Virtual System of Speaker Tracking by Camera Using an Audio-Based Source Localization

H. Sayoud, S. Ouamour and S. Khennouf

***Abstract*— In this research work we deal with the problem of automatic speaker tracking by camera. Such tracking systems do exist nowadays, but they suffer from a number of problems such as: the response time due to the system inertia, the disturbing motor noise and the mechanical oscillations of the mechatronic module. In order to overcome these problems, we thought to employ a virtual tracking system using a fixed camera that does not require any mechanical part. We have proposed and conceived a virtual tracking system which is able to ensure the required task by using only two cardioid microphones and a classic video camera. Hence, in this research work, we will present two main tasks: the first one deals with the problem of audio speaker localization with only two microphones and the second one deals with the problem of virtual camera orientation. However, in our virtual tracking system, the task of speaker tracking is ensured by the orientation of the ROI (Region Of Interest) of the camera towards the active speaker.**

Experiments of virtual speaker tracking by camera have been done off-line, in a small meeting room without echo cancelation. Results show the good performances of the proposed localization methods and a correct tracking by the new virtual technique.

***Index Terms*— Virtual speaker tracking, Automatic speaker localization, Virtual camera control, Artificial vision and perception, Signal processing, Artificial intelligence**

I. Introduction

In the context of automatic analysis of meetings, robust localization and tracking of active speakers is of fundamental importance, particularly for enhancement and recognition of speech in microphone-array based ASR (Automatic Speaker Recognition) systems.

The tracking process can be ensured by two means. The most common technique consists in moving a mobile camera towards the target [1]. Another possible technique uses fixed special cameras that possess large visibility coverage: ‘panoramic cameras’ [2], and proceeds to a selection of the wanted region of interest.

The mobile camera has several problems due to the noise produced by the driving motors, the mechanic oscillations and an important delay due to the mechanic response time of these ones.

The technique using panoramic cameras represents a good alternative to overcome these problems, but the main disadvantages of the panoramic cameras are their expensive cost and the distortion of the image, which is not a linear image.

Trying to simulate the localization faculty of the human ears with two opposite cardioids microphones and a fixed camera, and trying to avoid the mechanic problems of mobile cameras and the disadvantages of panoramic cameras, we have proposed a new tracking method which uses normal fixed camera, and where we use a virtual algorithm to automatically select the optimal region of interest.

II. Experimental Speech Database

We have built a speech database which we called DB11. The DB11 database contains several scenarios with different speakers, speaking alternatively in a natural manner and with different configurations.

We have used two cardioids microphones placed in opposition and separated by a fixed distance.

There are two general configurations: a stable configuration and a mobile configuration.

In the stable configuration, the speakers are seated at one of the three fixed positions specified between the two opposite microphones: Left position, Middle position or Right position in a same line.

In the mobile configuration, the speaker walks smoothly from one side to the other (eg. from the left microphone to the right microphone). He can utter specific phonemes or speak any continous sentence. The distance between the two microphones is 1m.

The number of scenarios is 11 and the number of speakers is 7 (4 female and 3 male speakers). The signals collected by the 2 cardioid microphones are sampled at a frequency of 44 kHz and quantified with 16 bits, by a stereophonic acquisition. The two channels of this acquisition will be used to control the orientation of a mobile camera (figure 1) toward the active speaker.

Experiments of virtual camera orientation have been done with two types of cameras: a JVC digital video camera and Kinstone webcam camera.

Manuscript received February 05, 2012; revised March 09, 2012.

H. Sayoud is a full professor at the USTHB University of Algiers (halim.sayoud@gmail.com).

S. Ouamour is an associate professor at the USTHB University of Algiers (siham.ouamour@gmail.com).

S. Khennouf is a Lecturer at the University of Msila (khenouf.salah@gmail.com).

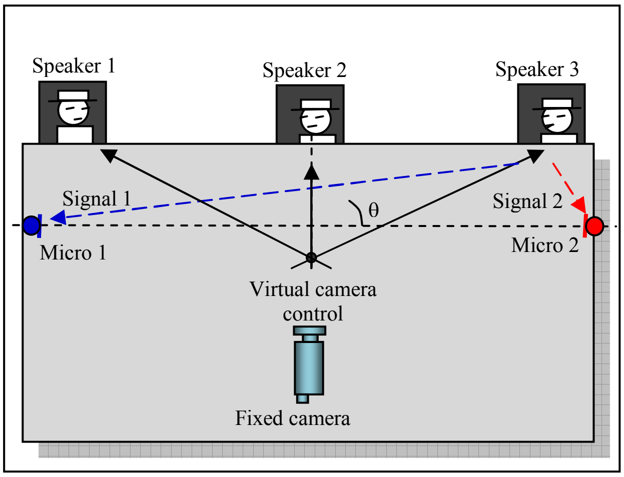


Fig. 1. Description of the meeting room configuration: the 2 signals collected from the 2 microphones are processed and used to control the virtual tracking of the active speaker.

A. Consent to Publish

By submitting your paper, authors are responsible for obtaining any security clearances and agree to follow the above copyright notice.

III. Proposed Methods for the Task of Speaker Localization

A. The Filtered Correlation based method (FCM) (new proposed method)

In our new method, the discrete correlation is computed on every pair of speech segments of 0.25s from the 2 microphones (signal x of the right microphone and signal y of the left microphone). Our filtered correlation scheme is described here below:

First, we compute the discrete cross-correlation:

$$\mathrm{Cor}_{xy}(\mathrm{m})=\sum_{i=-N}^{+N} x(i).y(i+m) \qquad (1)$$

But since the pitch range is limited for human beings [3], we have decided to apply a band-pass elliptic filter, with the following cut-off frequencies: 20Hz and 300Hz, to the correlation signal Cor_{xy}. The choice of this specific bandwidth is made in order to attenuate the noise with the harmonics and to preserve only the glottic signal, which has a limited frequency ranging (usually from 70Hz to 350Hz).

The filtered correlation obtained is denoted by: $FCor_{xy}$

Now a scanning algorithm will analyse all the filtered correlation and detect the position (m_{max} corresponding to a time T_{maxCor}) of its maximum.

B. The Energy Differential based method (EDM) (new proposed method)

The first order energy is computed on every speech segment of 0.25s for the 2 microphones (signal x of the right microphone and signal y of the left microphone), with the following manner:

$$E_x = \sum_{1=0}^{N} |x_i| \qquad (2)$$

$$E_y = \sum_{1=0}^{N} |y_i| \qquad (3)$$

Then, a Silence Activity Detection (*SAD*) is applied to distinguish the silence areas from the activity ones.

The energy differential is computed as follows:

$$DExy = \log\,(Ex/Ey) = \log(Ex) - \log(Ey) \qquad (4)$$

So it is easy, now, to estimate the relative position of the speaker with regards to the microphones positions. For instance, it is easy to deduce if the speaker is in the right side, left side or in the middle.

C. The Fusion technique

Since we have implemented two methods of speaker localization, namely the filtered correlation method and the energy differential method, and since the two detection decisions of those methods are not necessarily similar, we thought to fuse the two detection methods in order to get a more precise estimation of the speaker position.

The fusion in the broad sense can be performed at different hierarchical levels or processing stages. A very commonly encountered taxonomy of data fusion is given by the following three-stage hierarchy [5] [6]:

- Feature level where the feature sets of different modalities are combined. Fusion at this level provides the highest flexibility but classification problems may arise due to the large dimension of the combined (concatenated) feature vectors.

- Score (matching) level is the most common level where the fusion takes place. The scores of the classifiers are usually normalized and combined in a consistent manner [5].

- **Decision level** where the outputs of the classifiers establish the decision via techniques such as majority voting. Fusion at the decision level is considered to be rigid for information integration [6].

In our research work, we have used the third class of fusion, which we called FFC fusion (Fusion Favoring the Correlation method).

IV. Virtual Control of the Camera for the Task of Speaker Tracking

In this section, we will investigate the control of the fixed camera via a new technique of virtual region of interest [2].

The tracking process can be ensured by two means. The most common technique consists in moving a mobile camera towards the target. Another possible technique uses a special fixed camera, called ‘panoramic camera’, which possess large visibility coverage, and proceeds to a selection of the wanted region of interest.

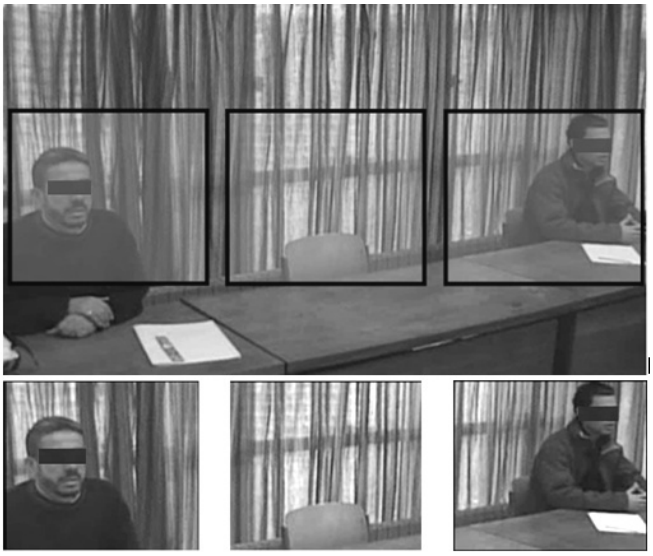
The mobile cameras have several problems due to the noise produced by the driving motors, the mechanical oscillations and an important delay due to the mechanical response time of these ones. The technique using panoramic cameras represents a good alternative to overcome these problems, but on the other hand, the panoramic camera based techniques suffer from two main disadvantages: the expensive cost of the cameras and the distortion of the captured image, which is not linear [7].

Trying to avoid the problems of mobile cameras and the disadvantages of panoramic cameras, we have proposed a new tracking method which uses normal fixed camera, and where we use a virtual algorithm to automatically select the optimal region of interest. We called it VROI technique (or Virtual Region Of Interest based technique).

The corresponding algorithm for the virtual camera control (figure 2) is given below:

```
Tuning of the scan speed: Introduction of the Virtual Pause.
At time 'j', the algorithm is defined as follows:

Moving=Estimated_Position(j)-Estimated_Position(j-1)
If Moving = 0
    Then Virtual Pause
Else if Moving > 0
    Then  For i = 1 to Moving
              Shift left : $ROI(X_0)= ROI(X_0)+1$
              Virtual Pause
          End
    End
Else if Moving < 0
    Then  For i = 1 to Moving
              Shift right : ROI(X0)= ROI(X0)-1
              Virtual Pause
          End
    End
End
```



Left virt. orientation   Medium virt. orientation   Right virt. orientation

Fig. 2. The virtual command permits the system to make a scan of the whole region (the great image above) and limit the visualization to a particular limited region (eg. the 3 small images). In this example, we can see the 3 images got with the 3 different virtual orientations: left, middle and right respectively.

V. Experiments and Results

In order to evaluate the proposed localization methods and the virtual technique of camera control, we have conducted several experiments which can be divided into 2 main series:

- Experiments of speaker localization;
- Experiments of virtual camera orientation.

A. First Experiment: Experiment of speaker localization

All these experiments are made with the two methods: FCM and EDM and also with the fusion technique FFC.

In the following figure we may refer to the FCM and EDM methods by the denotation "correlation" and "energy" respectively, for a purpose of simplification.

The comparison of the different techniques (figure 3) shows that the best detection performances are given by the fusion technique with a score of 88% for all the scenarios.

Figure 3 shows that the energy based method is a bit more precise than the correlation based method. But the fusion technique gives absolutely the best performances with an improvement of +1.51% over the EDM method.

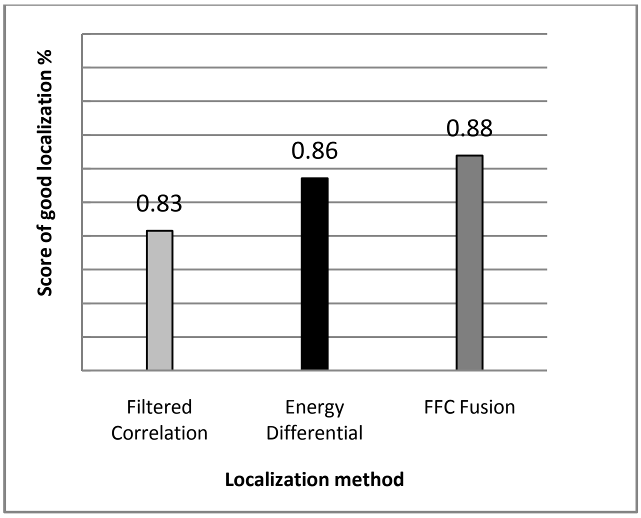


Fig. 3. Localization score in % of the 9 scenarios.

B. Second Experiment: Experiment of virtual Camera orientation

The use of a mobile camera meets several problems such as the motors noise, the mechanical oscillations and the important response time due to the system inertia. That is why we decided to undertake the experiment of virtual camera orientation.

So, in this experiment, we will describe the 2nd series of experiments which consists in ensuring the video speaker tracking by a fixed camera, and which employs the new VROI technique (see details in section 4)

During this experiment, we have noticed the following points:

- Advantages

The main advantages can be summarized by the following points:

- Excellent stability;
- No constraints at high speeds;
- Absence of oscillations even with brisk transitions;
- Response time equal to zero (real time);
- No noise;
- Working mode easily modifiable (by a simple program).

- Disadvantages

During the experiments, we have noticed some disadvantages with the camera that we can quote below:

- Image resolution reduced by 50%;
- Limited field of vision.

VI. General Discussion

This research work gave us the opportunity to investigate a multidisciplinary domain, requiring strong skills and thorough knowledge in both signal processing and artificial vision.

It is usually difficult to master such complex multidisciplinary projects, but it represents really a motivating and promising research work.

In the overall, we conceived two systems: a system for the localization of the active speaker and another system for the virtual control of the camera orientation in order to track that speaker.

By observing the whole results got with the three series of experiments, we have deduced the following conclusions:

- The methods of speaker localization show a detection precision of about 90% for the natural speech, which represents a good precision;
- By doing a comparison between the two localization techniques, we notice that the EDM method gives much better results for the fixed speakers;
- The video speaker tracking has been done correctly, in off-line mode, by the new virtual tracking technique;
- For this last task, the virtual speaker tracking with fixed camera seems very interesting for applications requiring real time processing or silent moving;
- However, the most important disadvantages of the virtual technique are the reduction of the image resolution and the limitation of the vision field.

As perspective and for improving the performances of our system, we suggest doubling the number of microphones for the vertical virtual movement, in order to ensure a 2D localization.

References

[1] Ouamour S. (2002), Système Automatique pour la Poursuite du Locuteur en Vue de la Commande d'une Caméra. Magister Thesis, USTHB University, 2002.

[2] Sun X., , Foote J., Kimber D., Manjunath B. S. (2005), Region of Interest Extraction and Virtual Camera Control Based on Panoramic Video. IEEE transactions on multimedia 2005, vol. 7, no5, pp. 981-990, 2005.

[3] Mennen I., Schaeffler F., and Docherty G. (2008), "A methodological study into the linguistic dimensions of pitch range differences between German and English," Speech Prosody Conference, Campinas, Brazil, May 6-9, 2008, pp. 527-530.

[4] Verlinde P. (1999), Contribution à la vérification multi-modale de l'identité en utilisant la fusion de décisions, Thèse de Doctorat, Ecole Nationale Supérieure des Télécommunications, MA, Bruxelles (Belgique), September 17th 1999.

[5] Jain A. K., Ross A., and Prabhakar S. (2004), An Introduction to Biometric Recognition. IEEE Transactions on Circuits and Systems for Video Technology Journal, Volume 14 (1), 4-20 January, 2004.

[6] Stylianou Y., Pantazis Y., Calderero F., Larroy P., Severin F., Schimke S., Bonal R., Matta F., and Valsamakis A. (2005). GMM-Based Multimodal Biometric Verification. Final Project Report 1, Enterface'05, Mons, July 18 - August 12, 2005.

[7] Thoby M. (2006), Fisheye lenses compared: Sigma 8mm f/4 Vs Nikkor 10,5mm f/2,8. *Updated on 17 June, 2006.* http://michel.thoby.free.fr.