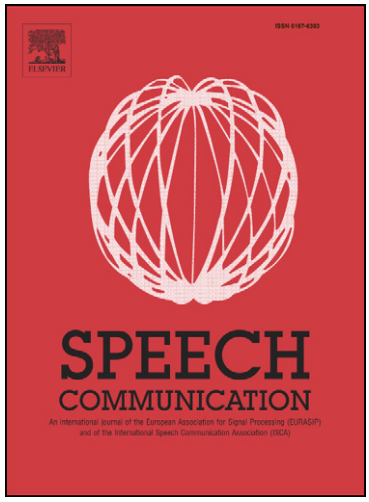
A Semi-Automatic Method for Extracting Vocal-Tract Movements from X-Ray Films

Julie Fontecave Jallon, Frédéric Berthommier



A Semi-Automatic Method for Extracting Vocal-Tract Movements from X-Ray Films

Julie Fontecave Jallon, Frédéric Berthommier

PII: S0167-6393(08)00099-X
DOI: 10.1016/j.specom.2008.06.005
Reference: SPECOM 1736

To appear in: *Speech Communication*

Received Date: 16 May 2008
Revised Date: 26 June 2008
Accepted Date: 27 June 2008

Please cite this article as: Jallon, J.F., Berthommier, F., A Semi-Automatic Method for Extracting Vocal-Tract Movements from X-Ray Films, *Speech Communication* (2008), doi: 10.1016/j.specom.2008.06.005

A Semi-Automatic Method for Extracting Vocal-Tract Movements from X-Ray Films

Julie Fontecave Jallon, Frédéric Berthommier

GIPSA-Lab, Department Speech and Cognition
Domaine Universitaire, Ensieg, BP 46, 38402 Saint Martin d'Hères, France

Julie.Fontecave@gipsa-lab.inpg.fr, Frederic.Berthommier@gipsa-lab.inpg.fr

Abstract

Despite the development of new imaging techniques, existing X-ray data remain an appropriate tool to study speech production phenomena. However, to exploit these images, the shapes of the vocal tract articulators must first be extracted. This task, usually manually realized, is long and laborious. This paper describes a semi-automatic technique for facilitating the extraction of vocal tract contours from complete sequences of large existing cineradiographic databases in the context of continuous speech production. The proposed method efficiently combines the human expertise required for marking a small number of key images and an automatic indexing of the video data to infer dynamic 2D data. Manually acquired geometrical data are associated to each image of the sequence via a similarity measure based on the low frequency Discrete Cosine Transform (DCT) components of the images. Moreover to reduce the reconstruction error and improve the geometrical contour estimation, we perform post-processing treatments, such as a neighborhood averaging and a temporal filtering. The method is applied independently for each articulator (tongue, velum, lips, and mandible). Then the acquired contours are combined to reconstruct the movements of the entire vocal tract. We carry out evaluations, including comparisons with manual markings and with another semi-automatic method.

Key words: Cineradiography, contour extraction, low frequency DCT components, vocal tract movements.

1. Introduction

The full sagittal view of vocal tract articulators during running speech, obtained by X-Rays, remains unsurpassed by modern imaging techniques, considering the great advantage allowed between temporal resolution and overall view. The amount of recorded data is sizeable but this is under-exploited due to the tedious hand tracing usually necessary for the analysis of such data. Obviously automatic extraction processes can circumvent this issue and allow the exploitation of these valuable data.

1.1. Relevance of the cineradiography and Databases

X-ray films are classically a reference technique to study speech production (Fant, 1960, Maeda, 1979, Mermelstein, 1973, Wood, 1979). Indeed, unlike imaging techniques such as

ultrasound (Akgul *et al.*, 1999), electropalatography (Hardcastle, 1972) or EMA (Perkell *et al.*, 1972), X-ray films provide a complete dynamic view of the entire vocal tract from the glottis to the lips. And although Magnetic Resonance Imaging (MRI) gives better resolved images (Badin *et al.*, 1998), since it reduces the problem of occlusion, like the superimposition of the mandible over the tongue, cineradiography allows the observation of movements with an optimal temporal resolution (about 50 im/s).

Due to ethical concerns, X-ray imaging technology is now rarely practiced. Since cineradiography has made the proof of its interest, it has become imperative to preserve and digitize the existing films and to make them available for the speech research community. In this framework, Munhall *et al.* (1995) have compiled the ATR "X-ray film database for Speech Research" from films contributed by Rochette (e.g., the Laval43 sequence, treated in this article) and by Perkell and Stevens (the MIT film, also considered in the following). This speech database is the largest one, with 25 different films offering 55 minutes and nearly 100000 images. Other digitized cineradiographic databases exist; e.g., a French database (Arnal *et al.*, 2000) has been elaborated by the Strasbourg Institute of Phonetics and the Grenoble Institute of Speech Communication and includes the Wioland sequence, which has been the support of our preliminary work (Fontecave & Berthommier, 2005).

1.2. Contour extraction

Thus digitized, those databases offer the possibility of a new look at old data. More knowledge about speech production processes might come out from the analysis of those sequences, provided that improvements in feature extraction methods are made. The exploitation of X-ray video sequences requires a preliminary extraction of articulators contours.

1.2.1. Manual vs. automatic extraction

In speech studies, geometrical data extraction from X-ray films is generally realized manually: configurations of the vocal tract are obtained image-by-image thanks to manual layouts (Badin *et al.*, 1995, Bothorel *et al.*, 1986, Maeda, 1979). Most often, contours are traced by hand from a projection of the picture onto a piece of paper in a dark room and then digitized by a scanner. This laborious hand treatment precludes the analysis of every frame for long sequences. For now, studies based on radiographic imaging usually concern very short sequences or limit their measurements to selected articulatory targets. Hence the access to the temporal dimension of 2D midsagittal tract changing shapes is hugely restricted, whereas it is one of the most relevant aspects of cineradiographic data.

In the context of exploiting large cineradiographic sequences yielding several thousands of images in each, the manual extraction task is too long to be considered and in the past, attempts of automatic methods have been proposed.

In 1994, Tiede and Bateson presented some ways to automatically process the X-ray images so as to facilitate the exploitation of the existing databases. A practical method for the extraction of tongue contours at the image level was proposed by Laprie and Berger (1996). Since the single "Snake" method (Kass *et al.*, 1987) introduced in computer vision to extract contours is unable to achieve the task, the authors make "Snakes" cooperate with an optical flow method where contours are not sufficiently isolated from spurious contours. But no evaluation was published.

Later, Thimm and Luettin (1999) achieved the automatic processing of a complete sequence of the ATR database (Laval43). Their method, detailed further for sake of comparison, is based on a contour approach and uses a representative and limited set of state images. But the quality of the contour estimations with such automatic methods is weak in comparison with the manual extraction. And it is readily noticeable that the automatic tongue extraction remains particularly difficult when obscured by superimposed structures like the teeth.

Thereby, in order to improve the result, we propose to reintroduce a part of human expertise. Our approach uses an existing algorithm (Berthommier, 2004), called "retro-marking", which consists in the temporal inference of geometrical marking from video data. The proposed method is semi-automatic and is made up of a human manual task followed by automatic video treatments.

1.2.2. Aim of the retro-marking algorithm

This algorithm, appropriate for off-line video processing, has been proposed for the extraction of geometrical features without using markers (make-up, balls…). This builds a transformation function of implicit parameters, extracted from the video signal, into explicit and controlled geometrical parameters. It associates a manual marking of geometrical features on a limited number of key images, and an automatic estimation for every frame of the sequence. The geometrical features can be contours or anchor points of the contours. The link between geometrical and video features depends on the low frequency structures contained in the images. The algorithm is schematized in Fig. 1.

In this first application (Berthommier, 2004), the algorithm was used for the extraction of lips geometry starting from a well-framed video database (Heckmann *et al.*, 2000) recorded without the traditional blue chroma key method. This method is usually practiced to prevent from marking the lips contours, by using a blue coloration of the lips (Guiard-Marigny *et al.*, 1996).

The geometrical features, for the retro-marking algorithm, were 8 points describing the mouth opening parameters. The video features were the 24*12*3 first components of the Discrete Cosine Transform (DCT). The geometrical extraction for each frame of the sequence was inferred thanks to an indexing based on a transformation function relating the lips parameters and the video data.

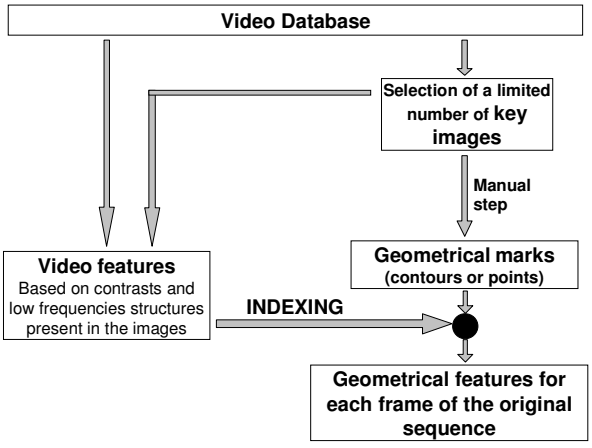


Figure 1: General principle of the retro-marking algorithm.

1.3. Objectives

The method proposed in this article adapts the retro-marking algorithm to long sequences of X-ray images. Indeed, the greater part of cineradiographic data is recorded without any marker, and usable anatomic markers, such as the teeth, are not sufficient. Moreover, due to the DCT phase shift-sensibility (Keckmann *et al.*, 2003), the precision of the video centering is crucial for the use of the retro-marking algorithm. Fortunately this stability of the video recording is a condition generally fulfilled for cineradiographic sequences: subjects are required not to move.

According to the retro-marking principle, a user interaction step is combined with an automatic reconstruction; this latter is based on the high redundancy of speech movements.

The method is applied sequence by sequence. The first film considered here is the Laval43 sequence from the ATR database. It has been extracted from the DVD provided by ATR. Laval43 is a video film (sentences read by a male native speaker of Canadian French) originally recorded in 1974 by Dr. Rochette at Laval University in Quebec, at the rate of 50 im/sec. A frame rate conversion was realized during digitalization; the digitized video film on the DVD is available at 29.97 im/s in NTSC format. This is converted in bitmap (BMP) sequences; the extracted images are 24 bit BMP images (720*480 pixels) of the vocal tract. For the Laval43 sequence, only 3973 images are usable.

The Laval43 sequence is well centered: no head motion correction is needed.

The retromarking method is applied articulator by articulator. First (in part 2), the principle is described and it is detailed for the tongue. Then part 3 extends the method to other articulators so as to reconstruct the complete vocal tract. Furthermore a quantitative evaluation is realized in part IV to adjust the parameters and to measure the reconstruction error of the technique. At last, comparative studies are carried out. First, we apply the method to the MIT sequence, another sequence of the ATR database, treated by Perkell in 1969, and we compare the error rate between manual and semi-automatic evaluated contours. Then we compare on the Laval43 sequence our proposed method with the semi-automatic method of Thimm and Luettin (1999).

2. Principle of the method illustrated with the tongue contour extraction

The method has 3 main steps: (1) the manual process applied for a small number of key images and defining the geometrical features, (2) an automatic indexing step of the full database according to these key images, based on the retro-marking algorithm and which allows the association of the geometrical marking for each frame and (3) some post-processing treatments in order to restore the continuity of the movements. We attempt to infer geometrical marking thanks to the retro-marking and we ground the inference process on associative properties.

2.1. Manual step applied on full-size images

The first step is a manual marking phase. We do not adopt a semi-polar grid, usually used for the vocal tract tracing (Bothorel *et al.*, 1986, Heinz & Stevens, 1964). Our method is based on an articulator by articulator decomposition and a grid is specified for each element.

The purpose of this manual step is to describe the articulator shape with a small number of points. But we actually want to limit the set of points to a set of degrees of freedom (*dof*). For most of the points, we aim at defining each point with one *dof*, i.e., one coordinate (X or Y). The choice of the fixed coordinate for a 1-*dof*-point is made such that for each frame of the sequence,

the point can always be marked on the articulator contour. A regular spacing between points is chosen to allow a realistic layout of the contour when connecting the points. The manual task for these points consists in posing a mark on the contour at the fixed coordinate and thus in determining the other coordinate, i.e., the degree of freedom.

Since it is not always possible to find a fixed coordinate intersecting the contour (especially at its boundaries) on every frame, some points are let free, outside the grid, with 2 *dof*. These free points allow taking into account more movement variability.

For the tongue contour in Laval43 (Fig. 2), the points 3 to 13 are defined thanks to horizontal and vertical lines of a grid. And two points are 2-*dof*-points, without any fixed coordinate; these are points 1 and 2.

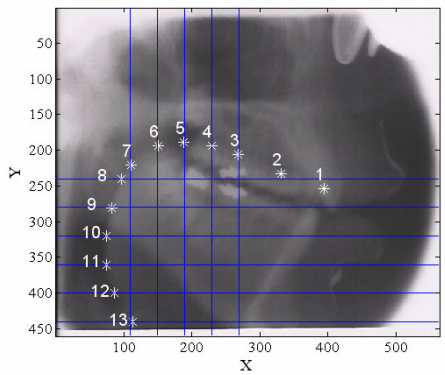


Figure 2: Degrees of freedom for the tongue contour in Laval43 sequence. The tongue contour, excluding the tongue tip, is marked using *-markers on the grid lines.

Thus, the tongue is defined by 13 points or 15 degrees of freedom: 11 *dof* describe the body and the root (coordinates Y for points 3 to 7 and X for points 8 to 13) and 4 *dof* describe the tongue tip.

The manual step concerns only a limited number of images, called key images. We randomly choose n key images $(K_i)_{i=1:n}$ among the N images (S_t) of the whole sequence. For the tongue in Laval43, the manual task is applied on 200 key images. The choice and the number of key images are motivated further in this article (§ 4.2.1. and 4.2.2.).

The manual tracing of the points specified above is carried out with great care by one of the authors for the n key images. Some rules are chosen for this marking. For example, when the tongue is bent, two contours appear for the tongue back profile (due to x-ray projection); we systematically choose to mark the most inner contour, i.e. the most contracted part of the tongue. The tongue tip (point 1) is marked as the most forward point of the contour; this point is visible or inferred thanks to the dynamic possibilities of the interface described below.

The manual marking is realized on images successively displayed on the computer screen. For each point, the mark is fixed by hand.

A sizeable effort has been made to design an ergonomic Matlab interface, as often described in digitized X-ray processing works (Roy, 2003, Tiede & Bateson, 1994). Besides the display of the static grid resulting from the *dof* choice, this interface allows dynamic observations thanks to a slider. It allows to take into account motion and thus to show the articulator in its context. The dynamic observation facilitates the visualization of contours, which can be either occluded on static images, especially owing to the superimposition of the jaw and the tongue, or badly

contrasted; in particular, the tip is often not visible.

This dynamic aspect is a fundamental point of the method, since it enables the extraction of a tongue contour, which is, in many cases, barely visible on the static key image. Beyond it is a main difference with extraction techniques such as Thimm & Luettin (1999) in which those occlusion events are partly responsible of the faintness of the method.

A linear interpolation of the contour is then obtained by connecting the points (13 points in the tongue case). After this manual marking step, we get the XY-coordinates of 13 points for the *n* key images $(K_i)_{i=1:n}$, corresponding to geometrical configurations $(G_i)_{i=1:n}$.

2.2. Automatic step applied on "framed" images

The main retro-marking step is the automatic indexing of the video sequence according to the key images. It allows an association between the geometrical features (the *dof* marked on the key images) and the video features.

The video features are the lowest frequency Discrete Cosine Transform (DCT) components of each image. As other linear image transforms, the DCT transforms the image pixel values into a lower dimensional space. It removes redundant information and codes only salient visual features. The DCT is a simple and computationnally efficient image transform (Potamianos *et al.*, 1998). Applied on cineradiographic images, the DCT has a strong energy compaction property (Rao & Yip, 1990): most of the signal information tends to be concentrated in a few low-frequency components. The DCT is similar to a Discrete Fourier Transform, but using only real numbers. It enables to follow the phase variations of the low-frequency component, which are related to movement.

The DCT components are calculated on "framed" images; i.e., the original images are resized (decimated with Adobe, in order to reduce the computation time), centered and cut out so as to focus only on the considered articulator, the tongue here. One framed image can be observed in Fig. 3a (image of 105*95 pixels). For each image, the DCT components are calculated on the averaged 8 bit RGB components of the image.

(a) (b)

Figure 3: (a) Framed image (105*95 pixels) considered for calculating the DCT components used for the automatic indexing.
(b) Image reconstructed by Inverse Discrete Cosine Transform from the 575 (24*24-1) lowest DCT components calculated on the (a) image.

The lowest frequency components are sufficient to follow the tongue movements, the highest frequencies can be considered as non-useful (redundant or noisy) information. We only keep 575 components: the 24*24 components in the left top corner of the DCT matrix, except the first one

related to the mean value. The number of DCT components to retain is motivated in part 4.2.3. In Fig. 3b, the image is obtained thanks to an Inverse Discrete Cosine Transform from these 575 lowest DCT components only (the matrix is completed with zeros). On this "low frequency" image, the tongue is well visible (here in a backward position).

The automatic indexing consists in a quantization of the video sequence according to the key images. For each image of the sequence, we look for the most similar key image, by using a similarity measure defined as the Euclidian distance between the lowest frequency DCT components of the images. For each image, the similarity is measured between the image and the *n* key images. The key image associated to the smallest distance allows defining for each image S_t of the sequence an index *j* corresponding to the number of this key image.

$$j = index_K(S_t) = \underset{i}{\operatorname{argmin}} \sqrt{\sum_{p=2}^{24*24} (DCT_p(S_t) - DCT_p(K_i))^2} \qquad (1)$$

Thus each frame of the sequence is assigned by the index of the nearest key image.

The second step of the retro-marking technique consists in a simple automatic geometrical marking of the original images *(S_t)*. This uses, thanks to the indexing, the geometrical information defined for the key images only.

To each index *j* correspond one key image K_j and one geometrical configuration G_j, i.e., in the case of the tongue, the 13 points manually marked to define the articulator position. This configuration is associated to every frame of the sequence assigned with the index *j*. A raw geometrical marking of the original sequence is thus realized.

Here, the geometrical information is restored thanks to video information; and at this stage, the articulator movements are only partly reconstructed. The superimposition of the tongue contour in the original video sequence allows observing significant jumps. This first estimation is presumably affected by multiple sources of errors (quantization effects, indexing errors). We aim at reducing significantly this baseline reconstruction error and enhancing the movement reconstruction by restoring the temporal continuity.

2.3. Post-processing treatments: Reconstruction of the geometrical information across time

Post treatments are proposed in this section to reduce the error. For sake of illustration and in order to observe the effect of these operators, we introduce some appropriate representations, i.e., Principal Component Analysis. Those PCA are calculated on video features (low frequency DCT components) and on geometrical features (*dof*) of the key images. For illustration, we keep the 2 first PCA components of each space and represent the data in a video 2D-plane (called video PCA plane) and in a geometrical 2D-plane (or geometrical PCA plane).

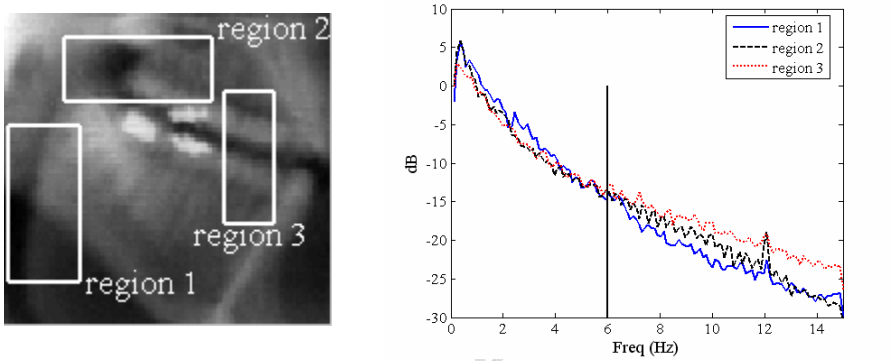
The motion reconstruction enhancement consists on one hand in reducing the quantization effects by temporal filtering of the geometrical features and on the other hand in compensating the irregularities of the relation between the 2 representations, by neighborhood averaging.

2.3.1. *Temporal filtering of the geometrical features*

To set up the suitable temporal filtering, a spectral analysis is carried out. For temporal images

sequences, the Power Spectral Density (PSD) on pixels is considered. The PSD is calculated (as in De Paula *et al.*, 2006) on selected regions of the framed images (Fig. 4a): the PSD is computed for each pixel of the 3 regions, and then we average for each region the PSD of the considered pixels. The regions are chosen to globally fit with the movements of the articulator (here the tongue). A small change of size or position of one region does not significantly affect the calculated PSD, since the variations related to the considered articulator are dominant in this region, and considering this region is fixed once for the full sequence.

We assume that the Power Spectral Density, along the temporal dimension, at the pixel level is significant of the tongue motion bandwidth. We observe (Fig. 4b) a low-pass distribution and we fix the video components cut-off frequency at 6 Hz. Up to this frequency, the curves are superimposed. Higher frequencies are considered as noise and therefore not taken into account.



(a) (b)

Figure 4: (a) Selected regions on framed image for spectral analysis.
(b) Mean PSD on pixels pooled from the 3 selected regions for a sequence of 1000 images and definition of a 6Hz cut-off frequency.

We also calculate the PSD on the 15 degrees of freedom of the tongue, i.e. we compute the PSD for each *dof* along the raw-indexed image sequence (generated by simple indexing) and then we average these 15 PSD. We check (Fig. 5) that this PSD provides nearly the same bandwidth as the one calculated on video data, i.e., attenuation of about 20 dB at 6 Hz. The 6 Hz cut-off frequency F_c is thus adopted for the geometrical data: a low-pass temporal filtering is applied on the sequence of geometrical features (Fig. 5). We choose a 0-phase filter of 8^{th} order in order to have zero phase distortion and to eliminate the distribution tail.

Notice that this 6 Hz frequency is linked to the video rate of the digitized images, i.e., 29.97 im/sec. In relation to the 50 im/sec rate of the original video sequence, the effective cut-off frequency is evaluated at 10Hz.

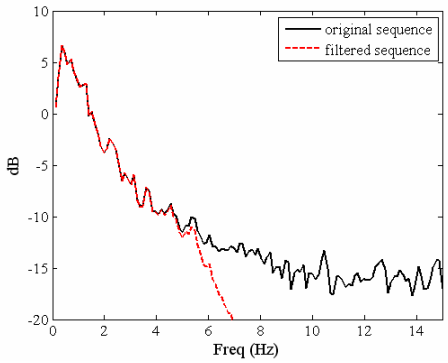
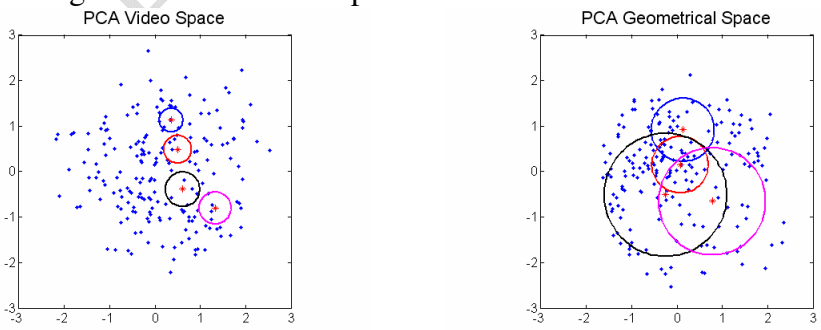


Figure 5: Mean PSD on the tongue 15 geometrical degrees of freedom for a 1000 images sequence with and without filtering (temporal low pass filter with cut-off frequency at 6 Hz).

2.3.2. *Neighborhood averaging*

We consider the neighborhood relationship of both video and geometrical representations. For sake of illustration, we observe the projection of video and geometrical data in the planes resulting from the 2 first PCA components. We check the distance between each point and its neighborhood in the 2 spaces.

For one point, corresponding to one key image, in the video PCA plane (PCA on the DCT components of the key images), we take into account its neighborhood, i.e., its k nearest neighbors. For a few examples and with e.g., k=10, we represent (Fig. 6a) the neighborhood of a point, with a circle centered on the considered point and whose radius is equal to the mean distance between this point and its 10 neighbors. An equivalent circle is drawn in the geometrical PCA plane (Fig. 6b). We observe an increase of the circles' radii in the geometrical PCA plane vs. the video PCA plane and a greater overlap of the circles in the geometrical PCA plane. By extrapolation of these observations in the PCA planes to the complete data spaces, we conclude that neighbors close in the video space (DCT components) are less close in the geometrical space (*dof*). The irregularity of the relation between the two representations is underlined through the disparities of the neighborhood relationship.

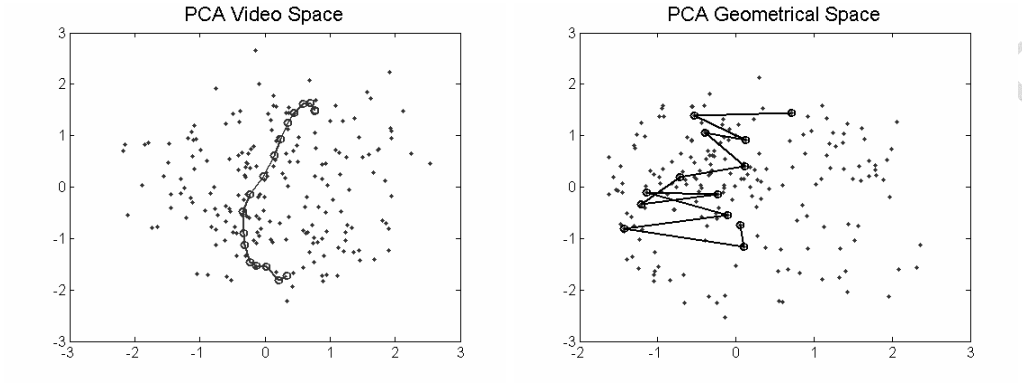


(a) (b)

Figure 6: By observing neighborhood circles in the video (a) and geometrical (b) PCA planes, we highlight the discontinuities of the relation between the 2 representations.

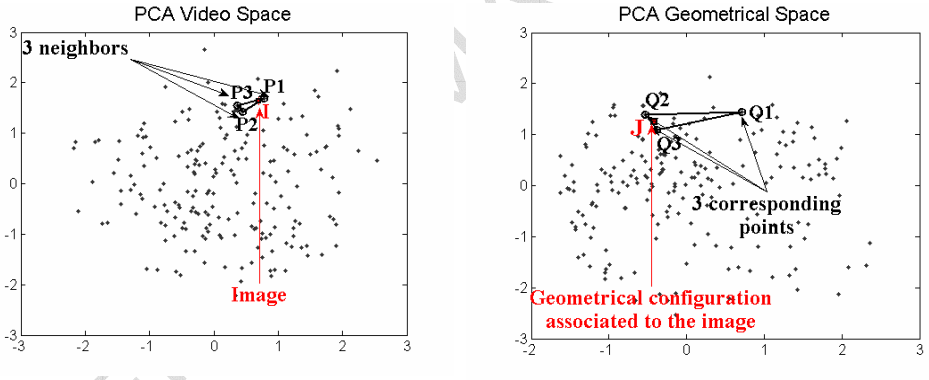
Additional evidence of these irregularities between the two spaces is observed in the temporal domain, by checking the trajectories. For sake of representation, these are projected again in PCA

planes. In the video PCA plane, a trajectory is generated by connecting the projected points, associated to the images of a short video sequence (Fig. 7a). The trajectory in the geometrical PCA plane is generated via the indexing by connecting the points associated to the successive key images (Fig. 7b). Two consecutive images close in the video space are not close in the geometrical space. The trajectory in the geometrical plane shows severe discontinuities we attenuate by averaging the geometrical configurations of neighbors taken in the video space (Fig. 8).



(a) (b)

Figure 7: Trajectories projected (a) in the video PCA plane and (b) in the geometrical PCA plane via the indexing.



(a) (b)

Figure 8: Schema of the neighborhood averaging in the PCA planes with 3 neighbors. Starting from one image I and its 3 neighbors, we get 3 geometrical configurations, which are averaged to get the configuration J associated to I.

The principle of neighborhood averaging is described here with 3 neighbors. For each image S_t, we find, among the key images, 3 closest neighbors K_{i1}, K_{i2}, and K_{i3} (closest in term of video features thanks to the similarity measure applied with 575 DCT components). The 3 vectors of geometrical configuration GK_{i1}, GK_{i2}, and GK_{i3} respectively associated to the key images K_{i1}, K_{i2}, and K_{i3} are averaged to calculate $\hat{GK}_t$. This is illustrated in PCA planes: in Fig. 8a, the image S_t and its 3 neighbors are projected in the video PCA plane (points I, P_1, P_2, and P_3) and in Fig. 8b, the 3 corresponding contours are represented in the geometrical PCA plane (points Q_1, Q_2, and Q_3). The projection on the geometrical plane of the new averaged configuration is the point J

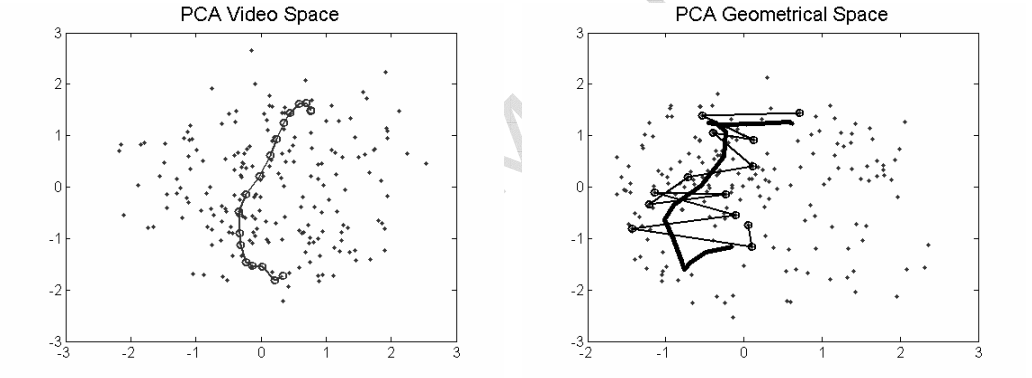
(Fig. 8b). We take into account a supplementary weighting, which is the inverse of the Euclidian distance calculated on the DCT components between S_t and K_{i1}, K_{i2} and K_{i3}.

This principle is applied with k neighbors. The formula for the averaged geometrical configuration is

$$\hat{GK}_t = \frac{\sum_{j=1}^{k} \frac{GK_{ij}}{d(S_t, K_{ij})}}{\sum_{j=1}^{k} \frac{1}{d(S_t, K_{ij})}} \qquad (2)$$

We show, by projecting the new geometrical trajectory on the geometrical PCA plane (Fig. 9b), that the irregularities are compensated, the trajectory is much smoother. As for now, we distinguish the multi-indexing or k-neighbor-indexing from the simple indexing, defined previously, and also named in the following the 1-neighbor-indexing.

Those two post-processing treatments, the temporal filtering and the neighborhood averaging, are mathematically independent, they are complementary and they can be applied successively. Thus, the multi-indexing can be followed by the temporal smoothing of the resultant series of geometrical configurations as well as for the 1-neighbor-indexing.



(a) (b)

Figure 9: Trajectories projected (a) in the video PCA plane and (b) in the geometrical PCA plane via the indexing (in thin line) and via the 3-neighbor-indexing (in thick line).

3. Reconstruction of the complete vocal tract

In this section, the method successfully developed on tongue movements is adapted for other articulators of the vocal tract. Considering the speech production process, the movements of articulators are more or less coordinated. For example, (1) the upper and lower lips are entirely coordinated, (2) the tongue tip movements depend on the tongue body movement, but not entirely, and the inverse is not true, (3) the lips and the tongue movements are coordinated, (4) the velum is independent. Nevertheless, we consider each articulator independent and the whole process is applied to each element separately with appropriate parameters.

A framing of the various articulators is then considered so as to get independent analyses of the geometrical configurations; each target articulator is dominant in its own frame, in terms of

signal. The benefit of this separate framing is to limit the number of key images to be marked in case of uncoordinated articulators, e.g., the tongue and the velum. A separate processing of each articulator minimizes interferences and avoids useless combinatorial analysis, in opposition to a joint analysis in which the codebook size should be increased. For example, instead of 50 manual sample tracings for the velum and 100 for the tongue for separate processing, one would need 50x100=5000 tracings for joint processing. In case of coordinated articulators, the relation between geometrical configurations is often complex, such as between the lips and the tongue. And the tongue movement signal would dominate the lips one. The case between the tongue body and the tongue tip is specific. We realize a global analysis of the tongue (including the root, the body and the tip), assuming that the tip does not significantly interfere on the tongue body. Then we correct this estimation with a separate analysis of the tongue tip, as explained below.

Thus, considering each articulator independent, the original images are first framed and cut out so as to only include each articulator for the whole sequence. We try to have the most restricted image as possible to avoid interferences. Then the parameters of the method, e.g. the number of key images, the points and degrees of freedom, the number of DCT components used for indexing are determined for each element independently.

Post-processing treatments are applied using the parameters defined for the tongue (4-neighbor-indexing and temporal filtering at F_c=6 Hz). The choice of this cut-off frequency is discussed again in §4.2.5.

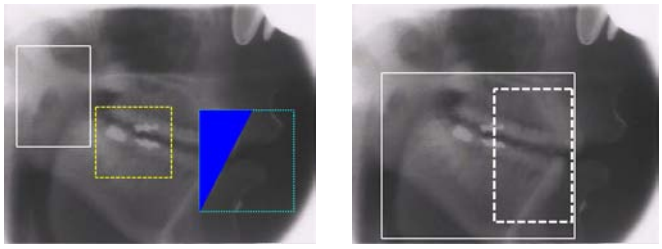
3.1. Independent treatment for each articulator

A separate estimation of all visible articulators is carried out: the tongue (as described above), the tongue tip, the lips, the velum, the mandible. Unfortunately the glottis is not visible. In this part, we describe the parameters applied for each individual element and some specificities of treatment. Some validations of the parameters values choice are detailed in part 4. Table 1 summarizes these values for each articulator. The two first columns specify the number of points and *dof*; these can be recovered from Fig. 11 for some articulators. The frames defining the regions of interest can be observed in Fig. 10.

Recall that the original image size is 720*480 and note that images have been previously decimated for the tongue and the tongue tip (noted *(dec)* in Table 1), but not for other articulators. The similarity measure for the indexing uses 575 low frequency DCT components for each articulator (except the mandible). This number choice for most articulators is motivated in §4.2.3.

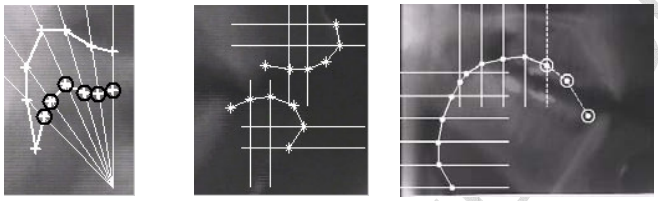
Articulator	**Parameters**				
	Points	**Degrees of freedom**	**Key images**		**DCT components**
			n	**Frame size**	
Tongue	13	15	200	105*95 *(dec)*	24*24 - 1
Tongue tip	3	5	200	48*75 *(dec)*	24*24 - 1
Velum	13	14	100	142*186	24*24 - 1
Upper lip	6	8	200	182*186	24*24 - 1
Lower lip	6	8			
Mandible	4	4	60	131*131	12*12 - 1

Table 1: Retro-marking parameters used for geometrical extraction of various articulators in Laval43 sequence



(a) (b)

Figure 10: Specific frames defined for each articulator in Laval43
(a) From left to right, each frame focuses on the velum, the mandible and the lips.
(b) For the tongue, the thin frame (also visible in Fig. 3a) is defined for decimated images (105*95 pixels). The position of the tip is estimated locally starting from a specific frame (the dashed one), smaller and also defined on decimated images.



(a) (b) (c)

Figure 11: Degrees of freedom defined for various articulators in Laval43
(a) The manual marks of the velum are pointed according to a polar grid. The black circles are related to the lower part of the velum.
(b) Lines allow the marking of 8 *dof* out of 16 on the lips.
(c) The 5 *dof*, corresponding to the 3 encircled points on the tongue, define the tip and are used for its specific estimation.

In the Laval43 sequence, the tongue tip is well visible and we observe that its movements are fast and sometimes relatively independent of the tongue body, due to vowel-consonant co-articulation effects. To improve the tracking of the tip, we complete the tongue estimation with a specific extraction of the tip movements. This consists in a double marking which associates the overall estimation of the 15 tongue *dof* (described above in part 2), and an independent estimation of the tip including 5 *dof* only. These degrees of freedom correspond to the 3 encircled points in Fig. 11c. The frame (the thick dashed one in Fig. 10b) focuses on the tip and delineates a reduced region at the front of the vocal tract. This frame is included within the tongue global frame, in order to take into account the dependency between the tongue body and the tongue tip movements, addressed in this section introduction. The same 200 marked key images are used for the tongue tip estimation. The fusion of the two estimates is carried out by substitution in the global estimation of the 5 *dof* related to the tip: we combine the 10 backward points globally estimated and the 3 forward points specifically estimated.

The velum, which is traditionally difficult to observe, is well visible on radiographic films. The framed image for the velum is presented in Fig. 10a. The marking of 13 points with 14 degrees of freedom is realized using a polar grid (Fig. 11a). The manual marking is, for the velum, a quite easy task. Moreover its movements' variability is reduced compared to the tongue and 100 key images are well enough to allow a good reconstruction of the velum shapes.

The lips' marking is defined similarly for the upper and lower lips (Fig. 11b). Horizontal and

vertical degrees of freedom allow the tracking of the available information, i.e., the opening and the protrusion of the lips. The frame for the DCT indexing (Fig. 10a) is located on the lips and on the front teeth but a mask is added to remove the teeth influence (some pixels are set to 0). The position of this mask is fixed for the whole sequence to hide the front teeth. The low contrast complicates the marking of the lips.

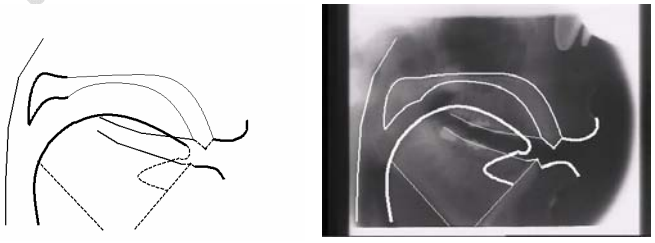
The front teeth are marked using the same frame without the black mask. At last, the mandible extraction is easily realized thanks to a few points on the molars and a frame focused on these well-contrasted teeth (Fig. 10a).

3.2. Complete vocal tract

In the purpose of recovering the geometry of the whole vocal tract, we must restore the continuity between the various elements. However recall that the glottis is not visible in the Laval43 sequence and in many other sequences. The recovery of the vocal tract has three main steps. First we complete the contours estimation, especially with the rigid parts that also need to be marked. Then a spline smoothing is considered for the various articulators. Finally a "connecting" step is required and consists in defining the junctions between elements.

Since the palate is fixed, its marking is done once and for all so as to fit most of the shapes observed on the whole sequence. By considering the lower jaw rigid, the mandible marking is completed starting from the estimated points on the well-visible molars. The pharynx is actually not entirely fixed: the retro-marking method is applied with 5 points (1 *dof* each, Y fixed) and a specific frame.

At this point, each articulator of the vocal tract has been marked separately and is defined by a few points or *dof*, for each frame of the sequence. For each articulator, by linearly connecting the considered points, we get a first representation that is quite irregular. To improve the geometrical representation of each articulator contour, we can perform a spline interpolation, frame by frame. This requires the use of polynomial curves, whose number and degree depend on the articulator. For example and without going into detail, in the Laval43 sequence, for the tongue, two 3rd order polynomials are combined to approach the whole contour. These spline interpolations do not significantly change the error estimation at the point level.



(a) (b)

Figure 12: (a) Combination of independently estimated contours (thick lines), fixed parts (thin lines) and reconstructed segments (dashed lines).
(b) Complete vocal tract contour for one image of the Laval43 sequence.

To perform the outline of our complete vocal tract contours, a final step consists in defining the junctions between articulators and we make our choices to achieve the reconstruction; e.g., the tongue tip is represented by a half circle. Other reconstructions have been proposed in the

literature, such as the CASY representation based on splines (Rubin *et al.*, 1996). Finally, the reconstructed junctions, the estimated smoothed contours and the fixed parts are combined (Fig. 12a) to get a full vocal tract shape for each image. One image is visible in Fig. 12b.

4. Evaluation of the marking error

At first, the results are visualized and qualitatively evaluated by displaying the superimposition of the geometrical vocal tract configurations in the original video sequence.

In this part, we evaluate quantitatively the geometrical marking error due to the automatic processing.

4.1. Geometrical error measurement

The error evaluation is realized articulator by articulator and it is based on a comparison on test images between our semi-automatic estimated marking and the manual marking, considered as the reference.

Those evaluations use a Jackknife technique. It consists in forming new sets of n key images (K_i) by omitting, in turn, a little proportion of n_2 images (T_j) of the original set of key images (all images have been marked by the same expert). On these n_2 omitted images (considered as test images), we quantify the deviation between the manual reference marks (GT_j) and the marks $(\hat{G}K_j)$ estimated from the method applied with the n key images. We first consider the reconstruction RMS (root mean square) error *dof* by *dof* on the test frames; this is noted $Edof_1(x)$, where x represents the considered *dof*. The final error *Edof* for one articulator is then the mean value of the $Edof_1$ error on the p degrees of freedom of this articulator and on the test frames. *Edof* is the geometrical error temporally integrated; it is expressed in pixels/*dof* of the full-size images (720*480).

$$Edof_1(x) = \sqrt{\frac{1}{n_2}\sum_{T_j}(\hat{G}K_j(x) - GT_j(x))^2} \qquad (3)$$

$$Edof = \frac{1}{p}\sum_{x=1}^{p} Edof_1(x) \qquad (4)$$

4.2. Parameters adjustment

Thanks to this error measurement, we can tune the main parameters of the method. Each value of *Edof* presented in the following is obtained by Jackknife technique over 10 simulations, each applied with different sets of key and test images among the 200 marked images. *Edof* is actually the mean value of *Edof* over these simulations.

4.2.1. Choice of key images

The strategy used to retain the key images for the manual marking task is a random choice among the database. This strategy does not need any a priori knowledge about the sequence. The treatment of the video information is carried out aside from the audio aspect. In other words, the phonetic information is not taken into account for the key images choice. This strategy is based on the temporal redundancy of the sequence; it aims at well representing the original distribution of the video data: the key images density is higher where the data density is important. Consequently the extreme positions, which are seldom in the sequence, are worse represented.

Therefore another strategy opposes this random sampling; this second strategy is based on the speech signal knowledge and favouring some speech segments, especially extreme articulatory gestures.

Considering the global marking error evaluated for the whole sequence, the choice strategy, favouring the extreme positions, will increase the error. Indeed, although this choice allows reducing the error for the extreme positions, it increases the small reconstruction errors in the dense part of the distribution, and in consequence the total error on the sequence.

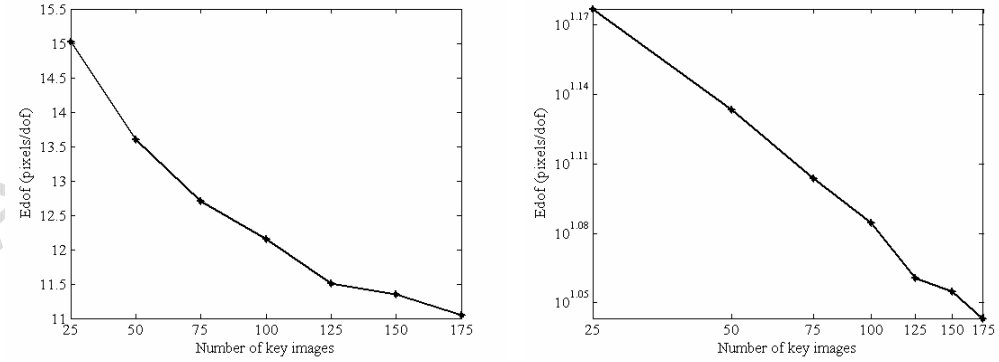
To justify the random approach, we pay special attention to the contours extraction of extreme apical constriction positions. Therefore the following comments take into account phonetic aspects. Many consonants articulations are at extreme positions of the tongue, especially of the tongue tip. We first remark that the extreme positions are better represented with the realized over-sampling, consisting in applying the method with 200 key images for the tongue. Increasing this number from 100 to 200 has a beneficial influence on extreme positions estimations.

We observe the video sequence, reconstructed with the estimated tongue contours, and in particular, segments corresponding to extreme articulatory gestures. This qualitative observation shows that these extreme positions are correctly tracked. Quantitatively, we then consider 72 realizations of the alveolar consonants [s, z, t, d] of the corpus. These consonants are called alveolar, due to their production place, i.e., contact between the tongue and the palate in the alveolar zone. Without going into detail, we evaluate that 65% (respectively 83%) of these consonants are detected with a constriction between the tongue tip and the palate lower than 5 pixels (respectively 10 pixels).

Note that other strategies of key images choice may be possible and compared to the random one; it implies, before the automatic processing, the manual marking of each new set of key images.

4.2.2. Number of key images

The random choice of key images being validated, we consider now the number of keys to retain. This choice is a compromise between the reconstruction error rate and the time cost of manual processing. The influence of the keys number *n* on the error rate *Edof* is shown (Fig. 13a) for the 15 tongue *dof* for the method including simple indexing and low-pass temporal filtering.



(a) (b)

Figure 13: Influence of the number of key images on the tongue *Edof* error calculated with temporal filtering and simple indexing ((a) linear scale – (b) loglog scale).

As expected, the error decreases with a higher number of marked key images. From 25 to 100,

Edof decreases by 3 pixels, whereas it only decreases by 1 pixel between 100 and 175 keys. With loglog scale (Fig. 13b), the relationship is nearly linear. By extrapolation, this relation allows evaluating at 305 the number of key images required to reduce the error by 1 pixel (from 11 to 10 pixels) in respect to the value obtained with 175 keys. We estimate that 200 key images is a good compromise between the manual marking step effort and the marking error.

Notice as well that besides the fact that retaining 200 key images for the tongue improves the estimation of extreme positions, it also covers enough configurations to allow the capture of the tongue movements independently of the jaw and dentition ones. Indeed configurations of open or closed jaw for various tongue positions are well tracked along the sequence.

4.2.3. Number of DCT components according to the considered articulator

The number of DCT components is fixed at 575, independently of the considered articulator (see Table 1). We aim here at validating this choice.

Some *Edof* measures are realized on the lips and the velum for various numbers of DCT components taken into account for the indexing. Fig. 14 shows that there is almost no influence for the velum, a little more for the lips. Using fewer DCT components, without changing consistently the results, could have reduced the computation time. Since this time cost is acceptable, we have kept 24*24 components.

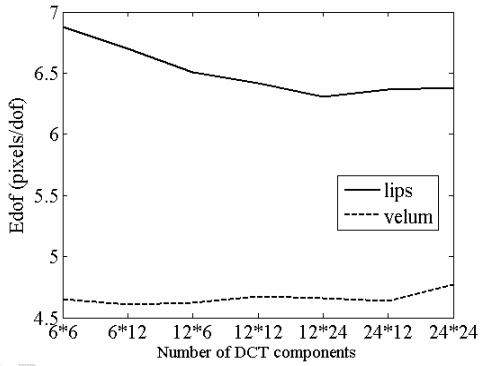
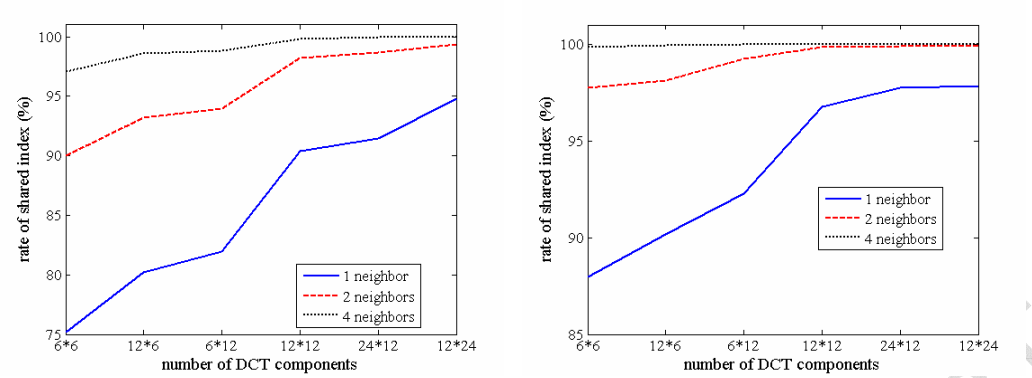


Figure 14: *Edof* error, calculated with simple indexing and without temporal filtering, for the lips (175 key images) and for the velum (75 keys).

The following figures (Fig. 15) show the similarities of indexing according to the number of DCT components, for the lips and the velum. Considering the indexing with 24*24 DCT components as the reference, we compare for the whole sequence the percentage of common indexes with other blocks of DCT components (the first component is always let aside). With a 12*12 DCT block, the simple-indexing is similar at 90% for the lips. When considering 2 neighbors without order, the indexing is 90% similar from 6*6 DCT components, for both the lips and the velum. With a 4-neighbor-indexing, similarity is reached with 6*6 components for the velum and 12*12 for the lips. We notice that the multi-indexing attenuates widely the influence of the number of DCT components. The neighboring indexes are identical; the residual difference is due to the distances taken into account in the averaged weighting.

These 2 series of measures (Fig. 14 and 15) validate the choice of 24*24 DCT components for all articulators.



(a) (b)

Figure 15: Knowing the index with 24*24 DCT components, percentage of frames indexed with the same key images according to the number of DCT components and the neighborhood size (a) Lips – (b) Velum.

4.2.4. Neighborhood size

As explained in §2.3.2 with 3 neighbors, multi-indexing instead of simple indexing provides an error reduction. We vary the neighborhood size from k=1 to k=10 and measure the global error *Edof* on the 15 tongue *dof* without temporal filtering and with 175 key images. Increasing the number of neighbors from 1 to 4 significantly decreases the error, but there is no supplementary gain for k>4 (Fig. 16).

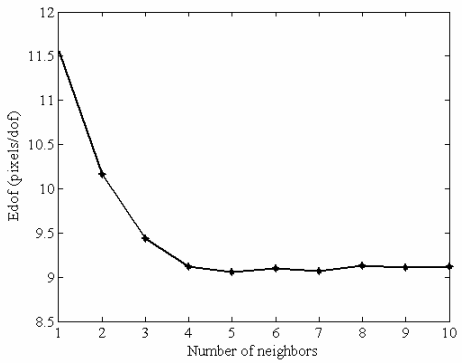


Figure 16: Influence of the neighborhood size on the tongue *Edof* error calculated without filtering and starting from 175 key images.

4.2.5. Temporal low pass filtering and choice of F_c

Another parameter of the method is the cut-off frequency F_c used for low-pass temporal filtering. For now, this filtering (8th order 0-phase filter) is applied for all articulators with F_c = 6 Hz. But is that value adapted to all articulators? We are particularly interested here in the tongue tip, whose movements are fast.

Two measures are realized so as to observe the F_c influence.

First (Fig. 17a), by way of a Jackknife, an *Edof* analysis with 175 key images, 4-neighbor-indexing and temporal filtering is realized on the 5 *dof* of the tongue tip, for various cut-off frequency values (from 3 to 12 Hz).

In parallel, tongue contours estimations are realized starting from the 200 key images and for various cut-off frequencies. We isolate some critical images, those corresponding to the 72

consonants [s, z, t, d] of the corpus and we measure for these frames (Fig. 17b) the constriction size between the palate and the estimated tongue tip.

We observe in both cases that, beyond 6 Hz, there is little influence of this frequency. *Edof* keeps stable and the constriction size for alveolar consonants is not significantly reduced. The observation of these 2 results motivates the choice of F_c=6 Hz also for the tongue tip.

A third measure is carried out, based on the alveolar consonant configurations previously considered. On one hand, we mark manually the effective constriction between the tongue and the palate, for these 72 frames. On the other hand, the tongue contour is estimated by the method applied with temporal filtering at 6 Hz, and an estimation of the tongue-palate constriction is evaluated for the alveolar configurations. For these images, we then compare the estimated constriction distance with the effective one and we observe that they are very close. For 90% of the alveolar consonant configurations, the deviation between the 2 distances is lower that 5 pixels (Fig. 17c). A deviation higher than 5 pixels is rare (7 configurations over 72) and the 5-pixel-threshold is estimated at about 1.25 mm. This mm-value is only indicative, since the calibrating information is not available on the Laval43 film.

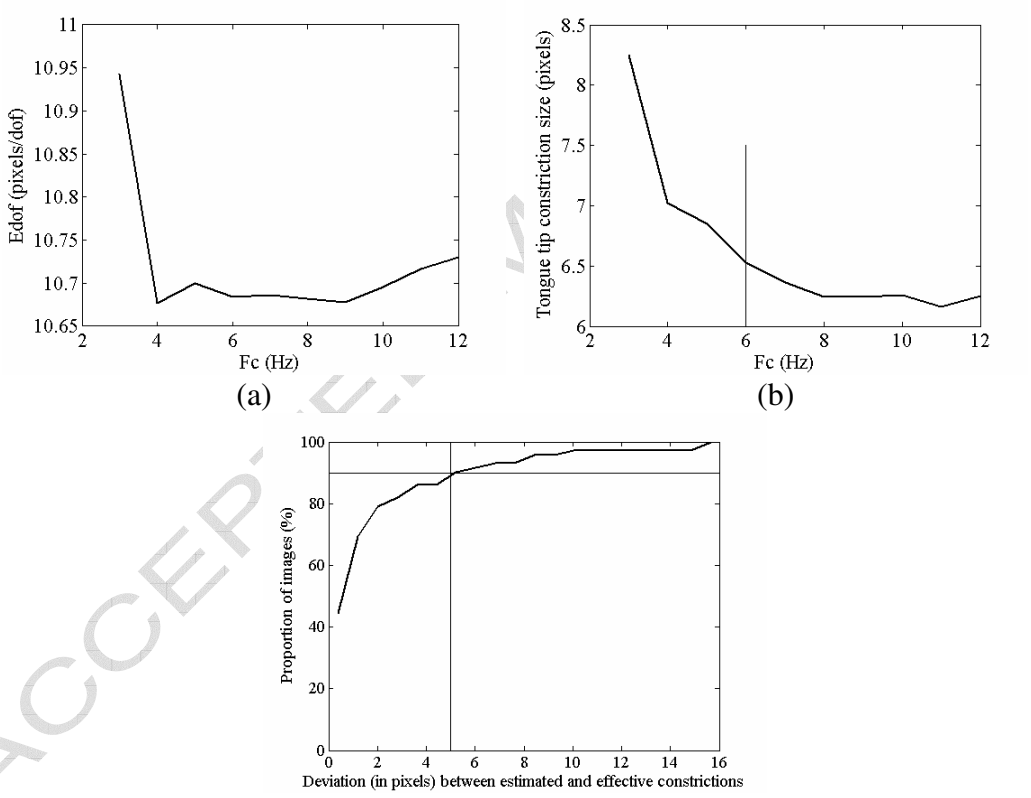


Figure 17: (a) *Edof* error, calculated with 4-neighbor-indexing and 175 key images, for the 5 *dof* of the tongue tip, for various lowpass temporal filtering.
(b) Mean constriction size at the tongue tip for alveolar consonants of the Laval43 corpus.
(c) Cumulative distribution of alveolar consonant configurations according to the deviation between estimated and effective constrictions.

4.2.6. Effect of post-processing treatments

At last, we quantify here the improvement on error due to the combination of post-processing treatments. We apply successively the error reduction methods, presented in §2.3. and mathematically independent. We reduce the *Edof* reconstruction error by more than 2 pixels (black line in Fig. 18) for the global estimation of the tongue contours. The temporal filtering improves the reconstruction only in the case of simple indexing. There is no noticeable improvement when the indexing takes into account 4 neighbors. The two operations are probably redundant. And the effect of neighborhood averaging seems to be more important: the error decreases with 2.5 pixels between 1-neighbor-indexing and 4-neighbor-indexing (without filtering), whereas it only decreases with 1 pixel between filtering or not (in the case of simple indexing).

Taking into account the tongue tip specific extraction described in §3.1 we observe a decrease (about 1 pixel) of *Edof* (dashed line in Fig. 18) on the tongue's 15 *dof*. The error *Edof* between the manual reference marking and the estimation, for the 5 *dof* of the tip only, is reduced from 12.5 to 10.3 pixels thanks to the tip specific indexing.

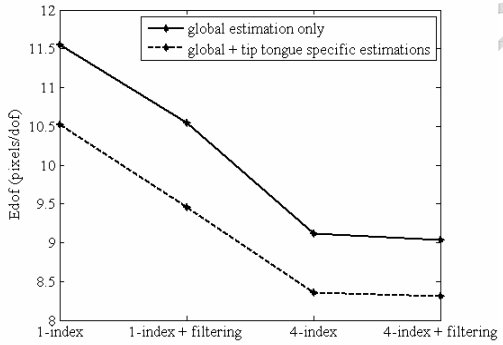


Figure 18: Contribution of error reduction post treatments observed with *Edof* error evaluated from 175 key images and for the 15 *dof* of the tongue.

4.3. Final marking error for the various articulators

For each articulator, the semi-automatic method is applied with a 4-neighbor-indexing and with temporal filtering at F_c=6 Hz. Other features are summarized in Table 1 and differ depending on the considered element. Table 2 details the marking error by *dof*, *Edof*, evaluated for various articulators with the chosen parameters, using a Jackknife technique over 10 simulations on the key images.

	Degrees of freedom	*Edof* (pixels/*dof*)	Jackknife n_2 / n
Tongue	15	8.3	25 / 175
Tongue tip	5	10.3	25 / 175
Velum	14	3.4	25 / 75
Upper lip	8	3.4	25 / 175
Lower lip	8	5	25 / 175

Table 2: *Edof* EVALUATION FOR VARIOUS ARTICULATORS (IMAGE SIZE: 720*480)

Concerning the lips, their estimation suffers from the very low contrast. Moreover we observe that the upper lip movement is better reconstructed than the lower lip, due to the highest

movement variability of the lower lip.

The velum is well estimated, the mean reconstruction error is evaluated at 3.4 pixels/*dof*. But we note that the error is not homogeneous for the 14 degrees of freedom: it is higher for the upper part of the velum than for the lower part (Fig. 11a), respectively 3.2 and 2.5 pixels/*dof* (when omitting, in both parts, the free point representing the velum tip). This is consistent with the higher distortion qualitatively observed for the upper part of this articulator.

5. Comparison studies

5.1. Testing the method on another sequence and comparing with manual extraction: the M.I.T. film of the ATR database

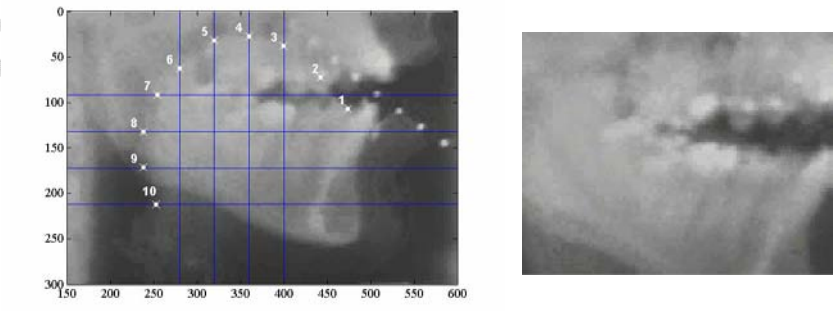
The ATR database contains one film designated as 'the M.I.T. film', which was filmed with a frame rate of 45 im/sec in 1962 at K.T.H. in Sweden under the direction of S. Öhman (Öhman & Stevens, 1963). K. Stevens (a male native speaker of Canadian English) was the subject. The film contains single repetition of 31 non-sense mono- and bisyllables, followed by two sentences. Painting their midline with barium adhesive enhanced the outlines of the lips and tongue. As for Laval43, the digitized M.I.T. film is available on the ATR database DVD at 29.97 im/sec. 1630 consecutive images of the sequence are considered.

For sake of testing the proposed semi-automatic method on another sequence, we consider the tongue only. We aim here at comparing semi-automatic estimated markings with manual ones.

5.1.1. Semi-automatic method applied to the tongue

The method is applied as described in section 2. First a manual marking step is considered for a limited number of key images. A fixed coordinate grid is defined, very similarly as for Laval43. Horizontal and vertical lines define 8 points (1 *dof* each) and 2 points are free for the tongue tip. The coordinate grid is shown in Fig. 19a.

According to the smaller size of the MIT sequence, we limit at 150 the number of key images and the 2 authors manually mark the same 150 images, randomly chosen. Note that unfortunately, the expected help by the barium adhesive on the tongue midline turns out disappointing for static images. It facilitates the manual marking of the contours only for a small proportion of these images.



(a) (b)

Figure 19: (a) Coordinate grid defined for the manual tongue marking of the MIT sequence.
(b) Framed image considered for the automatic indexing of the MIT sequence.

The automatic indexing requires a set of DCT components, calculated on framed images; one example is shown in Fig. 19b. For each framed image, the 24*24-1 lowest frequency DCT components are conserved.

The semi-automatic method is applied with 4-neighbor-indexing and with low-pass temporal filtering at F_c=6.7 Hz (in order to have the same effective 10 Hz cut-off frequency).

The method is applied twice, starting from each set of manual marks, providing 2 sets of estimated marks for the MIT sequence.

5.1.2. Evaluation of the marking error

The evaluation of the various markings is realized thanks to some *Edof* measures. *Edof* has already been defined and used for the Laval43 sequence. It is a geometrical error by degree of freedom. Moreover, to analyze the degree of mismatch at the frame level and the statistic of large deviations, we introduce, specifically for the comparison studies, a complementary error measure, which is an error by frame, noted *Efra*.

Efra is a RMS error calculated on the *p dof* of the considered articulator, i.e., here, the 12 *dof* defining the tongue. This allows a measure of the variance of the degrees of freedom by frame. This measure is realized on test images, between an estimated marking and a manual marking. With the same notations as (3) and (4), for one test frame *j*, $Efra_1(j)$ is given by formula (5). The final error *Efra* for one articulator is then the mean value of the $Efra_1$ error on the n_2 test images. *Efra* is an instantaneous error geometrically integrated; it is expressed in pixels/*dof*.

$$Efra_1(j) = \sqrt{\frac{1}{p}\sum_{x=1}^{p}(\hat{GK}_j(x) - GT_j(x))^2} \qquad (5)$$

$$Efra = \frac{1}{n_2}\sum_{j=1}^{n_2} Efra_1(j) \qquad (6)$$

As for *Edof*, the presented error values for *Efra* result from a Jackknife technique. This is applied over 25 simulations taking into account 120 key frames for the semi-automatic method and 30 test frames for the evaluation.

The following table evaluates the semi-automatic method by comparing the RMS errors (*Edof* and *Efra*), in the same way as Akgul *et al.*, 1999. We measure the errors between the 2 sets of manual marks for the same key images (manual vs. manual in Table 3), and between the semi-automatic estimated *dof* and the manually extracted ones (estimation vs. manual in Table 3). For the "manual vs. manual" column, errors *Edof* and *Efra* are evaluated by substituting, in (3) and (5), the semi-automatic estimated configurations by one manual marking and by comparing it to the other one. For the "estimation vs. manual" column, the error is first calculated, as defined above, between estimated and manual marks starting from the marking of each author. The error reported Table 3 is the mean value over the 2 authors.

In agreement with Perkell (1969), we estimate that 42 pixels measured on the images correspond to 1 cm in the midsagittal plane[1]. The effective error values in millimetres are evaluated in Table 3.

	Manual vs.	Estimation vs.

[1] The distance between 2 lead pellets (in the top right corner of the MIT frames) is evaluated at 30 pixels on the BMP images and this corresponds to one centimetre according to Perkell, 1969. An average magnification factor of 1.4 is taken into account to calculate actual midsagittal distances from the tracings. Thus 30 pixels match with 1/1.4 cm.

	Manual	Manual
Edof	4.5 pixels 1.07 mm	6.7 pixels 1.59 mm
Efra	4.73 pixels 1.13 mm	5.86 pixels 1.39 mm

Table 3: Comparison between manual and semi-automatic estimated markings for the tongue for the MIT sequence

According to the error values, the semi-automatic estimated contours are close to manual measurement variations.
The manual errors are comparable to the automatic ones: these values underline the toughness of the marking task (more than 1 mm even for a manual marking).

5.2. Comparing with another semi-automatic extraction method

From all sequences of the ATR database, the Laval43 sequence has the advantage of being completely treated by another semi-automatic method of contours extraction, the one of Thimm and Luettin at IDIAP in 1999. This provides a comparison between two semi-automatic estimations of the tongue contour.

5.2.1. The extraction method set up at IDIAP

Thimm and Luettin (1999) propose a direct extraction of the geometrical information, followed by a temporal tracking.

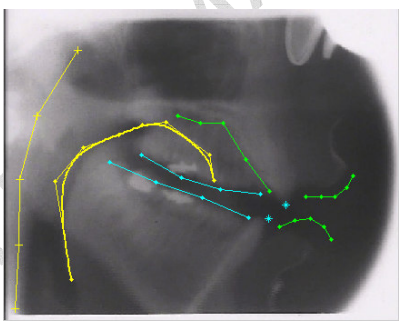


Figure 20: Contours estimated by the method of Thimm and Luettin, 1999, for one image of Laval43.

A preliminary treatment, based on histograms normalization, is applied on the images to reduce the illumination variations (Thimm & Luettin, 1998). Contours are detected in all normalized images using a Canny edge detector. Representative contours of the tracked articulator are extracted from these images and called “state images”. These are used in a matching procedure which searches the optimal score between these images and the original ones. To ensure good results, the edges used for the state images should be selected consistently. The number of images depends on the articulator and each contour is defined specifically, e.g., the tongue contour is considered from the lower point of the pharynx to the tongue tip, provided this latter is visible. The selection of a representative set of state images can be performed in an iterative manner (not detailed here), which introduces a part of manual intervention.

To complete this extraction procedure, the temporal information is used to reduce the global

error. Thimm and Luettin propose a contour tracking algorithm, applied to objects whose general position is known (or at least limited to a small number of positions) and that are subject to non-linear, very fast deformations. During this temporal tracking procedure, the transitions between states are limited to the small movements, determined by the distance between the splines defining the contours.

The results obtained by Thimm and Luettin with this method, noted TL, are digitally available in detail on the web and concern several articulators of the vocal tract. But they do not allow reconstructing its complete shape, especially because the tongue tip is often missed due to jaw occlusion.

5.2.2. Comparison of the estimation methods

We only focus here on the results concerning the tongue. To allow an objective comparison between the recovered TL results and our results, noted FB, the Laval43 images, as well as our estimated contours, are resized (images of 564*460, corresponding to the format used by Thimm *et al.*). Measures in pixels are somewhat rescaled by this transformation.

Our FB retro-marking method is applied for the tongue with 4-neighbor-indexing and temporal filtering and provides a set of *dof*, for each image of the sequence. In the TL estimation, these *dof* are not directly available. For each frame, the tongue contour is defined by a spline. Starting from these splines, we measure some TL *dof*, corresponding to ours. Because of the missing data for the tip (since its estimation is difficult with a contour approach), we limit the number of degrees of freedom. We discard the 5 *dof* related to the tip and also the 2 lower points of the pharynx. Thus we only consider the 8 points (8 *dof*) defining the body and the root of the tongue and characterized by vertical and horizontal lines in Fig. 21a. For the TL estimation, the *dof* are then measured as the intersection between these lines and their estimated splines.

To compare the 2 estimations, FB and TL, we consider both error measurements, *Edof* and *Efra*, previously introduced and calculated here for 8 *dof*. They both compare the deviation between the manual marking and each of the 2 estimates.

The *Edof* error by *dof* is calculated using a Jackknife on 25 test images. It gives an error of less than 8.5 pixels/*dof* for our estimation (starting from 175 key images) whereas the results with the estimated *dof* of Thimm and Luettin are around 15 pixels/*dof*, bearing in mind that the average length of this tongue section (body and root) is about 250 pixels.

Simulations using a Jackknife are also realized to evaluate the $Efra_1$ error by frame on a large number of test images and for each estimation (TL and FB). This allows a statistic analysis of large deviations between manual marking and estimations. With this RMS error on *dof* by frame, we highlight differences between the 2 estimations (Fig. 21c and 21d). Considering 10% of excursion (characterized with vertical and horizontal lines on the 2 figures), we find a threshold error at 14 pixels for the FB estimation and at 28 pixels for the TL estimation. The rate of high deviation between manual and estimated marks is then much higher with the TL estimation.

An error measure, similar to $Efra_1$, can be considered across the whole sequence Laval43 between the 2 estimations: we note it $Efra_2$. For one frame *j* of the sequence, $Efra_2(j)$ is given by formula (7), where $\hat{FB}_j(x)$ (resp. $\hat{TL}_j(x)$) is the geometrical estimation of *dof x* by the FB method (resp. the TL method).

$$Efra_2(j) = \sqrt{\frac{1}{p}\sum_{x=1}^{p}(\hat{FB}_j(x) - \hat{TL}_j(x))^2} \qquad (7)$$

Efra$_2$ is represented in Fig. 21e, for one sentence of Laval43. We observe an example of mismatch in the middle of the sentence: this high value of *Efra*$_2$ puts in evidence a high deviation between the 2 estimations. In this case, the tongue contour estimated by retro-marking is correct.

Remarkably, our approach preserves the contour on all images, even if it is not entirely visible. This is not always possible with the contour based approach proposed by Thimm *et al.*, and this is a penalty to estimate the tongue tip position.

6. Conclusion

After a limited manual processing step, the "retro-marking" method based on low frequency DCT video parameters is automatic and allows the geometrical extraction of each articulator of the vocal tract for speech sequences thousands of images in length. The manual step aims at being minimal but the quality of the marking is critical to ensure the success of the technique. A few days are necessary to mark the key images for the various articulators, and then the automatic treatment takes a few minutes on a Pentium IV.

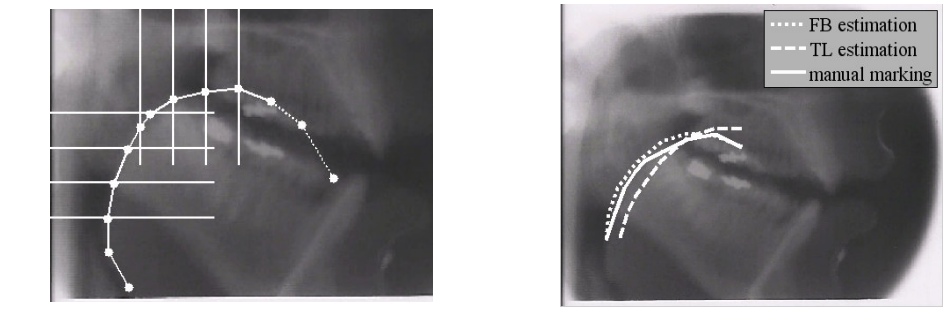
For the Laval43 and the MIT sequences, the contours are qualitatively well estimated. Results are observed for Laval43 on videos obtained by superimposition of the contours in the original video sequence.

The method has already been applied on 2 sequences of the ATR database, providing good quality contours, compared to a manual reference. An extension to the complete database can be envisaged, at least for a few articulators. The database is homogenous, and the proposed framework can be, almost directly, applied to process each sequence. Moreover, the associated audio information, available for the whole database, opens the possibility of video and audio joint treatments.

The algorithm is not x-ray specific; it may have applications to other imaging techniques. Among them, the MRI seems promising for the future, especially with the development of dynamic MRI (e.g. Narayanan *et al.*, 2004). In that area, the vocal tract contours are better visible and this could facilitate the marking task. Two methods might be combined to improve the complete treatment: an optimized technique for the static key images processing followed by our temporal treatment based on these key images.

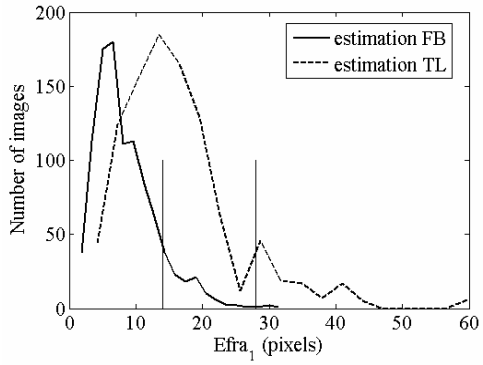
7. Acknowledgment

We thank K. Munhall and B. Burt for providing a DVD with a copy of the ATR "X-Ray film database for Speech Research," including the Laval43 and the MIT sequences. We thank P. Perrier and R. Sock for providing the Wioland sequence used for our preliminary developments.

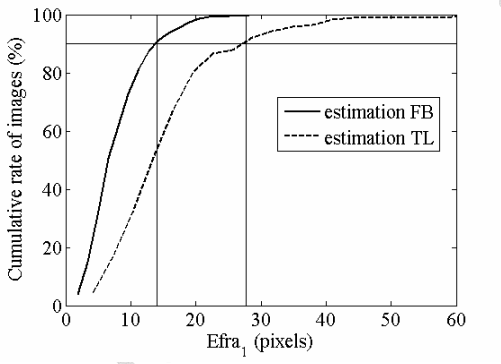


(a)

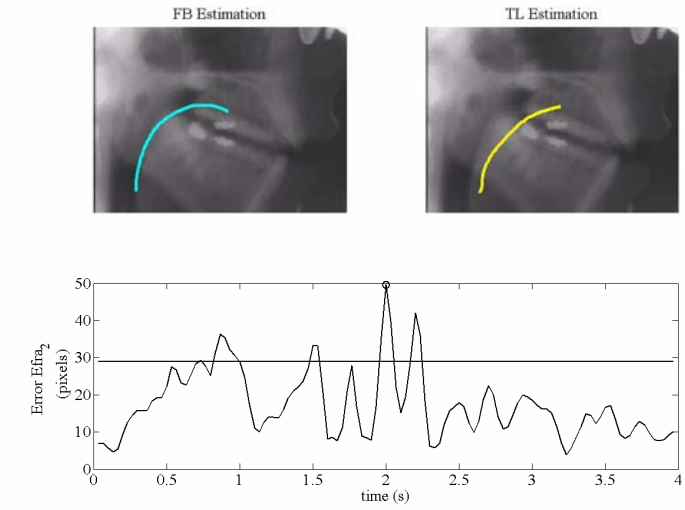
(b)



(c)



(d)



(e)

Figure 21: Comparing 2 tongue contour estimations on Laval43
(a) Only 8 *dof* (defined by horizontal and vertical lines) are taken into account in the comparison.
(b) Comparison on one test image of a tongue manual marking with 2 estimated markings.
(c) Distribution of the RMS error $Efra_1$ on 8 *dof* between manual and both estimated markings.
(d) Cumulative rate of images according to the previous distributions.
(e) Mismatch observed in the middle of one sentence and the estimated contours at this moment.

8. References

[1] Akgul, Y.S., Kambhamettu, C. and M. Stone. Automatic extraction and tracking of the tongue contours. In *IEEE Transactions on Medical Imaging*, 18, pages 1035-1045, 1999.

[2] Arnal, A., Badin, P., Brock, G., Connan, P.-Y., Florig, E., Perez, N., Perrier, P., Simon, P., Sock, R., Varin, L., Vaxelaire, B. and Zerling, J.-P. Une base de données cinéradiographiques du français. In *Proc. XXIII*^{èmes} *Journées d'Etudes sur la Parole*, Aussois, France, 2000.

[3] Badin, P., Bailly, G., Raybaudi, M. and Segebarth, C. A Three-dimensional linear articulatory model based on MRI data. In *Proc. Int. Conf. on Spoken Language Processing*, Sidney, Australia, 1998.

[4] Badin, P., Gabioud, B., Beautemps, D., Lallouache, T., Bailly, G., Maeda, S., Zerling, J.-P. and Brock, G. Cineradiography of VCV sequences: articulatory-acoustic data for a speech production model. In *Proc. Int. Conf. on Acoustics*, Trondheim, Norway, 1995.

[5] Berthommier, F. Characterization and extraction of mouth opening parameters available for audiovisual speech enhancement. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Montreal, Quebec, Canada, 2004.

[6] Bothorel, A., Simon, P., Wioland, F. and Zerling, J.-P. Cinéradiographie des voyelles et consonnes du français. *Travaux de l'Institut de Phonétique de Strasbourg*, 1986.

[7] Fant, G. Acoustic theory of speech production, The Hague: Mouton, 1960.

[8] Fontecave, J. Extraction semi-automatique des mouvements du tractus vocal à partir de données cinéradiographiques. *PhD dissertation*, Institute of Speech Communication, Grenoble Institute of Technology, Grenoble, France, 2006.

[9] Fontecave, J. and Berthommier, F. Quasi-automatic extraction method of tongue movement from a large existing speech cineradiographic database. In *Proceedings of the European Conference on Speech Communication and Technology*, Lisboa, Portugal, 2005.

[10] Fontecave, J. and Berthommier, F. Semi-automatic extraction of vocal tract movements from cineradiographic data. In *Proceedings of the International Conference on Spoken Language Processing*, Pittsburgh, USA, 2006.

[11] Guiard-Marigny, T., Adjoudani, A. and Benoît, C. A 3D model of the lips for speech synthesis. Progress of speech synthesis, Springer-Verlag, 1996.

[12] Hardcastle, W.J. The use of electropalatography in phonetic research. *Phonetica*, 25, pages 197-215, 1972.

[13] Heckmann, M., Berthommier, F., Savariaux, C. and Kroschel, K. Labeling audio-visual speech corpora and training an ANN/HMM audio-visual speech recognition system. In *Proceedings of the International Conference on Spoken Language Processing*, Beijing, China, 2000.

[14] Heckmann, M., Berthommier, F., Savariaux, C. and Kroschel, K. Effects of image distorsions on audio-visual speech recognition. In *Proceedings Audio Visual Speech Processing*, St Jorioz, France, 2003.

[15] Heinz, J.M. and Stevens, K.N. On the Derivation of Area Functions and Acoustic Spectra from Cineradiographic Films of Speech. *Journal of the Acoustical Society of America,* 36, 1964.

[16] Kass, M., Witkin, A. and Terzopoulos, D. Snakes: Active Contour Models. In *International Journal of Computer Vision*, 4, pages 321-331, 1987.

[17] Laprie, Y. and Berger, M.-O. Extraction of tongue contours in x-ray images with minimal user interaction. In *Proceedings of the International Conference on Spoken Language Processing*, Philadelphia, USA, pages 268-271, 1996.

[18] Maeda, S. Un modèle articulatoire de la langue avec des composantes linéaires. In *XX*^{èmes} *Journées d'Etude sur la Parole*, Grenoble, France, pages 152-164, 1979.

[19] Mermelstein, P. Articulatory model for the study of speech production. In *Journal of the Acoustical Society of America*, 53, pages 1070-1082, 1973.

[20] Munhall, K.G., Vatikiotis-Bateson, E. and Tohkura, Y. X-ray Film database for speech research. In *Journal of the Acoustical Society of America,* 98, pages 1222-1224, 1995.

[21] Narayanan, S., Nayak, K., Lee, S., Sethy, A. and Byrd, D. An approach to real-time magnetic resonance imaging for speech production. In *Journal of the Acoustical Society of America*, 115, pages 1771-1776, 2004.

[22] De Paula, H., Yehia, H.C., Shiller, D., Jozan, G., Munhall, K.G. and Vatikiotis-Bateson, E. Analysis of audiovisual speech intelligibility based on spatial and temporal filtering of visible speech information. Speech Production: Models, Phonetic Processes and Techniques, Harrington & Tabain (eds), Psychology Press, 2006.

[23] Perkell, J. Physiology of Speech Production, M.I.T. Press, Cambridge, M.A, 1969.

[24] Perkell, J., Cohen, M., Svirsky, M., Matthies, M., Garabieta, I. and Jackson, M. Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. In *Journal of the Acoustical Society of America,* 92(6), pages 3078-3096,1972.

[25] Potamianos, G., Graf, H.P. and Cosatto, E. An Image Transform Approach for HMM Based Automatic Lipreading. In *Proceedings of the International Conference on Image Processing*, Chicago, USA, 3, pages 173-177, 1998.

[26] Rao, K.R. and Yip, P. Discrete Cosine Transform: Algorithms, Advantages, Applications. Academic Press, Boston, 1990.

[27] Roy, J.-P. INTRIC, une interface de traitement d'images cinéradiographiques. *Travaux de l'Institut de Phonétique de Strasbourg*, pages163-177, 2003.

[28] Rubin, P., Saltzman, E., Goldstein, L., McGowan, R., Tiede, M. and Browman, C. Casy and extensions to the task-dynamic model. In *Proceedings of the 4*th *International Seminar on Speech Production,* Grenoble, France, 1996.

[29] Stevens, K.N. and Öhman, S. Cineradiographic studies of speech. STL-QPSR, 4(2), pages 009-011, 1963.

[30] Thimm, G. and Luettin, J. Illumination-robust pattern matching using distorted histograms. *IDIAP Research Report*, Martigny, Suisse, 1998.

[31] Thimm, G. and Luettin, J. Extraction of articulators in x-ray image sequences. In *Proceedings of the European Conference on Speech Communication and Technology,* Budapest, Hungary, pages157-160, 1999.

[32] Tiede, M.K. and Vatikiotis-Bateson, E. Extracting articulator movement parameters from a videodisc-based cineradiographic database. In *Proceedings of the International Conference on Spoken Language Processing*, 1994.

[33] Wood, S. A radiographic examination of constriction location for vowels. In *Journal of Phonetics*, 7, pages 25-43, 1979.