SPEAKER TRACKING MOBILE ROBOT AUDIO VISUAL SYSTEM FOR SURVEILLANCE APPLICATIONS

Alexander Bekiarski, Snejana Pleshkova-Bekiarska

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

Abstract

In the mobile robots for security and video surveillance applications exists an audio and visual tracker system. Combining visual and audio data from video cameras and microphones is a more realistic model similar to the human system of seeing and hearing. The proposed method for audio and visual speaker tracking used the results of human body shape determination and audio information from sound localization as direction of sound arrival. An appropriate filter is proposed to combining and integration of audio and visual information to efficiently apply the specifics of audio and visual features.

1. INTRODUCTION

The mobile robot perceives the speaker sounds, process them and determine the speaker localization or the direction of arrival (DOA) of sound [1]. Another 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 [2], [3]. To realize this task it is possible to merge the available audio and video information. For this purpose both audio and visual information must be available to avoid the limitations in terms of tracking performance. The sensors providing input data are a video camera for speaker position finding and separation and two microphones for sound from speaker localization. This combination gives as the result better localization and tracking of speaking persons or speakers if it is performed by a decision filter.

2. AUDIO VISUAL MOBILE ROBOT SYSTEM

2.1. System architecture

In the Fig.1 is presented general view of the proposed and tested audio visual robot system for speaker tracking purposes in a surveillance application.

Audio system consists of a microphone array with two microphones M1, M2 and Audio Processing block. Audio Processing block using a Direction of Arrival (DOA) estimation algorithm to determine the speaker sound wave direction in relation of robot position.

The video robot system consists of a single TV camera seeing the speaking person in area of robot observation and a Video processing block. The Video Processing block perform the decision of the speaker co-ordinates (x_v,y_v).

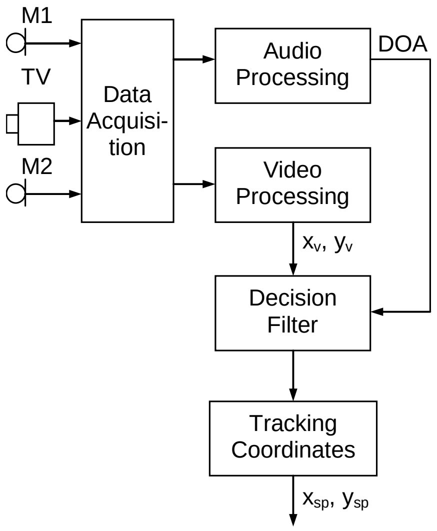


Fig. 1

The information for the speaker from Audio and Video Processing blocks is sent to a Decision Filter, in which it is decided whether or not the values of the pair DOA and (x_v,y_v) are right or correct speaker position.

2.2. Audio Processing

The microphone array gives an estimation of the audio source or speaker location in term of angle θ, which is explained in Fig. 2.

The speech signals $s_1(t)$ and $s_2(t)$ from two spatially separated microphones M1 and M2 are picked up from source signal $s(t)$ corresponding to the speech wave from the speaker in the presence of noise and can be mathematically written as:

$$s_1(t)=s(t)+n_1(t) \qquad (1)$$

$$s_2(t)=\alpha s(t+D)+n_2(t)\,, \qquad (2)$$

where:

$n_1(t)$ and $n_2(t)$ are added in the place of microphones M1 and M2, respectively;

α – amplitude coefficient of attenuation;

D – time delay between two microphone signals $s_1(t)$ and $s_2(t)$.

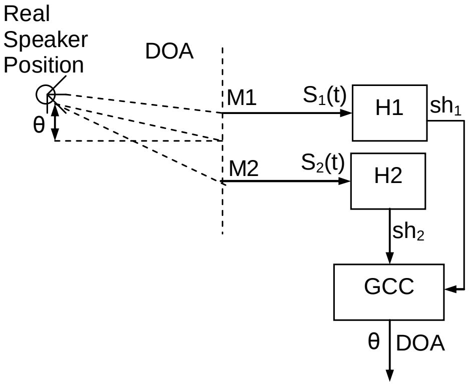


Fig. 2

Let suppose that all signals belong to a set R, the noise signals are random processes and source speech signal $s(t)$ is uncorrelated with noise $n_1(t)$ and $n_2(t)$:

$$s(t), n_1(t) \text{ and } n_2(t)\in R \qquad (3)$$

In most of the existing methods determination of the delay D or time offset τ is computed as the cross-correlation or expectation E between two microphone signals $s_1(t)$ and $s_2(t)$:

$$R_{s_1s_2}(\tau)=E[s_1(t)s_2(t-\tau)] \qquad (4)$$

For example, if the speech signals can be treated as ergodic processes, the equation (4) can be written as:

$$R_{s1s2}(\tau)=\frac{1}{T_2-T_1}\int_{T_1}^{T_2}s_1(t)s_2(t-\tau)dt\,, \qquad (5)$$

where:

T_2-T_1 is the observation time interval, which for joint audio and video speaker localization must be equal to the video frame period T_{Fr}, related to the frame frequency f_{Fr}:

$$T_2-T_1=T_{Fr}=\frac{1}{f_{Fr}}\,, \qquad (6)$$

With the pre-filters, H1 and H2, shown in Fig. 2 the equations for the cross-correlation can be modified as Generalization of Cross-Correlation (GCC):

$$R_{sh_1sh_2}(\tau)=\int_{-\infty}^{+\infty}\psi_g(f)G_{s_1s_2}(f)e^{j2\pi\tau}df\,, \qquad (7)$$

where:

sh_1 and sh_2 are the microphone signals after pre-filters H1 and H2;

$G_{s_1s_2}(f)$ - cross power spectral function between $s_1(t)$ and $s_2(t)$;

$\psi_g(f)$ - general frequency weighing:

$$\psi_g(f)=H_1(f)H_2^*(f) \qquad (8)$$

and $H_1(f), H_2(f)$ are the response of the two pre-filters in frequency area.

From the equation (8) is seen, that $\psi_g(f)$ depend on pre-filter $H_1(f), H_2(f)$ transform. Different pre-filtering transforms exists, but one of them phase transform (PHAT) is most used in the applications of DOA estimation and speaker localization. For PHAT transform $\psi_g(f)$ is described as:

$$\psi_g(f)=\frac{1}{\left|G_{s_1s_2}\right|}=\frac{1}{\left|S_1(f)S_2(f)\right|}\,. \qquad (9)$$

The described Audio Processing block show the necessary operations to achieve the satisfactory precision of the DOA estimation from the robot audio system.

2.3. Video Processing

In the similar way the speaker position (x_{sp},y_{sp}) is calculated from the TV camera information. This calculation start with the motion detection, separation of binary image mask of the speaker and ended with the pair of co-ordinates (x_v,y_v) calculation as the center of gravity of human body .All of these operations are included in the Video Processing block in the Fig.1.

2.4. Decision filter

Both results (DOA or θ) and (x_v,y_v) from Audio and Video Processing blocks are shown in Fig.1 as input information of Decision Filter, which is chosen as a particle filter. A particle filter represents the unknown probability density function by a set of m-random samples $sp_1, ..., sp_m$. In the case of a speaker tracking system in a 2D area of robot observation each particle sp_i is represented as a hypotheses of speaker location (x_i,y_i). The particles are considered as the vectors in state space associated with an individual weight W_i, and combined with the real audio and video observation DOA or θ and (x_v,y_v) in a way to decide and estimate the more probably connected observation to the real speaker co-ordinates (x_{sp},y_{sp}). The decision particle filter performs two mains steps:

- the prediction step to generate new particles from the set of particles in the previous time instance;
- the measurement step to adjust the weights W_i of new particles with respect of current observations DOA_i or θ_i and $\left(x_v^i, y_v^i\right)$.

The continuous execution of these steps is accompanied with calculations of current weights:

$$w_i = c_A.p(A_t / sp_i) + c_V.p(V_t / sp_i), \quad (10)$$

where:

c_A and c_V are dynamic mixture weights;

A_t and V_t – the current audio and video observations, i.e. DOA or θ and (xv,yv), respectively.

The dynamic mixture weights c_A and c_V can be interpreted as confidence measures for the audio and video robot system:

$$c_A = 1/\sqrt{\left(\sigma_x^A\right)^2 + \left(\sigma_y^A\right)^2} \quad (11)$$

$$c_V = 1/\sqrt{\left(\sigma_x^V\right)^2 + \left(\sigma_y^V\right)^2}, \quad (12)$$

where:

$\sigma_x^A, \sigma_x^V, \sigma_y^A$ and σ_y^V denote the standard deviation of the particle set's x- or y- components, weighted with audio or video scores, respectively.

In order to generate the final speaker tracking output (x_{sp},y_{sp}) it is used a m x m window to scan the x,y area of observation and to find the highest accumulated particle scores, which is chosen as final decision of the speaker position (x_{sp},y_{sp}).

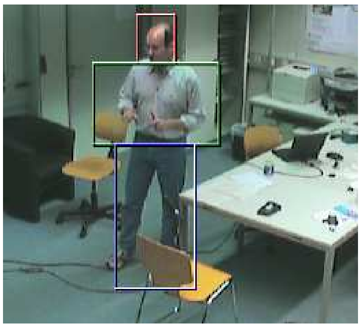
3. RESULTS OF SIMULATIONS AND CONCLUSION

The proposed algorithm is simulated and tested with a set of audio and video sequences chosen to reveal the properties, advantages and errors with using only video, only audio or combined video and audio information. The results are summarized in comparative way in Table1. and on the Fig.3 with calculated speaker co-ordinates (x_{sp},y_{sp}).

The analysis of the results gives the reason to conclude that the proposed algorithm of speaker tracking robot audio visual system work with an appropriate efficiency and gives little percentage of miss classification.

Table 1

Tracking mode	Missed speakers classification
Video only	14,3 %
Audio only	22,6%
Video + Audio	5.1%



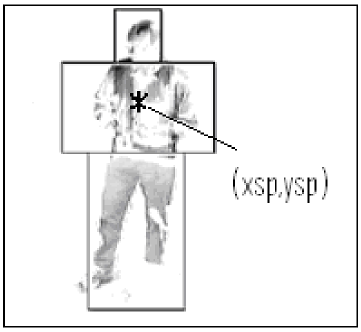


Fig. 3

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] H. Savada, R. Mucai, S. Araki, "Direction of arrival estimation for multiple source signals using independent component analysis", Proc. ISSPA, Paris, France, 2007.

[2] 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, 2006.

[3] Master, "Speech spectrum modeling from multiple sources", Master Thesis, Cambridge University, Engineering Department Cambridge, England, 2008.