Quasi-automatic extraction of tongue movement from a large existing speech cineradiographic database

Julie Fontecave and Frédéric Berthommier

Institut de la Communication Parlée
INPG, Grenoble, France
{fonte, bertho}@icp.inpg.fr

10.21437/Interspeech.2005-430

Abstract

Automatic analysis of tongue movement in large existing cineradiographic databases can provide valuable information to understood speech production. We describe here a method for semi-automatic extraction of articulatory information from video observation in order to derive quasi-automatically a geometrical parameterization of the vocal tract movements. The algorithm starts with a limited manual processing step consisting in marking 10 points (12 degrees of freedom) on 100 chosen key images. The treatment on the whole sequence is then automatic thanks to a retro-marking method. At first, the whole database is indexed via a similarity measure performed with the key images. Then, we associate on the original images the geometrical information recovered on the key images via this indexing. Different complementary error reduction methods are also proposed. Averaging geometrical configurations of a neighborhood, temporal filtering and spline interpolation allow to reduce the reconstruction error to about 10 pixels for a tongue contour of average length of 260 pixels.

1. Introduction

The analysis of non visible vocal tract movements is a classical problem. Different imaging techniques had been proposed : cineradiography, ultrasounds [1], MRI [2],[3], electromagnetic midsagittal articulography EMMA [4]. But most of these visualization techniques are not suitable for large sequence analysis. In all studies (as in [5]), the cineradiographic data are exploited after a laborious manual step. The quantitative information about the vocal tract configuration is extracted by drawing by hand image per image. We describe here a semi-automatic method which allows the exploitation of a large amount of such data. After a limited manual step (training) the geometrical information is extracted in the full database. Remarkably, we infer geometrical information from video observation without the direct use of markers and without a contour extraction technique either.

The cineradiographic datatabase [6] we use is composed of 5673 images (490*480 pixels) recorded at 25 im/sec, from 64 concatenated video sequences. These 64 sentences are pronounced by a single female French speaker. The manual processing of this database for extracting the tongue contour is known to be difficult because of the very poor contrast and the occlusion. We assume that semi-interactive methods, such as active contour models [7], of image by image processing cannot be applied. In fact, motion and context images must be taken into account to extract geometrical features from a given image. On the other hand, the time redundancy of the movements is high, because vocal tract gestures are pseudo-periodic. This challenging task is a good support for the development of new algorithms for the capture of biological movement without the use of markers. The paper is divided in 2 parts, the presentation of an adaptation of the "retro-marking" algorithm [8] and the development of improvements for reconstruction of the geometrical information, which are tested quantitatively.

2. Algorithm

2.1. Retro-marking

A new method called "retro-marking" (Fig. 1) is implemented. It consists in processing manually, with an interface, a small number of key images $K = (K_i)_{i=1:n}$ $(n = 100)$, $(K_i) \in \{1..N\}$, and then indexing automatically the frames of the full database $S = (S_t)_{t=1:N}$ $(N = 5673)$ according to these key images. The 100 key images (K_i) are chosen randomly among the whole database (1.75% of the database).

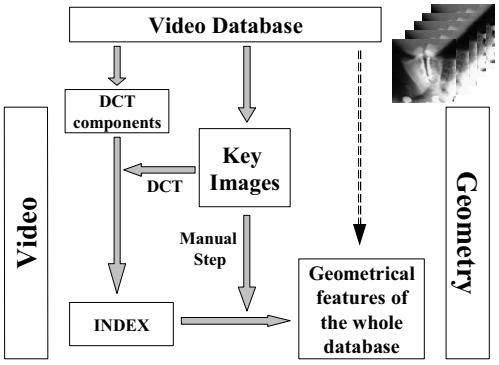
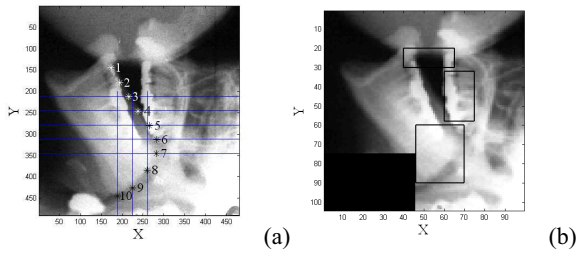


Figure 1 : Principle of retro-marking technique for derivation of geometrical features from DCT parameters through a video database

The manual process applied for the key images only aims to describe with only a few points (10 points with 12 degrees of freedom, d.o.f.) the position of the tongue contour. The 10 points on the tongue are defined in Figure 2a. Horizontal and vertical lines have been set to limit at one the number of d.o.f. for 8 of these points. The points 1 and 2 (tongue tip) have 2 d.o.f. each.

A raw and sometimes irregular figure of the tongue contour is obtained by connecting the 10 points. After this manual step of marking, we have the XY-coordinates of 10 points for the 100 (K_i) images, corresponding to a geometrical configuration G_i

associated to each key image. Let remark that motion and adjacent images are also taken into account during this manual step, thanks to the use of a slider for showing the context sequence. The quality of the final result is constrained by the quality of the manual marking, which is carefully performed by an expert.



(a) (b)

Figure 2 : (a) Out of the tongue tip, the manual marks are pointed at the intersect between the tongue contour and the horizontal or vertical lines. For one point (e.g., 6), once the line has fixed one coordinate (Y), the degree of freedom is on the other coordinate (X).
(b) Smaller images used to calculate the DCT components for the automatic indexing and selected regions for spectral analysis (§*2.2.1.*)

The main retro-marking step is the automatic indexing of the full database S according to the K images. It allows an association between the geometrical features and the video features. The video features are the 575 (24*24 - 1) lowest frequency DCT (Discrete Cosine Transform) components of each image. These components have been calculated on smaller images (Fig. 2b : 104*99 pixels) resized, centered and framed to remove some artifacts (as the piston hidden on the bottom left corner). An index i is calculated for each S_t by assigning the index of the nearest key image. The similarity measure is the Euclidian distance between the DCT components (out of the first one).

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

The second step of the retro-marking consists in a geometrical marking of the original images (S_t) by association via the index of the geometrical information available for the key images only. At this stage, the tongue movements are partly reconstructed, but we observe significant jumps.

2.2. Reconstruction of the geometrical information across time

We aim to reduce significantly the baseline reconstruction error by restoring continuity. To observe the effect of error reduction operators, we introduce some intermediate representations, i.e. Principal Component Analysis or PCA. Those PCA are applied on video features (DCT components) and on geometrical features (marked points coordinates) of key images only. 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 mapping between the 2 representations.

2.2.1. Temporal smoothing of the geometrical features

The direct observation of the tongue motion bandwidth (via a spectral analysis) has shown that the video components frequency is about 3.75 Hz. To be systematic, we have observed the Power Spectral Density (PSD) of DCT components, on selected pixels and also on the 2 first components of the video PCA. The PSD on pixels are calculated on selected regions (as in [9]) of the images (Fig. 2b). This bandwidth observed for the video data is the same for the geometrical information. Therefore, a low-pass temporal filtering (4th order 0-phase lowpass filter) with a cut-off frequency at 3.75 Hz is applied on the sequence of geometrical features.

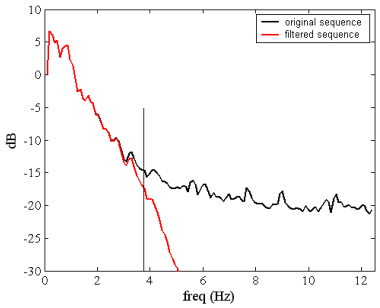


Figure 3 : Mean PSD on pixels pooled from the 3 regions of Fig. 2b for a sequence of 1000 images without and with filtering (temporal low-pass filter with cut-off frequency at 3.75 Hz)

2.2.2. Neighborhood averaging

The observation of trajectories by projection in the 2 principal plans of video and geometrical features reveals the irregularity of the relationship between video and geometry (Fig. 4 a,b). Two consecutive images which are close in the video space (Fig. 4a) are not necessarily close in the geometrical space (Fig. 4b). The trajectory in the geometrical space generated via the indexing shows severe discontinuities we attenuate by averaging the geometrical configurations of the 3 neighbors taken in the video space (Fig. 4 c,d).

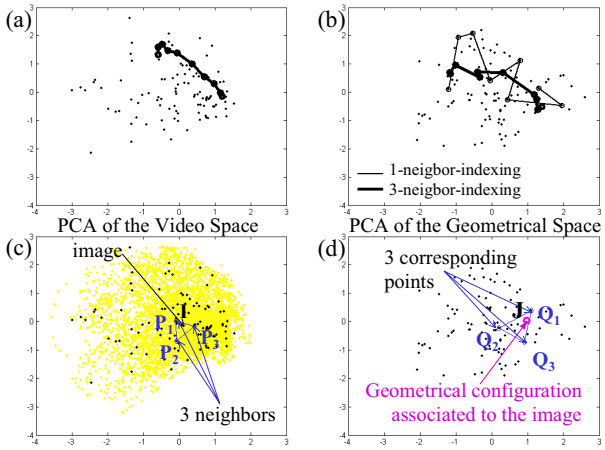


Figure 4 : By averaging geometrical configurations of 3 neighbors instead of 1 (d), the discontinuities of the trajectory in the geometrical space (b) are reduced.
(a), (b) : visualization of trajectories
(c), (d) : principle of neighborhood averaging

For each image S_t, we find 3 closer neighbors K_{i1}, K_{i2} and K_{i3} (closer in term of video features thanks to the similarity measure applied with 575 DCT components) among the 100 key images. Figure 4 c,d, the image S_t and its 3 neighbors are

projected in the video plane (points I, P_1, P_2 and P_3) and in the geometrical plane (points Q_1,Q_2 and Q_3). 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 generate $\tilde{GK}_t$. The projection of this new point on the principal plane is the point J (Fig. 4d). A supplementary weighting takes into account the Euclidian distances d calculated in the video space (DCT components) between S_t and K_{i1}, K_{i2} and K_{i3}.

$$\tilde{GK}_t = \frac{\frac{GK_{i1}}{d(S_t,K_{i1})} + \frac{GK_{i2}}{d(S_t,K_{i2})} + \frac{GK_{i3}}{d(S_t,K_{i3})}}{\frac{1}{d(S_t,K_{i1})} + \frac{1}{d(S_t,K_{i2})} + \frac{1}{d(S_t,K_{i3})}} \quad (2)$$

This 3-neighbor-indexing method can be followed by the temporal smoothing of the resultant series of geometrical configurations as well as for the 1-neighbor-indexing.

2.2.3. *Spline interpolation*

The contour connecting the points $\tilde{GK}_t$ is quite irregular, and a spline interpolation by a 5-degree polynomial $\tilde{SK}_t$ that fits the points of $\tilde{GK}_t$ in a least-squares sense improves the frame by frame estimate.

2.2.4. *Practical implementation*

A graphical user interface has been implemented for the manual step of marking. The manual processing of 100 key images, including 10 points each, lasts at least 2 hours. After the manual step, the computation time for applying the method for the whole database (5673 images) lasts a few minutes, starting from the smaller images.

3. Evaluation

At first, the result was visually evaluated by displaying the superimposition of the geometrical configurations (geometrical features defined with the XY-coordinates of 10 points) in the original video sequence.

3.1. Error measurement

For having a quantitative error measure, the reference is a set of new images marked by the same expert in the same conditions, so that the estimates generated from the key images can be compared directly with the expert marks. A second procedure is proposed, with a comparison of the two sequences generated from two different sets. Practically, the expert has carried out the manual marking on 200 images, and for each simulation, among them, 100 are considered as key images $(K_i) \in \{1..N\}$, giving a set (GK_i) of marks and the 100 others $(T_j) \in \{1..N\}$ are considered as test images, giving a set (GT_j) of marks.

We evaluate the reconstruction RMS (root mean square) error among the pair of geometrical features, the key one and the test one. $Edof_1$ and $Etot_1$ are calculated by comparing on the 100 test frames only, the marks estimated from the key images (as an external reference) with the marks of the corresponding test images. $Edof_2$ and $Etot_2$ are calculated by comparing on all frames the marks estimated from the 2 different sets of key images.

($Edof_1$) error d.o.f. by d.o.f. on the 100 test frames (T_j) between the marks (GT_j) of the test images and the marks $(\tilde{GK}_j)$ estimated from the key images (K_i)

$$Edof_1(x) = \sqrt{\frac{1}{n}\sum_{T_j}(\tilde{GK}_j(x) - GT_j(x))^2} \quad (3)$$

($Edof_2$) error d.o.f. by d.o.f. on the whole sequence (S_t) between the marks $(\tilde{GT}_t)$ estimated from the test images (T_j) and the marks $(\tilde{GK}_t)$ estimated from the key images (K_i)

$$Edof_2(x) = \sqrt{\frac{1}{N}\sum_{t}(\tilde{GK}_t(x) - \tilde{GT}_t(x))^2} \quad (4)$$

($Etot_1$) mean value of the ($Edof_1$) error on the 12 d.o.f., on the 100 test frames (T_j)

$$Etot_1 = \frac{1}{12}\sum_{x=1}^{12} Edof_1(x) \quad (5)$$

($Etot_2$) mean value of the ($Edof_2$) error on the 12 d.o.f., on the whole sequence (S_t)

$$Etot_2 = \frac{1}{12}\sum_{x=1}^{12} Edof_2(x) \quad (6)$$

The error can also be calculated after spline interpolation. In this case, the values of the 12 d.o.f. are derived from the interpolated contour.
For all these measures, the unit is the pixel of the 490*480 images.

3.2. Simulations

Thanks to this error measurement, we can tune the main parameters and combine optimally the previous methods.

3.2.1. *Number of key images*

Taking 100 key images only is a compromise between the reconstruction error rate and the cost of the manual processing. The influence of the keys number n on the global error rate ($Etot_1$) is shown (Fig. 5) for the model including simple indexing and temporal filtering. From 25 to 100 keys, the error decreases of 3 pixels, whereas it only decreases of 1 pixel between 100 and 200 keys. With loglog scales, the relationship is linear with slope $p = -0.1$. This means that to reduce the error by 10% only, the number of keys must be multiplied by 2.5. For carrying out these simulations, we have used an additional set of 100 marked images, as test images, in order to keep up to 200 images in our initial set.

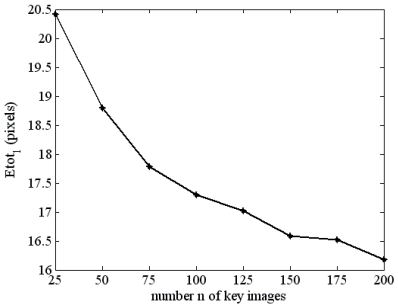


Figure 5 : Influence of the keys number on the ($Etot_1$) error calculated after 1-neighbor-indexing and temporal smoothing.

3.2.2. *Neighborhood size*

As shown in § *2.2.2*, taking 3 neighbors instead of 1 provides a great error reduction. We have varied the neighborhood size from k=1 to k=10 and measured the global error ($Etot_1$) without temporal filtering. Increasing the number of neighbors from 1 to 4 significantly decreases the error, but there is no supplementary gain for k>4 (Fig. 6).

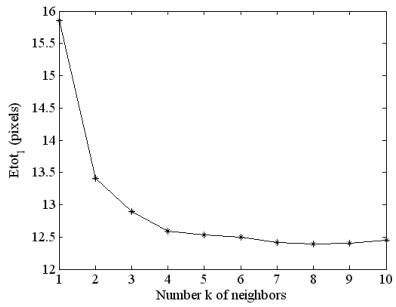


Figure 6 : Influence of the neighborhood size on the error ($Etot_1$)

3.2.3. *Cumulative improvement*

Five models M1..5 are built using combinations of the 3 error reduction methods (Fig. 7c). By applying those error reduction methods we are able to reduce gradually the reconstruction error $Etot_1$ by nearly 5 pixels (Fig. 7d). The error $Etot_2$ is more optimistic (reduction by 9 pixels) since the 2 sequences are processed similarly. This cumulative effect shows that the 3 methods are complementary. On Figure 7b, the error ($Edof_2$) for M4 is shown d.o.f. by d.o.f., in order to attest this is uniformly distributed along the tongue contour.

4. Conclusion

After a limited manual processing step, the "retro-marking" treatment is automatic and takes a few minutes. The manual step aims at being minimal but the quality of the marking is critical to ensure the success of the technique.

The superimposition of the tongue contour in the original video sequence allows observing that we retrieve the tongue movement well. Thanks to the quantitative evaluation applied on different combinations of error reduction methods, we have shown that the best global error $Etot_1$ is evaluated at about 11 pixels and $Etot_2$ at about 8 pixels, bearing in mind that the tongue contour has an average length of 260 pixels.

After extensions of this approach to recover the lips, the larynx and the soft palate and then to obtain complete vocal tract configurations, the impact of the error will be evaluated with speech synthesis tests. Inversion and correlation between lips and tongue are other outlooks for future work.

Acknowledgments : We would like to thank Pascal Perrier and Rudolph Sock for providing the digital video in the context of the CNRS project of valorization of cineradiographic data [6].

5. References

[1] Akgul Y.S., Kambhamettu C. and Stone M., "Extraction and tracking of the tongue surface from ultrasound image sequences", *Proc. of IEEE Computer Vision Pattern Recognition*, Santa Barbara, California, June 1998.

[2] Davis E.P., Douglas A.S. and Stone M., "A Continuum Mechanics Representation of Tongue Deformation", *Proc. of Int. Conf. on Spoken Language Processing - ICSLP'96*, Philadelphia, USA, Oct. 1996.

[3] Engwall O., "A 3D tongue model based on MRI data", *Proc. of Int. Conf. on Spoken Language Processing - ICSLP'00*, Beijing, China, Oct. 2000.

[4] Hoole P., "On the lingual organization of the German vowel system", *J. Acoust. Soc. Amer.*, Vol. 106, Aug. 1999.

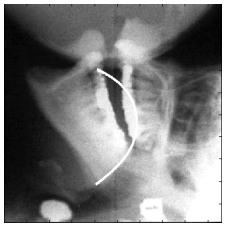
[5] Beautemps D., Badin P. and Bailly G., "Linear degrees of freedom in speech production : Analysis of cineradio- and labio-film data and articulatory-acoustic modeling", *J. Acoust. Soc. Amer.*, Vol. 109, May 2001.

[6] Wioland F., "Faits de jointure en français. Implications aux niveaux articulatoire et acoustique. Incidences sur le plan des fonctions linguistiques", Doctorat d'Etat, IPS, UMB, Strasbourg, France, 1985.

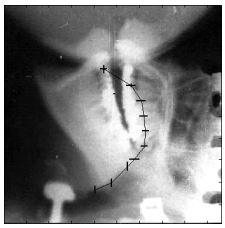
[7] Laprie Y. and Berger M.-O., "Extraction of Tongue Contours in X-Ray Images with Minimal User Interaction", *Proc. of Int. Conf. on Spoken Language Processing - ICSLP'96,* Philadelphia, USA, Oct. 1996.

[8] Berthommier F., "Characterization and extraction of mouth opening parameters available for audiovisual speech enhancement", *Proc of Int. Conf. on Acoustics, Speech and Signal Processing – ICASSP'04*, Montreal, Quebec, Canada, May 2004.

[9] De Paula H., Yehia H.C., Shiller D., Jozan G., Munhall K.G. and Vatikiotis-Bateson E., "Linking production and perception through spatial and temporal filtering of visible speech information", *Speech Production : Models, Phonetic Processes and Techniques. Harrington & Tabain (eds) Psychology Press (to appear),* 2005.



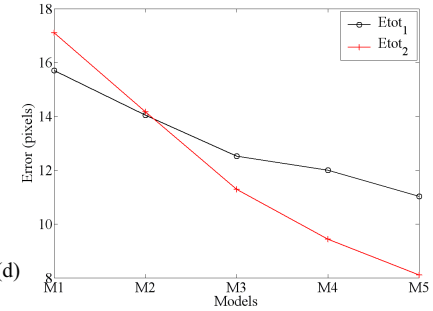
(a)



(b)

Models / Methods	M1	M2	M3	M4	M5
Neighborhood size	1	1	4	4	4
Temporal filtering	-	+	-	+	+
Spline Interpolation	-	-	-	-	+

(c)



(d)

Figure 7 : (a) The result of M5 can be seen on a video which is the superimposition of the geometrical estimates in the original video sequence : <http://www.icp.inpg.fr/~bertho/m2p/eur05/video_tongue.wmv>
(b) M4 error bars ($Edof_2$) superimposed on one key image (image size: 490*480 pixels)
(c) Error reduction models M1..5
(d) Cumulative contribution of error reduction models observed with ($Etot_1$) and ($Etot_2$) error evaluation