INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Tracking contours of orofacial articulators from real-time MRI of speech

Mathieu Labrunie [1, 2], Pierre Badin [1, 2], Dirk Voit [4], Arun A Joseph [4], Laurent Lamalle [3], Coriandre Vilain [1, 2], Louis-Jean Boë [1, 2], Jens Frahm [4]

[1] Univ. Grenoble Alpes, GIPSA-Lab, F-38000 Grenoble, France
[2] CNRS, GIPSA-Lab, F-38000 Grenoble, France
[3] Inserm US 17 - CNRS UMS 3552 - Univ. Grenoble Alpes & CHU de Grenoble, UMS IRMaGe, France
[4] Biomedizinische NMR Forschungs GmbH am Max-Planck-Institut für biophysikalische Chemie, Göttingen, Germany

(mathieu.labrunie, pierre.badin)@gipsa-lab.grenoble-inp.fr

Abstract

We introduce a method for predicting midsagittal contours of orofacial articulators from real-time MRI data. A corpus of about 26 minutes of speech has been recorded of a French speaker at a rate of 55 images / s using highly undersampled radial gradient-echo MRI with image reconstruction by nonlinear inversion. The contours of each articulator have been manually traced for a set of about 60 images selected – by hierarchical clustering – to optimally represent the diversity of the speaker articulations. The data serve to build articulator -specific Principal Component Analysis (PCA) models of contours and associated image intensities, as well as multilinear regression (MLR) models that predict contour parameters from image parameters. The contours obtained by MLR are then refined, using the local information about pixel intensity profiles along the contours' normals, by means of modified Active Shape Models (ASM) trained on the same data. The method reaches RMS of predicted points to reference contour distances between 0.54 and 0.93 mm, depending on articulators. The processing of the corpus demonstrated the efficiency of the procedure, despite the possibility of further improvements. This work opens new perspectives for studying articulatory motion in speech.

Index Terms: Real-time MRI; orofacial articulators; speech articulation; automatic contour tracking; multiple linear regression; Active Shape Models.

1. Introduction

The considerable progress accomplished in real-time dynamic MRI in the last decade (*cf.* [1] or [2]) has made this medical imaging technique extremely interesting for the study of the movements of orofacial articulators for speech [3] or swallowing [4] tasks, by allowing the acquisition of voluminous corpora of midsagittal images at a rate of 30-60 images per second with a resolution of the order of 1.5 mm/px. To characterize and model these data, it becomes necessary to develop automatic methods to track the contours of articulators from these images with a quality as precise and reliable as the traditional semi-manual ones (*cf.* [5]).

This article describes our approach to develop such a method for all the orofacial articulators involved in speech production. Two types of structures are considered: the rigid bony structures and the deformable articulators. The skull is a rigid structure that must be tracked to monitor the involuntary head movements of the subjects. Other interesting rigid structures are the jaw and the hyoid bone which have specific movements in speech. The deformable articulators comprise the lower and upper lips, the tongue, the epiglottis, the velum (soft palate), the pre-epiglottic fat pad, and the naso- oro-pharyngeal posterior wall (see examples in Figure 1d). Note that we consider the tongue contour to extend from the apex extremity to the junction with the epiglottis, excluding the lower side of the apex and the mouth floor.

2. Previous work

Before describing our approach, we present a review of the literature on articulator tracking from MRI data based on the very detailed work of Silva & Teixeira [3].

Bresch & Narayanan [6] proposed a contour fitting method in the Fourier space of the images. A contour model was made of three major vocal tract regions (above the hard palate, below the tongue and behind the pharyngeal wall) delimited by polygonal boundaries. A gradient descent minimized the distance between the image and the vocal tract region models in their Fourier spaces. Though no quantitative evaluation was provided, the examples of contours determined for each articulator testify to a reasonable quality of the method.

Eryildirim & Berger [7] developed a tongue segmentation algorithm from static MRI images similar to the Active Shape Models [8] based on a shape model built by Principal Component Analysis from contours manually traced for 38 images. The detection of the tongue contour extremities was improved by the use of a non-rigid registration method. They obtained an average reconstruction error of 1.6 mm.

Finally, Silva & Teixeira [3] recently proposed a modified Active Appearance Model (AAM, [9]) for tracking articulatory contours from dynamic MRI images. They used two AAM models, one built from contours manually traced on 30 images of non-nasal articulations, and the other one from 21 nasal articulations. They found that their approach was faster and converged better than the traditional AAM. Note that every articulator was clearly identified. The errors were expressed in terms of the Dice coefficient of similarity, which reflects the difference of the number of pixels on both sides of closed contours. This measure is however very sensitive to the

http://dx.doi.org/10.21437/Interspeech.2016-78

total area of the contours, whereas the issue of interest is rather the distance between the ground truth and the extracted contours, as provided for instance by RMS distance errors. This is particularly important for contour regions close to other contours, such as tongue tip close to hard palate.

In a preliminary study [10], we tested several contour tracking methods based on multilinear regression models (MLR), demon registration, and ASM, for very small corpora (460 to 3200 images) of different types of MRI images. We found that demon-based elastic registration methods were less accurate than MLR ones, and that the most accurate method was *articulator-specific* MLR prediction followed by a correction procedure inspired from ASM, leading to RMS point-to-contour errors of 0.59 – 0.65 mm.

The aim of the present work is to extend this exploratory work to a much larger corpus. We describe the implementation of this method which is based on a training corpus containing manually traced images, and we present evaluation results, as well as first articulatory models made from these data.

3. Real-time MRI and speech corpus

For studying different aspects of articulation, we instructed a French speaker to utter again the corpus that he had previously recorded by electromagnetic articulography ([11]).

This corpus consisted of a set of (1) two repetitions of 266 nonsense vowel-consonant-vowel (VCV) sequences, where C is one of the 19 French consonants / semivowels and V is one of 14 French oral and nasal vowels; (2) two repetitions of 109 pairs of CVC real French words, differing only by a single cue (the French version of the Diagnostic Rhyme Test [12]); (3) 68 short French sentences and 9 longer phonetically balanced French sentences [13]; and (4) 11 long arbitrary sentences.

Real-time MRI recordings were conducted at the *Max-Planck-Institut für biophysikalische Chemie, Göttingen*, using a 3 Tesla *Siemens Prisma Fit* MRI System equipped with a 64-channel head coil and following the procedure described in [2]. Foam pads were inserted on each side of the subject's head in order to minimize its yaw and roll movements, leaving room only for pitch movement in the midsagittal plane that were compensated later at post-processing. MRI acquisitions involved a low-flip angle gradient-echo sequence with radial encodings and a high degree of data undersampling. In more detail, experimental parameters were: field of view 192×192 mm², acquisition matrix 136×136, pixel size 1.41 mm/px and depth 12 bits/px, slice thickness 8 mm, flip angle 5°, echo time 1.28 ms, repetition time 2.02 ms, 9 radial projections per frame, and 18.18 ms measuring time per frame. Serial image reconstructions were defined as solution to a nonlinear inverse problem which jointly estimates all receive coil sensitivity profiles together with the desired image. Technically, this is accomplished with use of an iteratively regularized Gauss-Newton method as described in [1]. The mathematical treatment of the ill-conditioned problem largely benefits from a temporal regularization to the previous frame which is most suitable for dynamic image series as in real-time MRI. Typically, the procedure allowed for real-time movies in a midsagittal position of the speaker at a rate of 55 images / second. Speech sound was synchronously recorded by means of an optical microphone. After manual labelling of the data based on audiovisual films from the serial images and corresponding sound tracks, about 87,000 images were retained, representing a total of about 26 minutes of speech without pauses. Note that all images were oversampled by a factor 2 and denoised by means of type 1 Daubechies wavelets before further processing.

4. Head movement compensation by template matching

As it is impossible to immobilize completely the head of the subject during MRI data acquisition, the movements of the skull must be determined and compensated. Rigid template matching was used to track the 2D translation and rotation movements of the speaker's skull in the midsagittal plane. A mask that encloses steady regions while excluding non-characteristic or variable tissues was built for a reference image. Next, an optimization procedure found, for each image, the 2D translation and rotation that minimizes its distance to the reference image in terms of the RMS difference of the pixel intensities inside the mask. This allowed to align all images to the reference image, and thus to compensate for the speaker's head movements in the sagittal plane.

5. Registration based on learning methods

Standard non-rigid image registration methods [14] determine the optimal transformation that maps the source image into the target image and apply it to the contours known on the source image to produce the desired contours on the target image. These methods take into account the properties of the images, but ignore the properties of the contours themselves. Having access to manually traced contours for the articulators of interest on a corpus representative of all the data offers thus the possibility of introducing relevant information about the target contours: this would allow to considerably improve the results by providing the optimization algorithm with complementary constraints. We describe below our approach: the selection of the images for the training corpus, the manual tracing of the contours, the training of *appearance* and *shape* models from the selected corpus and the tracking procedure.

5.1. Selection of the training corpus

To build a model general enough to represent all articulations of interest, the training corpus must cover as exhaustively as possible the diversity of the articulations produced by the speaker, while minimizing the number of images of which contours need to be manually traced. To build this set, we distributed the images of the corpus in n_{cl} classes by ascending hierarchical clustering, using the RMS of the differences of pixels intensities between two images as a metric, assuming that this metric is well correlated with the Euclidian distances between the associated contours (we found a posteriori correlations greater than 0.85). As the images in the whole corpus were too many to be entirely processed at once, the procedure was applied in two stages. Each sentence was first clustered into n_{cl} classes, n_{cl} being approximately one tenth of the number of images of the sentence. The representative of every class was then chosen as the element of the class the most distant from the elements of all other classes, to make sure that the periphery of the space was also well represented. All these representatives were then pooled together and the procedure was repeated to yield the final set. Various tests showed that $n_{cl} = 59$ was a good compromise which retained the richness of the corpus. In addition, this metric was found to produce a coherent dendrogram in the sense of the cophenetic correlation coefficient.

5.2. Manual tracing of the contours

Following the procedure described in [15], the contours of the various deformable articulators were manually traced by means of cubic spline curves from the midsagittal image, while the rigid contours were manually positioned by 2D rototranslation.

5.3. Multiple Linear Regression (MLR)

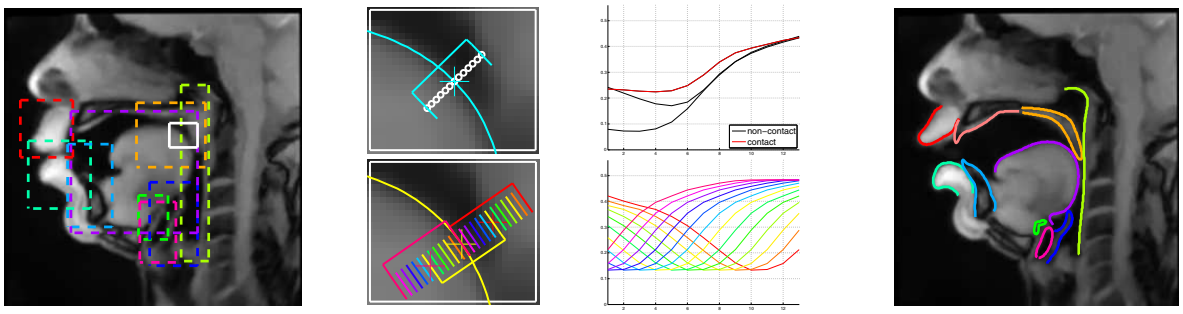
The simplest model of prediction of contours as a function of images is the one which predicts each contour coordinates as a linear combination of the intensities of the pixels of a region of interest. A rectangular bounding box encompassing its shape for all articulations in the training corpus was thus determined for each articulator (see Figure 1a). The dimensionality of the vectors of pixel intensities representing these sub-images was reduced by the use of PCA M_{int} models, retaining n_{int_arts} ($1{:}n_{art}$) components for each articulator. Similarly, the contour coordinates of each articulator were modelled by PCA individual models M_{arts} with n_{cnt_arts} ($1{:}n_{art}$) components. The numbers of components were chosen in order to minimize the prediction errors. A relatively high number allows the models to fit the articulator contours at best. Note that the components of the different articulators are partially correlated between each other, which brings additional redundancy and thus more flexibility. The MLR_{arts} models of prediction of the contours from the images were finally obtained by multiple linear regression of the contours control parameters as a function of the intensities control parameters over all the n_{cl} training data, articulator by articulator.

5.4. Modified Active Shape Models (ASMM)

Active Shape Models [8] aim to determine the contours of a given object in a set of images. They rely on two complementary types of information related to the object, namely the image local intensities and the contour, which are modeled based on a set of exemplars. In order to initiate the phase of model training, a set of images representative of the object's possible shapes is determined, and the object's contour (or shape) is manually established by an expert for each image. The contour is then represented by a shape model, i.e. a linear model of contour made of a small number of PCA components extracted from the training set. In complement, the profiles of pixel intensities along the normals to the contour in each point are modeled by appearance models of various sorts. Next, an iterative segmentation stage starts with an initialized contour. The optimal position of each contour point along its normal is determined by minimizing a criterion using its appearance model. The contour made of these points is finally regularized by the shape model in order to limit the influence of image noise and of outliers. This procedure is repeated in an iterative and a multi-scale scheme, and the object contour is eventually determined.

In our implementation, which we shall call *modified ASM* (ASMM), we use the shape models M_{arts} described above. Appearance models are built for each point of each articulator, at two levels of images sampling (r_{scale} of 2, 1) in the following way. For each point of each contour, an intensity profile is sampled by interpolation on $n_{pfl} = 13$ points distributed along a normal segment centered on the point by steps of one pixel (see Figure 1b, top, n_{pfl} points marked by white circles). Instead of being modeled by PCA as in the traditional ASM, these intensity profiles are associated with classes. Two main classes are determined. The "no contact" class contains the profiles for which the distance of the contour point to the nearby articulators along the normal is superior to a threshold of $thr_{cont} = 2$ pixels; the "contact" class contains those for whom the distance between nearby contours is lower than thr_{cont}. As several sorts of profiles were obtained for each class, due to the variability of the normals orientations and of grey levels of tissues, a finer classification was necessary. Each class was thus divided by a k-means algorithm into an optimized number of subclasses (up to 5 in practice). Each subclass is finally represented by its average profile (see two examples in Figure 1c, top).

For each articulator, the registration procedure begins with the initialization of the contour using the adequate articulator-specific MLR model. Next, the appearance is explored for each point of the contour (see Figure 1b, bottom) by sampling the intensity profiles at n_{pfl} points along the normal and varying the position i_{ctr} of the profile center (see Figure 1c bottom) by steps of 1 pixel over a range of n_{search} pixels (n_{search} is chosen between 4 and 8 depending on the articulator). The distances of all these profiles (Figure 1c bottom) to the average profiles (Figure 1c top) of all the subclasses are then calculated. If the minimal distance corresponds to the "no contact" class, the point of index i_{ctr} is considered as the corrected contour point. In the opposite case, it is impossible to determine exactly its position, and it is thus ignored in the following stage. An exception to this rule is made for tongue points near the palate: an exploration of the intensity profiles along the palate allocates each point to a



a b c d

Figure 1: (a) preprocessed image with articulator-specific bounding boxes; (b, top) zoom on the white bounding box of (a) with the manually drawn tongue contour and the normal segment carrying the intensity profile for one point; (b, bottom) zoom with the predicted contour and an illustration of the normal segments delimited by color lines for the exploration of profiles; (c, top) examples of average profiles representing one "contact" subclass and two "no contact" subclasses; (c, bottom) profiles corresponding to (b, bottom) ; (d) examples of predicted contours.

"contact" / "no contact" class. If a point of the tongue could not be determined, the closest point of the palate marked as "contact" is used as the tongue point, which ensures that the tongue contour is indeed in contact with the palate in this case. Moreover, if a point is inside a nearby articulator whose contour has already been determined, it is shifted back to the boundary of this articulator; note that articulators are processed in the following order, which gives a priority to the previous articulators over the next ones: jaw, pre-epiglottic fat pad – found more reliable than tongue root –, tongue, upper lip, lower lip, pharyngeal wall, velum, epiglottis, and hyoid bone.

The next stage of the procedure consists in fitting the shape models to the points determined to the previous stage. Note that points not determined in the regions of contact are reconstructed (extrapolated) by the shape models during this procedure. This procedure is iteratively executed for the two r_{scale} resolutions, from the coarsest to the finest, the result of one being used as the initialization of the next one.

Besides, as extremities are especially difficult to track [7], and since reconstruction errors were higher near extremities [10], a length variation was applied to the reconstructed contours near tongue tip and *submentale* (dip between the lower lip and the chin) and the final contour was selected as the one with the lowest reconstruction error.

Note that the same procedure was applied to both deformable articulators and rigid articulators (jaw and hyoid bone), since we found [10] that methods based on template matching had a lower performance than the MLR method corrected by the ASMM for the rigid articulators (note that the PCA models of these articulators have 3 components only).

6. Evaluation

All evaluations were obtained by a leave-one-out cross-validation procedure (*cf.* [16]) which computes the estimation error for every test element from models established from the $n_{cl} - 1$ other elements. We used two metrics to measure the error for each contour: the Root Mean Square (RMS) of the point-to-point distances between contours (p2p) computed over all points and all observations, and the RMS of the distances of the predicted points of contours to the drawn contours (p2c). Detailed results are given in **Erreur ! Source du renvoi introuvable.**.

The articulator-specific contour models M_{arts} using n_{cnt_arts} components each explain more than 90% of the variance of each articulator and lead to a mean reconstruction error of 0.30 mm for p2p (0.26 mm for p2c) over all articulators. This performance can be considered as the baseline of the contour tracking procedure. Similarly, the articulator-specific MLR_{arts} models that predict the contours of each articulator through the M_{arts} models from the n_{int_arts} components of the images cropped for each articulator produce mean errors of 2.07 (1.25) mm over all articulators.

Articulator	n_{int_art}	n_{cnt_art}	M_{arts} p2p	MLR p2p	ASM p2p	M_{arts} p2c	MLR p2c	ASM p2c
upper lip	3	8	0.28	1.52	1.08	0.23	1.09	0.54
velum	5	9	0.26	1.41	1.30	0.23	0.88	0.79
pharynx	4	3	0.50	1.92	1.81	0.48	0.94	0.75
hyoid	9	3	0.04	2.18	1.52	0.03	1.26	0.85
lower lip	7	7	0.40	2.19	1.32	0.34	1.26	0.56
jaw	7	3	0.01	1.51	1.28	0.01	0.76	0.54
epiglottis	7	7	0.35	2.14	1.58	0.31	1.34	0.62
tongue	8	9	0.55	3.74	2.83	0.48	2.41	0.93
pre-epigl	6	6	0.31	2.01	1.38	0.27	1.27	0.67
Sum/Mean	**67**	**55**	**0.30**	**2.07**	**1.57**	**0.26**	**1.25**	**0.69**

Table 1. Number of components for M_{arts} and M_{int} models; various errors (in mm, obtained by cross-validation). Last line gives the total number of components or the mean errors over all articulators.

Finally, we have found that the ASMM corrections of the contours predicted by the M_{arts} models reduce the mean error to 1.57 (0.69) mm. An example of resulting contour is displayed in Figure 1d. Note the RMS error estimations are averages and conceal disparities between phonemes and locations on articulators: errors are larger on the tongue tip for instance.

Due to image artefacts, the procedure described above could significantly fail for some images, especially in fast transitions. The prediction parameters associated with the contours found were thus low-pass filtered for each sentence in order to add an additional temporal constraint. The overall improvement is small, but more significant for tongue and lower lip extremities (up to 0.19 mm). The final results are illustrated in short movies provided as supplementary material (IS_78_*.mp4, in slow motion at 10 images / s).

7. Conclusion and perspectives

We developed a new method for the prediction of midsagittal contours of the main orofacial articulators involved in speech from real-time MRI images. This method is based on MLR articulator-specific models that predict contours from pixel intensities in the region of the articulators and are further refined with modified ASM models using the local information of the intensity profiles along the normals of the contours. The performance of this method is clearly higher than that described in articles which give explicit figures. It is reached at the cost of the manual contour edition of a training corpus of about 60 images. This inconvenience is however small if one wishes to handle corpora of a 100,000 images for the same speaker. These results are close to those obtained by [10] on a much smaller corpus of similar images, which confirms the value of the method. The method can still be improved: interpenetration of contours could be dealt with after the ASMM stage, and various image filtering methods could also be more extensively explored. It will be also useful to analyze the distribution of errors along the contours of the various articulators, and to try reducing the errors in critical regions such as that of tongue tip.

This method opens the way to fruitful perspectives. The large quantity of contours that can be obtained will allow to establish more elaborate articulatory models and to analyze more finely coarticulation and articulatory variability in speech. The quantities of possible data will also allow the implementation of machine-learning methods for the transformation of articulation, for example, in order to map articulations between speakers.

The proposed tracking method will be also tested on swallowing data, though it is expected to be more challenging due to frequent contacts between articulators and food.

8. Acknowledgements

This work has been partially funded by the French *Agence Nationale pour la Recherche*: grant ANR-13-TECS-0011 "e-SwallHome – Swallowing & Respiration: Modelling & e-Health at Home" in the "Technologies pour la Santé" program.

9. References

[1] M. Uecker, S. Zhang, D. Voit, A. Karaus, K.-D. Merboldt, and J. Frahm, "Real-time magnetic resonance imaging at a resolution of 20 ms," *NMR in Biomedicine* vol. 23, pp. 986–994, 2010.

[2] A. Niebergall, S. Zhang, E. Kunay, G. Keydana, M. Job, M. Uecker, and J. Frahm, "Real-Time MRI of speaking at a resolution of 33 ms: Undersampled radial FLASH with nonlinear inverse reconstruction," *Magnetic Resonance in Medicine*, vol. 69, pp. 477-485, 2013.

[3] S. Silva and A. Teixeira, "Unsupervised segmentation of the vocal tract from real-time MRI sequences," *Computer Speech & Language*, vol. 33, pp. 25-46, 2015.

[4] A. Olthoff, S. Zhang, R. Schweizer, and J. Frahm, "On the physiology of normal swallowing as revealed by magnetic resonance imaging in real time," *Gastroenterology Research and Practice*, vol. 2014, pp. 10, 2014.

[5] A. Serrurier and P. Badin, "A three-dimensional articulatory model of the velum and nasopharyngeal wall based on MRI and CT data," *Journal of the Acoustical Society of America*, vol. 123, pp. 2335-2355, 2008.

[6] E. Bresch and S. Narayanan, "Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images," *IEEE Transactions on Medical Imaging*, vol. 28, pp. 323-338, 2009.

[7] A. Eryildirim and M.-O. Berger, "A guided approach for automatic segmentation and modeling of the vocal tract in MRI images," presented at 19th European Signal Processing Conference (EUSIPCO 2011), Barcelona, Spain, 2011.

[8] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models -Their training and application," *Computer Vision and Image Understanding*, vol. 61, pp. 38-59, 1995.

[9] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 681-685, 2001.

[10] M. Labrunie, P. Badin, L. Lamalle, C. Vilain, L.-J. Boë, J. Frahm, and P. Birkholz, "Suivi de contours d'articulateurs orofaciaux à partir d'IRM dynamique," presented at 31èmes Journées d'Etude de la Parole, Paris, France, 2016.

[11] A. Ben Youssef, P. Badin, G. Bailly, and P. Heracleous, "Acoustic-to-articulatory inversion using speech recognition and trajectory formation based on phoneme hidden Markov models," presented at Interspeech 2009, Brighton, UK, 2009.

[12] J. Peckels and M. Rossi, "Le test de diagnostic par paires minimales. Adaptation du français du "Diagnostic Rhyme Test" de W.D. Voiers," *Revue d'Acoustique*, vol. 27, pp. 245-262, 1973.

[13] P. Combescure, "20 listes de dix phrases phonétiquement équilibrées," *Revue d'Acoustique*, vol. 56, pp. 34-38, 1981.

[14] D.-J. Kroon and C. H. Slump, "MRI Modality transformation in demon registration," in *IEEE International Symposium on Biomedical Imaging, ISBI '09*. Boston, MA: IEEE Signal Processing Society, 2009, pp. 963-966.

[15] P. Badin and A. Serrurier, "Three-dimensional modeling of speech organs: Articulatory data and models," presented at IEICE Technical Report, Kanazawa, Japan, 2006.

[16] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," pp. 40-79, 2010.