



**HAL**  
open science

# 3D dynamic spatiotemporal atlas of the vocal tract during consonant-vowel production from 2D real time MRI

Ioannis K Douros, Yu Xie, Chrysanthi Dourou, Karyna Isaieva, Pierre-Andre Vussoz, Jacques Felblinger, Yves Laprie

► **To cite this version:**

Ioannis K Douros, Yu Xie, Chrysanthi Dourou, Karyna Isaieva, Pierre-Andre Vussoz, et al.. 3D dynamic spatiotemporal atlas of the vocal tract during consonant-vowel production from 2D real time MRI. *Journal of Imaging*, 2022, Special Issue Spatio-Temporal Biomedical Image Analysis, 8 (9), pp.227. 10.3390/jimaging8090227. hal-03808325

**HAL Id: hal-03808325**

**<https://inria.hal.science/hal-03808325>**

Submitted on 10 Oct 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 **3D dynamic spatiotemporal atlas of the vocal tract during consonant-vowel production**  
2 **from 2D real time MRI**

3

4 Ioannis K. Douros<sup>a,b</sup>, Yu Xie<sup>c</sup>, Chrysanthi Dourou<sup>d</sup>, Karyna Isaieva<sup>b</sup>, Pierre-Andre´  
5 Vuissoz<sup>b</sup>, Jacques Felblinger<sup>e,a</sup>, Yves Laprie<sup>a</sup>

6

7 <sup>a)</sup> Universite´ de Lorraine/CNRS/Inria LORIA, 54000 Nancy, France

8 <sup>b)</sup> Universite´ de Lorraine/INSERM U1254 IADI, 54000 Nancy, France

9 <sup>c)</sup> Department of Neurology, Zhongnan Hospital of Wuhan University, 430071 Wuhan, China

10 <sup>d)</sup> School of ECE, National Technical University of Athens, Athens 15773, Greece

11 <sup>e)</sup> Universite´ de Lorraine INSERM 1433, CIC-IT, CHRU de Nancy, F-54000 Nancy, France

12

13

14

15 Corresponding Author: Ioannis K. Douros

16 Corresponding E-mail: [ioandouros@gmail.com](mailto:ioandouros@gmail.com)

17

## 18 **ABSTRACT**

19 In this work we address the problem of creating a 3D dynamic atlas of the vocal tract that  
20 captures the dynamics of the articulators in all three dimensions in order to create a global  
21 speaker model independent from speaker specific characteristics.

22 The core steps of the proposed method are temporal alignment of the real-time MR images  
23 acquired in several sagittal planes and their combination with adaptive kernel regression. As a  
24 preprocessing step, a reference space was created to be used in order to remove anatomical  
25 information of the speakers and keep only the variability in speech production for the construction  
26 of the atlas. The adaptive kernel regression makes the choice of atlas time points independently  
27 of the time points of the frames that are used as an input for the construction.

28 The evaluation of this atlas construction method was made by mapping two new speakers to the  
29 atlas and by checking how similar the resulting mapped images are. The use of the atlas helps in  
30 reducing subject variability.

31 Results show that the use of the proposed atlas can capture the dynamic behavior of the  
32 articulators and is able to generalize the speech production process by creating a universal-  
33 speaker reference space.

34

35

36 Keywords: spatiotemporal atlas, generic speaker model, adaptive gaussian kernel

37

## 38 1. INTRODUCTION

39 The differences in anatomy and articulatory strategy between speakers lead to a very large  
40 variability of MRI images of the vocal tract, which prevents the creation of a unique 3D model  
41 that can represent any speaker. The creation of a generic approach and model that incorporates  
42 this variability starting from its construction is thus crucial. In the medical field, a popular  
43 approach to represent inter-subject image variability is the use of one or several atlases. In  
44 particular, this approach is very often used in brain studies for tasks like automatic region  
45 segmentation, region labeling, etc. For instance, several atlases built from data of adults have  
46 been used to automatically label and segment the brain regions of young prematurely born  
47 children (Gousias, 2008). Each of the adult atlases was registered to the target child image and  
48 the final labeling and segmentation were based on a combination of the registration results. Such  
49 approaches facilitate the creation of automatically labeled atlases for young children by taking  
50 advantage of the availability of specific adult atlases and adapting them to the case of children  
51 for whom it is more difficult to acquire data.

52 There are several techniques to create an atlas or tackle the various issues that can appear  
53 during the creation process. One method to construct a brain atlas is to use affine registration to  
54 generate the anatomy-free reference space and then use non rigid registration to create the  
55 “average brain” template (Seghers, 2004). Apart from creating a population specific brain atlas,  
56 one can create a subject specific brain atlas (Ericsson, 2008). The main idea is that the similarity  
57 (in terms of image, gender, age etc.) between the target subject and each subject of the rest of the  
58 population is computed and this information is used as a weighting factor when creating the atlas  
59 of the target subject.

60 Another type of issue could appear during the use of the atlas, and more specifically during

61 the registration process of a new image to the atlas in order to extract atlas information for the  
62 specific subject. In order to map brain slices with severe histological artifacts to brain atlases,  
63 one can use an automatic method to identify the regions of artifacts and keep only the edge of the  
64 “correct” brain perimeter (Agarwal, 2016). The estimated edge is then sampled, and these points  
65 are used as landmarks for point to point image registration with the atlas. The other possibility  
66 consists of mapping histological slices of the brain without brain reconstruction prior to  
67 registration since it can create artifacts (Xiong, 2018). The main problem that needs to be solved  
68 is how to find out the orientation used to acquire brain slices. In this approach every histological  
69 slice is mapped to the atlas independently. The overall similarity is checked, and the atlas is  
70 rotated until the angle providing the maximal mapping similarity is found. This method is  
71 claimed to have similar or even better accuracy than previous algorithms for this task.

72 Even though these works are mainly focused on the static brain anatomy, there is also interest  
73 regarding the dynamics of the brain and how it evolves across time. For example, an anatomical  
74 dynamic brain atlas of the mouse was built by using brain scans of six mice at seven time points.  
75 The resulting dynamic atlas has the ability to provide a static atlas at those predefined time points  
76 (Chuang, 2011). The idea of predefined time points was further extended in (Calabrese, 2013)  
77 where a multidimensional atlas is presented that includes various contrast levels for every time  
78 point in addition to the baseline dynamic information at the predefined time points.

79 However, using predefined time points during atlas construction can be a limiting factor not  
80 only in the data acquisition process but also when studying the brain evolution. To bypass this  
81 issue, a method is proposed in (Davis, 2010) which uses kernel regression to synthesize samples  
82 at any arbitrary time points by using all samples that are close the target time point. Other  
83 methods have been proposed like the one in (Liao, 2012) where first a dynamic model is built for

84 each subject before combining all these models to create the final dynamic atlas space.

85     Apart from creating anatomical atlases, these methods can be used to create probabilistic  
86 atlases to estimate prior probabilities for automatic brain segmentation like in (Kuklisova-  
87 Murgasova, 2011) where a 4-dimensional atlas is created based on affine transformations and  
88 gaussian kernels. Using kernels solves the problem of the dependency between data and atlas  
89 time points with the drawback that the resulting atlas time points could have been synthesized  
90 from a variable number of data. This may result in differences in consistency and smoothness  
91 across the atlas time points. One solution is to improve the normal kernel method and use  
92 adaptive kernels instead, as proposed in (Serag, 2012) which allow the same amount of data  
93 samples per synthesized atlas time point to be used.

94     Given the advancements and the flexibility in the atlas construction techniques, atlas could be  
95 a powerful tool for investigating speech production. Earlier studies of speech articulators and  
96 especially the tongue, used to be based on histological analyses (Takemoto, 2001) or tagged cine-  
97 MRI of multiple subjects (Stone, 2001, Parthasarathy, 2007). Later however, some works  
98 exploited the atlas idea to create a motion field atlas of the tongue (Xing, 2017, Woo, 2019) for  
99 analyzing the correlation between the tongue muscles activities (Xing, 2019).

100     Dynamic atlases could provide valuable assistance in the study of speech production because  
101 by construction they involve the static (linked to the speaker anatomy) and dynamic (linked to  
102 the articulatory strategy) variabilities. The second aspect corresponds to rapid geometrical  
103 changes, and consequently changes in the area function which have a strong acoustic impact  
104 (Skordilis, 2017, Takemoto, 2006). In the same conditions atlas techniques could also improve  
105 speech imaging techniques (Fu, 2016) as it would allow low quality images to be captured at  
106 very high frame rate and the acquired image resolution to be increased by registering a high-

107 resolution atlas to them. Indeed, spatio-temporal atlases are usually based on cine MRI to capture  
108 the 3D geometry of the vocal tract and its temporal evolution (Woo, 2015a, Woo, 2015b, Woo,  
109 2018). Such approaches rely on the repetition of a specific sentence to create the atlas. The  
110 underlying hypothesis is that the subject repeats the same sentence several times in exactly the  
111 same way, which requires prior training to speak by following a metronome. Additionally, the  
112 resulting atlas frame rate is fully dependent on the cine MRI acquisition frame rate.

113 In the present work, we propose a method for constructing 3D dynamic atlases of the vocal  
114 tract using real time MRI (rtMRI) of parallel sagittal planes at a high frame rate, without  
115 requiring prior training. The main question addressed is whether it is possible to reduce  
116 speakers' inter- and intra- variability by using the atlas space as a standard generic speaker.  
117 One of the contributions of our work is to employ the histological atlas creation approach  
118 (Xiong, 2018) to collect the 3D information, using rtMRI to acquire data, which offers a high  
119 frame rate and reduces the amount of repetitions required by other techniques like cineMRI. Such  
120 an approach is new for vocal tract atlases.

121 Another contribution is the use of the adaptive Gaussian kernel technique to create the atlas  
122 samples (Serag, 2012) with the advantage of making the atlas frame rate independent from the  
123 rtMRI frame rate. The proposed method thus gives more flexibility to control the resulting atlas  
124 parameters. Therefore, the same data can be used to create various atlases with different  
125 parameters without the need for new data acquisition every time. Finally, and this is a  
126 determining advantage in studying speech production, the atlas built with this method can be  
127 used as a reference speaker to reduce the variability between and within subjects.

128 Indeed, many works devoted to the production of speech from a general point of view are  
129 based on the implicit assumption that an articulatory model built from a single speaker, which is

130 often the case of the famous Maeda articulatory model (Maeda, 1990), is valid for all  
131 speakers. This is a simplification that reduces the scope and validity of many studies. In our  
132 approach, on the contrary, we have introduced the variability into the construction of the atlas  
133 itself, which therefore effectively covers a large speaker variability, provided that the speakers  
134 used are sufficiently diverse. Throughout the paper the atlas thus refers to a specific model for  
135 a population of 3D (2D on parallel planes) vocal tract dynamic images.

136 In this work a dynamic vocal tract atlas is generated from rtMRI using the new proposed  
137 algorithm and a 4-fold cross validation with histogram matching is used to evaluate whether the  
138 atlas space is a valuable generic speaker model in order to reduce variability between speakers.

139

## 140 2. METHOD

141 Our method for constructing dynamic atlas consists of the following steps:

142 1) **Acquire** 2D dynamic rtMRI parallel sagittal planes of the vocal tract during the  
143 production of several CVs.

144 2) **Create** a subject independent anatomical space based on a silent articulatory configuration  
145 .

146 3) **Use this space** to remove subject's specific anatomical information from the dynamic  
147 images.

148 4) **Combine** the previously created "anatomical neutral" dynamic images to create the  
149 dynamic atlas.

150

### 151 2.1 Subjects

152 Subjects used in this study were four male and four female native speakers of French without



153 any speaking or hearing problems. The average age was 27.25 years with a standard  
154 deviation of 4.23 years.

155

## 156 **2.2 Data acquisition**

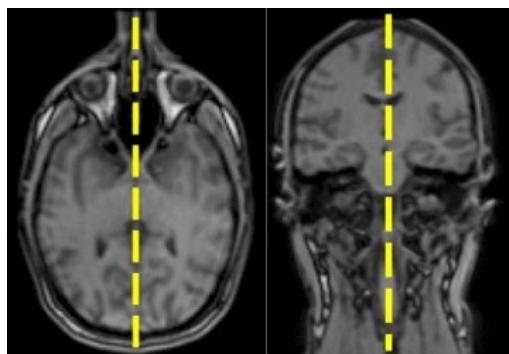
157 The data were acquired on Siemens Prisma 3T scanner (Siemens, Erlangen, Germany) located  
158 in Nancy Central Regional University Hospital under the approved ethical protocol  
159 “METHODODO” (ClinicalTrials.gov Identifier: NCT02887053). For the vocal tract  
160 measurements, 3D data was recorded using a multi-slice 2D T2 turbo spin echo (TR = 4610  
161 ms, TE = 100 ms, flip angle = 15 degrees). The thickness of scan slices is 2 mm, and pixel  
162 bandwidth is 445 Hz/pixel. Subjects were imaged while having the mouth closed and  
163 breathing through the nose. For acquiring dynamic data, we used a 2D rtMRI sequence. Even  
164 though there are 3D dynamic sequences (Lim, 2019), 2D still offers better spatial and  
165 temporal resolutions. In our approach, we used radial RF-spoiled FLASH sequence (Uecker,  
166 2010) with TR = 2.22 ms, TE = 1.47 ms, FOV = 19.2×19.2 cm<sup>2</sup>, spatial resolution  
167 1.41×1.41 mm<sup>2</sup>, flip angle = 5 degrees, and slice thickness is 8 mm. Pixel bandwidth is  
168 1670 Hz/pixel. The number of radial spokes is 9, and the resulting image resolution is  
169 136×136. The acquisition time was 44 sec. Images were recorded at a frame rate of 50 frames  
170 per second with the algorithm presented in (Uecker, 2010), using a 64-channel head-neck  
171 antenna.

172 To capture 3D information with the 2D rtMRI sequence, we relied on the approach  
173 employed to construct brain histological atlases. Since the maximum width of the studied  
174 vocal tracts was 40 mm, we used 5 sagittal planes in total, the mid-sagittal one, two on the left  
175 and two on the right, with 0 mm frame spacing between them. For each subject 5 contiguous

176 sagittal planes (R2, R1, Mid, L1, L2) were acquired covering the whole vocal tract. For each  
177 slice the subject repeated the 12 CV syllables at a natural speed as instructed. To help the  
178 subject to reproduce the CVs in an identical way through the 5 repetitions, the text of the  
179 syllables was projected in the MRI for the duration of the acquisition.

180 As described in (Xiong, 2018) a major issue when dealing with slices is their orientation,  
181 which should be the same for all the speakers. Care was taken, to ensure the exact sagittal  
182 alignment of the midsagittal slice for each subject to avoid misalignment problems previously  
183 reported (Xiong, 2018). A way to solve this issue could have consisted of mapping the slices  
184 to an atlas and correct them afterwards. However, to the best of our knowledge, there does  
185 not exist such an atlas. Therefore, instead of correcting slices, we tackled this issue one step  
186 before, during the real time acquisition step, by using an MRI acquisition protocol designed to  
187 be as strict as we could make it to ensure that every time the target sagittal plane (i.e. R2, R1,  
188 Mid, L1, L2) was exactly the one being acquired.

189 The acquisition protocol was chosen to be as short as possible, keeping in mind that it  
190 should include a periodic check of the subject's initial orientation and correct midsagittal  
191 positioning. The midsagittal plane was defined as the plane which passes in the middle of C2-  
192 C3 (in the coronal view) and separates the 2 brain hemispheres (in the axial plane). An  
193 overview of the midsagittal plane definition can be seen in Fig. 1 and Fig. 2 gives the  
194 overview of the acquisition algorithm.



195

196 *Figure 1: Definition of the midsagittal plane using axial (right) and coronal (left) view*

---

**Algorithm 1** Acquisition scheme

---

**Run** a 3D localizer sequence after having comfortably installed the subject in the machine.

targetPlane  $\leftarrow$  Mid

**setMidPlane** :

**Acquire** 3 groups of 3 slices of the vocal tract. Groups are chosen on perpendicular planes. The midsagittal plane is then defined and a short rtMRI sequence on several perpendicular planes is carried out to verify that the plane is correct.

**Acquire** multislice 2D images used for measuring the vocal tract.

**loop**:

**Acquire** rtMRI data in the targetPlane

**Acquire** a 3D localizer.

if movement is detected between the localizers goto setMidPlane

targetPlane  $\leftarrow$  next(targetPlane)  $\triangleright$  The order of planes is Mid, L1, L2, R1, R2.

if targetPlane  $\leq$  R2 goto loop

---

197

198

*Figure 2: Algorithm for Acquisition*

199

200 This study focused on 12 CV syllables with  $C=\{f, p, s, t\}$  and  $V=\{i, a, u\}$ , i.e. /fi/, /fa/,

201 /fu/, /pi/, /pa/, /pu/, /si/, /sa/, /su/, /ti/, /ta/, /tu/. The choice of these syllables was made so

202 that we have two types of consonants, i.e. stops (/p/, /t/) and fricatives (/f/, /s/), two places of

203 articulation, i.e. labials (/f/, /p/) and alveolars (/s/, /t/), in the context of the cardinal vowels

204 (/i/, /a/, /u/). At this point it is important to note that initially we planned to include also the

205 plosive /k/ in order to cover the three main places of articulation. However, probably due to

206 the supine position in the MRI machine and the force of gravity, some subjects randomly  
207 pronounced either /k/ or /q/ during the acquisition even after proper instructions about the  
208 place of articulation. Given the difficulty of some subjects to accurately produce /k/ through  
209 all the repetitions, we decided to exclude it.

210 To prevent co-articulation effects from previous random vocal tract positions, subjects  
211 were instructed to close the mouth and breath from the nose before articulating every CV so  
212 as to impose the same initial silence position every time. Additionally, the subject was  
213 instructed to finish every CV with /p/ so as to impose a minimal anticipatory coarticulation  
214 effect onto the vowel.

215 We chose /p/ because lips are the closest articulators to the head coil. The signal is thus  
216 stronger, and the image quality is very good for this articulator. Consequently, the contact  
217 between lips which is used as a temporal landmark can be detected with a very good  
218 accuracy. Therefore, in practice, subjects uttered /sil//C//V//p/.

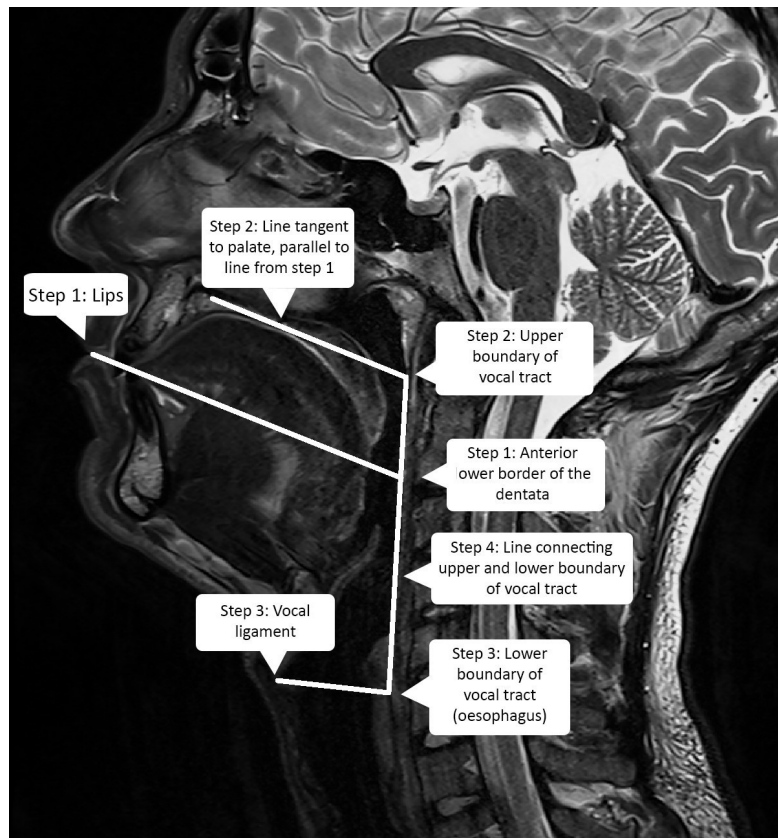
219

### 220 **2.3 Vocal tract measurements**

221 A practical way to increase the probability that subjects have different vocal tract sizes,  
222 without measuring it directly, is to measure their height before including them in our  
223 experimental protocol (Roers, 2009). The smallest subject was 160 cm while the tallest was 187  
224 cm (average 174 cm).

225 In order to assess ability of the atlas to be used as a standard generic speaker model we  
226 measured vocal tract dimensions of included subjects to ensure that there is enough variability in  
227 the dataset. Although several methods have been proposed, for instance by using relative vocal  
228 tract/head position (Perry et al.2017) or automatic articulatory landmark extraction (Eslami,

229 2020) there is no standard method for measuring the vocal tract in terms of height, length and  
230 depth since there is no strict definition of those measures due to the complexity of the vocal tract  
231 shape, which depends on the position, the articulated phoneme, etc. Therefore, we proposed the  
232 following method to measure the length and height of the vocal tract. It uses the midsagittal  
233 plane and the first step is to draw a line from the outer touching point of the lips towards the  
234 anterior lower border of the body of the axis vertebra (Fig. 3).



235

236

*Figure 3: Vocal tract measurements algorithm*

237

238 The segment from the lips up to the intersection with the pharyngeal wall is defined as the  
239 length of the buccal cavity. The second step is to draw a line, parallel to the previous one and  
240 tangent to the palate. The intersection point between this line and the pharyngeal wall is defined  
241 as the upper boundary of the vocal tract. The third step is to draw a line from the platform of the

242 vocal folds until the esophagus. This point at the upper part of esophagus is defined as the lower  
243 boundary of the vocal tract. The height of the vocal tract is defined as the distance between its  
244 lower and upper boundaries (Fig. 3). To estimate the width of the vocal tract all the sagittal  
245 planes are scanned and the number of planes where the vocal tract is visible at the bottom of the  
246 pharyngeal cavity gives the width of the vocal tract. Table I shows the measurements for our  
247 group of subjects. The difference between the shortest and longest measure is 22 mm ( $\sigma = 6.5$   
248 mm) for the buccal cavity length and 25 mm ( $\sigma = 8.6$  mm) for the height, i.e. more than 25 %  
249 of these dimensions approximately. For the purpose of our task we thus consider that these sizes  
250 exhibit sufficient variability (Roers, 2009). Fig. 4 shows the “silence” frames from all the  
251 speakers in the dataset.

252

253

255

256

257

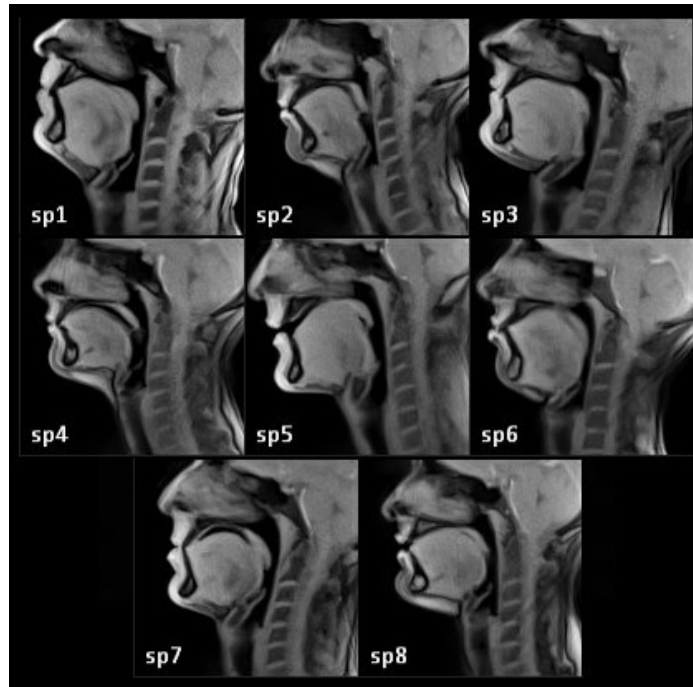
258

259

260

261

262



263 *Figure 4: Midsagittal (M) frames for silence for all speakers (sp1-sp8 left to right, top down).*

264

*sp{odd} are male and sp{even} are female speakers*

## 265      **2.4 Atlas construction**

266      The acquired dynamic films were manually labeled in order to achieve a better temporal  
267      segmentation. Image labelling was done by a person with around 5 years of experience working  
268      with this type of image and were then checked by an expert with more than 15 years of  
269      experience in the field. For every /sil//C//V//p/ we only kept the /C/ and the /V/ part.

270      The stop onset is the first image where there is a contact between the tongue tip and teeth for  
271      /t/, contact between lips for /p/ and negligible lip movement for /f/ and negligible tongue tip  
272      movement for /s/. The vowel onset is the first image where the constriction is released, i.e. there  
273      is no more contact between the tongue tip and teeth for /t/, and no more contact between lips for  
274      /p/, or the first image where there is increased lip movement for /f/ or the tongue tip for /s/. The  
275      vowel offset corresponds to the first image where lips are in contact because the subjects were  
276      instructed to articulate a /p/ after the second vowel. The average duration (number of frames at  
277      50 Hz and in ms) per phoneme across all planes and speakers is given in Table II.

278      The proposed construction algorithm relies on three hypotheses. First, all the slices are in the  
279      expected plane. For instance, all the central slices are in the mid-sagittal plane and all the other  
280      sagittal slices are shifted from the mid-sagittal plane accordingly. This is a direct consequence of  
281      the very strict acquisition protocol we designed, and the anatomical position we chose. As a  
282      consequence, images of one given plane and speaker can be compared and mapped with the  
283      corresponding images of all the other speakers. Anatomical differences between speakers could  
284      potentially affect this hypothesis all the more since a potential error can stack as one moves  
285      further from the midsagittal plane. However, we expect this error not to be significant because  
286      we moved just two slices away at most from the mid-sagittal plane and the slice thickness was  
287      big enough so that the outer parts of the vocal tract (in the sagittal direction) will lie within the R2

288 and L2 planes for all subjects.

289 The second hypothesis is that the order of events is the same for all the speakers, which is  
290 expected and reasonable at the scale of an isolated CV.

291 Third, due to the frame rate of 50 Hz, small piece-wise linear extensions or compressions of  
292 the images in time are not affecting significantly the dynamics of articulation.

293 For describing the construction of the atlas silence space, we will refer to the midsagittal  
294 plane for simplicity unless it is specified differently. The process presented below for the  
295 midsagittal plane is repeated for all the other planes. Before every image transformation or  
296 averaging in this work, histogram matching is performed to transform the histogram of the  
297 moving image to the one of the reference images. This is intended to compensate for intensity  
298 differences between images (Seghers, 2004).

299 The atlas construction process can be divided in four major steps:

- 300 1) **Create** the anatomically-free reference space.
- 301 2) **Make** dynamic data anatomically free.
- 302 3) **Align** data temporarily.
- 303 4) **Synthesize** the atlas samples.

304 The objective of step 1 is to make the data anatomically neutral. By anatomically neutral we  
305 mean that data are independent of anatomical variability and correspond to a virtual neutral  
306 speaker. For this purpose, we used a silence frame during breathing, at a resting position before  
307 speakers start recording the CV (as described in the protocol, i.e. breathing from the nose with  
308 closed mouth and without any visible articulatory movement) from all N speakers in order to  
309 create the reference anatomically free space. The average histogram was computed and all the  
310 images' intensities were transformed so that their histogram will match with it (Rueckert, 1999).



311 For image registration, the transform used ( $T(x,y)$  with  $x,y$  being the image coordinates) is  
 312 composed of two parts, the global and the local one.

$$313 \quad T(x, y) = T_{global}(x, y) + T_{local}(x, y) \quad (1)$$

314 In our case an affine transformation was used for  $T_{global}(x,y)$  and a cubic B-spline tensor  
 315 product on control point grid transformation for  $T_{local}(x,y)$  (Lee, 1997). Therefore

$$316 \quad T_{local}(x, y) = \sum_{l=0}^3 \sum_{m=0}^3 B_l(u) B_m(v) \phi_{i+l, j+m} \quad (2)$$

317 where  $\phi_{i,j}$  are the control points with  $\delta_x, \delta_y$  the spacing between them

$$318 \quad i = \lfloor x/\delta_x \rfloor - 1 \quad (3)$$

$$319 \quad j = \lfloor y/\delta_y \rfloor - 1 \quad (4)$$

$$320 \quad u = x/\delta_x - \lfloor x/\delta_x \rfloor$$

$$321 \quad (5)$$

$$322 \quad v = y/\delta_y - \lfloor y/\delta_y \rfloor \quad (6)$$

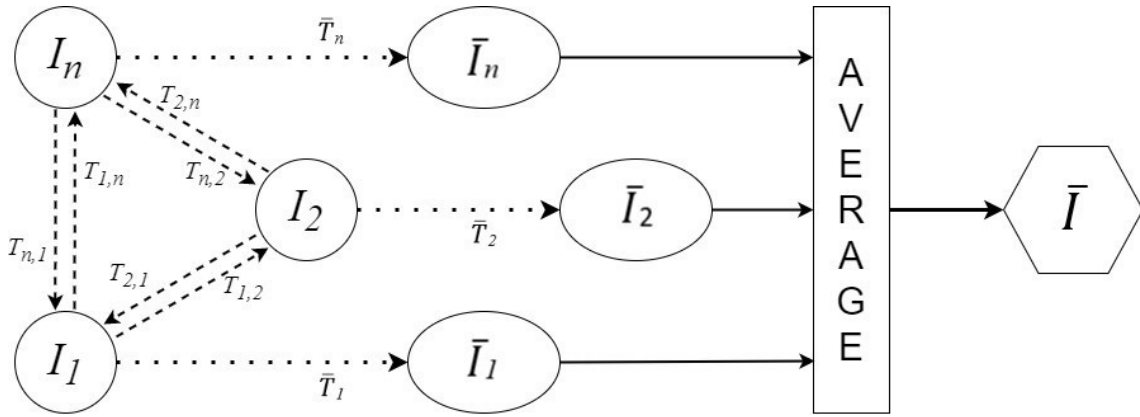
323 and  $B_l, B_m$  is the  $l^{th}$  and  $m^{th}$  B-spline base function (Lee et al.1996). Each image was  
 324 registered to all other N-1 speakers' images using the described non-rigid B-spline based  
 325 transformation using the image\_registration function of the MATLAB toolbox "B- spline Grid,  
 326 Image and Point based Registration" (Kroon, 2019).

327 This toolbox was used for all the transformations performed in this work. For every  
 328 image we get N-1 transforms. The average transformation (without any further weighting)  
 329 is computed for every image and this average transformation is applied to the corresponding  
 330 image to produce the anatomical free version which is image dependent. Finally, all the N image  
 331 dependent anatomical free spaces are truly averaged to create the final reference space (image  
 332 independent, anatomically neutral).

333 More precisely, for the  $i^{\text{th}}$  silence image from the set of silent images  $\{I_{1..n}\}$  the  
 334 transformations  $T_{i,j}, i \neq j$  are computed and averaged to give the average transformation

335 
$$\bar{T}_i = \frac{1}{N-1} \sum_{j=1..n, i \neq j} T_{i,j}$$
 Finally, the final reference space is created by applying the  $\bar{T}_i$  transforms

336 to the corresponding images and averaging them  $\bar{I} = \frac{1}{n} \sum_{i=1..n} \bar{T}_i(I_i)$  with  $\bar{T}_i(I_i) \approx \bar{I}_i$ . A visual  
 337 representation can be seen in Fig. 5.

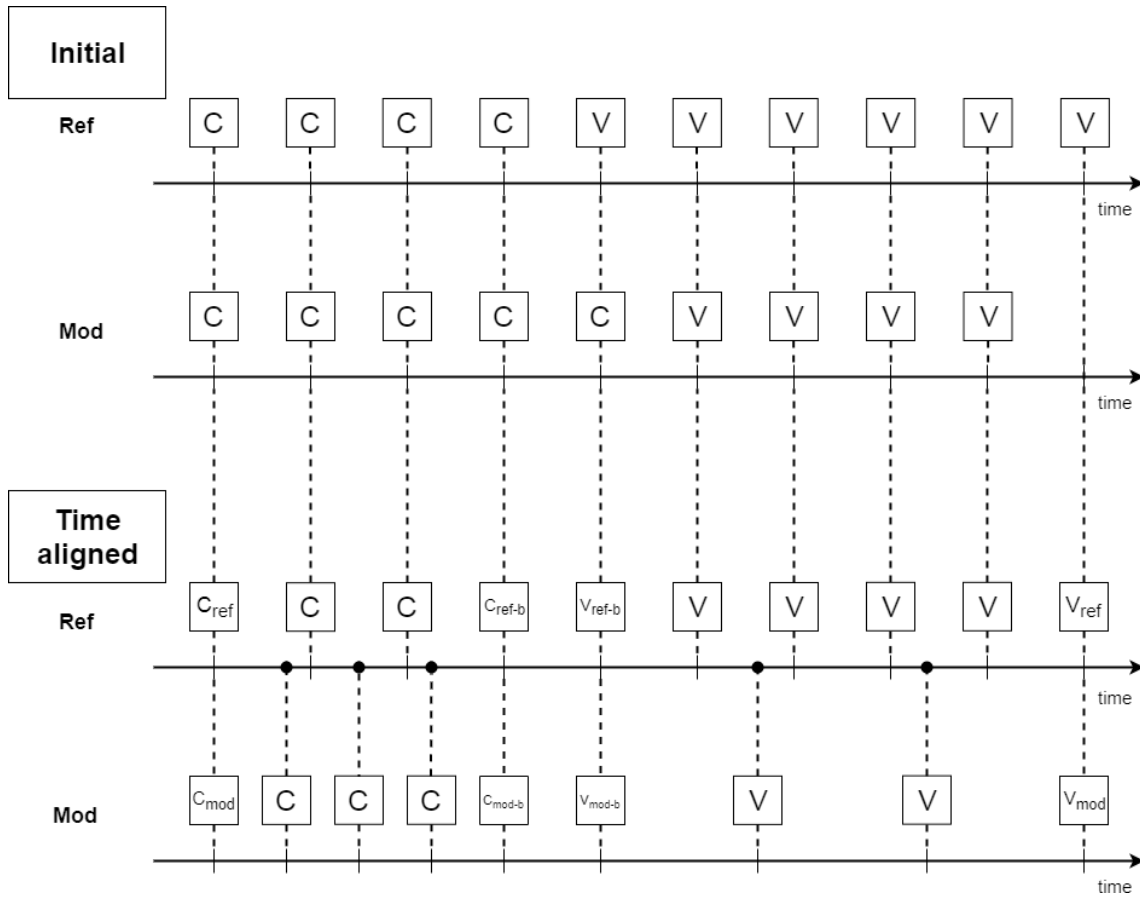


338  
 339 *Figure 5: Creating the reference space. Every  $i^{\text{th}}$  silence image is registered to all others,*  
 340 *the computed transformations are averaged to give  $\bar{T}_i$  and applied to the  $i^{\text{th}}$  image to get  $\bar{I}_i$ .*  
 341 *The resulting images are averaged to get the final reference space image  $\bar{I}$ .*

343 Step 2 is intended to make the data anatomically free. First, the images' histogram of all  
 344 the CVs is matched with the histogram of the reference and all the CV images are then  
 345 transformed to the reference space using only an affine transformation (one for each image of all  
 346 the CV images of all the speakers) computed with the same MATLAB function as in Step 1  
 347 because it transforms the anatomy of the data to the reference anatomy but keep the vocal tract  
 348 position variability, i.e. the position of the articulators (Kuklisova-Murgasova, 2011).

349 Step 3 is intended to process the anatomical free data for applying the adaptive kernel  
350 technique. For each CV, all the planes of all the speakers were used to specify the corresponding  
351 average C and V duration. These values are set as the time reference durations for each of the C  
352 and V of the atlas. Data are then piece-wise linearly aligned to those CV time duration values  
353 using rtMRI frame rate to pass from the frame space to the time domain in order to compute the  
354 global time.

355 For example, in order to align a CV to be modified to a reference CV, the C and V parts of  
356 the modified CV are independently and linearly extended or compressed until the duration of  
357 both C and V of the modified CV match with those from the reference CV. This alignment  
358 technique (see Fig. 6) is intended to achieve time alignment so as the duration of the modified  
359 (Mod) CV is that of the reference (Ref) CV, but not to map each frame of the reference CV to  
360 one of the current CV. In practice, this procedure creates one anatomical free image series for  
361 each of the 12 CVs from the image series of all speakers for the same CV, by putting all frames  
362 in a global time scale based on the time stretching or compressing defined by the piece-wise  
363 linear alignment. It should be noted that the resulting series may have multiple frames at one  
364 time point and that samples are not homogeneously distributed across time.

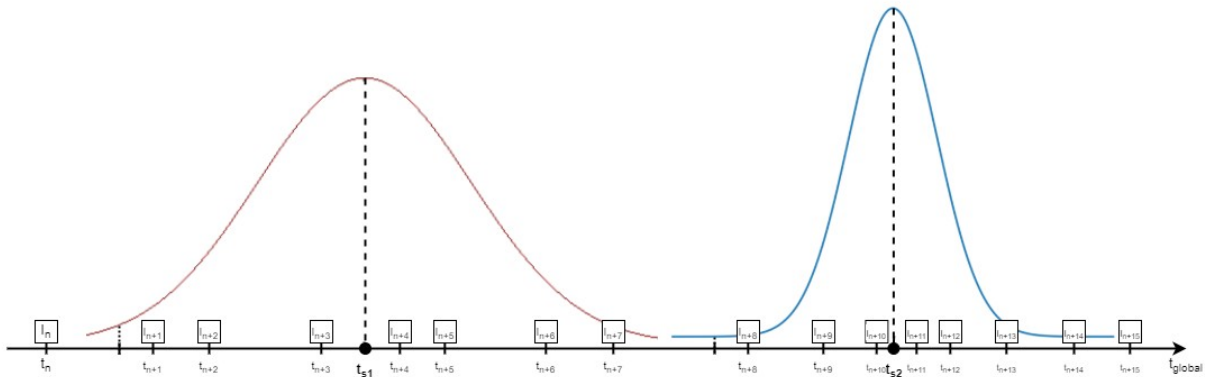


365

366 *Figure 6: Piece-wise time alignment. Mod is the CV which duration is to be modified in order to*  
 367 *match the duration of the reference (Ref) CV. On the top are both CVs before time alignment*  
 368 *(Initial) and on the bottom the time aligned version of the Mod CV with the Ref CV*

369 Step 4 consists of synthesizing the atlas images from the global series of images, i.e. the 12  
 370 CVs involved in this work, by using the adaptive Gaussian kernel method (Serag, 2012). The  
 371 word “adaptive” refers to the width of the Gaussian kernel so that the same number of samples  
 372 will be used every time. The core idea is to generate the atlas image at a given target time point  
 373 from  $k$  images in the global series located in the vicinity of the target time point.  $k$  is a pre-  
 374 specified number of samples to choose the closest relevant samples and the resulting image is the  
 375 Gaussian weighted average of the  $k$  samples. This way, the resulting synthesized images are  
 376 sharper and less blurry.

377 The advantages are that the atlas frame rate is independent of the data acquisition frame rate  
378 and that the atlas sampling may not be regular since the time points can be chosen freely.  
379 Theoretically, the initial sampling rate has some influence, but the initial frame rate is high  
380 enough to study all common speech tasks (Lingala, 2016). However, the number of samples used  
381 to synthesize the images and the parameters of the Gaussian weights should be tuned. In (Serag,  
382 2012) the number of samples was chosen as a function of the number of subjects available in the  
383 vicinity of the target time point and could vary substantially, i.e. from 3 to 25, because the  
384 number of subjects recorded depended on time and the phenomenon monitored was much slower.  
385 Thus, when many subjects were available the gaussian was sharp, and conversely wider when  
386 fewer subjects were available. In our case the number of subjects is constant, i.e. 6, and  
387 consequently the number of samples available is almost constant if we consider that the dynamic  
388 variability is limited. We tested several choices and set  $k$  to 7 atlas samples within a window of  
389 20 ms, which is the recording period and is expected to be sufficient for our study (Lingala,  
390 2016). The Gaussian weighting was designed so that its mean value is the selected time point  $\tau$  to  
391 be synthesized and the standard deviation was tuned so that the weight of the farthest  $k$  sample  $\tau_f$   
392 from the center is 0.35 of the maximum value of the Gaussian distribution. Therefore, the  
393 parameters of the Gaussian distribution are  $\mu = \tau$  and  $\sigma = \sqrt{-(\tau - \tau_f)^2 / (2 * \ln(0.35))}$  (Serag et  
394 al.2012). This approach is illustrated in Fig 7.



395

396 *Figure 7: Adaptive Gaussian kernel technique. The width of the Gaussian is adapted based on*  
 397 *the distance between the desired synthesis time points ( $ts_1, ts_2$ ) with the available samples  $I_i$ . The*  
 398 *number of the samples contributing to frame generation is stable*

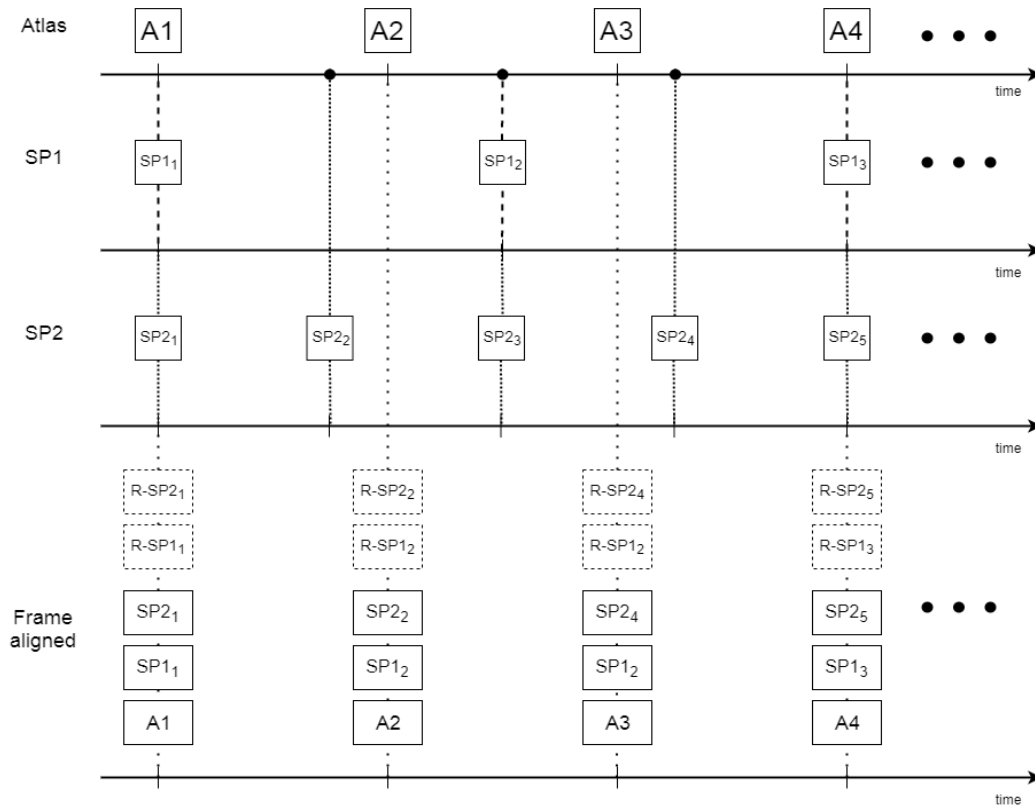
399

### 400 3. VALIDATION

401 To evaluate the results 4 fold cross validations were carried out using 6 subjects for training  
 402 and 2 subjects for testing for every fold. In every fold the two test subjects were chosen to be of  
 403 different gender to get results for both genders. Both of the test CVs are piece-wise linearly  
 404 temporally aligned with the corresponding atlas CV. For each frame of each atlas CV the  
 405 temporally closest frame of the corresponding test CV is selected. It is thus possible for a test  
 406 frame to be used more than once while some others may not be used at all. At this point, for each  
 407 CV each atlas frame is linked to two frames of the corresponding CV, i.e. one for the two test  
 408 subjects.

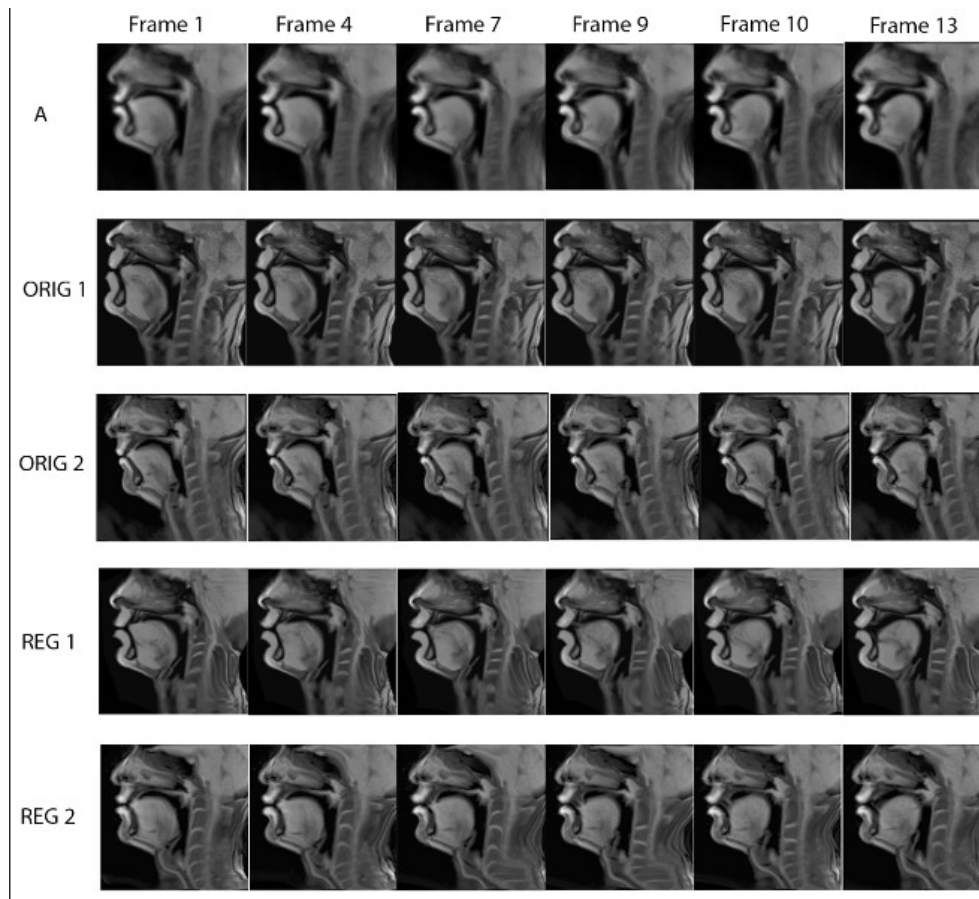
409 All the frames linked with the same atlas frame form a stack of images as seen in Fig. 8. Each  
 410 stack includes an atlas image and the corresponding images of: **(i)** speaker 1 image without  
 411 registration, **(ii)** speaker 2 image without registration, **(ii)** speaker 1 image after registration, **(iv)**  
 412 speaker 2 image after registration. Examples of every stack of images in the midsagittal plane

413 can be seen in Fig. 9. Histogram matching is applied so that the histograms of the linked images  
 414 with one atlas frame fit its histogram. Test images are mapped to the atlas image using the B-  
 415 spline non-rigid transformation (the same technique as that used for construction). For example,  
 416 images of row A from Fig. 9 are the reference images of the atlas. Images from row ORIG 1 and  
 417 ORIG 2 are mapped to those of row A and the resulting images are shown in rows REG 1 and  
 418 REG 2. The similarity between the original images (ORIG 1 and ORIG 2) and those of row A for  
 419 all frames of all planes is computed. The similarity between the transformed images (REG 1 and  
 420 REG 2 rows) is calculated as well to check that the similarity increased after registration.



421  
 422 *Figure 8: Frame alignment used for tests. A represents the atlas frames and  $SP_{ij}$  original frames*  
 423  *$j$  for speaker  $i$  and  $R-SP_{ij}$  the registered framed within the atlas space.*

424



425

426 *Figure 9: The midsagittal frames of the atlas with the corresponding test subject frames before*  
 427 *and after transformation with the atlas.*

428

429 The idea of this procedure is to transform any given image of a target speaker CV as close as  
 430 possible to the corresponding atlas image. We use cross-correlation as a similarity measurement  
 431 between images mapped from the atlas and original images (Serag et al.2012). The cross-  
 432 correlation value is normalized by the auto-correlation of the atlas frame. More precisely, for  
 433 each stack of images A is an atlas image, O1 and O2 the original images of speaker 1 and  
 434 speaker 2, and R1 and R2 the corresponding registered images to the atlas. All images represent  
 435  $M \times W$  matrices of pixel density values with M, W being image dimensions. Before registration  
 436 with the atlas (BA) the similarity (with zero-padding) is defined as:



437 
$$BA = \frac{\max_{m=0}^{M-1} \sum_{w=0}^{W-1} O_1(m, w) O_2(m-k, w-l)}{\max_{m=0}^{M-1} \sum_{w=0}^{W-1} A(m, w) A(m-f, w-g)}$$

438 With

439 
$$-(M-1) \leq k, f \leq M-1$$

440 
$$-(W-1) \leq l, g \leq W-1$$

441 After registration The similarity (with zero-padding) is defined as:

442 
$$AA = \frac{\max_{m=0}^{M-1} \sum_{w=0}^{W-1} R_1(w, n) R_2(w-t, w-c)}{\max_{m=0}^{M-1} \sum_{w=0}^{W-1} A(m, w) A(m-f, w-g)}$$

443 With

444 
$$-(M-1) \leq t, f \leq M-1$$

445 
$$-(W-1) \leq c, g \leq W-1$$

446 These measurements are averaged across space and time in order to produce Table III.

447 Columns 2 and 4 are the averages of BA and AA respectively and column 3 and 5 are the  
448 corresponding standard deviations.

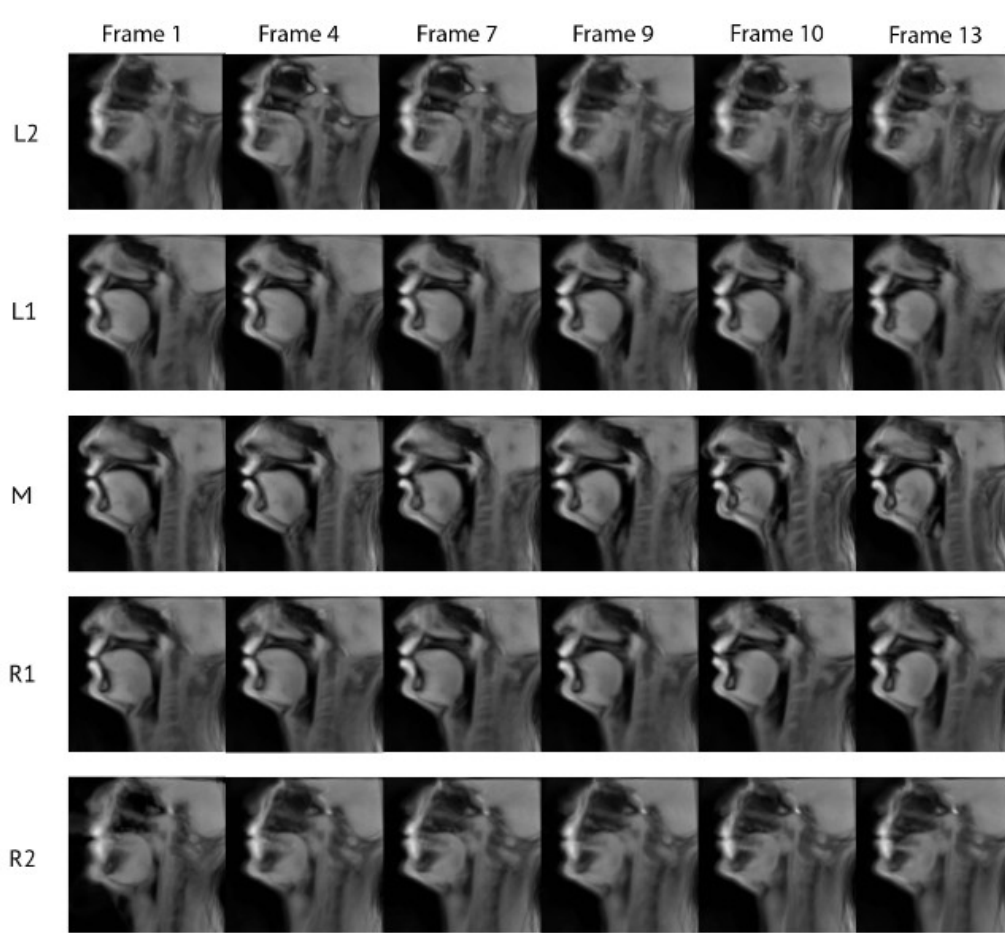
449

#### 450 4. RESULTS

451 The methods presented above regarding the atlas construction were applied to the acquired  
452 data on all 5 planes. During the atlas construction process, small time variations appeared  
453 during the various registration processes due to the fact that by nature some speakers are  
454 anatomically more similar/different from each other. Fig. 10 present examples of frames from all  
455 sagittal planes in the atlas space for /tu/. The visual assessment confirms that the synthesized  
456 images represent the natural vocal tract position with the expected dynamics. This is further

457 quantitatively supported by the numerical results of Table III. As it can be seen from Table III,  
458 the average similarity between the images after applying the atlas is increased while the  
459 standard deviation decreases (col. 4, 5) compared to the similarity and the standard deviation  
460 without the atlas (col. 2, 3). Fig. 9 shows the midsagittal frames of the atlas with the  
461 corresponding frames of the test subjects before and after atlas transformation. The places of  
462 articulation are clear for both /t/ and /u/.

463



464

465 *Figure 10: Frames 1, 4, 7, 9, 10, 13 of the atlas planes without sp5, sp6 for /tu/*

466

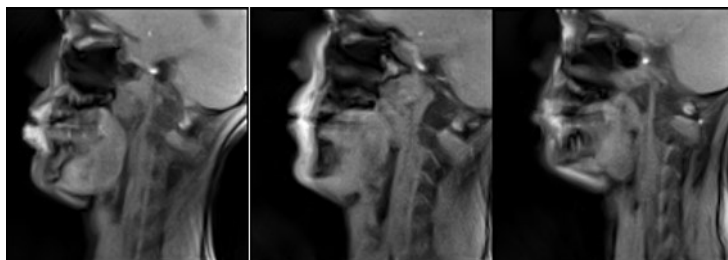
467 We can see the dynamics of the tongue starting from the very beginning of /t/ where the

468 tongue presses the alveolar region up until the end where the tongue tip is lowered for the  
469 production of /u/. Fig. 10 show the temporal evolution of the articulator positions in the five  
470 planes. For example, by visually comparing the tongue position between midsagittal and adjacent  
471 planes (e.g. frame 9), one can notice that the tongue is lower in the midsagittal plane near the  
472 teeth region. Additionally, for most of the images of R1 and L1 planes lips are almost closed, in  
473 contrast to the midsagittal plane where they are clearly open. This information cannot be derived  
474 from the midsagittal frames alone. The results of the normalized image similarity before and after  
475 the application of atlas are presented in Fig. 9.

476

## 477 5. DISCUSSION

478 Images of the R2 and L2 planes are blurrier compared to the other planes due to the fact that  
479 the original images of the speakers at that plane (Fig. 11) suffer from a “partial volume effect”  
480 (Ballester, 2002). Indeed, the slice thickness is 8 mm and when moving away from the  
481 midsagittal plane, the volume of one pixel may correspond to a mixture between more than one  
482 type of tissue (muscles, fat, teeth) and air, which give rise to some blurring (see Fig. 10 row 5).  
483 However, one can still extract useful information about the movement of articulators like the  
484 tongue body.



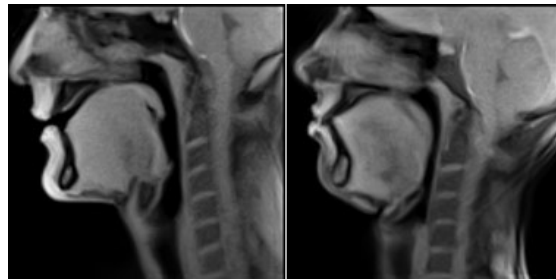
485

486 *Figure 11: Original L2 frames during /u/ for speakers 6-8 (left to right). One can notice that*  
487 *images in this plane are a bit more blurry compared to the midsagittal plane (Fig. 10 row 5)*

488

489 By comparing the atlas images against the individual subject's images, one can notice that  
490 atlas images are less sharp. This could be due to histogram matching that took place before every  
491 image transformation, or to the initial histogram matching of all the silence frames with their  
492 average histogram. It could also be due to the interpolation kernel during the spatial transform or  
493 because of the image averaging procedure both during silence creation and during the atlas  
494 sample generation. Additionally, another reason is that at step 2 of the atlas construction process  
495 (when the subject independent anatomical space is created) there is some loss of sharpness due to  
496 anatomical and head posture differences (Fig. 12). Even though the reference silence image does  
497 not look strongly connected with the final atlas synthesized images, any loss of sharpness could  
498 further propagate. Indeed the silence frame was used as a reference to match the histograms and  
499 was also used to transform all the dynamic data of all subjects in order to remove subjects'  
500 anatomical information and create "anatomically neutral" dynamic data.

501



502

503 *Figure 12: Silence frames for two speakers. One can see that more vertebra are visible for*  
504 *speaker 5 (left) compared to speaker 6 (right)*

505

506 The second noticeable point is that the spine is not very sharp in some cases for two reasons.  
507 This region is also affected by the general loss of sharpness, but the main reason is that posture  
508 and anatomical differences between subjects, especially between males and females result in that

509 more vertebra are visible for some subjects and less for others (see Fig. 9 row 3). This probably  
510 affects the transformation algorithm since these extra vertebrae have no place to be directly  
511 mapped. They are therefore compressed, or extended in the opposite case, within the spine.  
512 However, we can see that the main articulators like the tongue are not strongly affected. Even if  
513 there is no objective criterion that specifically focuses on the articulators since every image was  
514 treated as a whole this behavior was expected because all the images contained the whole vocal  
515 tract and thus the impact of moving articulators is indirectly stronger on the transformations  
516 computed compared to that of some vertebra (C6) that sometimes appears and sometimes not.  
517 Additionally, the similarity criterion that was used for image registration (Rueckert, 1999) is  
518 mutual information which further supports the visual observations.

519 The first use of atlas concerns the highlighting of average or speaker-specific articulatory  
520 strategies. The measurement of the similarity between the speaker's images registered on the  
521 atlas and the atlas images is a way to detect these articulatory strategy deviations. The second  
522 potential use concerns the study of the dynamic 3D area function (Takemoto, 2006) since it  
523 allows the use of one representative subject, i.e. the atlas, instead of one random subject. The  
524 advantage is that one could use the method proposed by those authors directly on the atlas  
525 in order to get generic results, preventing us from having to extract area functions from  
526 several subjects and then combine them, which is the common strategy so far.

527 Another use of the atlas concerns the transformation of 2D rtMRI videos into 3D dynamic  
528 videos (Douros, 2019, Douros, 2020) since the atlas incorporates the real 3D dynamic  
529 information that occurs during the production of continuous speech, and not just estimates it from  
530 static 3D and midsagittal rtMRI. By using the atlas one can directly extract the 3D shape of the  
531 vocal tract by using the stacks of the parallel sagittal images and use them to calculate

532 transformations from the midsagittal plane to the parasagittal planes. They can be used to find  
533 estimations of the 3D dynamic shape of the vocal tract by using only the midsagittal plane. Such  
534 videos would allow the complex tongue constriction events to be investigated in depth(Lim,  
535 2019).

536 Automatic tracking of the vocal tract contours (Labrunie, 2018, Takemoto, 2019) could also  
537 take advantage of the atlas to map a specific subject data whose data have to be delineated. The  
538 main advantage is that once the atlas is created, it could be used to process new rtMRI data  
539 without requiring every time data pre-processing, retraining models etc. Finally, the main  
540 contribution of this work is that the atlas is a true golden speaker which embodies speaker  
541 independent articulatory gestures.

542

## 543 **6. CONCLUSION**

544 To summarize, this paper presents a method for creating a dynamic 3D atlas of the vocal tract  
545 that can be used as a reference space for studying speech production. 2D rtMRI data on parallel  
546 planes were combined using piece-wise linear alignment and adaptive Gaussian kernel method to  
547 synthesize the images of the final atlas. The main contribution is to incorporate the speaker  
548 variability directly in the construction of the atlas. This approach almost removes inter-speaker  
549 variability of the resulting space, therefore providing a generic speaker model. Since any speaker  
550 can be “projected” onto this generic speaker a direct extension will consist in transforming one  
551 speaker into another using the atlas as a pivot with the anatomical adaptation on one hand and  
552 the temporal adaptation, i.e. finer articulatory strategy aspects, on the other hand. This could be  
553 particularly useful to exploit resources which do exist for one or a few speakers only. For  
554 instance, when 3D area functions have been acquired for one speaker the mapping between this

555 speaker and the generic speaker gives a mapping that can then be used for any speaker by using  
556 the generic speaker as a pivot. This solution gives a more robust mapping than what could be  
557 done for each pair of speakers independently. Another application would consist of investigating  
558 language specific articulatory strategies by exploiting atlases built for several languages. The  
559 comparison of the language atlases would enable invariant articulatory features imposed by  
560 anatomy to be separated from language specific strategies.

561 A limited number of CVs was involved in this study and an ambitious perspective would be  
562 to incorporate all the phonetic contexts of a language, i.e. all VCVs, CVs, CCVs..., in order to be  
563 able to exhaustively cover the articulation of the target language. The recording of all the  
564 contexts required for 8 speakers, 5 planes, together with the corresponding fine temporal  
565 annotations required to build the global atlas is unrealistic. A perspective thus would consist of  
566 defining a minimal set of sequences used to build an atlas which would nevertheless be able to  
567 cover exhaustively the articulation of the target language, and provide efficient coarticulation  
568 modeling as well.

569        **ACKNOWLEDGMENTS**

570    Research supported by the project ArtSpeech of ANR (Agence Nationale de la Recherche),  
571    France, CPER "IM2MP", "LCHN" and FEDER. We thank Agne`s Basile, research technician  
572    at CIC-IT, Nancy, France for her help with the construction of the acquisition protocol,  
573    Athanasios Katsamanis for the inspiring discussions and Valerie Parmentier, operator of the MRI  
574    system at CHRU Brabois Hospital.

575



576 **REFERENCES**

- 577 Agarwal, N., Xu, X., and Gopi, M. (2016). Robust registration of mouse brain slices with  
578 severe histological artifacts. In *Proceedings of the Tenth Indian Conference on Computer  
579 Vision, Graphics and Image Processing*, page 10. ACM.
- 580 Ballester, M. Á. G., Zisserman, A. P., & Brady, M. (2002). Estimation of the partial  
581 volume effect in MRI. *Medical image analysis*, 6(4), 389-405.
- 582 Calabrese, E., Badea, A., Watson, C., and Johnson, G. A. (2013). A quantitative  
583 magnetic resonance histology atlas of postnatal rat brain development with regional  
584 estimates of growth and variability. *Neuroimage*, 71:196–206.
- 585 Chuang, N., Mori, S., Yamamoto, A., Jiang, H., Ye, X., Xu, X., Richards, L. J., Nathans,  
586 J., Miller, M. I., Toga, A. W., et al. (2011). An mri-based atlas and database of the  
587 developing mouse brain. *Neuroimage*, 54(1):80–89.
- 588 Davis, B. C., Fletcher, P. T., Bullitt, E., and Joshi, S. (2010). Population shape regression  
589 from random design data. *International journal of computer vision*, 90(2):255–266.
- 590 Douros, I., Tsukanova, A., Isaieva, K., Vuissoz, P.-A., and Laprie, Y. (2019). Towards a  
591 method of dynamic vocal tract shapes generation by combining static 3d and dynamic 2d  
592 mri speech data. INTERSPEECH 2019
- 593 Douros I., Kulkarni A., Xie Y., Dourou C., Felblinger J., Isaieva K., Vuissoz P.-A., and  
594 Laprie Y. (2020). MRIvocal tract sagittal slices estimation during speech production of  
595 cv, in 28th European Signal Processing Conference (EUSIPCO 2020), 2020
- 596 Ericsson, A., Aljabar, P., and Rueckert, D. (2008). Construction of a patient-specific  
597 atlas of the brain: Application to normal aging. In *2008 5th IEEE International*

598 *Symposium on Biomedical Imaging: From Nano to Macro*, pages 480–483. IEEE.

599 Eslami, M., Neuschaefer-Rube, C., and Serrurier, A. (2020). Automatic vocal tract landmark  
600 localization from midsagittal MRI data. *Scientific Reports*, 10(1):1–13.

601 Fu, M., Woo, J., Liang, Z.-P., and Sutton, B. P. (2016). Spatiotemporal-atlas-based  
602 dynamic speech imaging. In *Medical Imaging 2016: Biomedical Applications in*  
603 *Molecular, Structural, and Functional Imaging*, volume 9788, page 978804.  
604 International Society for Optics and Photonics.

605 Gousias, I. S., Rueckert, D., Heckemann, R. A., Dyet, L. E., Boardman, J. P., Edwards, A.  
606 D., and Hammers, A. (2008). Automatic segmentation of brain MRIs of 2-year-olds into  
607 83 regions of interest. *Neuroimage*, 40(2):672–684.

608 Kroon Dirk-Jan (2019). B-spline Grid, Image and Point based Registration  
609 ([https://www.mathworks.com/matlabcentral/fileexchange/20057-b-spline-grid-image-](https://www.mathworks.com/matlabcentral/fileexchange/20057-b-spline-grid-image-and-point-based-registration)  
610 [and-point-based-registration](https://www.mathworks.com/matlabcentral/fileexchange/20057-b-spline-grid-image-and-point-based-registration)), MATLAB Central File Exchange. Retrieved May 15, 2019.

611 Kuklisova-Murgasova, M., Aljabar, P., Srinivasan, L., Counsell, S. J., Doria, V., Serag, A.,  
612 Gousias, I. S., Boardman, J. P., Rutherford, M. A., Edwards, A. D., et al. (2011). A dynamic 4d  
613 probabilistic atlas of the developing brain. *NeuroImage*, 54(4):2750–2763.

614 Labrunie, M., Badin, P., Voit, D., Joseph, A. A., Frahm, J., Lamalle, L., Vilain, C., and Boë, L.-  
615 J. (2018). Automatic segmentation of speech articulators from real-time midsagittal MRI based  
616 on supervised learning. *Speech Communication*, 99:27–46.

617 Lee, S., Wolberg, G., Chwa, K. Y., & Shin, S. Y. (1996). Image metamorphosis with scattered  
618 feature constraints. *IEEE transactions on visualization and computer graphics*, 2(4), 337-354.

619 Lee, S., Wolberg, G., & Shin, S. Y. (1997). Scattered data interpolation with multilevel B-

620 splines. *IEEE transactions on visualization and computer graphics*, 3(3), 228-244.

621 Liao, S., Jia, H., Wu, G., Shen, D., Initiative, A. D. N., et al. (2012). A novel framework for  
622 longitudinal atlas construction with groupwise registration of subject image sequences.  
623 *NeuroImage*, 59(2):1275–1289.

624 Lim, Y., Zhu, Y., Lingala, S. G., Byrd, D., Narayanan, S., and Nayak, K. S. (2019). 3d dynamic  
625 MRI of the vocal tract during natural speech. *Magnetic resonance in medicine*, 81(3):1511–1520.

626 (Lingala et al.2016) Lingala, S. G., Sutton, B. P., Miquel, M. E., and Nayak, K. S. (2016).  
627 Recommendations for real-time speechMRI. *Journal of Magnetic Resonance Imaging*,  
628 43(1):28–44.

629 Maeda, S. (1990). Compensatory articulation during speech: Evidence from the analysis and  
630 synthesis of vocal-tract shapes using an articulatory model. In *Speech production and speech*  
631 *modelling*, pages 131–149. Springer.

632 Parthasarathy, V., Prince, J. L., Stone, M., Murano, E. Z., and NessAiver, M. (2007). Measuring  
633 tongue motion from tagged cine-mri using harmonic phase (harp) processing. *The Journal of the*  
634 *Acoustical Society of America*, 121(1):491–504.

635 Perry, J. L., Kuehn, D. P., Sutton, B. P., and Fang, X. (2017). Velopharyngeal structural and  
636 functional assessment of speech in young children using dynamic magnetic resonance imaging.  
637 *The Cleft Palate- Craniofacial Journal*, 54(4):408–422.

638 Roers, F., Mu¨rbe, D., and Sundberg, J. (2009). Voice classification and vocal tract of singers: A  
639 study of x-ray images and morphology. *The Journal of the Acoustical Society of America*,  
640 125(1):503–512.

641 Rueckert, D., Sonoda, L. I., Hayes, C., Hill, D. L., Leach, M. O., and Hawkes, D. J. (1999). Non-

642 rigid registration using free-form deformations: application to breast mr images. *IEEE*  
643 *transactions on medical imaging*, 18(8):712–721.

644 Seghers, D., D’Agostino, E., Maes, F., Vandermeulen, D., and Suetens, P. (2004).  
645 Construction of a brain template from mr images using state-of-the-art registration and  
646 segmentation techniques. In *International Conference on Medical Image Computing and*  
647 *Computer- Assisted Intervention*, pages 696–703. Springer.

648 Serag, A., Aljabar, P., Ball, G., Counsell, S. J., Boardman, J. P., Rutherford, M. A.,  
649 Edwards, A. D., Hajnal, J. V., and Rueckert, D. (2012). Construction of a consistent  
650 high-definition spatio-temporal atlas of the developing brain using adaptive kernel  
651 regression. *Neuroimage*, 59(3):2255–2265.

652 Skordilis, Z. I., Toutios, A., Tölgner, J., and Narayanan, S. (2017). Estimation of vocal  
653 tract area function from volumetric magnetic resonance imaging. In *2017 IEEE*  
654 *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages  
655 924–928. IEEE.

656 Stone, M., Davis, E. P., Douglas, A. S., NessAiver, M., Gullapalli, R., Levine, W. S.,  
657 and Lundberg, A. (2001). Modeling the motion of the internal tongue from tagged cine-  
658 images. *The Journal of the Acoustical Society of America*, 109(6):2974–2982.

659 Takemoto, H. (2001). Morphological analyses of the human tongue musculature for  
660 three-dimensional modeling. *Journal of Speech, Language, and Hearing Research*.

661 Takemoto, H., Goto, T., Hagihara, Y., Hamanaka, S., Kitamura, T., Nota, Y., and  
662 Maekawa, K. (2019). Speech organ contour extraction using real-time mri and machine  
663 learning method. *Proc. Interspeech 2019*, pages 904–908.

664 Takemoto, H., Honda, K., Masaki, S., Shimada, Y., and Fujimoto, I. (2006).  
665 Measurement of temporal changes in vocal tract area function from 3d cine-MRI data.  
666 *The Journal of the Acoustical Society of America*, 119(2):1037–1049.

667 Uecker, M., Zhang, S., Voit, D., Karaus, A., Merboldt, K.-D., and Frahm, J. (2010).  
668 Real-time MRI at a resolution of 20 ms. *NMR in Biomedicine*, 23(8):986–994.

669 Woo, J., Lee, J., Murano, E. Z., Xing, F., Al-Talib, M., Stone, M., and Prince, J. L.  
670 (2015a). A high-resolution atlas and statistical model of the vocal tract from  
671 structural MRI. *Computer Methods in Biomechanics and Biomedical Engineering:  
672 Imaging & Visualization*, 3(1):47–60.

673 Woo, J., Xing, F., Lee, J., Stone, M., and Prince, J. L. (2015b). Construction of an  
674 unbiased spatio-temporal atlas of the tongue during speech. In *International Conference  
675 on Information Processing in Medical Imaging*, pages 723–732. Springer.

676 Woo, J., Xing, F., Lee, J., Stone, M., and Prince, J. L. (2018). A spatio-temporal atlas and  
677 statistical model of the tongue during speech from cine-MRI. *Computer Methods in  
678 Biomechanics and Biomedical Engineering: Imaging & Visualization*, 6(5):520-531.

679 Woo, J., Xing, F., Stone, M., Green, J., Reese, T. G., Brady, T. J., Wedeen, V. J., Prince, J. L.,  
680 and El Fakhri, G. (2019). Speech map: A statistical multimodal atlas of 4d tongue motion during  
681 speech from tagged and cine mr images. *Computer Methods in Biomechanics and Biomedical  
682 Engineering: Imaging & Visualization*, 7(4):361–373.

683 Xing, F., Prince, J. L., Stone, M., Wedeen, V. J., El Fakhri, G., and Woo, J. (2017). A four-  
684 dimensional motion field atlas of the tongue from tagged and cine magnetic resonance imaging.  
685 In *Medical Imaging 2017: Image Processing*, volume 10133, page 101331H. International

686 Society for Optics and Photonics.

687 Xing, F., Stone, M., Goldsmith, T., Prince, J. L., El Fakhri, G., and Woo, J. (2019). Atlas-based  
688 tongue muscle correlation analysis from tagged and high- resolution magnetic resonance  
689 imaging. *Journal of Speech, Language, and Hearing Research*, 62(7):2258–2269.

690 Xiong, J., Ren, J., Luo, L., and Horowitz, M. (2018). Mapping histological slice sequences to the  
691 allen mouse brain atlas without 3d reconstruction. *Frontiers in neuroinformatics*, 12:93.

692

693 **TABLES**

694

695

TABLE I: VT measurements

Speaker	Length (mm)	Height (mm)	Width (mm)
SP1	97	92	40
SP2	77	76	32
SP3	99	81	40
SP4	89	69	34
SP5	94	86	36
SP6	87	81	32
SP7	88	90	38
SP8	87	67	34
Mean	89.8	80.3	35.8
SD	6.5	8.6	3.1

696

697

698

699

700

701

702

703

704

705

706

707

708

709

710

TABLE II: Average phoneme duration (in number of frames at 50 fps)

syllable	C	V	CV
fi	9	5.65	14.65
fa	8.175	6.475	14.65
fu	7.525	6.9	14.425
pi	6.55	7.275	13.825
pa	7.475	8.55	16.025
pu	6.6	7.625	14.225
si	8.775	5.875	14.65
sa	8.9	6.05	14.95
su	9.025	5.2	14.225
ti	7.6	6.825	14.425
ta	6.85	6.7	13.55
tu	7.025	4.85	11.875

711

712

713

714

715

716

717

718

719

720

721

722

723

724

725

726



727

TABLE III: Cross validated results

phoneme	Mean (before)	SD (before)	Mean (after)	SD (after)
fi	0.872	0.044	0.975	0.014
fa	0.876	0.047	0.976	0.014
fu	0.869	0.043	0.974	0.015
pi	0.874	0.044	0.976	0.015
pa	0.874	0.046	0.975	0.014
pu	0.873	0.040	0.974	0.015
si	0.872	0.044	0.975	0.014
sa	0.870	0.044	0.974	0.019
su	0.873	0.045	0.976	0.016
ti	0.873	0.046	0.974	0.016
ta	0.877	0.048	0.976	0.016
tu	0.874	0.044	0.975	0.021

728

729

730

731

From left to right: CV, average similarity score before the use of atlas, standard deviation of the average similarity before the use of atlas, average similarity after the use of atlas, standard deviation of the average similarity after the use of atlas