EXTRACTION OF ARTICULATORS IN X-RAY IMAGE SEQUENCES

G. Thimm[*] *and J. Luettin*

IDIAP
CP 592, 1920 Martigny, Switzerland
mgeorg@ntu.edu.sg, luettin@idiap.ch

ABSTRACT

We describe a method for tracking tongue, lips, and throat in X-ray films showing the side-view of the vocal tract. The technique uses specialized histogram normalization techniques and a new tracking method that is robust against occlusion, noise, and spontaneous, non-linear deformations of articulators. The tracking results characterize the configuration of the vocal tract over time and can be used in different areas of speech research.

1. INTRODUCTION

Several authors have suggested that more knowledge about the speech production process (*e.g.* co-articulation, dynamics, inter/-intra speaker differences) might lead to improved feature extraction methods. X-ray films provide still the best dynamic view of the whole vocal tract and many important research results have been based on such data. In those studies, quantitative information of the articulators was extracted by hand, which restricted the analysis in both the number of samples and in the detail of measurements. The automatic extraction of articulatory information by image analysis techniques could circumvent these shortcomings.

Previous work that has addressed the segmentation of X-ray films include [3] and [1], however, no final results were published and the tracking did not include the lips and the jaws. Other methods, besides X-ray, to gain quantitative information about articulators include the use of MRI and tags [2] and ultrasound [5] [6].

We propose 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 (*i.e.* very fast) deformations. The approach is very robust to 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 object deformations with state transitions. During the tracking procedure, state transitions are restricted to those associated with small movements, which are determined by calculating the distance between splines approximating the contours. Spontaneous deformations are dealt with by joining the state sequences obtained by tracking the image sequence forward and backward in time.

The ATR X-ray film database [4] is probably the largest database of its kind and represents to date the main source of time-varying vocal tract information. It contains 25 films with the recordings of 14 persons and has a total duration of about 30 minutes. We have digitized and transferred the database to QuickTime format that includes the synchronized sound and separate images in GIF format (approx. 100'000 images). We have applied the contour tracking algorithm to track and extract articulators in such films. One film (Laval 43) with 3944 frames of the ATR X-ray database was completely analyzed, the results are publicly available[1].

2. LOCALIZING THE TEETH

Some articulators like the teeth are more distinct in the images du to higher contrast and are therefore located first. We then use the benefits of image normalization and constraints on relative positions to track other articulators.

The upper front teeth are located first by a pattern matching algorithm using distorted gray-level histograms. Similarly, a reference point in the rear upper teeth is tracked. The position of the head in images is then normalized using these two coordinates. Finally, the position of the lower front teeth and a reference point of the lower teeth is determined.

The X-ray films are affected by a variable illumination, caused by either instable X-ray energy or varying shutter speed of the camera. This effect can not be eliminated by standard linear histogram normalization. A standard pattern matching algorithm based on the difference between gray-values yields 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 1(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 ob-

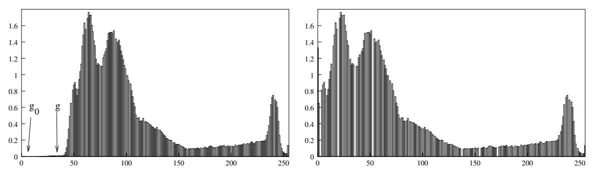
[*]Now at Nanyang Technological University, Singapur

[1]http:www.idiap.ch/vision

tain the same brightness for an object in all images. The cut-off value g is chosen for each image according to the formula:

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

In this formula, $\text{hist}(i)$ is the number of pixels with gray-value i, N is the number of pixels in the image, g_0 is chosen close to zero and to the left of the nearly empty part of the gray-level histogram, and q is the fraction of pixel values that are allowed in the interval $[g_0, g]$. Figure 1(b) shows the histogram of the resulting modified image.



(a) Original histogram (b) Modified histogram.

Figure 1: Histogram modification.

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. The illumination changes are represented by a parameter which changes the shape of the histogram mapping function (for more details see [7]).

3. 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 2(a)). In a first step, edges are detected in all normalized images. From these edge images, representative edges that correspond to a certain object are extracted (figure 2(b); how to select representative edges is discussed in section 4). 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 2(c)). Both images are further associated with the same state which is proper to them.

(a) Canny-filtered image. (b) State image. (c) Matching image.

Figure 2: Creation of a state image.

The matching images $\mathcal{S}_i$ 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 $i = \text{argmax}_k(\text{score}(\mathcal{S}_k, \mathcal{X}))$ with respect to some image $\mathcal{X}$ is defined as the optimal state and written as $\mathcal{S}(\mathcal{X})$.

Although equation (2) evokes a rather high computational complexity, the implementation can be made 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 5 limits the number of matching images for which the score has to be calculated to a small subset.

A simple tracking procedure would consist of calculating the optimal state for each image and an association with the contour of the corresponding optimal state image. As this procedure does not yield satisfactory results, temporal information is used to reduce the number of errors (see section 5).

4. SELECTION OF 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 3.

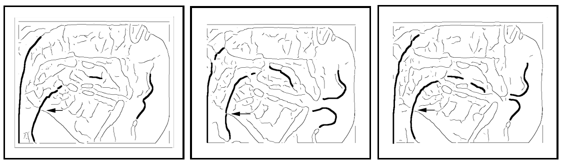


Figure 3: The bold lines are examples for the selected edges. From the points marked by the flash downwards, the edge of the tongue is also the edge of the front throat.

The selection of a representative set of state images can be performed in an iterative manner: first, some edge images are selected randomly. Then, the image sequence is tracked using the corresponding set of matching images. If the feature is not well localized in some images, some of the respective edges are added to the set of state images and the whole procedure is iterated.

5. ADDING TEMPORAL INFORMATION TO THE TRACKING PROCEDURE

The basic matching procedure described in section 3 can be disturbed by edges not belonging to the tracked object. The tracking procedure can be improved by using temporal information. *I.e.* under the assumption that the deformation of a feature between consecutive frames is small, the states that are reachable from a given state is a small subset of the whole training set. For this approach it is necessary to possess information about possible state transitions. Whether or not a certain transitions is possible, is here estimated by calculating some distance between states and then using only a percentage p of the transitions with the smallest distances.

Supposed that the edges are approximated by cubic splines. Then 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. 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, how to chose these parts is explained in [8].

To obtain the state transition matrix $\mathcal{T}$, a minimal limit $\mathcal{L}_i$, that is proper to each state $\mathcal{S}_i$, is searched, so that for p percent of the transitions $\mathcal{L}_i > \mathcal{D}_{i,j}$ is true. Then, $\mathcal{T}_{i,j}$ is defined as 1, if $\mathcal{L}_i > \mathcal{D}_{i,j}$ and 0 otherwise. Typically, $p = 30\%$ was chosen. Furthermore, $\mathcal{T}$ is augmented by a row $\mathcal{T}_{0,j} = 1$, corresponding to an initial state $\mathcal{S}_0$.

The tracking procedure is an iterative process, in which the selection of a set of possible 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:

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

For the first image $\mathcal{X}_1$ in the sequence, the preceding state is defined as S_0 and the optimal state $\overrightarrow{\mathcal{S}}(\mathcal{X}_t)$ is defined as the state with the maximal score.

6. JOINT FORWARD-BACKWARD TRACKING

One important assumption in section 5 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 high-velocity movements, the velocity and acceleration of the respective articulators is low and fulfills 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 5.
2. Calculate the backward tracking sequence $\overleftarrow{\mathcal{S}}$ in a similar manner, except for 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 $\overleftrightarrow{\mathcal{S}}$:

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

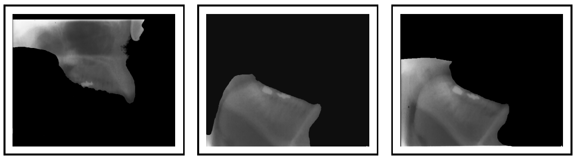
This approach can be used for the lips as well as for the rear and front throat. The tracking of the tongue is discussed in section 7.

7. TRACKING THE TONGUE

The tongue causes another problem: it is often hidden by the jaws, which means that the contour of the tongue 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 background subtraction. The background subtraction is, however, a little bit more complex than in standard applications. As the upper jaw is not moving, it can be directly subtracted (figure 4(a)). However, 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 4(b) and 4(c)) due to the following reason: as the background image contains the tongue, 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 front throat, which can be tracked more easily, one of the two different background images of the lower jaw are used.

Figure 5 shows an example: the jaws are subtracted from the original image 5(a), resulting in image 5(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 5(c)).

The set of state images is created as described in section 3. The tracking procedure is similar, with exception that possible state transitions are also limited by the

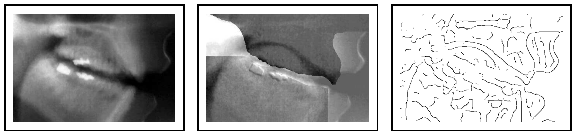


(a) Upper jaw.

(b) Lower jaw, tongue advanced.

(c) Lower jaw, tongue drawn back.

Figure 4: Background images subtracted from images before the edge detection for the tongue.



(a) Normalized X-ray image.

(b) Upper and lower jaw subtracted.

(c) Corresponding edge image.

Figure 5: Subtraction of the jaws followed by edge detection to track the tongue.

distance between the position of the front throat and the contour of the tongue in the state image. The distance between all front throat state images and tongue state images is estimated by the horizontal distance between the edges of the tongue and the throat. During the match of a state with an edge image, the assumption of slow motion is used and counterbalanced by joining forward and backward tracking results. Furthermore, the distance of the position of the previously detected front throat is used to restrict the matched states for the tongue to typically 30% of all states for the tongue.

8. CONCLUSION

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 the contours of the lips, the front, and rear throat are located in more than 98% with a sufficient precision. Similarly, the position of the tongue should be sufficiently precise in 95% of the frames. We further estimate the error for the position of the teeth below 7 pixel (approx. 2mm on images of 565 X 460 pixels).

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.

9. ACKNOWLEDGEMENTS

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

10. REFERENCES

[1] M.-O. Berger and Y. Laprie. Tracking articulators in X-ray images with minimal user interaction: example of the tongue extraction. In *Proc. IEEE Int. Conf. Image Processing*, 1996.

[2] E. P. Davis, A. S. Douglas, and M. Stone. A continuum mechanics representation of tongue deformation. In *Int. Conf. Spoken Language Processing*, pages 788–792, 1996.

[3] Y. Laprie and M. Berger. Towards automatic extraction of tongue contours in x-ray images. In *Proc. Int. Conf. Spoken Language Processing*, pages 268–271, 1996.

[4] K. Munhall, E. Vatikiotis-Bateson, and Y. Tokhura. X-ray film database for speech research. *J. Acoust. Soc. Am.*, 98(2):1222–1224, 1995.

[5] 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):3107–3112, 1995.

[6] M. Stone and L. Lundberg. Three-dimensional tongue surface shapes of english consonants and vowels. *J. Acoust. Soc. Am.*, 99(6):1–10, 1996.

[7] 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, 1999. To appear.

[8] G. Thimm. Segmentation of X-ray image sequences showing the vocal tract. IDIAP-RR 99-01, IDIAP, January 1999.