Tracking Articulators in X-ray Movies of the Vocal Tract*

Georg Thimm

IDIAP, CP 592, CH-1920 Martigny, Switzerland

Abstract. Tongue, lips, palate, and throat are tracked in X-ray films showing the side-view of the vocal tract. Specialized histogram normalization techniques and a new tracking method that is robust against occlusion, noise, and spontaneous, non-linear deformations of objects are used. Although the segmentation procedure is optimized for the X-ray images of the vocal tract, the underlying tracking method can be used in other applications.
Keywords: contour tracking, edge template, joined forward-backward tracking

1 Introduction

Although speech and speaker recognition systems achieve nowadays high performances in laboratory conditions, their performance is still unsatisfying under real-life conditions. Many researchers advocate, that more knowledge about the speech production process (*e.g.* co-articulation, dynamics, inter/-intra speaker differences) lead to improved feature extraction methods. Therefore, we attempted to extract the position and shape of articulators in the ATR X-ray films [6], showing a side-view of the vocal tract. We have investigated contour modeling techniques proposed in the literature [5, 2, 7], but obtained only insufficient results. Furthermore, we observed that the optical flow fields obtained with be best performing algorithms in [1] do not give good indications on the motion of the articulators. Another approach to segment the ATR X-ray films is described in [4]. But no final results were published and lips or jaws were not tracked. Note that other methods to gain quantitative data on the motion of articulators exist: MRI and tags are used in [3], and ultra sound in [8, 9].

2 Overview of the Segmentation

Some articulators are more distinct in the images than others. Therefore, these are located first, and then, using the benefits of image normalization and constraints on relative positions, other articulators are tracked (compare figure 1):

* This work has been performed with financial support from the Swiss National Science Foundation under Contract No. 21 49 725 96.

1. The X-ray images are filtered with a Gaussian filter and their histograms are then zero-normalized (section 3).
2. The upper front teeth are located by the means of a pattern matching algorithm using distorted grey-level histograms (section 3).
3. Similarly, a reference point in the rear upper teeth is tracked. Teeth fillings are very robust objects for this purpose.
4. The images are normalized for the position of the upper teeth and their grey-levels.
5. and **6.** The position of the lower teeth is determined.
7. The images obtained in **4.** are filtered with a *Canny edge detector*.
8., **9.**, and **10.** The edges corresponding to the lips and the rear throat are tracked in the edge images extracted in **7.**
11. A modified Canny filter, that neglects negative gradients in x-direction, is applied to the images obtained in **4.**
12. The front throat is tracked in the edge images obtained in **11.**
13. Image parts representing the upper and lower jaw with the tongue in an advanced and lowered, as well as in a back and high position, are subtracted.
14. These two series of images are filtered with a Canny edge detector.
15. The tongue is tracked. Possible positions are restricted by temporal information as well as the location of the front throat.

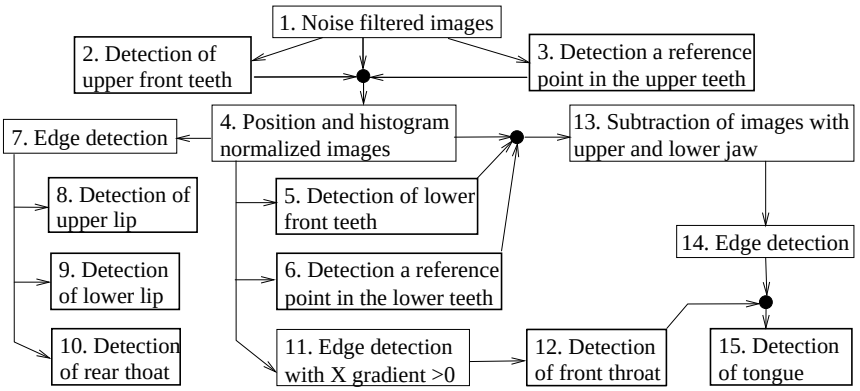


Fig. 1. Tracking of teeth, lips, front and rear throat, as well as the tongue

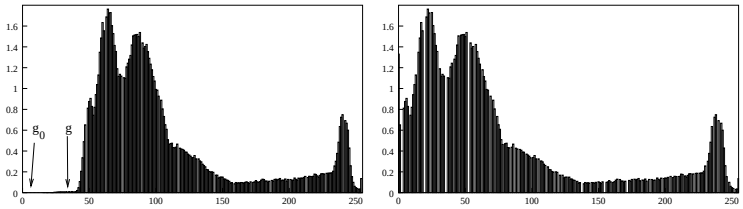
3 Illumination Normalization

The X-ray films are affected by a variable illumination, caused by either an instable X-ray energy or a varying shutter speed of the camera. This effect can not be eliminated by standard linear histogram normalization. Standard pattern matching algorithms yield therefore unsatisfactory results.

A first step to overcome this problem is to remove parts of the histogram: black parts with gray-values smaller than some value g_0 correspond to noise and parts of the image that are of no interest. Furthermore, gray-values in the interval $[g_0, g]$ are almost not occurring, as shown in figure 2(a). Consequently, all pixels with gray-values smaller than g are set to g and the image is normalized to cover the whole range of gray-levels. Although modified images have a higher contrast, the main benefit is to obtain the same brightness for an object in all images. The cut-off value g is chosen for each image according to the formula:

$$g = \max\left\{ K \mid \sum_{i=g_0}^{K} \mathrm{hist}(i) < N \cdot q \quad \text{and } K \in [g_0, 255] \right\}. \tag{1}$$

In this formula, $\mathrm{hist}(i)$ is the number of pixels with gray-value i, N is the number of pixels, g_0 is chosen close to zero in the left of the nearly empty region of the gray-level histogram, and q is the fraction of pixel values that are allowed in the interval $[g_0, g]$. Figure 2(b) shows the histogram of the resulting image.



(a) Original histogram (b) Modified histogram.

Fig. 2. The lower part of the histogram of the images is removed.

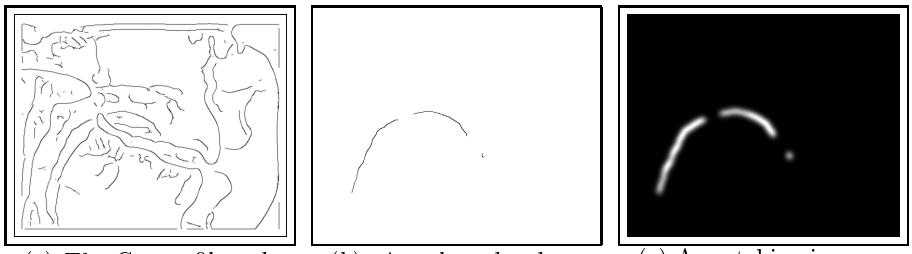
The histogram of the images are also subject to non-linear distortions. We compensated for this in the pattern matching process: the histogram of the template is modified during a comparison with some image location by the means of a flexible histogram mapping function (for more details see [11]).

4 Tracking using Edge Templates

4.1 Edge-based Template Matching

This section describes the basic matching procedure for edge templates with an edge image. The approach assumes that the object (*e.g.* the tracked articulatory feature) 1. is exactly once present in an image, 2. it is invariably at the same place and has the same orientation and size, respectively all possible places, orientations, and sizes of the object are represented in the training data. The procedure is, however, robust against **small** deformations, translations, and rotations, as well as occlusion and noise. It does not require that the edges corresponding to the object are connected.

The matching procedure uses edge images, as produced by a Canny edge detector (figure 3(a)). In a first step, edges are detected in all normalized images. From these edge images, representative edges are extracted (figure 3(b)). Such images are called *state images* in the following. These state images are inverted and blurred by a Gaussian filter, resulting in so-called *matching images* $\mathcal{S}_i$ (figure 3(c)). Both images are further associated with the same state which is proper to them.



(a) The Canny-filtered image.

(b) A selected edge: the tongue.

(c) A matching image.

Fig. 3. Creation of a state (image). Image (a) shows a side view of the vocal tract with lips (right), jaws, and throat (left).

The matching images $\mathcal{S}_i$ (the figure background is encoded as 0, the foreground as positive values) are used in the matching procedure. The score of a matching image $\mathcal{S}_i$ with respect to an image $\mathcal{X}$ is calculated as

$$\text{score}(\mathcal{S}_i, \mathcal{X}) = \sum_{x,y} \mathcal{X}(x,y) * \mathcal{S}_i(x,y) \tag{2}$$

The matching image $\mathcal{S}_i$ with $i = \text{argmax}_k(\text{score}(\mathcal{S}_k, \mathcal{X}))$ is defined as the optimal state and written as $\mathcal{S}(\mathcal{X})$. Although equation (2) evokes a rather high computational complexity (per frame n multiplications of matrices in the size of the images, if n is the number of possible states), an implementation can be efficient: only non-zero parts of the matching image need to be considered in equation (2) which permits considerable optimizations. Furthermore, the tracking procedure described in section 4.3 limits the number of matching images for which the score has to be calculated to a small subset.

4.2 Selecting State Images

In order to obtain good results with the matching procedure, the edges used for the state images should be selected consistently. In particular, the size of the selected edges and cut-off points should be similar. Example choices are given in figure 4 by the bold lines.

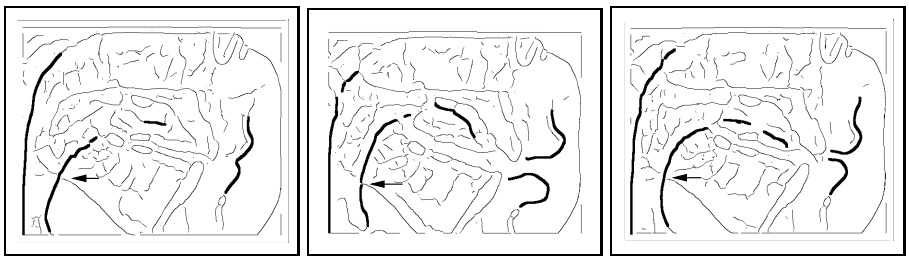


Fig. 4. The bold lines are examples for selected edges. From the points marked by the flash downwards, the edge of the tongue is also the edge of the front throat.

4.3 Adding Temporal Information to the Tracking Procedure

The performance of the basic matching procedure described in section 4.1 can be improved by using temporal information. *I.e.*, if the deformation and/or displacement of a feature is small between consecutive frames, a prediction of possible next states can be done on the base of the current state. Whether or not certain transitions are possible, is here estimated by calculating a heuristic distance between contours and then permitting only transitions corresponding to the smallest distances.

These distances would ideally be defined as the mean traveling distance of contour points. As such a strong correspondence between contour points can not be determined, the distance $D_{i,j}$ between two edges i and j is defined as the ratio of surface delimited by the splines to the mean length of the splines approximating the contours. Note, that generally the endpoints of the splines do not correspond to the same points of the feature. The splines to be used for this calculation are therefore only parts of the splines which approximate the edges (compare [10]).

Then, for each state $\mathcal{S}_i$, the element $\mathcal{T}_{i,j}$ of the state transition matrix, is defined as 1, if $\mathcal{D}_{i,j}$ is among the p percent smallest elements of vector $\mathcal{D}_i$ and 0 otherwise. Furthermore, $\mathcal{T}$ is augmented by a row $\mathcal{T}_{0,j}=1$, corresponding to an initial state $\mathcal{S}_0$. For this project, $p=30\%$ was chosen. A transition from state i to state j is possible if, and only if, $\mathcal{T}_{i,j}=1$.

4.4 Tracking a Feature

The tracking procedure is an iterative process, in which the selection of a set of possible next states by means of the transition matrix $\mathcal{T}$ alternates with the calculation of the optimal state with respect to this selection. More precisely, the score of $\mathcal{S}_i$ with respect to matching image $\mathcal{X}_t$ and the optimal state $\overrightarrow{\mathcal{S}}(\mathcal{X}_{t-1})$ for the previous frame is calculated using the following formula:

$$\overrightarrow{\text{score}}(\mathcal{S}_i,\mathcal{X}_t)=\begin{cases}\text{score}(\mathcal{S}_i,\mathcal{X}_t) & \text{if } \mathcal{T}_{j,i}=1 \text{ with } \overrightarrow{\mathcal{S}}(\mathcal{X}_{t-1})=\mathcal{S}_j\\ 0 & \text{otherwise.}\end{cases} \tag{3}$$

Whereas for the image $\mathcal{X}_1$, the preceding state is defined as S_0. Then, the optimal state $\overrightarrow{\mathcal{S}}(\mathcal{X}_t)$ is defined as the state with the maximal score.

5 Joined Forward-Backward Tracking

One important assumption in section 4.3 is, that objects move slowly. Although this is true most of the time, there are exceptions: tongue and lips can move so fast so that they assume almost opposite extreme positions in consecutive frames.

However, before and after those spontaneous, high-velocity movements, the velocity and acceleration of the respective articulators are low or even zero and fulfill for a certain time the assumption of slow movements. The following approach exploits these observations and reduces the tracking errors that are due to fast movements.

1. Calculate the forward tracking sequence $\overrightarrow{\mathcal{S}}$ as in section 4.3.
2. Calculate the backward tracking sequence $\overleftarrow{\mathcal{S}}$ in a similar manner, but using a score that restricts the states in a backward manner.
3. Join state sequences $\overrightarrow{\mathcal{S}}$ and $\overleftarrow{\mathcal{S}}$ to form the forward-backward sequence $\overleftrightarrow{\mathcal{S}}$:

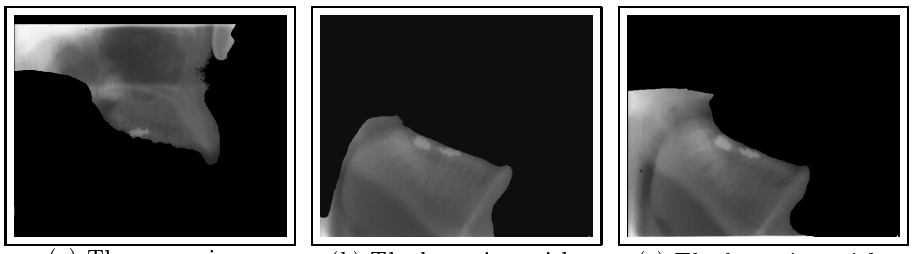
$$\overleftrightarrow{\mathcal{S}_i} = \begin{cases} \overrightarrow{\mathcal{S}_i} & \text{if } \overrightarrow{\text{score}}(\overrightarrow{\mathcal{S}_i}) \geq \overleftarrow{\text{score}}(\overleftarrow{\mathcal{S}_i}) \\ \overleftarrow{\mathcal{S}_i} & \text{otherwise} \end{cases} \tag{4}$$

This approach can be used for the lips as well as for the rear and front throat.

5.1 Tracking the Tongue

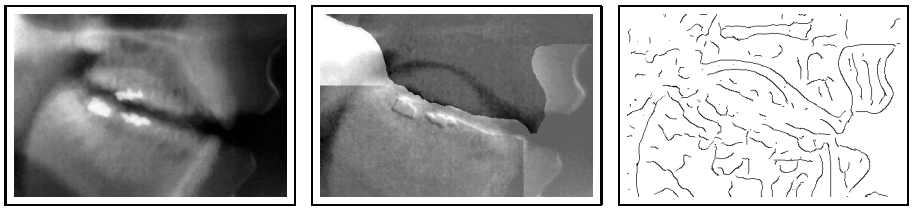
The tongue is often hidden by the jaws, especially the teeth, which means that its contour is not or only hardly visible. Sometimes even a human observer is unable detect the precise location of the tongue. The tracking procedure is consequently augmented by a specialized version of background subtraction. As the upper jaw is not moving, its image can be directly subtracted (figure 5(a)). The background image with the lower jaw has to be oriented according to the current position of the jaw, which is known from the tracking of the lower teeth. Furthermore, two background images of the lower jaw with different tongue positions are required (figures 5(b) and 5(c)): as no background image without the tongue is available, the contour of the tongue will disappear if the tongue in the image is at the same position as in the subtracted background image. Therefore, according to the position of the more easily tracked front throat, one of the two different background images of the lower jaw are used.

Figure 6 shows an example: the jaws are subtracted from the original image 6(a), resulting in image 6(b). It can be seen that the image region corresponding to the tongue is more uniform and the fillings in the upper teeth disappeared. In consequence, the edge of the tongue in the region of the mouth is nicely detected by a Canny edge detector (image 6(c)).



(a) The upper jaw (b) The lower jaw with the tongue advanced (c) The lower jaw with the tongue drawn back

Fig. 5. Background images subtracted from images before the edge detection for the tongue.



(a) A part of a X-ray image. (b) Background images subtracted. (c) The resulting edge image.

Fig. 6. An example for the subtraction of the jaws followed by an edge detection as used for the tracking of the tongue.

The tracking procedure is similar, with exception that possible state transitions are also limited by the estimated horizontal distance of the contours of the front throat and the tongue. The cut-off values for *tongue–tongue* and *tongue–front throat* distances are tuned in order to restrict the number of states images that are actually matched with an image to approximately 30%.

6 Results

The film Laval 43 with 3944 frames showing the vocal tract was completely analyzed and the results are made available on the WWW page of the IDIAP vision group (URL: `http:www.idiap.ch/vision`). The data includes the position of the front teeth, some position in the rear upper and lower teeth, as well as contours in the form of splines for the lips, tongue, and throat. We assume that the precision of the tracking procedure is sufficient for speech analysis purposes, although a quantization of the error is infeasible in practice. However, we estimate that a human would have located the contours of lips and throat in more than 98% very similarly. For the position of the tongue we estimated this figure to be above 95%. Further, the maximal error for the position of the teeth is likely to be below 2mm in real world dimensions.

7 Conclusion

We proposed a contour tracking algorithm that can be applied to objects of which the general position is known (or limited to a small number of positions), but that are subject to non-linear, spontaneous deformations. The approach is very robust against noise and occlusion, and is based on the assumption that deformations, with the exception of rarely occurring spontaneous deformations, are slow. The approach associates contours with states, and motion as well as object deformations with state transitions. During the tracking procedure, the state transitions are restricted to those associated with small movements. Spontaneous deformations are dealt with by joining the state sequences obtained by tracking the image sequence forward and backward in time.

With respect to the obtained results, the low quality of the X-ray database, and the difficulties proper to this type of data, the method can be considered to be very robust.

We expect the obtained results for the film Laval 43 of the ATR database to be sufficient for speech research purposes.

References

1. J. Barron, S. Beauchemin, and D. Fleet: On optical flow, in Int. Conf. on Artificial Intelligence and Information-Control Systems of Robots, (1994) 3–14.
2. T. Cootes, A. Hill, C. Taylor, and J. Haslam: Use of active shape models for locating structures in medical images, Image and Vision Computing 12 (1994) 355–365.
3. E.P. Davis, A.S. Douglas, and M. Stone: A continuum mechanics representation of tongue deformation, in Proc. of Int. Conf. on Spoken Language Processing (Bunnell and Idsardi, eds.) 2, New Castle, Delaware, Citation Delaware (1996) 788–792.
4. Y. Laprie and M. Berger: Towards automatic extraction of tongue contours in x-ray images, in Proc. of Int. Conf. on Spoken Language Processing 1, Philadelphia, USA (1996) 268–271.
5. J. Luettin and N.A. Thacker: Speechreading using probabilistic models, Computer Vision and Image Understanding 65:2 (1997) 163–178.
6. K. Munhall, E. Vatikiotis-Bateson, and Y. Tokhura: X-ray film database for speech research, J. Acoust. Soc. Am. 98:2 (1995) 1222–1224.
7. L.H. Staib and J.S. Duncan: Boundary finding with parametrically deformable models, IEEE Trans. on Pattern Analysis and Machine Intelligence 14 (1992) 1061–1075.
8. M. Stone and E. Davis: A head and transducer support system for making ultrasound images of tongue/jaw movement, J. Acoust. Soc. Am. 98:6 (1995) 3107–3112.
9. M. Stone and L. Lundberg: Three-dimensional tongue surface shapes of english consonants and vowels, J. Acoust. Soc. Am. 99:6 (1996) 1–10.
10. G. Thimm: Segmentation of X-ray image sequences showing the vocal tract, IDIAP-RR 1, IDIAP, CP 592, CH-1920 Martigny, Switzerland (1999).
11. G. Thimm and J. Luettin: Illumination-robust pattern matching using distorted color histograms, in Lecture Notes in Computer Science (5th Open German-Russian Workshop on Pattern Recognition and Image Understanding), Springer Verlag (1998). To appear.