A Spatial-Temporal technique of Viseme Extraction: Application in Speech Recognition ◊

Salah Werda [1]
Walid Mahdi [1]
Abdelmajid BEN Hamadou [1]

[1] *LARIM, Institut Supérieur d'Informatique et du Multimédia de Sfax, Tunisie*
salah.werda@edunet.tn; {walid.mahdi, benhamadou.abdelmajid} @isims.rnu.tn

Abstract

Speech recognition is a basic component in several research projects nowadays. However, to understand a speech, hearing is not enough, it is sometimes necessary to see it. Indeed perspective studies proved that visual information brought by the interlocutor's face in a degraded communication condition, contributed largely to the improvement of speech-intelligibility. In fact several domains are concerned with the use of visual information such as e-learning, Human-Machine interaction, etc.

This paper presents a method allowing to carry out a spatial-temporal tracking of some points of interest in the speaker's face and to indicate the different configuration of the mouth through visemes. Later on these visemes will be associated to relatively precise physical measures like the spreading of the lips and mouth height, in order to establish a correlation between the phoneme and the viseme. The results of our experiment show that we can describe the whole French phonemes by the visemes.

◊ *This work is subscribed among the CMCU Project undertaken in LIRIS Laboratory, CNRS, Ecole Centrale de Lyon, France, and in collaboration with Pr. Liming CHEN director of the laboratory and Mr. Mohsen ARDABILIAN FARD Assistant Professor in Ecole Centrale de Lyon, France .*

1. Introduction

Numerous works, on the automatic recognition of the audiovisual speech from the oldest [1] until the most recent ones [2] and [3], proved that the visual canal carries out useful information for speech recognition. Phenomena such as the perceptive illusions of McGurk [4] (speakers confronted to an auditory stimuli /ba/ and visual stimuli /ga/ perceive the stimuli /da/) or the effect "Cocktail party " (centered attention on a special speaker surrounded by multiple speakers discoursing at the same time) show the importance of visual information in the perception of the speech and that the lack of consistency between the auditory and visual sources can generate a bad interpretation of the speech.

It is around this thematic that the present work appears, and it concerns precisely the first phase of the recognition system: the extraction of precise and pertinent visual features for the categorization of the visemes (Figure1).

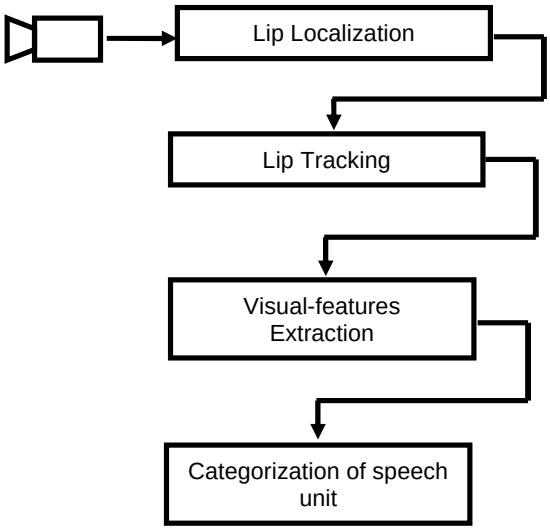


Figure 1. Lip Features Extraction

2. Methods of Labial Segmentation:

Many research works stressed their objectives on the research of automatic and semiautomatic methods for the extraction of visual indices necessary to recognize visual speech. In fact, two types of approaches have been used for the extraction of the labial information from a speaker's face:

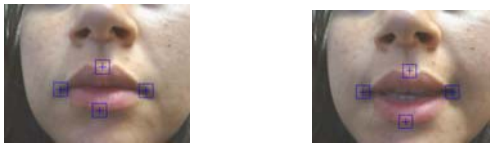
- The low-level approach (Image approaches) [5] and [6], controlled by data, which use directly the image of mouth region. This approach supposes that the lips pixels have features different from that of skin pixels. Theoretically, the segmentation can therefore be done while identifying and separating the classes lips and skin. In practice, methods of this type allow rapid locations of the interest zones and make some very simple measures of it (width and height of the lips, for example). But they do not permit to carry out a precise detection of lips' edge.
- The high level approach (Model approaches) [7], [8], [9] and [2], directed by physical extraction distance, based on the model usage. For example we can mention the active edges, which were widely used in lips segmentation. They also exploit the pixel information of the image, but they integrate regularity constraints. The big deformability of these techniques allow them to be easily adapted to a variety of forms. This property is very interesting when it is a matter of segmenting objects that we cannot foresee the form in advance (sanguine vessels, clouds...), but appears more as a handicap when the object structure is already known (mouth, face, hand...).

Potamianos [10] and Matthews [11] showed that image approaches allowed better performances in terms of recognition rates than model approaches in different conditions. For him Evno [12] [13], the most promising methods in labial segmentation, are the ones of high level because they are based on lip models, but it is necessary to wonder on the interest of the extraction of the labial contours in their entirety for the recognition stage. According to speech specialists, the pertinent parameters of verbal communication expression are: the heights, widths and inter-labial surface. From this interpretation we notice that it will be judicious to opt for an extraction method of these parameters based on the detection and the tracking of some points of interest (POI) that will be sufficient to characterize labial movements.

Thus, the problem of the labial segmentation is to define POI on the lips and to track them throughout the speech sequence (The problem of immobility of the head can be resolved by the use of fixed camera on the face guaranteeing a fixed plan of the zone of the lips [14]). In the following parts, we will present a brief view on the Sequential Similarity Detection Algorithms (SSDA), and we will propose in another part a new spatial-temporal approach for the movement tracking which is based on the different directions of Freeman coding and on the vote techniques.

3. Algorithm of movement tracking:

For us the problematic of POI tracking movement (that are defined by blocks of w * w) on lips contours, comes back to detect in successive pictures of a video sequence, the blocks that present the maximum similarity with the blocks manually defined on the first image of the sequence (Figure2).



(a) (b)

Figure 2. Example of tracking POI on Successive image: (a) First initialized image, (b) Detection of the different POI on an image of the same sequence

Therefore, it is well now a matter of a detection problem and of block research that have the maximum similarity with the pattern defined on the first image of the sequence. For this subject several algorithms and similarity distances have been presented, but we notice that it is difficult to adapt them to our problematic due to the complex nature of the movements of the lips. In the following party we are going to present a general view on these algorithms and we will develop our own approach of POI tracking.

3.1. Brief view on algorithms of detection of sequential similarity (SSDA)

In the literature we discover several methods for the research or tracking on « Pattern» in an image. One of the most adequate for our problematic is the « Template Matching » [15], [16].

It consists in discovering and searching for an arbitrary model in level of gray g (x, y) in a given image f (x, y) and this by using some measures and similarity distances:

$$\int_G (f-g)^2 \quad \text{Or} \quad \int_G |f-g|^2 \quad \text{Or} \quad \max_G |f-g|$$

The minimum for these measures will be the best model to choose. In the case of the Euclidean Distance:

$$\int_G (f-g)^2 = \int_G f^2 + \int_G g^2 - 2\int_G f.g$$

the maximum of : $2\int_G f.g$ is the best match, the other terms being constant. This ``cross-correlation"

yields a result only if the integral is computed over the whole area G.

In the discrete case, this takes the form:

$$R(i,j) = \sum_m \sum_n f(i+m, j+n).g(m,n)$$

If the variation in the energy of the image f can be ignored. Otherwise the Normalization of Cross-Correlation (NCC) has to be used:

$$R(i,j) = \sum_m \sum_n f(i+m, j+n).g(m,n) / \sqrt{\sum_m \sum_n f(i+m, j+n)^2}$$

It takes the same amount of computing time for any $g \in G$.

We can distinguish several disadvantages by the use of the NCC for the Template Matching:

- If the energy of the image $\sum f^2(x, y)$ varies with the position, the research of the model can fail. For example, if the correlation between the regions looked for and the model is well inferior to the correlation between the model and the effect of luminance.
- The measure depends enormously on the size of the model.
- Sensitive to the ubrut variation of luminance that can occur in an image sequence.
- This measure is enough abstract since in the case of small model size, the measures of this method are very sensitive to the noise.

Therefore we clearly notice that this research method of model is applicable in several cases, but we judge that it is not sufficient and convincing in our problematic, where the details are very important, the movements are rapid and well defined and the variations of luminance are very susceptible especially for the POI that concerns the inferior lip. In what follows we will present our new approach for tracking lip movements, from the POI, in a sequence of images.

3.2. Algorithm of POI tracking

As we explained it in section 2, the originality of our technique of labial-movement tracking lies in the case of being limited to a set of POI. In addition at the time of POI tracking it is useless to sweep the entire image; there are algorithms of tracking based on the principle of the coding of Freeman to predict all the possible spatial movements of a set of points.

Our POI movement tracking approach bases itself thus:

- On the different directions of the coding of Freeman to localize the points candidate describing the potential movements of a POI;
- On a vote technique to identify among all the points candidate, the one that corresponds better to the origin POI.

In what follows details of these two techniques are given.

3.2.1. The Freeman coding of directions: In order to follow better the movements of the lips, we supposed in the present work that the head of the speaker is immobile in comparison to the camera. We limit ourselves thus to the track of the movements of the lips that are generally complex, quick movements and especially in all directions. These different directions can be easily represented by the directions of the coding of Freeman whose principle consists in presenting the connection of the successive pixels according to figure 3.

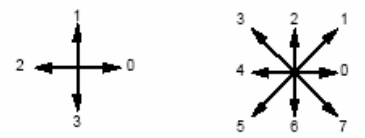
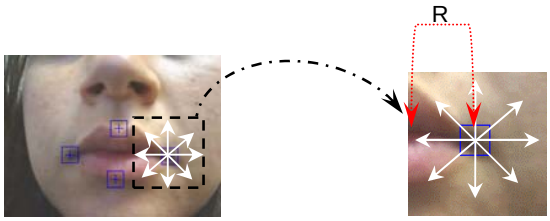


Figure 3. Different directions of Freeman coding

So if P_n is POI and Pm is a point describing the potential movement of P_n, the displacement of P_n to P_m cannot therefore be done only in the 8 possible directions of the coding of Freeman.

Nevertheless, in order to increase the precision of our technique of predictions of POI movements, we spread our research of the candidate points of POI to the set of the Pm points being located in one of the direction of the coding of Freeman on a ray of R points (R the number of block for every direction). In our experimentation R is set up to the value 10. Figure 4 illustrates this principle.



R: number of block for every direction

Figure 4. Different research directions according to the coding of Freeman on a ray of R points

3.2.2. Technical research and voting principle: Our research approach can be summarized in two parts:

– The first step consist in calculating the list of the distances ($\mathbf{D}\{\mathbf{b}_n\ (x, y)\}$) in level of gray one for n blocks candidates with the model to search for in every potential point. All the distance values will be protected in accumulator that will be treated in a second part of our approach (Figure 5).

Accumulator			
$D\{\boldsymbol{\delta}_1(x_1, y_1)\}$	…	$D\{\boldsymbol{\delta}_1(x_{w*w-1}, y_{w*w-1})\}$	$D\{\boldsymbol{\delta}_1(x_{w*w}, y_{w*w})\}$
⋮		⋮	⋮
$D\{\boldsymbol{\delta}_n(x_1, y_1)\}$	…	$D\{\boldsymbol{\delta}_n(x_{w*w-1}, y_{w*w-1})\}$	$D\{\boldsymbol{\delta}_n(x_{w*w}, y_{w*w})\}$

w: Block Size

Figure 5. 'Accumulator': Table of luminance distance for the different candidate Blocks

– In the second step we add to our accumulator new columns where we will stock the number of voices for each block and we will vote at the end for the block having maximum voices (Figure 6).

➔ Voices will be attributed according to the following algorithm (w: Block Size):

```
For i = 1 To w*w
 Do
        Min←Acc[i][1]
 // Initialization of the minimal distance on a column
// Research of the minimal distance for the column i
// 8 the number of different directions of Freeman coding
// R the number of block for every direction

        For j = 2 To 8 * R
            Do
              If (Acc[i][j]<min)
                 Min← Acc[i][j]
               EndIf
            EndDo

     // Attribution of the voices for blocks having the
     minimal distance for the column I for every direction

        For j = 2 To 8 * R
            Do
               If (Acc[i][j]=min)
              Acc[w*w+1][j]← Acc[w*w+1][j]+1
             EndIf
            EndDo

EndDo
```

Accumulator			Nb. voice
$D\{\mathbf{b}_1(x_1, y_1)\}$	…	$D\{\mathbf{b}_1(x_{w*w}, y_{w*w})\}$	13
$D\{\mathbf{b}_m(x_1, y_1)\}$	…	$D\{\mathbf{b}_m(x_{w*w}, y_{w*w})\}$	80
$D\{\mathbf{b}_n(x_1, y_1)\}$	…	$D\{\mathbf{b}_n(x_{w*w}, y_{w*w})\}$	26

Block elected for the research model

Figure 6. 'Accumulator': Table of luminance distances with the vote principle

➔ To calculate the distances of luminance we can use any measure of similarity (SSD, NCC …). For our case we calculate it as follows:

$$D\{\mathbf{b}_n(x, y)\} = |\ \mathbf{g}_i(x, y) - f_{i-1}(x, y)\ |$$

Where:

– $D\{\mathbf{b}_n(x, y)\}$ **:** Distance of luminance for the pixel (x, y) of block $\mathbf{b}_n$ of the POI n.
– i **:** Number of image of the sequence.
– **g** (x, y) **:** Luminance of the model looked for.
– f (x, y) **:** Luminance of the image in which we make the research of the model.

➔ With this method:
- We highlight all the details of the model to look for,
- The result is independent of the size of the block,
- We diminished the noise effect; the latter will not influence the entirety of the candidate block.

In fact thanks to our vote principle the noise will affect only some voices, and it will not affect the whole result.

4. Experimental results:

In this part we will present some experimental results for our approach of model researches while comparing it with other approaches.

4.1. Comparison between our approach detection and other approaches:

SIMI°MatchiX is image processing software for automatic movement tracking. The pattern matching algorithm can be utilized with video clips automatically tracks user-defined patterns [17]. We use this software to situate our 'vote algorithm' according to other model research techniques. The experimental results are the following ones (Figure 7):

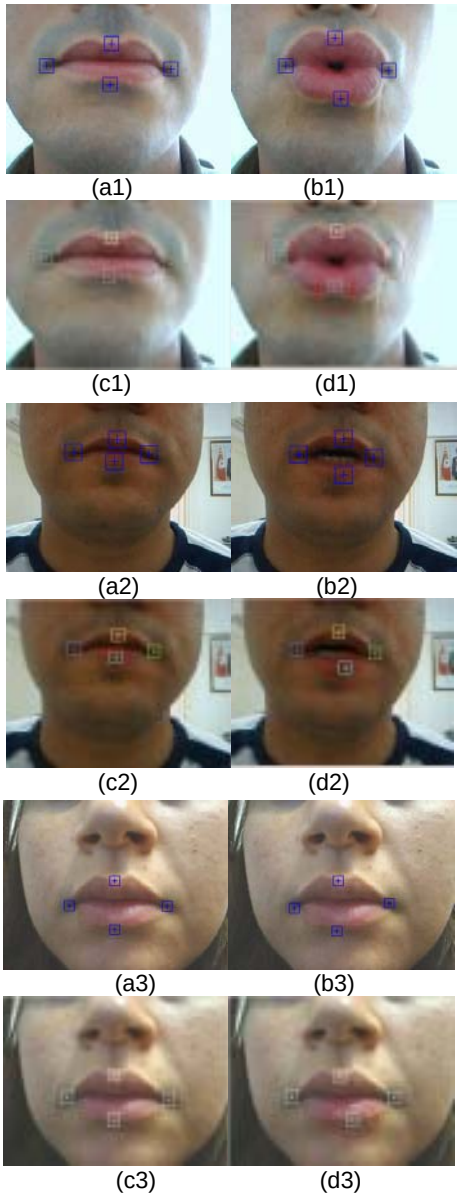


Figure 7. Experimental Results for different speakers: (a) (b) followed by vote algorithm and (c) (d) followed by software 'SIMI°MatchiX'

In these results we notice deficiencies of 'SIMI°MatchiX' system especially for the movements of the lower lip due to the variations of luminance that are very important in this zone.

4.2. Result of POI tracking on our audio-visual corpus:

After POI tracking, we tried to clear and to track the variation of some physical distances, vertical opening (DV) and horizontal opening (DH) (Figure 8).

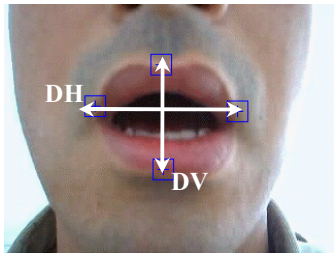
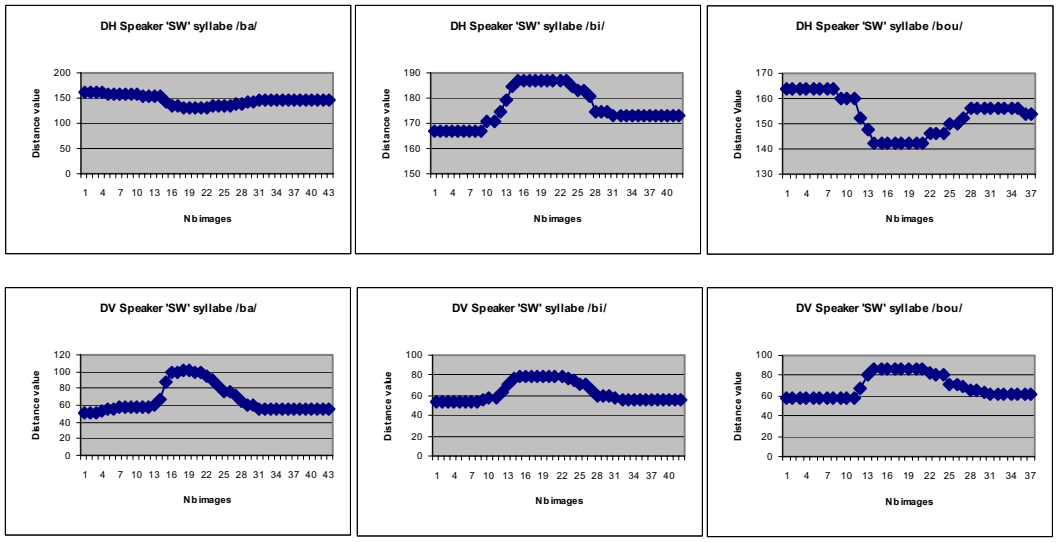


Figure 8. Tracking Distances for the recognition

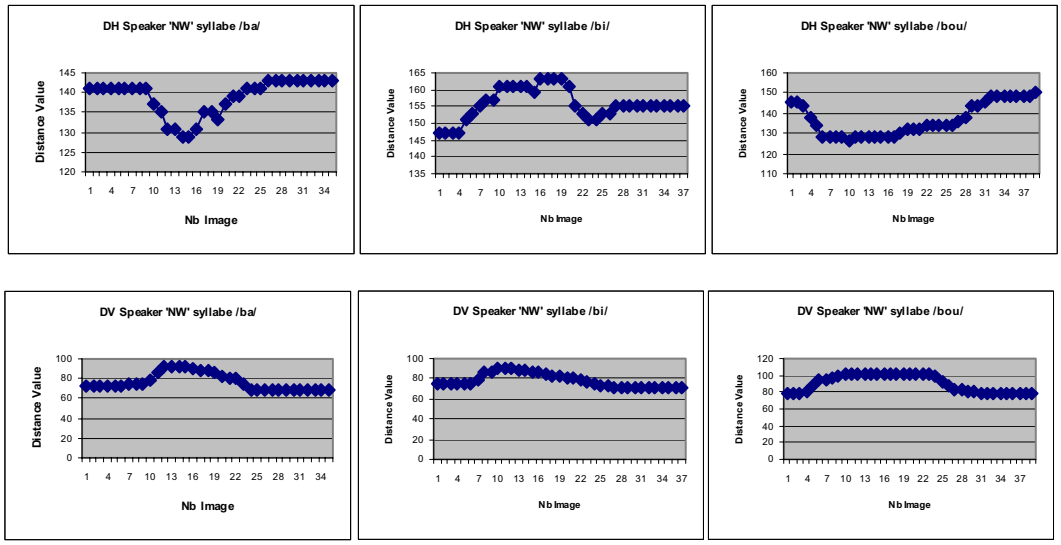
The chosen sequences for distances tracking are sequences of phrases for syllables that are visually differentiable: /ba/, /bi/ and /bou/. The experimental results are given by the following curves for different speakers (Figure 9). We see on these curves that there are some differences between the variations of the distances and every viseme. These variations are well justified and well identified with the visual features of every syllable. We also notice that these variations have the same pace for different speaker; therefore one can think about a multi-speaker recognition system.

5. Conclusion

In this paper we tried to present a new spatial-temporal approach of movement tracking. Our technique is applied to the problematic of lips tracking for speech recognition; it is bases on two principles: the one of the different directions of the coding of Freeman and the one of the vote. This approach has been tested with success on our audio-visual corpus, for the tracking of characteristic points on lips contours. We notice also that the tracking of the variations of the distances with different syllables was discriminatory and reflects well the reality of the movement of the lips. As perspective to this work we propose to add other descriptors in addition to the opening distances of the lips (especially descriptors of the internal features of the mouth) and to conceive a system for the classification of these features and the recognition of visemes.



(a)



(b)

Figure 9. Tracking of the variations of the Vertical Distances (DV) and Horizontal (DH) with several syllables /ba/, /bi/ and /bou/ for the speaker “SW” (a) and “NW” (b)

6. References

[1] Petajan, E. D., Bischoff, B., Bodoff, D., and Brooke, N. M., "*An improved automatic lipreading system to enhance speech recognition*," CHI 88, pp. 19-25, 1988.

[2] Philippe Daubias, *Modèles a posteriori de la forme et de l'apparence des lèvres pour la reconnaissance automatique de la parole audiovisuelle*. Thèse à l'Université de Maine France 05-12-2002.

[3] Roland Goecke, *A Stereo Vision Lip Tracking Algorithm and Subsequent Statistical Analyses of the Audio-Video Correlation in Australian English*. Thesis Research School of Information Sciences and Engineering. The Australian National University Canberra, Australia, January 2004.

[4]McGurck et John Mcdonald. *Hearing lips and seeing voice*. Nature, 264 : 746-748, Decb 1976.

[5] Iain Matthews, J. Andrew Bangham, and Stephen J. Cox. *Audiovisual speech recognition using multiscale nonlinear image decomposition*. Proc . 4th ICSLP, volume1, page 38-41, Philadelphia, PA, USA, Octob 1996.

[6] Uwe Meier, Rainer Stiefelhagen, Jie Yang et Alex Waibe. *Towards unrestricted lip reading*. Proc 2nd International conference on multimodal Interfaces (ICMI), Hong-kong, Jan 1999.

[7] Prasad, K., Stork, D., and Wolff, G., "*Preprocessing video images for neural learning of lipreading*," Technical Report CRC-TR-9326, Ricoh California Research Center, September 1993.

[8] Rao, R., and Mersereau, R., "*On merging hidden Markov models with deformable templates*," ICIP 95, Washington D.C.,1995.

[9] Patrice Delmas*, Extraction des contours des lèvres d'un visage parlant par contours actif (Application à la communication multimodale)*. Thèse à l'Institut National de polytechnique de Grenoble, 12-04-2000.

[10] Gerasimos Potamianos, Hans Peter Graft et eric Gosatto. *An Image transform approach For HM based automatic lipreading*. Proc, ICIP, Volume III, pages 173-177, Chicago, IL, USA Octb 1998.

[11] Iain Matthews, J. Andrew Bangham, and Stephen J. Cox. A *comparaison of active shape models and scale decomposition based features for visual speech recognition*. LNCS, 1407 514-528, 1998.

[12] N.Eveno, "*Segmentation des lèvres par un modèle déformable analytique*", Thèse de doctorat de l'INPG, Grenoble, novembre 2003.

[13] N. Eveno, A. Caplier, and P-Y Coulon, *"Accurate and Quasi-Automatic Lip Tracking"* , IEEE Transaction on circuits and video technology, parution en mai 2004.

[14] Nicolas Eveno, Patrice Delmas, Pierre-Yves Coulon : *"Vers l'extraction automatique des lèvres d'un visage parlant"*, GRETSI 01, pp 193-196, Toulouse, septembre 2001.

[15] Rosenfeld76, A. Rosenfeld and A.C. Kak, *Digital Picture Processing*, Computer Science and Applied Mathematics, Academic Press, New York, 1976.

[16] Pratt78, W.K. Pratt, *Digital Image Processing*, Wiley, New York, 1978.

[17] SIMI Reality Motion Systems, GmbH, D-85705 Unterschleissheim , Germany. www.simi.com/matchix