Liu et al. EURASIP Journal on Image and Video Processing (2017) 2017:88
DOI 10.1186/s13640-017-0233-y

EURASIP Journal on Image and Video Processing

RESEARCH Open Access

Inner lips feature extraction based on CLNF with hybrid dynamic template for Cued Speech

Li Liu[1*], Gang Feng[1] and Denis Beautemps[2]

Abstract

In previous French Cued Speech (CS) studies, one of the widely used methods is painting blue color on the speaker's lips to make lips feature extraction easier. In this paper, in order to get rid of this artifice, a novel automatic method to extract the inner lips contour of CS speakers is presented. This method is based on a recent facial contour extraction model developed in computer vision, called Constrained Local Neural Field (CLNF), which provides eight characteristic landmarks describing the inner lips contour. However, directly applied to our CS data, CLNF fails in about 41.4% of cases. Therefore, we propose two methods to correct the *B* parameter (aperture of inner lips) and *A* parameter (width of inner lips), respectively. For correcting the *B* parameter, a hybrid dynamic correlation template method (HD-CTM) using the first derivative of smoothed luminance variation is proposed. HD-CTM is first applied to detect the outer lower lips position. Then, the inner lower lips position is obtained by subtracting the validated lower lips thickness (VLLT). For correcting the *A* parameter, a periodical spline interpolation with a geometrical deformation of six CLNF inner lips landmarks is explored. Combined with an automatic round lips detector, this method is efficient to correct *A* parameter for round lips (the third vowel viseme made of French vowels with a small opening). HD-CTM is evaluated on 4800 images of three French speakers. It corrects about 95% CLNF errors of the *B* parameter, and total RMSE of one pixel (i.e., 0.05 cm on average) is achieved. The periodical spline interpolation method is tested on 927 round lips images. The total error of CLNF is reduced significantly, which is comparable to the state of the art. Moreover, the third viseme is properly distributed in the parameter *A* and *B* plane after using this method.

Keywords: CLNF, Luminance variation, HD-CTM, Periodical spline interpolation, Inner lips contour parameters, Cued Speech, Visemes

1 Introduction

Lips detection is an active research topic since lips (especially inner lips) hold significant information speech production, and it plays an important role in speech recognition based on lips visual features. In 1967, Cornett [1] developed Cued Speech (CS), which is a complement of lipreading to enhance speech perception from visual input including lips and hand. This system was adapted from American English to French in 1977. In French CS, which is named *Langage Parlé Complété* (*LPC*) [2], five hand positions are used to encode the vowels and eight hand configurations to encode the consonants. It is often used by deaf people or hearing people when they communicate with deaf orally educated. This paper extracts inner lips contour in the CS case, in which the lips may be occluded by hand. Moreover, it can also be used in a non-CS case.

Several approaches to extracting lips contour in audiovisual speech processing have been investigated in the literature. One of the most widely used techniques is model-based lips detection. Active Shape Model (ASM) [3] and Active Appearance Model (AAM) [4] were proposed to segment lips contour. The shape and appearance of lips are learned from training data with manually annotated faces, and lips configurations are described by a set of model parameters. Bandyopadhyay [5] investigated a lips feature extraction technique combined with ASM and used the contrast between the lips and face to segment lips contour. In the early time, for

* Correspondence: li.liu@gipsa-lab.grenoble-inp.fr
[1]University Grenoble-Alpes, GIPSA-lab, F-38040 Grenoble, France
Full list of author information is available at the end of the article

model-based techniques, a large training set and good initial condition are necessary even if it may not be efficient. Stillitano et al. [6] used both active contours and parametric models for lips contour extraction. This method needs prior knowledge of lips shape. Another technique is based on segmentation in color spaces [7, 8]. Color-based clustering assumes that there are only two classes, i.e., skin and lips, and this technique may not be efficient if facial hair or teeth exist. Currently, deep learning is very popular in the feature extraction field. Hlavac presented a Convolutional Neural Network (CNN) for lips landmark detection in [9], which achieves a sub-pixel accuracy in landmark detection, but some errors remain since no robust features around the chin can be locked.

In 2013, Baltrusaitis et al. [10] proposed the Constrained Local Neural Field (CLNF), which is robust for facial landmark detection in the general case. Note that CLNF is a novel instance of the Constrained Local Model (CLM) [11] that deals with the issues of feature detection in the complex scene. CLNF learns the variation in the appearance of a set of template regions surrounding individual feature landmarks. It replaces the SVR patch expert of CLM by the Local Neural Fields (LNF) and uses Non-Uniform Regularized Landmark Mean-Shift as the new optimization method. In CLNF, the neural network layer and convolution kernel are used to capture non-linear relationships between pixel values and the output responses. CLNF is trained on about 4000 faces from independent databases HELEN, LFPW, and Multi-PIE. The experiments in [10] show that CLNF is more accurate than ASM, AAM, and CLM, and it is especially robust to occlusions, rotated face, and different lighting conditions. Therefore, we investigated an automatic inner lips tracker based on CLNF. However, directly applied to our visual database of CS data, CLNF failed in about 41.4% of cases. This work aims at improving CLNF performance using post-processing methods. Another possibility is to adjust CLNF by retraining only lips images, but large training is needed, and efficient features around the lips may not be enough to track the inner lips [9].

This paper presents the following contributions. We deal with the extraction of inner lips from video without using any artifices. Recall that the B parameter is the aperture of the inner lips and the A parameter means the width of the inner lips (see Fig. 1). Two post-processing methods for inner lips parameter extraction based on CLNF are presented. We named these methods as Modified CLNF in this paper. Figure 2 is an outline showing the logic structure of this work.

- For correcting the B parameter error of CLNF, a novel hybrid dynamic correlation template method (HD-CTM) using the first derivative of smoothed luminance variation is proposed to detect the outer lower lips. Then, inner lower lips position can be estimated by subtracting the validated lower lips thickness (VLLT) from the outer lower lips position. This method is robust for any lips shape.
- For correcting the A parameter, a periodical spline interpolation is proposed based on six CLNF inner lips landmarks for round lips (like [o] and [y]). The round lips are selected by an automatic round lips detector based on Discrete Cosine Transformation (DCT).

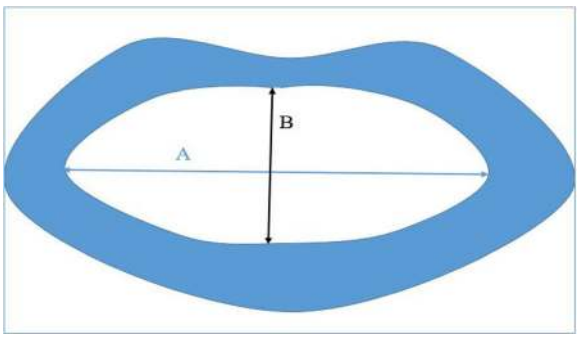

Fig. 1 B parameter is the aperture of the inner lips and A parameter the width of the inner lips

The method for correcting the B parameter is evaluated on 4800 images on three subjects. It corrects 95% CLNF errors, and total RMSE of one pixel (0.05 cm on average) is reached, instead of four pixels (0.2 cm on average) when using the original CLNF. The evaluation of the periodical spline interpolation is carried out on 927 round lips images. The total error is reduced from 2.12 cm to 0.35 cm.

2 Relation with prior works

In the previous study of CS lips feature extraction, Heracleous et al. [12] and Aboutabit et al. [13] extracted the A and B parameters by painting blue color on the subject's lips. Firstly, the gray level image was subtracted from the blue component of the RGB image. Then, a threshold was applied to the resulted image to segment the blue lips. To get rid of the blue lips, a dynamic correlation template method in our previous work [14] was introduced to estimate B parameters of inner lips. In this paper, we improve this method for correcting the B parameter by a HD-CTM, which is robust for tracking any lips shape. More precisely, compared with [14], three improvements are proposed to enhance the robustness of our method. They contain a new initial condition for the first image, an automatic tracking of the starting position for template searching region, and a novel dynamic hybrid template. Meanwhile, we focus on correcting the A

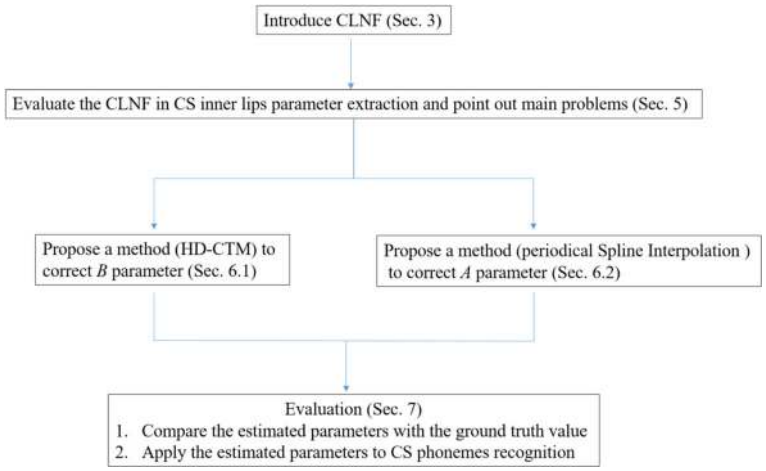

Fig. 2 An outline showing the logic arrangement of this work

parameter using the periodical spline interpolation with an automatic round lips detector to filter the third viseme. Above all, a complete inner lips tracking model which corrects the A and B parameters of CLNF for multi-subjects is proposed.

3 CLNF model

CLNF contains three main parts: a point distribution model (PDM) [15] of 68 points for facial contour, LNF patch expert, and Non-Uniform RLMS [16]. In these 68 points, 12 of them are dedicated to describing the outer lips contour and 8 for the inner lips contour. By applying PCA to the data, lips can be estimated using the sum of the mean shape and the variation part.

3.1 Local neural fields

In CLNF, LNF is used as the patch expert. It is an undirected graphical model, which models the conditional probability of a continuous valued vector y (the probability that a patch is aligned) depending on continuous x (the pixel intensity values in the support region). It shows the relationships between pixels (neighboring and long distance) by learning both similarity and long distance sparsity constraints. LNF also includes a neural network layer and a convolution kernel that can show non-linear relationships between pixel values and the output responses. LNF gives a conditional probability distribution with probability density:

$$P(y|X) = \frac{\exp(\Psi)}{\int_{-\infty}^{+\infty} \exp(\Psi) dy}, X = \{x_1, ..., x_n\}, y = \{y_1, ..., y_n\} \tag{1}$$

where X is the observed input and y is the predicted output (expected response maps). Ψ is the potential function in linear combination of vertex features, edge features with different coefficients. All the parameters are estimated by maximizing condition log-likelihood of LNF.

$$\left(\overline{\alpha}, \overline{\beta}, \overline{\gamma}, \overline{\Theta}\right) = \arg\max_{\alpha,\beta,\gamma,\Theta}(L(\alpha, \beta, \gamma, \Theta)) \tag{2}$$

$$L(\alpha, \beta, \gamma, \Theta) = \sum_{q=1}^{M} \log P\left(y^{(q)}, |x^{(q)}\right) \tag{3}$$

Given training data $\left\{x^{(q)}, y^{(q)}\right\}_{q=1}^{M}$ of M patches, we want to pick the α, β, γ, and Θvalues that maximize the conditional log-likelihood of LNF on the training sequences.

3.2 Non-Uniform RLMS

Non-Uniform RLMS is used as the new optimization method in CLNF. RLMS [16] is used in CLM which is a local approach and relies on an initial parameter estimate.

If we have an initial estimate p_0, we want to find a parameter update Δp to get closer to a solution $p^* = p_0 + \Delta p$ (where p^* is the optimal solution). Hence, the iterative fitting objective is as follows:

$$\arg\min_{\Delta p}\left[\mathfrak{R}(p_0 + \Delta p) + \sum_{i=1}^{n} D_i(x_i; I)\right] \tag{4}$$

Considering the different weight of each patch expert for each landmark, Non-Uniform RLMS takes into account the patch expert reliability compared with RLMS. The aim is to minimize the following objective function:

$$\arg\min_{\Delta P}\left(\|P+\Delta P\|^2_{A^{-1}}+\|J\Delta P-v\|^2_{W}\right) \quad (5)$$

where P is a vector of output prediction parameter obtained from PDM and LNF. v is the mean-shift vector over the patch response. J is the Jacobian of the landmark location for P. r is the matrix describing the prior on$P$.

4 Experiment dataset

The database contains videos of 50 French words made of numbers and daily words (such as "Bonjour" and "Rendez-vous"). The corpus is uttered 10 times by three French subjects: one female CS speaker and two male speakers (see Fig. 3). Video images of the speaker's upper body (720×576 RGB images, 50 fps) are recorded in a soundproof booth in GIPSA-lab, France. Images are obtained every 20 ms. Words and vowels are annotated with Praat based on speech sound signals. We use the first repetition of the three speakers containing 4800 images corresponding to all types of lips shape to evaluate the B parameter. As for evaluating the A parameter, 927 round lips images are used. To evaluate the performance of the original CLNF model and our model, the ground truth inner lips contour is extracted manually by an expert placing several landmarks on inner lips.

When studying lips parameter extraction problem, three visemes (lips shape) are often considered which correspond to the 13 French vowels [17] shown in Table 1. The first viseme correspond to opened vowels while the second viseme describe opened round vowels. The third viseme is characterized by small-opened round vowels. In this work, we use the visemes mentioned in [18].

Table 1 Three visemes of vowels

Vowels	
Viseme	Phonemes
V1	a,ɛ,ɛ̃,e,i
V2	ã,ɔ,œ
V3	o,ø,ɜ̃,y,u

5 Performance of CLNF on our data

CLNF is first straightforward applied to all video of each annotated word in the database. Among 68 facial points, 8 landmarks (6 of them for inner lips and 2 endpoints) are used to describe inner lips contour. Recall that main advantages of CLNF are its robustness to variable lighting conditions, in the presence of occlusion, and head movements.

A whole inner lips contour can be first generated using the eight inner lips landmarks. A linear interpolation [19] is used for the upper inner lips contour, and a spline interpolation [19] is applied to lower inner lips. Figure 4a shows one example of the excellent performance of CLNF, and the green curve shows interpolated inner lips. The A and B parameters are then calculated from the inner lips contour using the classic method in [20].

However, some mistaken cases remain. On the one hand, the landmarks of the lower lips are often poorly placed in a much higher position while no error is presented for upper lips (Fig. 4b). It causes a wrong B parameter. This phenomenon can be explained by the fact that the CLNF is based on a dictionary of training images. If the lips shape and appearance are not properly taken into account during the training phase, it may lack the template during the optimization step. In fact, the lower inner lips detection is challenging since the lips area is often very complex (tongue and teeth may be visible) and lighting condition is alterable.

On the other hand, the two endpoints of the inner lips may be poorly placed (Fig. 4c). It causes a mistaken A parameter. Indeed, from a "geometrical" perspective, the two endpoints of the inner lips are not false because, in this case, the inner contour can achieve these two endpoints. However, in an articulatory-acoustic point of view, these two points do not define the proper A parameter of the inner lips.

Concerning the error of the B parameter, a comparison between the original CLNF and ground truth values shows that CLNF only obtains about 58.6% accuracy on

Fig. 3 The first one is the CS speaker and other two are non-CS speakers. Note that in this work, color marks on the CS female speaker front and hand are not taken into account

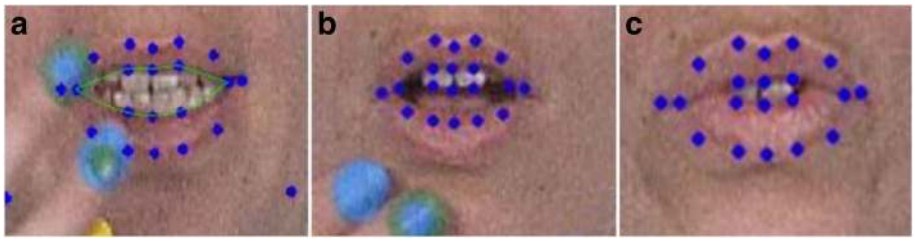

Fig. 4 Examples of 20 CLNF landmarks placed in the full lips region. Eight points describe the inner lips contour. **a** Good inner lips contour of CLNF with hand occlusion. The *green curve* is the inner lips contour using interpolation. **b** Mistaken CLNF landmarks in the case of the *B* parameter). **c** Mistaken CLNF two end landmarks for round open inner lips (mistaken *A* parameter)

average (75.2% for speaker 1, 52.2% for speaker 2, and 48.5% for speaker 3). We consider the errors as mistaken when they are larger than two pixels. In Fig. 5, we observe that most of the error appears negative. It means the mistaken CLNF lower inner point is often placed above the ground truth points. Since there is a large proportion of *B* parameter error and inner lips height plays a very crucial role in speech production, we pay particular attention to the *B* parameter correction.

To see the CLNF performance concerning the *A* parameter, three visemes are plotted using the first repetition of the CS speaker in the *A* and *B* parameter plane (Fig. 6). The distribution of each vowel is presented by a Gaussian ellipse. Figure 6a corresponds to the original CLNF landmarks and Fig. 6b to the ground truth. We observe that three visemes of the original CLNF are totally mixed compared with the distribution of the ground truth.

6 Proposed methods and experiment details

6.1 Parameter *B* correction based on HD-CTM and back-subtracting of VLLT

The proposed method is based on the luminance variation along the middle CLNF landmarks of lips. A suitable spline smoothing is first applied to this luminance variation and also the first derivation curve. The smoothing degree is carefully controlled so that the noise can be removed without losing useful information. A smoothing coefficient of $p = 0.01$ [21] is used for a good compromise. We may expect to determine the inner lips position by searching the local limit point in the first derivative of the smoothed luminance variation curve. But, this is not always feasible since there are many local limit points (see Fig. 7) without being given a searching interval. Or, even given a searching interval, the local limit position may be fuzzy or uncertain. Moreover, it is sensitive to noise and unable to guarantee coherent results for adjacent images. To overcome the above problems, we proposed to search the limit point using a correlation method with a dynamic template which is called HD-CTM in this paper.

However, inner lower lips detection still remains challenging since this area may be fuzzy, and several different cases have to be considered. For example, the luminance variation from the teeth to lower lips is different with that from the tongue to lower lips. In Fig. 7a, one can see that luminance decrease from the teeth to lower lips corresponds to a local minimal point. However, in Fig. 7b, c, it becomes complicated to find one particular local minimal point corresponding to the lower inner lips boundary. By contrast, in the region of the outer lower lips, the luminance variation is less complicated than in the inner lower lips region. In fact, the middle landmark (in the vertical sense) of the lower lips is more enlightened and corresponds to a high luminance variation. When the luminance goes down, it decreases rapidly as the chin (the lower part of lower lips) is darker. Also, we can see that the luminance value varies on the position. The first derivative of the luminance curve consequently shows a significant "V" shape corresponding to the luminance variation. Therefore, as a proposed solution, the HD-CTM is first applied to detect the outer lower lips position. Then, the inner lower lips position is estimated by subtracting the outer lower lips position from the validated lower lips thickness (VLLT) which will be introduced in detail in the next section. This method is illustrated by Fig. 8.

6.1.1 Determination of the lower outer lips position

This procedure contains the following three steps.

❖Definition of the template

Instead of directly finding a minimal local value in the first derivative curve, which is often difficult due to its great sensitivity to noise, a hybrid dynamic template corresponding to a typical derivative variation in the region around the outer lower lips position is first established. The template is obtained by training some derivative curves reflecting different lips shapes except for closed lips. In this approach, template length (L_M) is a key parameter. A very small template length makes results sensitive to noise while a very large length reduces detection precision. If it is badly chosen, it will not be sufficiently pertinent to indicate the "V" shape of the

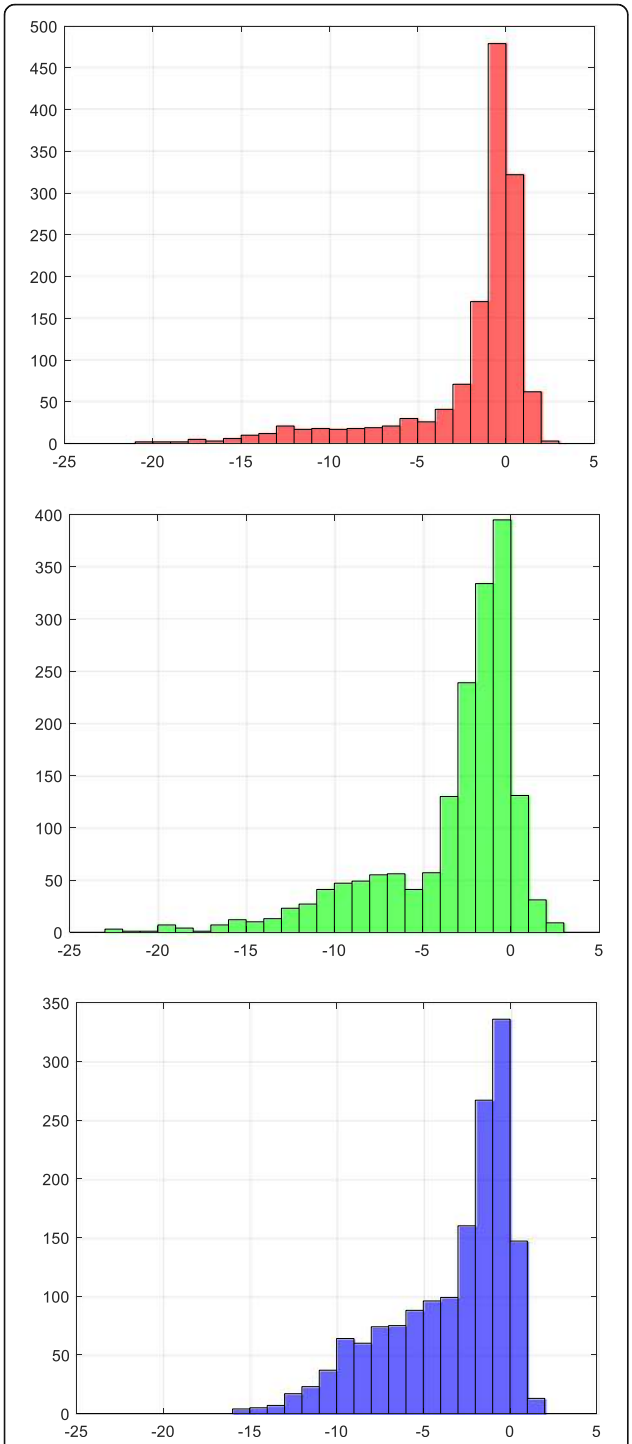

Fig. 5 Three figures correspond to three speakers. *Red*, *green*, and *blue* histograms represent the CLNF error of the three speakers, respectively. The abscissa is the error (larger than two pixels are considered as the mistake) and the *Y*-axis is the frequency of these errors

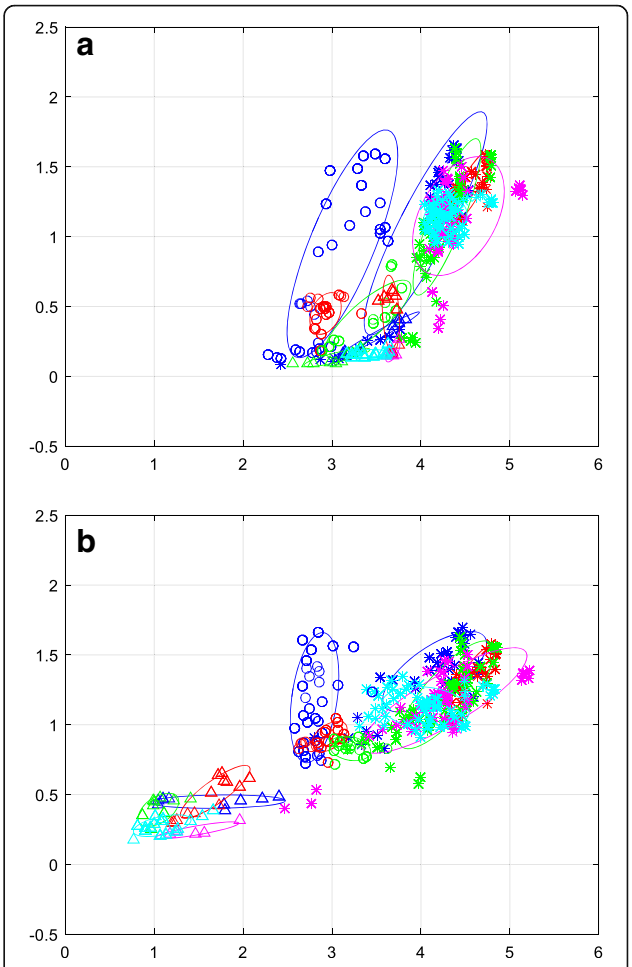

Fig. 6 a, **b** Stars correspond to the first viseme, circles to the second viseme, and triangles to the third viseme. The color order is *blue*, *red*, *green*, *magenta*, and *cyan* for each vowel of viseme. They are corresponding to the vowel order shown in Table 1

derivative curve. This length is set to 20 pixels experimentally so that a rapid sudden change of the outer lower lips position could be taken into account. An example of the template is illustrated in Fig. 9 by a magenta curve with circle. The template is not necessarily symmetric in our case. In fact, a symmetric shape was firstly tried, but it gave slightly larger errors.

In order to increase the capacity of the template to follow the variation the derivative curve, we use a hybrid dynamical template $m(i)$:

$$m(i) = \alpha m_0(i) + (1-\alpha) m_v(i) \tag{6}$$

where $m_0(i)$ denotes the fixed part of the above template. We denote by $v^{n-1}(i)$ the derivative curve of luminance variation for the previous lips image. The variable part of the template is defined as:

$$m_v(i) = v^{n-1}(i) \ \text{ for } \ i \in \left[k_{\text{opt}}^{n-1} + 1, \ \ k_{\text{opt}}^{n-1} + L_M\right] \tag{7}$$

where k_{opt}^{n-1} is the optimal position of the template for the previous lips image. α is the weight for the fixed part

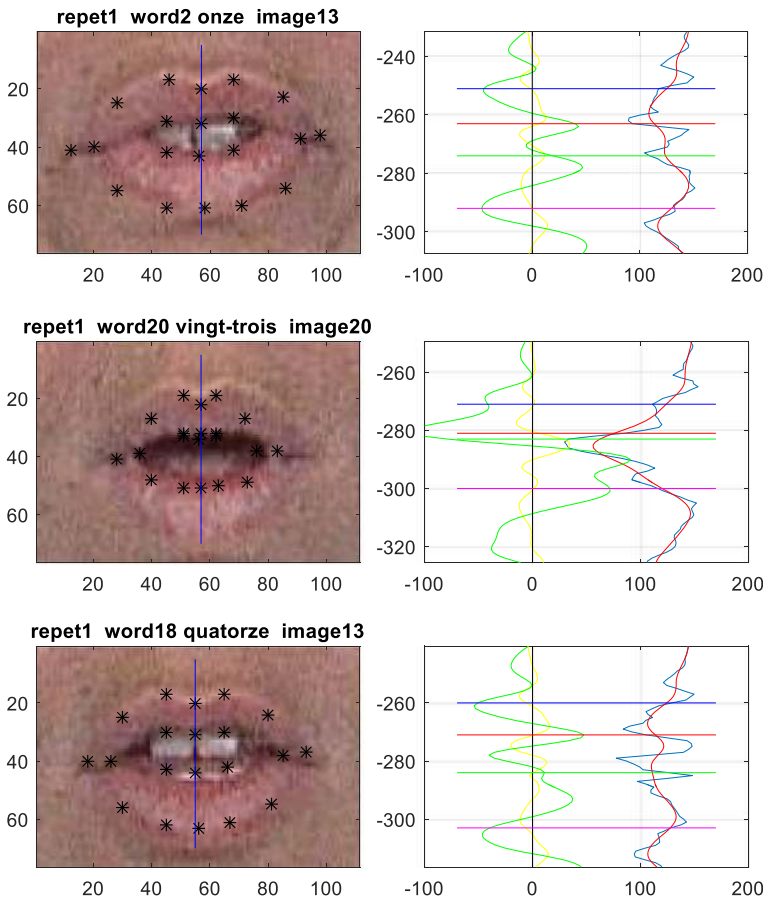

Fig. 7 The *left* figure shows lips ROI with CLNF lips landmarks (20 *black stars*). The *blue line* in the *left* figure corresponds to the *middle point* of the *inner lower lips*, and all the curves in the *right* figure are plotted according to this *blue line*. On the *right* figure, the *blue* curve is the original luminance variation. The *red* curve is the smoothed luminance. The *green* curve is the first derivative of the smoothed luminance. Four lines with *blue*, *red*, *green*, and *magenta* colors correspond to four middle CLNF landmarks around the *blue middle line* in the *left* figure. In (**a**), the luminance decreases from the teeth to lower lips which corresponds to a local minimal point in the *right* figure. In (**b**, **c**), it becomes complicated to find the local minimal point corresponding to *lower inner lips* boundary. However, the *outer lower lips* boundary is always corresponding to a local minimal point on the local "V" curve in the *right* figure for these three cases

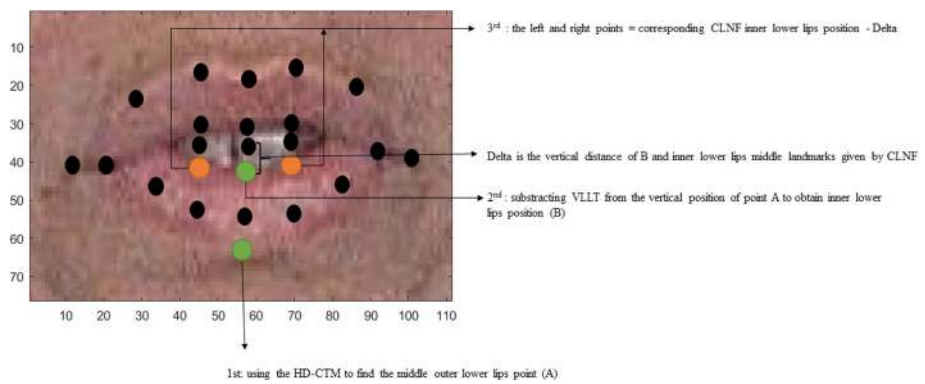

Fig. 8 Twenty *black lips* landmarks are given by the CLNF with 8 inner lips landmarks. The *green point* A is the *middle outer lower* lips position obtained by the HD-CTM and the *green point* B is the *middle inner lower* lips position by subtracting the VLLT from the position of A. Two *orange points* are the *left* and *right* key landmarks determined by the third step

Fig. 9 The figure is plotted along the middle point of the *upper inner lips*. *Blue curve*: original luminance variation. *Red curve*: the smoothed luminance. *Green curve*: the smoothed first derivative. *Red curve* with stars: correlation values between the template and the first derivation curve in function of the position of the template. *Magenta curve* with stars: the proposed hybrid dynamic template. Vertical *black line* around the 301 position corresponds to the initial searching position. The bold *red line* around position 307 located in the maximum correlation value corresponds to the estimated *lower outer* lips position

which is set to be 0.75 experimentally. Note that the performance of the proposed method is not very sensitive to this value; a range of α between 0.7 and 0.9 gives comparable results.

❖Determination of optimal outer lower lips

In order to determine an optimal position of outer lower lips, a correlation value is calculated between the current derivative curve $v(i)$ and the template $m_v(i)$ when the template scans through the searching interval. This method using the correlation with the template reduces the influence of noise and gives a more consistent result for the adjacent image.

The correlation is defined as:

$$c(k) = \sum_{i=1}^{L_M} v(i+k)m(i) \tag{8}$$

The optimal position of the template is determined by

$$k_{\text{opt}} = \arg\max_{k}(c(k)) \tag{9}$$

$k \in [k_0 - \delta_1, k_0 + \delta_2]$, which is the searching interval. k_0 is an initial searching position. δ_1 and δ_2 are two parameters which define the length of the searching interval.

❖Determination of the searching interval

- The searching region for the first image

For non-open lips, k_0 is the original CLNF outer lower middle lips position, while for opened lips, k_0 is defined as:

$$k_0 = A + \text{VLLT} \tag{10}$$

where A denotes the original CLNF inner lower lips position and VLLT is the validated lower lips thickness which is explained in detail in the following section. In this case, we take $\delta_1 = 3$ and $\delta_2 = 18$ experimentally.

- The searching region for images after the first image

The optimal initial search position is estimated from the previous image, so that a continuous tracking can be achieved. k_0 is defined as:

$$k_0 = k_{\text{opt}}^{n-1} + \Delta k \tag{11}$$

where k_{opt}^{n-1} denotes the optimal outer lower position for the previous image. Δk is an estimated translation of the current outer lower lips position with respect to the previous image. To calculate Δk, we take the previous derivation curve $v^{n-1}(i)$ in the interval $\left[k_{\text{opt}}^{n-1} - 10, k_{\text{opt}}^{n-1} + 10\right]$ and also the current derivative curve $v(i)$ in the same interval. After calculating the cross-correlation between these two derivative curves, a searching of its maximal value permits to determine Δk. This interval length is reduced when using this automatic tracking method. In this case, we take $\delta_1 = 3$ and $\delta_2 = 6$ experimentally. The details of the HD-CTM are shown by an example in Fig. 9.

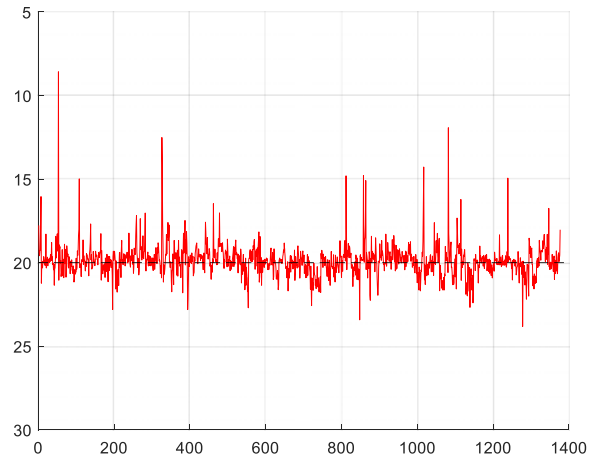

Fig. 10 The *red curve* is the distance between the *inner lower* lips position determined by the ground truth and the *outer lower* lips position obtained by the HD-CTM. Some evident errors come from our detection step

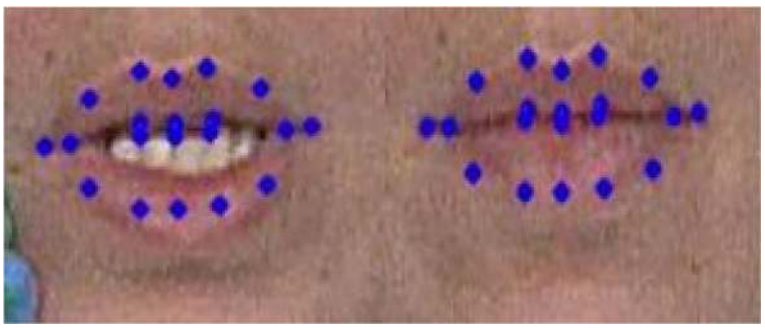

Fig. 11 Mistaken CLNF landmarks (left). Correct CLNF landmarks for closed lips (right). Note that the CLNF landmarks are very similar in these two cases

6.1.2 Determination of the lower inner lips position

To estimate inner lower lips, we first study the distance between the inner lower lips position given by the ground truth and the outer lower lips position obtained by the proposed method. It has been found that this distance is almost a perfect uniform distribution (Fig. 10) except for some errors from the detection error using HD-CTM. More importantly, the distance distribution is invariant whatever the variation of the lips shape. This distance floats slightly around a constant for each speaker. The distance is 19.9 ± 0.97 pixels for the female speaker and 19.7 ± 0.89 and 16.6 ± 0.83 pixels for the other two male speakers. The mean value of this distribution can be regarded as a validated lower lips thickness, which is called VLLT. For a given speaker, the VLLT can be estimated by training their data. For our three subjects, VLLT is set to be 20, 20, and 17 pixels, respectively. The inner lower lips position can then be estimated by subtracting the outer lower inner lips position from the VLLT.

Someone could think of using the "lower lips width" estimated by original CLNF landmarks instead of the VLLT. In fact, by comparing, we found that the "lower lips width" is poorly estimated especially when CLNF gives mistaken lips landmarks. Moreover, the performance of evaluation shows that using "lower lips width," we obtain higher RMSE (1.49) than using the VLLT (RMSE is 1.0).

It should be mentioned that, if the inner lower lip's middle position value obtained by Modified CLNF is less than that estimated by original CLNF, the initial value is kept. A parallel translation with the same distance as the inner middle lips point is proposed to locate two other inner lower points, which are the left and right points of the middle point (see Fig. 8).

6.1.3 Closed lips filter based on DCT analysis

When the upper and lower inner lips points are close to each other, the lips may be confused with true closed lips (Fig. 11). In fact, CLNF performs perfectly for closed lips for which the dynamic template is not suitable. To eliminate closed lips and keep the good result of the original CLNF, a closed lips detector based on DCT coefficients is developed.

Firstly, lips ROI is determined by the 20 landmarks of CLNF which efficiently delimit the lips region and determine a precise center of this region. A suitable-sized ROI is determined according to this center (Fig. 12). The ROI size is 110×75 pixels in our case.

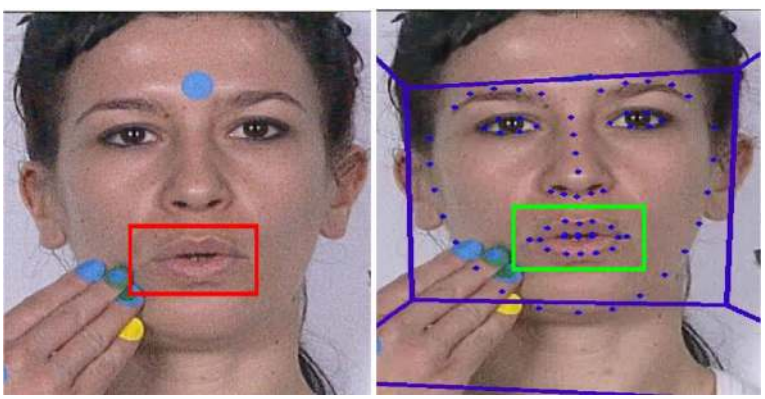

Fig. 12 Lips ROI determined by a blue mark in the front of the speaker (*left*). Lips ROI (the same size) determined by a center point estimated from CLNF landmarks (*right*)

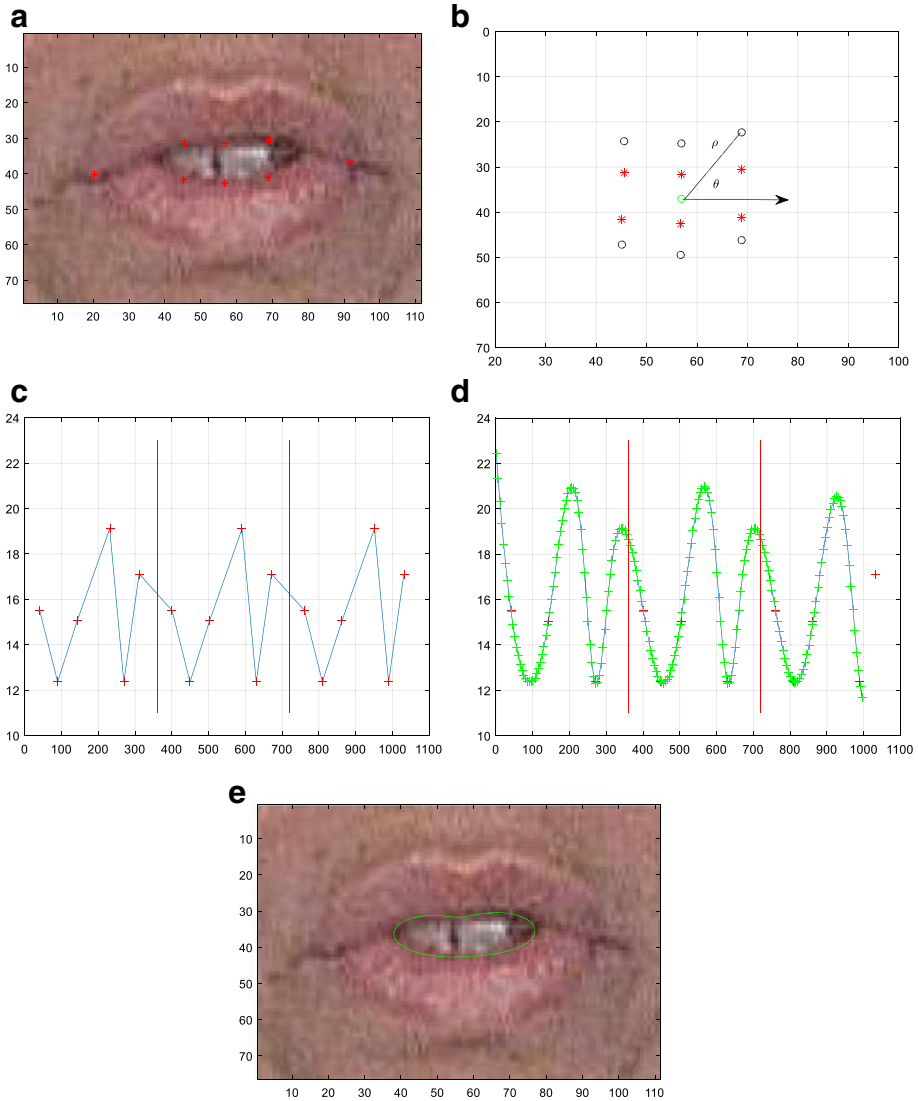

Fig. 13 **a** Speaker's lips with CLNF original landmarks (note that two *red* endpoints of the inner lips contour are mistakenly placed). **b** Six center points are plotted with *red stars* which are dilated in the vertical scale to form a square in polar coordinates. **c** The vertical axis presents the ρ value and the horizon axis θ. The six points are repeated three times. **d** Periodical spline interpolation is realized, and only the period inside two red lines is used to return to Cartesian coordinates. **e** The whole interpolated inner lips contour

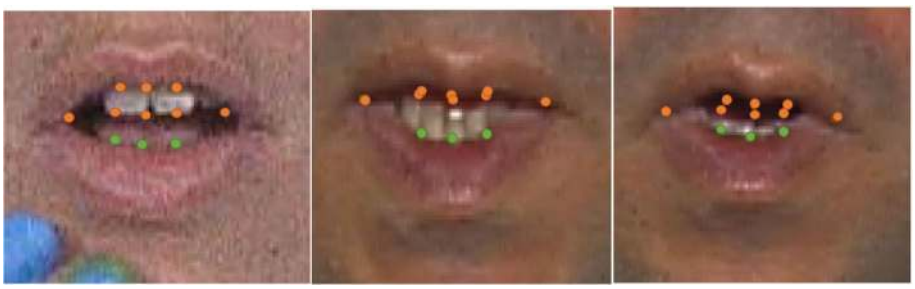

Fig. 14 Examples showing initial mistaken CLNF landmarks (*orange points*) and corrected landmarks using the proposed method (*green points*) for the inner lower lip

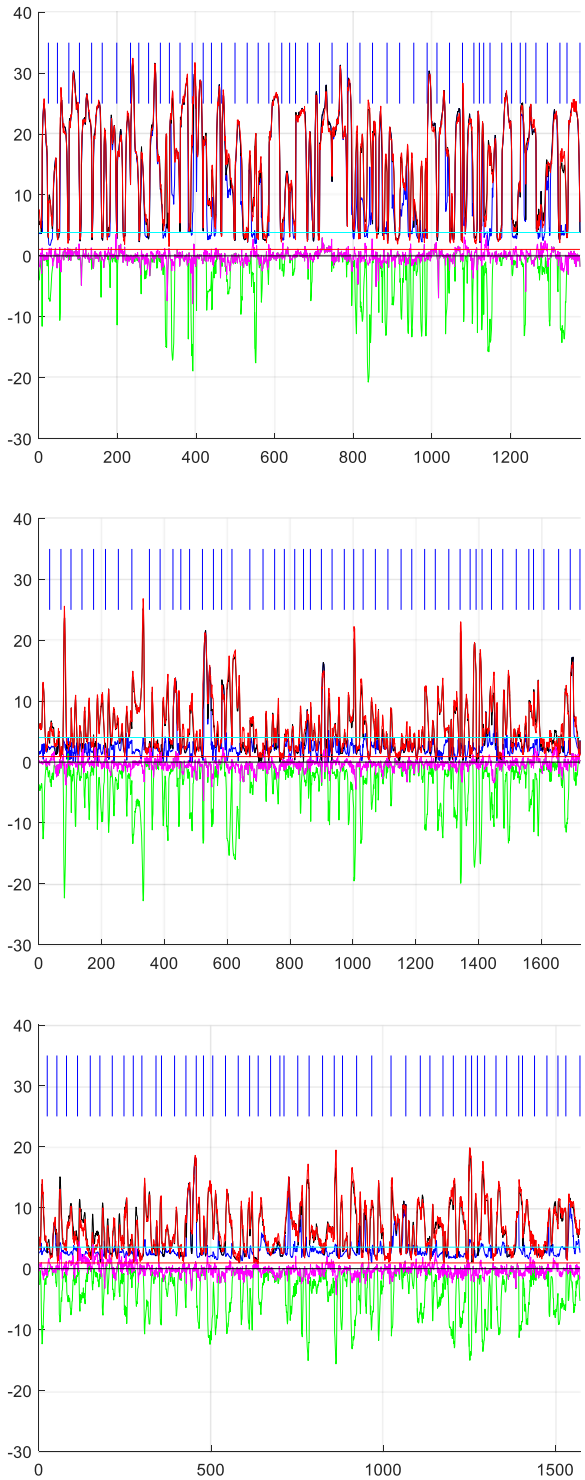

Fig. 15 *Red curve*: *B* parameter (expressed in pixels) determined by the ground truth value. *Blue curve*: *B* values by original CLNF landmarks. *Black curve*: *B* parameter by proposed method. *Green curve*: errors of original CLNF. *Magenta curve*: errors of Modified CLNF. The figure shows the result of three speakers from the top to the *bottom*. The *blue short lines* at the top are the segmentation for 50 words. The *cyan line* is the RMSE value

By large observations, we find that our detected lips ROI is as precise as the ROI determined by the blue mark on speaker's front in most of the cases. However, when the speaker's head rotates or shifts, the ROI is not accurate using the blue mark method, while our proposed method gives an accurate ROI centered of lips. In fact, the proposed method benefits from CLNF which is robust to the rotation or shift of the speaker's head.

DCT coefficients are calculated from lips ROI, and 10×10 coefficients in the low-frequency region are retained. The mean vector of these DCT coefficients is considered as a model for closed lips. By applying this model to all images, a threshold that distinguishes closed lips and open lips is obtained. For each new image, DCT coefficients of ROI are first calculated and then compared with the model. Images with distance less than a threshold are considered to be closed lips and thus will be skipped by the HD-CTM.

6.2 Parameter *A* correction based on periodical spline interpolation

For round lips, the two endpoints determined by CLNF are mistakenly placed from the acoustic point of view. In this case, the six points from CLNF (three upper points and three lower points) are assumed to be correctly determined (Fig. 13a). We propose to estimate the inner lips contour using the periodical spline interpolation based on these six points. They are firstly dilated in the vertical scale to form a square (see Fig. 13b) in order to obtain a regular repartition of these points in the polar coordinate. Cartesian coordinates of these points are then converted into polar coordinates. The center of the polar coordinates is situated in the middle of the two middle landmarks of CLNF inner lips. A spline interpolation is applied to the polar coordinates (Fig. 13c). In order to take into account the initial condition of the endpoint, the six points are periodized three times to prepare for a periodical spline interpolation (Fig. 13d). Finally, by returning to the original scale, a full contour interpolation can be obtained (Fig. 13e).

To apply this method, an automatic round lips detector is necessary to select round lips from the image sequence which contains all kinds of lips shapes. A similar method with closed lips filter in the above section is applied to detect round lips. Firstly, a round lips DCT template is trained from several round lips images. DCT coefficients are calculated from lips ROI. Secondly, this template scans through all the images, and the distance between the current image and template is computed. Lips are considered as round lips if its distance is less than a threshold which is determined experimentally. The performance of the automatic round lips detector is evaluated on 3184 lips images (10 repetitions of the first

Table 2 RMSE values for original CLNF model and for Modified CLNF (expressed in pixels and in cm)

RMSE	Speaker 1	Speaker 2	Speaker 3	Total
CLNF	3.84 (0.20 cm)	4.02 (0.21 cm)	3.53 (0.18 cm)	3.81 (0.20 cm)
Modified CLNF	1.06 (0.06 cm)	0.90 (0.05 cm)	0.94 (0.05 cm)	0.99 (0.05 cm)

speaker). Only 42 round lips are mistaken among 467 round lips images (about 9% error rate).

7 Results and discussions

To evaluate the performance of the proposed method, the A and B parameters estimated by the proposed methods are compared with the ground truth, and the statistic numerical results are given. Furthermore, the proposed methods are also compared with the baseline CLNF in Section 7.3.

7.1 Evaluation of the *B* parameter

The HD-CTM combined with the back-subtracting of VLLT efficiently corrects B errors of CLNF, which is visually shown in Fig. 14. The corrected B parameter is drawn by the black curve and the ground truth is drawn by the red curve in Fig. 15. Since they are very close to each other in most of the cases, we cannot distinguish them clearly. By contrast, the original CLNF B parameter drawn in blue has an evident difference with the ground truth, especially for the second and third speakers. The residual error between our estimated B parameter and the ground truth value is drawn magenta. One can see that errors are significantly reduced with respect to the original CLNF errors (green curves). RMSE of these errors is shown in Table 2. Total RMSE of the final errors is reduced to one pixel (0.05 cm), instead of four pixels (0.20 cm) when using the original CLNF. It outperforms the result in [22] (RMSE 0.1 cm).

7.2 Evaluation of the *A* parameter

The periodical spline interpolation method efficiently corrects A parameter errors of CLNF, which can be visually shown in Fig. 16. The evaluation of the A parameter is based on round lips images which are selected using the DCT filter (Section 6.2). A total 927 round lips images are selected (222 images for speaker 1, 396 images for speaker 2, and 309 images for speaker 2). In Fig. 17, the error between these methods and the ground truth is shown. We can see that the error is much less using the proposed method than the CLNF.

To further measure the error, we calculate the statistic error (Table 3). We observe there is huge bias in terms of mean value for CLNF error which is far greater than the standard deviation. From Fig. 17, therefore, we calculate the total error (E_{total}) to measure the precision of the A parameter, and it is calculated using the following formula.

$$E_{\text{total}} = \sqrt{v^2 + \sigma^2} \tag{12}$$

where v is the mean error (bias) and σ^2 is the variance of the error.

We can see the huge total error of CLNF is significantly reduced which is comparable to the state-of-the-art work [22, 23].

Concerning the error of the A parameter, the precision is not as demanding as the B parameter from the speech production point of view, and the estimation of the A parameter is less precise in practice. Meanwhile, comparing the error of the B parameter with the error of the A parameter, we can see the error of the B parameter is less than the A parameter, which is coherent with the result in [22].

To evaluate the joint performance of the A and B parameters, we explore the distribution of three visemes in the A and B parameter plane. In Fig. 18, three visemes are plotted for the first repetition of the speaker one. The distribution of each vowel is presented by a Gaussian ellipse. We recall that three visemes of original CLNF are mixed, especially for the third viseme (Fig. 6a). After the B parameter is corrected by the dynamic

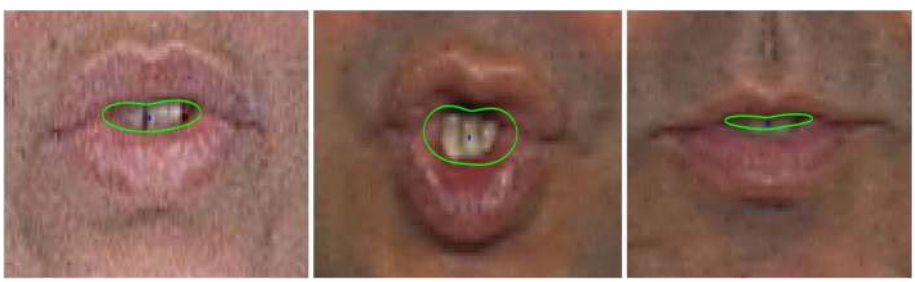

Fig. 16 The *green curve* is the whole inner lips contour for round open lips by the proposed periodical spline method

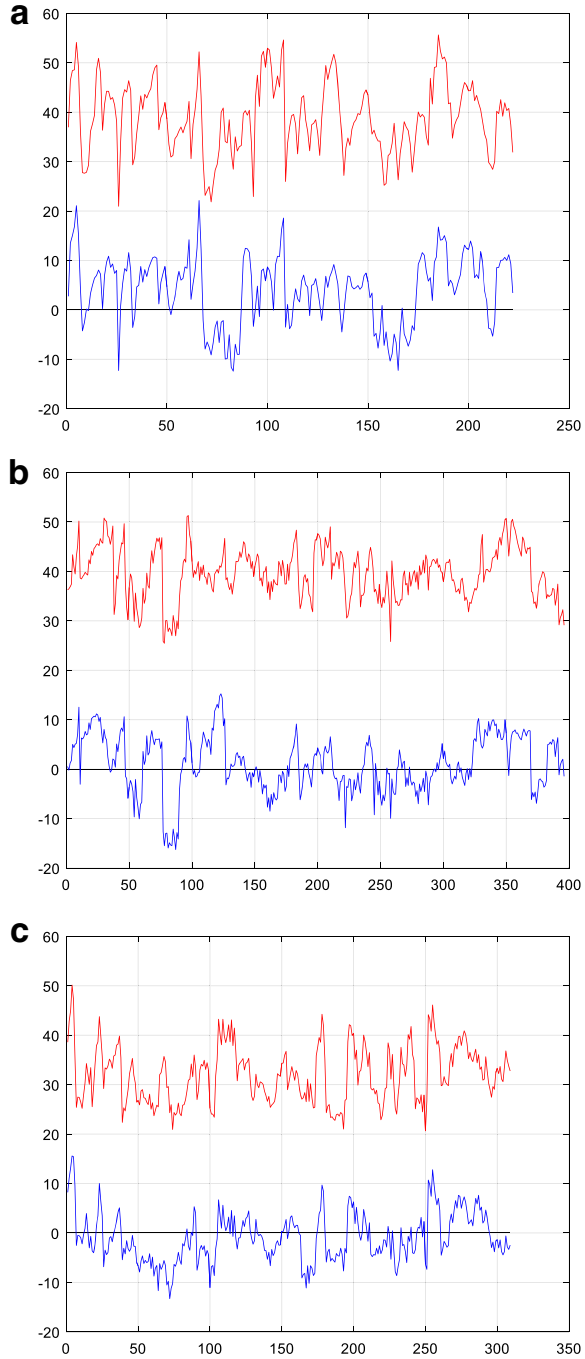

Fig. 17 *Y*-axis represents the pixel number. *Red curve*: the error of the *A* parameter (expressed in pixels) between the values obtained by the CLNF and the ground truth value. *Blue curve*: the error of the *A* parameter (expressed in pixels) between the values obtained by the periodical spline method and the ground truth value. (**a**–**c**) correspond to the result of three speakers

correlation template method, these visemes are well-distributed in the axis B (Fig. 18a). After the *A* parameter is corrected by the periodical spline interpolation, the third viseme is correctly distributed corresponding to the axis A (Fig. 18b). One can see that distribution of three visemes correspond correctly [18] to the ground truth distribution in Fig. 6b.

7.3 Application of the estimated lips parameters to CS phoneme recognition

In order to further evaluate the performance of the estimated inner lips parameters, French CS recognition based on 13 vowels is carried out using a HMM-GMM recognizer. We use the corpus of the first CS speaker with 10 repetitions. Eighty percent of the data (randomly chosen) are used for training and the remaining 20% used for testing (without overlap between the training and testing sets). A HMM-GMM decoder is built with a standard HMM configuration: context-independent, three-state, left-to-right, no-skip phoneme. It is trained with maximum likelihood estimation based on the EM algorithm. The labial features (inner lips *A* and *B* parameters) are modeled together with their first derivatives. At the decoding stage, the most likely image sequence of vowels is estimated by decoding the HMM-GMM state posterior probabilities using the Viterbi algorithm. 61.8% accuracy on average is obtained for the recognition of the 13 French vowels in CS. This result is comparable with the state of the art [12] in automatic CS recognition. It is another validation of our method for automatic tracking of inner lips parameters.

7.4 Discussions

- The periodical spline interpolation method is based on CLNF lips landmarks. When the real inner lips contour is inside the six inner lips landmarks of CLNF, we cannot expect the proposed method to give a satisfactory inner lips contour (about 5% of this case).
- In HD-CTM, there are several parameters that need to be optimized by training their data for each subject. Ongoing work is to reduce the subject-dependency of these parameters in this method.

8 Conclusions

This paper presents a new automatic approach to tracking the inner lips contour based on CLNF which is robust for facial landmark detection in the general case. However, the original CLNF presents mistakes in about 41.4% of cases for inner lips tracking. This paper aims at correcting CLNF errors by post-processing procedure. We propose two methods to correct the *B* parameter and the *A* parameter, respectively. In the case of the *B* parameter, an efficient method named HD-CTM based on the correlation with a hybrid dynamic template is investigated first to detect the outer lower lips position. Then, the inner lower lips position is determined by the

Table 3 Total error values for original CLNF model and the periodical spline interpolation method (expressed in pixels and cm)

RMSE	Speaker 1	Speaker 2	Speaker 3	Total
CLNF	39.64 (2.24 cm)	39.94 (2.26 cm)	32.55 (1.84 cm)	37.40 (2.12 cm)
The proposed method	8.03 (0.45 cm)	5.84 (0.33 cm)	5.07 (0.28 cm)	6.11 (0.35 cm)

back-subtracting of the VLLT. The evaluation of this method on about 4800 images of three speakers confirms its performance. In fact, RMSE is reduced from four pixels (0.2 cm) to one pixel (0.05 cm). For the *A* parameter, the periodical spline interpolation based on the dilated six CLNF inner lips points is used to estimate the *A* parameter for round lips. An automatic round lips detector based on the DCT coefficient of lips ROI is used to select the third viseme. This method is tested on 927 round lips images. The total error is reduced from 2.12 cm using CLNF to 0.35 cm with the proposed method. The remaining errors come from the mistaken six inner lips landmarks of the original CLNF. Our future work will aim at reducing the subject-dependency of the proposed method.

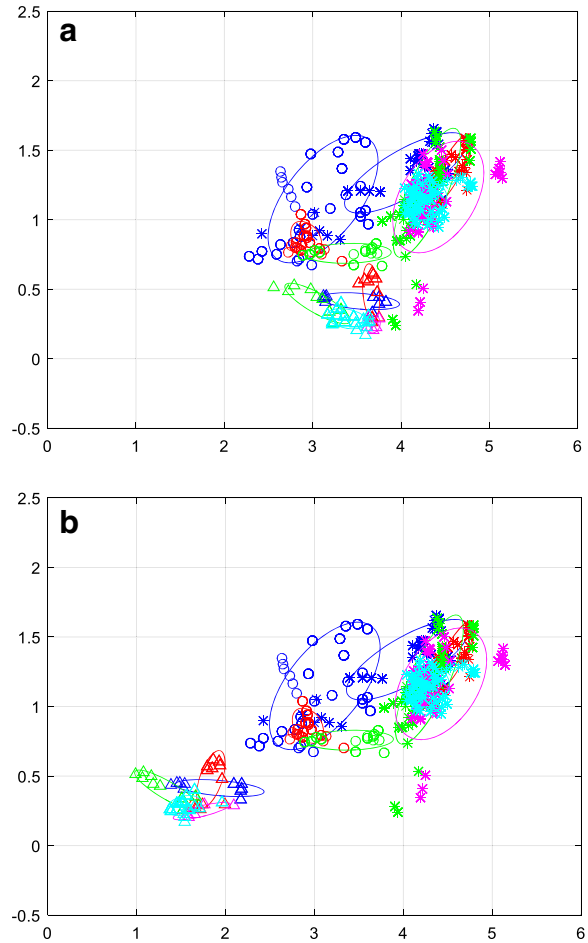

Fig. 18 The figures are plotted in the parameter *A-B* plane with all the vowels in one repetition for the first subject. **a** CLNF with corrected *B* parameter but non corrected *A* parameter for the round lips. **b** CLNF with both corrected *B* and *A* parameters for the third viseme with automatic detection of the third viseme; stars correspond to the first viseme, *circles* to the second viseme, and *triangles* to the third viseme. The color order is *blue*, *red*, *green*, *magenta*, and *cyan*. They are corresponding with the vowel order shown in Table 1

Acknowledgements
The authors would like to thank the volunteer speakers for their time spent on Cued Speech data recording and Christophe Savariaux (CNRS, GIPSA-lab, F-38000 Grenoble, France) for his help in the experimental setup.

Funding
We do not have any funding during this work. The publication fee is funded by the CNRS "PEPS-LGV" grant .

Authors' contributions
This work is realized in the framework of LL's PhD under the supervision of GF and DB. The main ideas of the research, experiment implementation, and development of the algorithm are performed by LL and GF. DB took part in the preparation of the Cued Speech data. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]University Grenoble-Alpes, GIPSA-lab, F-38040 Grenoble, France. [2]CNRS, GIPSA-lab, F-38040 Grenoble, France.

Received: 12 April 2017 Accepted: 24 November 2017
Published online: 19 December 2017

References

1. RO Cornett, Cued Speech. Am. Ann. Deaf **112**, 3–13 (1967)
2. V Attina, D Beautemps, MA Cathiard, M Odisio, A pilot study of temporal organization in Cued Speech production of French syllables: rules for a Cued Speech synthesizer. Speech Comm. **44**, 197–214 (2004)
3. T Cootes, An introduction to active shape models, model-based methods in analysis of biomedical images in image processing and analysis, ed. by R Baldock. (England: Oxford University Press, 2000), p. 223–248.
4. TF Cootes, GJ Edwards, GJ Taylor, Active appearance models. IEEE Trans. Pattern Anal. Mach. Intell. **23**, 681–685 (2001)
5. SK Bandyopadhyay, Lip contour detection techniques based on front view of face. J Global Res Comput Sci **2**(43–46) (2011)
6. S Stillitano, V Girondel, A Caplier, Lip contour segmentation and tracking compliant with lip-reading application constraints. Mach. Vis. Appl. **24**(1–18) (2013)
7. E Skodras, N Fakotakis, *An Unconstrained Method for Lip Detection in Color Images, in Proceeding of the International Conference Acoustic, Speech and Signal Processing* (2011), pp. 1013–1016
8. JM Zhang, LM Wang, DJ Niu, YZ Zhan, Research and implementation of a real time approach to lip detection in video sequence, in IEEE Conference on Machine Learning and Cybernetics, Xi'an, p. 2795–2799, 2003.
9. M Hlavac, Lips Landmark Detection Using CNN, in Studentská vědecká konference, 2016.

10. T Baltrusaitis, L.P Morency, P Robinson. Constrained local neural fields for robust facial landmark detection in the wild, in Proceeding of Computer Vision Workshops, Sydney, p. 354–361, 2013.
11. D Cristinacce, T Cootes, Feature detection and tracking with constrained local models, in Actes de British Machine Vision Conference, Edinburgh, p. 1–10, 2006.
12. P Heracleous, D Beautemps, N Aboutabit, Cued Speech automatic recognition in normal hearing and deaf subjects. Speech Comm. **52**, 504–512 (2010)
13. N Aboutabit, D Beautemps, O Mathieu, L Besacier, Feature adaptation of hearing-impaired lip shapes: the vowel case in the Cued Speech context, in Proceeding of the Interspeech, Brisbane, 2008.
14. L Liu, G Feng, D Beautemps, Automatic tracking of inner lips based on CLNF, in Proceeding of the International Conference Acoustic, Speech and Signal Processing, New Orleans, p. 5130–5134, 2017.
15. TF Cootes, CJ Taylor, DH Cooper, J Graham, Training models of shape from sets of examples, in Proceeding of the British Machine Vision Conference, Leeds, p. 9–18, 1992.
16. J Saragih, S Lucey, J Cohn, Deformable model fitting by regularized landmark mean-shift. Int. J. Comput. Vis. **91**, 200–215 (2011)
17. C Benoît, T Lallouache, T Mohamadi, C Abry, A set of French visemes for visual speech synthesis. Talking Machines: Theories, Models and Designs, ed. by G Bailly, C Benoît (Amsterdam: Elsevier Science Publishers B.V., 1992), p. 485–504.
18. N Aboutabit, Reconnaissance de la Langue Francaise Parlée Complétée (LPC): Décodage phonétique des gestes main-lèvres, INPG, Gipsa-lab, Université Grenoble Alpes in Grenoble, France, 2007.
19. C Bregler, S Omohundro, Surface learning with applications to lipreading, In Advances in Neural Information Processing Systems, vol 6, ed. by In J.D Cowan, G Tesauro, J Alspector (San Francisco, CA: Morgan Kaufmann Publishers, 1994), p. 43–50.
20. T Lallouache, Un poste Visage-Parole: Acquisition et traitement des contours labiaux, in Actes des Journées d'Etudes de la Parole, Montréal, 1990.
21. G Feng, Data smoothing by cubic spline filters. IEEE Trans. Signal Process. **46**, 2790–2796 (1998)
22. L Reveret, C Benoit, A new 3D lip model for analysis and synthesis of lips motion in speech production, in Proceeding of the ESCA workshop on Audio-visual speech processing, Australia, 1998.
23. L Liu, G Feng, D Beautemps, Inner lips parameter estimation based on adaptive ellipse model, in Proceeding of the International Conference Audio Visual Speech Processing, Stochholm, 2017.