Semi-Automatic Extraction of Vocal Tract Movements from Cineradiographic Data

Julie Fontecave and Frédéric Berthommier

ICP – Institut de la Communication Parlée
INPG, 46 avenue Félix Viallet, 38000 Grenoble, France
fonte, bertho@icp.inpg.fr

Abstract

Since high speed X-ray films still provide the best dynamic view of the entire vocal tract, large existing databases have been preserved and are available for the speech research community. We propose a new technique for facilitating the extraction of the vocal tract shape and the movements of the articulators from complete sequences of these databases. The method was first developed for the extraction of the tongue movements in "Wioland". It has been adapted to a sequence of the ATR database, Laval43 (provided by Rochette). The method, based on the retromarking algorithm, combines the human expertise applied for marking a small number of key images, and the automatic processing of the video data. It has been extended to other articulators (lips, velum) in order to obtain the shape and the sections of the complete vocal tract. Quantitative evaluations of the estimate error and a comparison with Thimm and Luettin (1999) are achieved.

Index terms: semi-automatic extraction, cineradiographic data processing, X-ray film database, vocal tract contour extraction

1. Introduction

Among all imaging techniques, X-ray films still provide the best dynamic view of the entire vocal tract. Although modern techniques, as the MRI, give better images, cineradiography allows the observation of movements with a good temporal resolution (around 60 im/sec). Due to ethical concerns, X-ray imaging technology is not practiced anymore. Since cineradiography has made the proof of its interest, it has become imperative to preserve those films and to make them available for the speech research community. In this framework, Munhall and al. [1] have compiled the ATR "X-ray film database for Speech Research" from films contributed by Perkell, Stevens and Rochette. In the same way, in the context of a CNRS project of valorization of cineradiographic data, the Strasbourg Institute of Phonetics and the Grenoble Institute of Speech Communication have elaborated a French cineradiographic database, including the Wioland and Flament sequences [2].

In most studies the cineradiographic data is exploited after a laborious manual step. The quantitative information about the vocal tract configuration is extracted by drawing by hand image per image. In [3], Tiede and Bateson presented some ideas for processing the X-ray images automatically. A practical method for the extraction of tongue contours at the image level was proposed by Laprie and Berger [4]. Later, Thimm and Luettin [5] achieved the automatic processing of a complete sequence of the ATR database (Laval43).

But the quality of the contours evaluated with this method is weak in comparison with the manual extraction, which is traditionally useful for grounding applications as articulatory synthesis. To improve the result, we propose to reintroduce the human expertise, and we recently developed a semi-automatic method [6] which can be applied film by film. It associates the manual marking of a limited number of key images and the automatic reconstruction of movements for the full sequence. The technique presented here is based on an adaptation of the "retro-marking" algorithm [7]. This method builds a transformation function of implicit parameters, extracted from the video signal, into explicit and controlled geometrical parameters. Let notice that the retro-marking may become completely automatic when it is possible to extract the geometrical data from the key images (an example is available for the hand movements, in http://www.icp.inpg.fr/~bertho/m2p/icslp06/hand.wmv). But contouring manually the vocal tract on static images is difficult even for a human expert and this task also requires some compromises and an aid provided by an interface.

In a pilot study [6], the method was developed for extracting tongue movements in the "Wioland" database. It has been extended for the treatments of other databases and other articulators. By applying our method on the sequence Laval43, we have been able to compare our tongue contours with those obtained by Thimm and Luettin [5]. Moreover, we show that we are able to generate sections which are compatible with standard articulatory synthesis models.

2. Principle of the method and application on tongue movements extraction

The cineradiographic sequence "Wioland" was recorded in 1977 and digitized recently. This, recorded at 66 im/sec, is composed of 5673 images of the vocal tract, obtained by concatenation of 64 sentences pronounced by a single female French speaker.
The method [6] 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, which allows the association of the geometrical marking for each frame and (3) some post-processing treatments in order to restore the temporal continuity.

The manual process aims to describe, for 100 key images chosen randomly among the whole database, the position of the tongue contour with 10 points: 8 at the intersection between the tongue contour and horizontal or vertical lines, and 2 unrestricted points for the tip, i.e. 12 degrees of freedom (dof). Motion and adjacent images are taken into account during this manual step,

using a manual slider for showing the context; it allows us in many cases to extract a tongue contour, barely visible on the static key image. The choice of the marking points and lines is made to avoid missing data. At this stage, for each key image, a figure of the tongue contour is obtained by connecting the 10 points.

The main retro-marking step is the automatic indexing of the full database. For each frame, an index is calculated by assigning the index of the nearest key image. The similarity measure is the Euclidian distance between the lowest frequency DCT (Discrete Cosine Transform) components of the 2 images. The video DCT components are calculated on a reduced frame. The images are resized, centered and framed to focus on the concerned articulator and to remove some artifacts. The second step of the retro-marking technique consists in a geometrical marking of the original images. This builds an association between the frames and the geometrical information available for the key images only. This first estimation is affected by quantization effects and some indexing errors. Subsequent treatments, temporal filtering and averaging of neighboring configurations (multi-indexing), are complementary and allow reduction of the reconstruction error. A spline smoothing is also applied on the estimated points.

Superimposing the tongue contour on the original video sequence (http://www.icp.inpg.fr/~bertho/m2p/icslp06/tongue-wioland.html) lets us qualitatively verify the retrieval of the tongue movement.

A quantitative error measure is realized thanks to a Jackknife technique with a second set of 100 marked key images. It consists in forming new sets of n_1 key images by omitting, in turn, a little set of n_2 images of the original set of key images. The $Etot_1$ error is evaluated on the omitted images thus considered as test images. We evaluate the reconstruction RMS (root mean square) error dof by dof on the 100 test frames, between the marks estimated from the quasi-automatic method and the manual marks. The final error $Etot_1$ is the mean value of the error on the 12 dof. It is evaluated at about 11 pixels (bearing in mind that the tongue contour has an average length of 350 pixels and that the full images are 720*540). It is estimated at about 3 mm, but this value is only indicative, since the calibrating information is not available on the film.

This method successfully applied on tongue movements can be adapted for other articulators of the vocal tract, especially on the lips and on the velum and can be exported on other cineradiographic databases, as it will be shown in the next part.
Considering each articulator independent, a specific treatment is applied for each articulator; i.e. the process is the same but uses parameters adapted to the considered element. First the original images are framed and cut out so as to only include the considered articulator for the whole sequence and to avoid interferences. The features of the method (number of key images, points and degrees of freedom, number of DCT components for the indexing…) are then chosen for each element. The $Etot_1$ evaluation is realized to quantify the deviation between the manual marking and the estimated one on test images, using a Jackknife technique. Videos are made by superimposing the estimated marking on the original images in order to evaluate the movement reconstruction.

3. Movements estimation of the whole vocal tract on Laval43

The ATR cineradiographic database is the largest one available for speech research, with 25 different films offering 55 minutes and nearly 100000 images. We have extracted the images from the DVD provided by ATR (Japan). For now, it is not possible to carry out the treatment on the whole database because of the manual marking step, which is specific to each film; we then concentrate on one sequence, Laval43. This sequence recorded in 1974 is composed of 3973 images (720*480 pixels) of the vocal tract, from a video film (sentences read by a male native speaker of Canadian French) recorded at 50 im/sec.

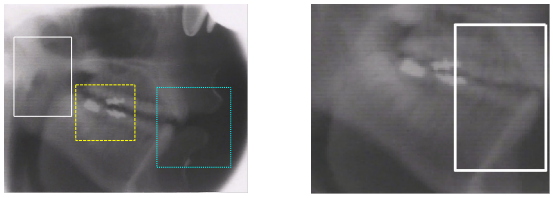
3.1. Separate extraction for vocal tract articulators

Starting from this sequence, we have independently analyzed the movements of various articulators thanks to the retro-marking technique: the tongue with a specific treatment of the tip, the velum and the lips. The parameters, adapted to each articulator, are summarized in Table 1. The frames defining the regions of interest can be observed in Figure 1 and the degrees of freedom are described in Figure 2.

Table 1: *Retro-marking parameters used for various articulators in Laval43*

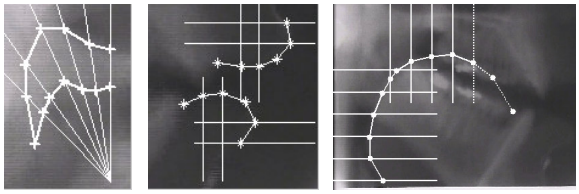
Articulator	**Parameters**				
	Points	**dof**	**Key Images**		**DCT Comp.**
			Number	**Frame Size**	
Tongue	13	15	200	105*95	24*24
Velum	13	14	100	142*186	24*24
Upper lip	6	8	200	182*186	24*24
Lower lip	6	8			

Let note that images have been decimated just for the tongue. The similarity measure for the indexing uses 575 (24*24-1) low frequency DCT components for each articulator but this number could have been reduced for some articulators according to the size of the key images.



(a) (b)

Figure 1: *Specific frames defined for each articulator* - (a) *Each frame focuses on the concerned articulator (from left to right: velum, mandible, lips).* - (b) *The position of the tip is estimated locally starting from a specific frame.*



(a) (b) (c)

Figure 2: *Degrees of freedom defined for various articulators in Laval43* - (a) *The manual marks of the velum are pointed according to a polar grid.* - (b) *Lines allow the marking of 8 dof out of 16 on the lips.* - (c) *Solid horizontal and vertical lines define the root and back of the tongue, 5 dof define the tip.*

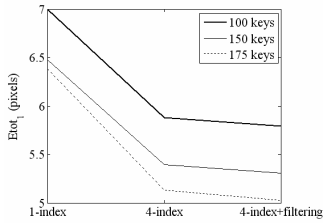
The $Etot_1$ reconstruction error depends on the number of key images and on the post-treatments. An example of $Etot_1$ evolution is given in Fig.3b for the lips. Let notice that there is little difference without or with temporal filtering, compared to the results obtained with the tongue in Wioland [6]. The upper lip moves less than the lower one, which might explain that the upper lip movement is better reconstructed than the lower (http://www.icp.inpg.fr/~bertho/m2p/icslp06/lips-laval43.html). This is confirmed by the $Etot_1$ error (Fig.3a).

The tongue tip is more visible in this film than in Wioland. We observe that its movements are fast and sometimes relatively independent. That's why a double marking of the tongue has been carried out to better capture the tip movements. This double marking associates an overall estimation of 15 dof for the complete tongue, and a specific estimation of the tip including 5 dof only. This specific estimation is calculated starting from a frame focused on the tip (Fig.2b). The fusion of these two estimates is carried out by substitution in the global estimation of the 5 dof related to the tip. Limiting ourselves to these 5 dof, the mean deviation $Etot_1$ between the manual marking of the tip and its estimation is reduced from 12 to 10 pixels thanks to the specific indexation. Taking now into account the 15 dof the global $Etot_1$ error for the tongue is less than 9 pixels (http://www.icp.inpg.fr/~bertho/m2p/icslp06/tongue-laval43.html).

The velum, which is traditionally difficult to record well, is visible in this film and this is a source of original data. The superimposition of the velum contour in the original video sequence shows the good quality of this result (http://www.icp.inpg.fr/~bertho/m2p/icslp06/velum-laval43.html).

	$Etot_1$ (pixels)	Jacknife n_2 / n_1
Tongue	8.8	25 / 175
Velum	3.6	25 / 75
Upper lip	4	25 / 175
Lower lip	6	25 / 175

(a)



(b)

Figure 3: *$Etot_1$ evaluation* - (a) *Results for the different articulators (full images are 720*480).* - (b) *$Etot_1$ error on the 16 dof of the lips with various posterior treatments (simple or multiple indexing, with filtering or not)*

The teeth and jaw positions are also extracted using the same method, but this is not detailed here (notice that we use a frame based on the high contrast of the teeth, the dashed one on Fig.1a).

3.2. Complete contouring of the vocal tract

To recover the geometry of the whole vocal tract, the rigid parts (palate, pharynx) are also marked. Since these parts are fixed, the marking is done once and for all so as to fit most of the shapes observed on the whole sequence.

At this point, all the articulators of the vocal tract have been marked separately. The interpolation between the points provided by retromarking is achieved with a spline smoothing specific for each articulator, and these are combined to get the full contour visible Fig. 4.

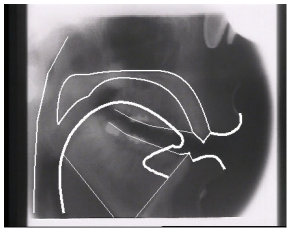


Figure 4: *Contours of the whole vocal tract in Laval43.* (http://www.icp.inpg.fr/~bertho/m2p/icslp06/Vtract-laval43.html)

4. A comparison study

4.1. The extraction method set up at IDIAP

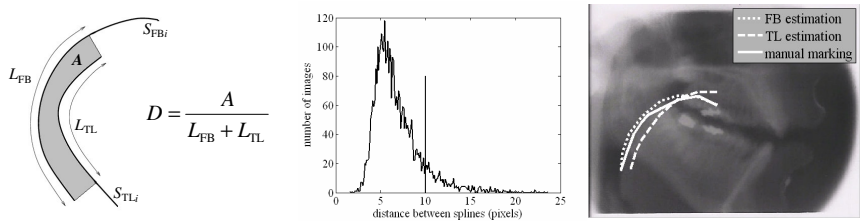
The Laval43 sequence was completely analyzed by Thimm and Luettin at IDIAP [5]. In brief, the results, available in detail on the site http://www.idiap.ch/machine_learning.php?project=64 were performed by using histograms normalization, contour extraction and a tracking method adapted for each articulator. This method, noted TL, also makes use of key images randomly selected, in which the tongue contour is extracted by a Canny edge detector and a restoration of the continuity. The tracking procedure uses the temporal information. The results concern several articulators of the vocal tract, but do not allow to rebuild its complete shape, especially because the tongue tip is often missed. We focus here on the results concerning the tongue. The Laval43 film and the TL results were recovered and conditioned to allow an objective comparison (our results are noted FB). For each image a spline defines the tongue contour, it will further be noted S_{TLi}. The comparison is made on images of 565*460, corresponding to the format used by Thimm and Luettin[5].

4.2. Comparison of methods

Our retro-marking method applied on the tongue has provided a set of splines, noted S_{FBi} (we omit the 2 lower points of the pharynx). To compare the 2 estimations starting from the 2 sets of splines S_{FBi} and S_{TLi}, 2 types of measurements are considered:

- The relative measure *D*, which calculates the distance between the 2 splines, proposed by Thimm [8].
  It corresponds to the area between the 2 splines, divided by the sum of their lengths (Fig.5a).
- The $Etot_1$ measure based on test images, which is the deviation between the manual marking and the 2 estimates.
  For this measure, because of the missing data for the tip (since its estimation is difficult with a contour approach), we have only taken into account the 8 points defining the body and the root of the tongue (Fig.5c).

The average error *D* between the 2 estimations is evaluated at 6.8 pixels. The distribution of these errors (Fig.5b) shows that for 10% of the base, there is a handover between the 2 methods (distance between the splines higher than 10 pixels). For these images, the average deviation is evaluated at 12.7 pixels.



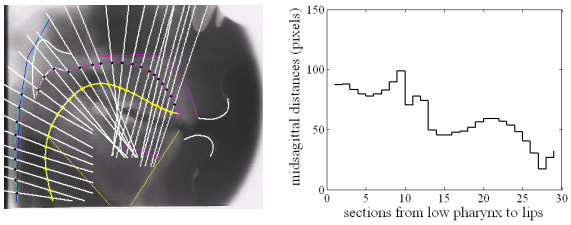
(a) (b) (c)

Figure 5: *Comparing 2 estimations*
(a) *Measuring the distance between 2 splines.*
(b) *Distribution of the difference D between the 2 estimations, and definition of a handover threshold D>10.*
(c) *Comparison on one test image of a manual marking of the tongue with 2 estimated markings (the tip is excluded from this comparison).*

The $Etot_1$ measure for our estimation (calculated with the best parameters) gives an error of less than 8 pixels whereas the results with the estimated contours of Thimm and Luettin are around 20 pixels, bearing in mind that the average length of this tongue section is about 250 pixels.
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, and this is a penalty for the tip.

5. From contours to mid-sagittal sections

In speech production studies, the mid-sagittal sections are one of the most convenient representations of articulatory data emerging from such analysis. For each image, starting from the contours described above, the mid-sagittal distances are measured along a grid [9]. Our grid has been elaborated line by line. Lines were first defined perpendicular to the palate or to the pharynx. Then the orientation of these lines has been slightly corrected in average along the midline in the mid-sagittal view to be as orthogonal as possible to the tongue as well. Corrections image by image have been carried out using procedures of Yehia (2002) [10] or Beautemps et al. (1995) [11]. Along the vocal tract, 29 midsagittal distances are measured, 27 thanks to the lines defined from the low pharynx to the alveolar zone, 1 between the upper and lower front teeth and 1 between the lips, for the whole sequence http://www.icp.inpg.fr/~bertho/m2p/icslp06/sections-laval43.html.



(a) (b)

Figure 6: *Midsagittal sections in Laval43.*
(a) *Sections grid superimposed on the estimated contours.*
(b) *Midsagittal distances measured from the grid.*

6. Conclusions

We have shown that the retro-marking method based on low frequency DCT video parameters allows the extraction of geometry and movements of the vocal tract. But the question still remains to know if our measurements are enough precise to be associated with speech temporal and spectral features in order to analyze the dynamic aspects of the relation between geometrical configurations and acoustics. Particularly, the knowledge of the area function leads to the formants estimation, and a comparison between effective and estimated formants is feasible. Temporally, the tracking of the tip tongue movements allows the detection of the tongue contact events, which can then be related to the production of consonants.

Acknowledgements: We would like to thank Pascal Perrier for providing the Wioland video, digitized in the context of the CNRS project of valorization of cineradiographic data and Kevin Munhall for providing a DVD with a copy of the ATR database.

7. References

[1] Munhall K.G., Vatikiotis-Bateson E. and Tohkura Y., "X-ray Film database for speech research", *J. Acoust. Soc. Amer.*, vol. 98, pages 1222-1224, 1995.

[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", *XXIIIèmes Journées d'Etude sur la Parole*, pages 425-428, 2000.

[3] Tiede M.K. and Vatikiotis-Bateson E., "Extracting articulator movement parameters from a videodisc-based cineradiographic database", *Proc. Int. Conf. on Spoken Language Processing*, pages 45-48, 1994.

[4] Laprie Y. and Berger M.-O., "Extraction of Tongue Contours in X-Ray Images with Minimal User Interaction", *Proc. Int. Conf. on Spoken Language Processing,* vol. 1, pages 268-271, 1996.

[5] Thimm G. and Luettin J., "Extraction of articulators in X-ray image sequences", *Proc. Eur. Conf. on Speech Communication and Technology*, pages 157-160, 1999.

[6] Fontecave J. and F. Berthommier., "Quasi-automatic extraction method of tongue movement from a large existing speech cineradiographic database", *Proc. Eur. Conf. on Speech Communication and Technology*, pages 1081-1084, 2005.

[7] Berthommier F., "Characterization and extraction of mouth opening parameters available for audiovisual speech enhancement", *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 3, pages 789-792, 2004.

[8] Thimm G., "Segmentation of X-ray image sequences showing the vocal tract", *IDIAP Research Report*, IDIAP, Suisse, 1999.

[9] Heinz J.M. and Stevens K.N., "On the Derivation of Area Functions and Acoustic Spectra from Cineradiographic Films of Speech", *J. Acoust. Soc. Amer.*, vol. 36, S4, page 1037, 1964.

[10] Yehia H.C., "A study on the speech acoustic-to-articulatory mapping using morphological constraints", *PhD Thesis, Nagoya University, Graduate School of Engineering*, 2002.

[11] Beautemps D., Badin P. and Laboissière R., "Deriving vocal-tract area functions from midsagittal profiles and formant frequencies : A new model for vowels and fricative consonants based on experimental data", *Speech Communication*, vol. 16, pages 27-47, 1995.