direction of arrival of sounds from these...
The beamforming basic data model is presented in Fig. 1.

It consist from the M number of microphone signals, which are the delayed of time delay $\tau_m(\theta)$ versions $y_m[k]$ or $Y_m(\omega, \theta)$, for $m=1,2,…,M$, of the speaker or other sound source signal, represented in spectral $s[k]$ or $S(\omega)$ in time and frequency space, respectively:

$$y_m[k] = s[k - \tau_m(\theta)] \quad Y_m(\omega, \theta) = e^{-j\omega \tau_m(\theta)}, \quad (1)$$

where $\tau_m(\theta)$ is time delay of each of $m = 1,2,…,M$ microphones in the microphone array:

$$\tau_m(\theta) = \frac{d_m \cos \theta}{c} f_s; \quad (2)$$

$d_m$ - the fixed distance between the first $M_1$ and each microphone $M_m$ for $m = 2,3,…,M$ in microphone array;

$f_s$ - sample frequency;

c - speed of the sound waves.

All microphone signals can be stack in a vector:

$$\mathbf{Y}(\omega, \theta) = \mathbf{d}(\omega, \theta) \cdot S(\omega), \quad (3)$$

where $\mathbf{d}(\omega, \theta)$ is “steering vector” or vector of the values of time delay between the received sound signals from each of two adjacent microphones in the microphone array:

$$\mathbf{d}(\omega, \theta) = [1 \quad e^{-j\omega \tau_2(\theta)} \quad … \quad e^{-j\omega \tau_M(\theta)}]. \quad (4)$$

The output signal vector, after beamforming as is shown in Fig. 1. $\mathbf{Z}(\omega, \theta)$ is:

$$\mathbf{Z}(\omega, \theta) = \sum_{m=1}^{M} \mathbf{F}_m^*(\omega)Y_m(\omega, \theta) = \mathbf{F}^H(\omega)\mathbf{Y}(\omega, \theta). \quad (5)$$

From the equation (5) it is possible to define the basics of beamforming as spatial directivity pattern with
The corresponding ‘transfer function’ for sound source signal at angle $\theta$ and frequency $\omega$:

$$H(\omega, \theta) = \frac{Z(\omega, \theta)}{S(\omega)} = F^H(\omega) e^{-j\omega \tau} = F^H(\omega) d(\omega, \theta)$$

(6)

The methods of fixed beamforming differ from the following operations of processing the incoming from microphones sound signals:
- delay-and-sum beamforming;
- weighted-sum beamforming;
- near-field beamforming.

The first type of beamforming utilizes the method of delay-and-sum processing of the sound signals from the microphones in the microphone array. The signals from the microphones are delayed and summed together and the array can be virtually steered to angle $\psi$ as it is shown in Fig.2.

![Figure 2.](image)

The equations describing the method of delay-and-sum processing, shown in Figure 2 are:

$$z[k] = \frac{1}{M} \sum_{m=1}^{M} y_m[k + \Delta_m]$$

(7)

$$F_m(\omega) = \frac{e^{-j\omega \tau}}{M}$$

(8)

$$\Delta_m = \frac{d_m \cos \psi}{f_s}$$

(9)

Angular selectivity is obtained, based on constructive (for $\theta = \psi$) and destructive (for $\theta \neq \psi$) interference:
- for $\theta = \psi$, this is referred to as a “matched filter”:
$$F(\omega) = \frac{d(\omega, \psi)}{M}$$

(10)
- for uniform linear array :
$$d_m = (m - 1)d$$

(11)

$$\Delta_m = (m - 1)\Delta$$

(12)

The simulations and testing of the described delay-and-sum beamforming method are carried out for many different situations in the concrete interaction between mobile robot and a speaking person. An example of the determined and used in these experiments microphone array beamforming is represented in the Figure 3.

![Figure 3.](image)

The parameters, which are used in this example, are following:
- $M=5$ microphones;
- $d=3$ cm inter-microphone distance;
- $\psi=60^\circ$ steering angle;
- $f_s=5$ kHz sampling frequency.

It is possible to use and to analyze another existing beam forming method like weighted-Sum beam forming, which is described with the following equations:

$$z[k] = \sum_{m=1}^{M} \omega_m y_m[k + \Delta_m]$$

(14)

In the field of near beamforming the characteristics of this method are:
- far-field assumptions not valid for sources close to microphone array;
- spherical wavefronts instead of planar waveforms;
- include attenuation of signals;
- 3 spherical coordinates $\theta, \phi, r$ (= position q) instead of 1 coordinate $\theta$;
- different steering vector.

The advantages and disadvantages of weighted-sum beamforming method are:
- can find the sound source location to very accurate positions;
- highly sensitive to initial position due to local maxima;
- high computation requirements and is unsuitable for real time applications;
- in presence of reverberant environments highly co-related signals therefore making estimation of noise infeasible.

The simulations and testing of the described weighted-sum beamforming method also are carried out for many different situations in the concrete interaction...
between mobile robot and a speaking person. An example of the determined and used in these experiments microphone array beamforming is represented in the Figure 4.

![Figure 4](image)

**3 Results and Conclusion**

The analysis of the beamforming, represented in the Figure 3 and Figure 4 follow to the decision, that there are two main directions in circle diagram. It is shown and can be made a comparison between the results of sound localization and mobile robot movement to the speaking person in the Figure 5 and Figure 6, to conclude, that the method of weighted-sum beamforming (Figure 6) gives a more precise robot movement control in comparison with the delay-and-sum beamforming method (Figure 5). These results from the comparison and the decision of existing of two main directions in the circle diagrams for the two beam forming methods are proposed to use together with the similar results, derived from the visual mobile robot system for calculation of the co-ordinates or centre of gravity of the same speaking person. In this way it is possible to integrate both results for sound localization, using beam forming, and the results from visual processing as co-ordinates of the speaking to the robot person. This means that the human robot interface work as an integrated audio visual system and can expected an significant improvement of the precision of both sound localization and speaking person co-ordinates, derived from audio and visual robot system.

**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**

