Loria
Laboratoire lorrain de recherche en informatique et ses applications

An articulatory model of the complete vocal tract from medical images

Yves Laprie
LORIA, Nancy

Electronic Speech Signal Processing 2017

Overview

A. Context
B. A brief history of articulatory models
C. Data for creating an articulatory model
D. Dimensionality reduction
E. Construction strategy
F. Evaluation and adaptation
G. Conclusion

Loria
Laboratoire lorrain de recherche en informatique et ses applications

A) Context

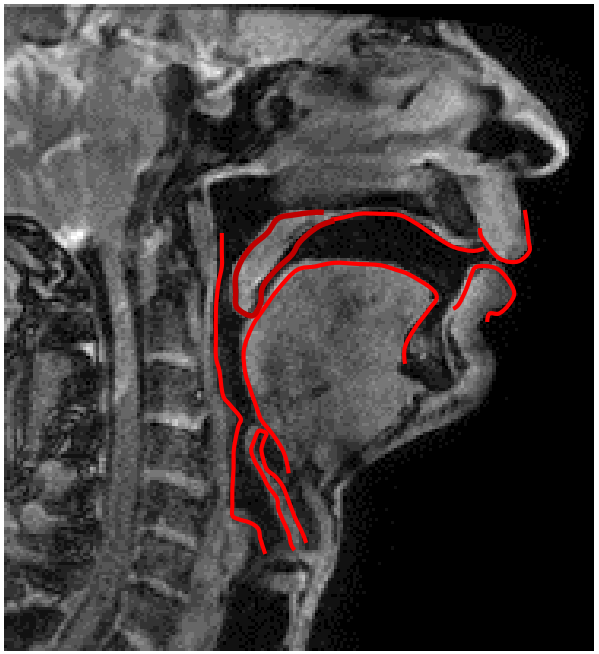
Why articulatory models?

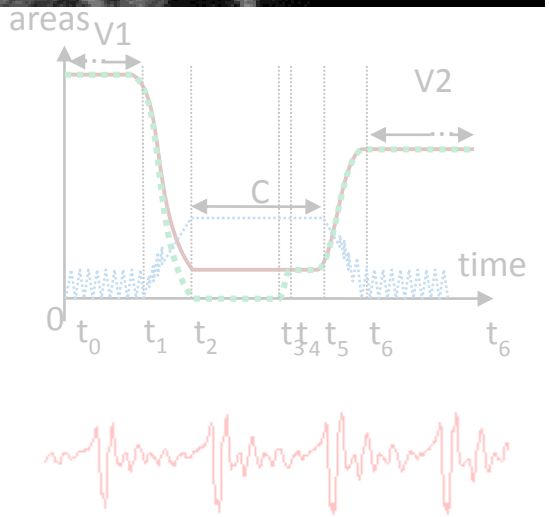
- Direct answer: for computing the geometrical vocal tract shape used as an input of articulatory synthesis.
- The goal is to represent the vocal tract articulators in a concise form while retaining as much as possible the variability of the tongue shape and vocal tract.

Articulatory synthesis

- Generate a sequence of vocal tract shapes by using articulatory and coarticulation models.
- From contours to the 3D shape and area function.
- Moving to the acoustic simulation
- Temporal coordination scenario

→synthetic speech signal

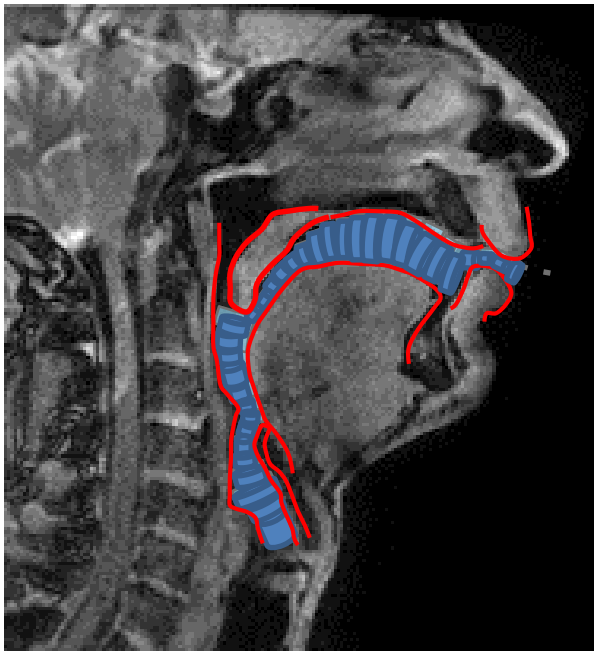


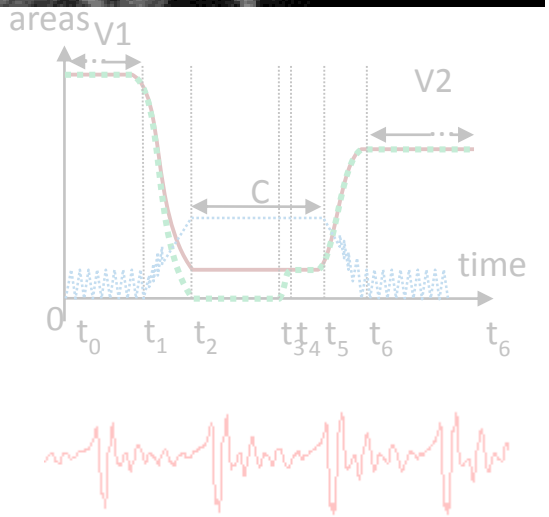


Articulatory synthesis

- Generate a sequence of vocal tract shapes by using articulatory and coarticulation models.
- From contours to the 3D shape and area function.
- Moving to the acoustic simulation
- Temporal coordination scenario

→synthetic speech signal

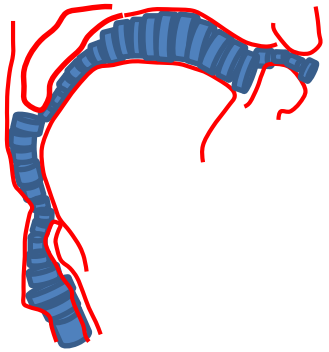


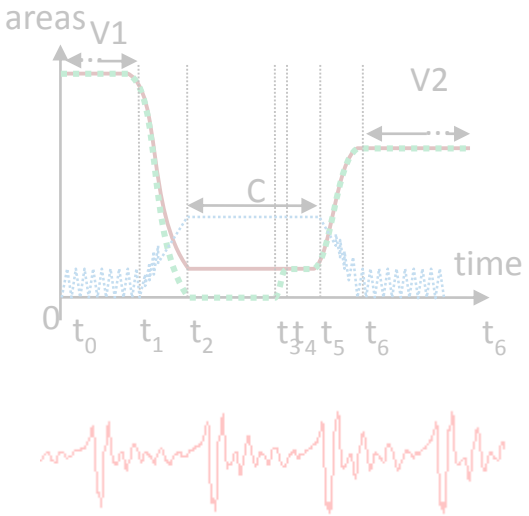


Articulatory synthesis

- Generate a sequence of vocal tract shapes by using articulatory and coarticulation models.
- From contours to the 3D shape and area function.
- Moving to the acoustic simulation
- Temporal coordination scenario

→synthetic speech signal

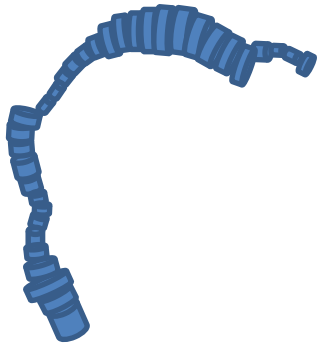


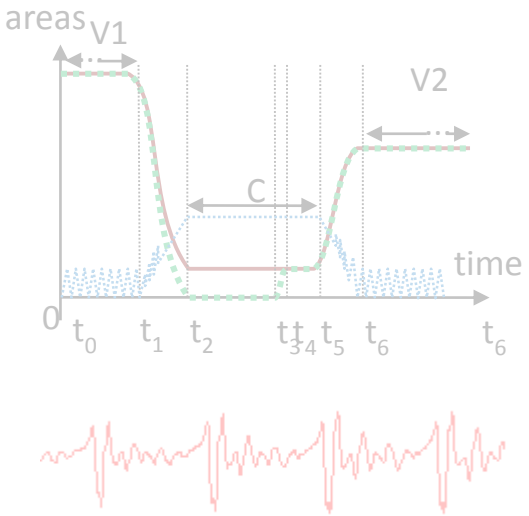


Articulatory synthesis

- Generate a sequence of vocal tract shapes by using articulatory and coarticulation models.
- From contours to the 3D shape and area function.
- Moving to the acoustic simulation
- Temporal coordination scenario

→synthetic speech signal

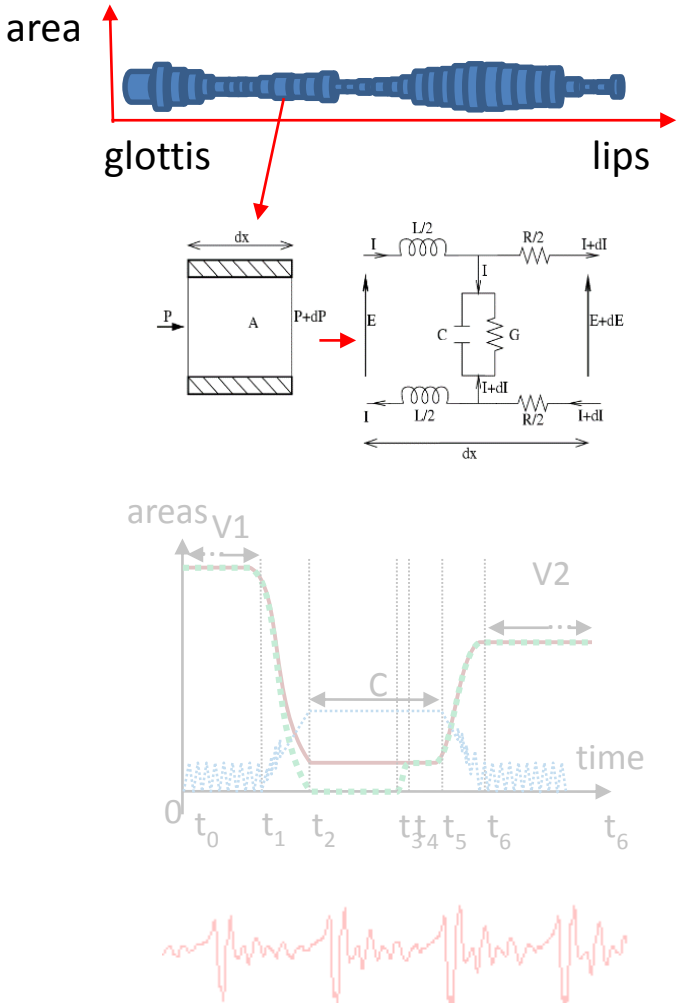




Articulatory synthesis

- Generate a sequence of vocal tract shapes by using articulatory and coarticulation models.
- From contours to the 3D shape and area function.
- Moving to the acoustic simulation
- Temporal coordination scenario

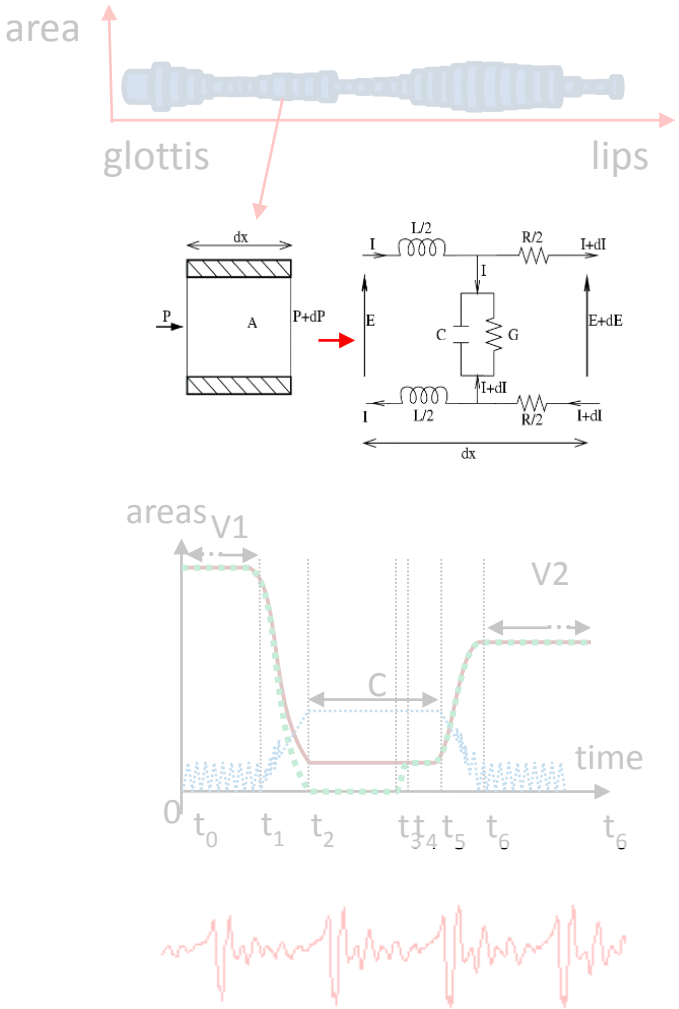
→synthetic speech signal



Articulatory synthesis

- Generate a sequence of vocal tract shapes by using articulatory and coarticulation models.
- From contours to the 3D shape and area function.
- Moving to the acoustic simulation
- Temporal coordination scenario

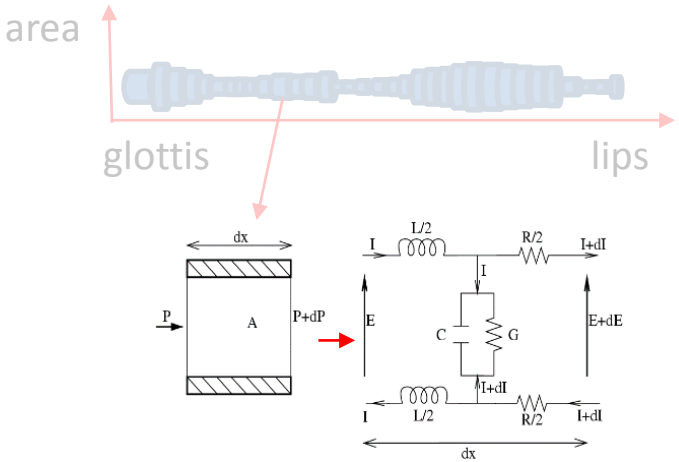
→synthetic speech signal

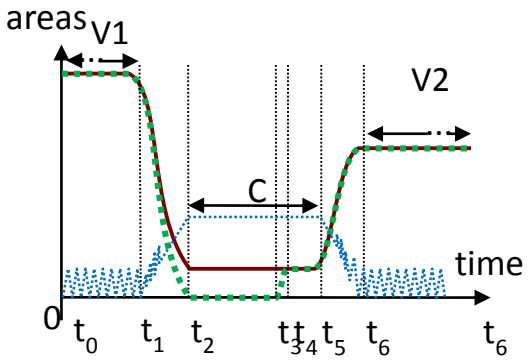


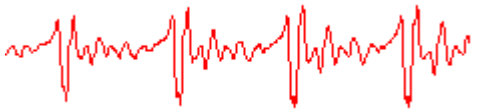
Articulatory synthesis

- Generate a sequence of vocal tract shapes by using articulatory and coarticulation models.
- From contours to the 3D shape and area function.
- Moving to the acoustic simulation
- Temporal coordination scenario

→synthetic speech signal



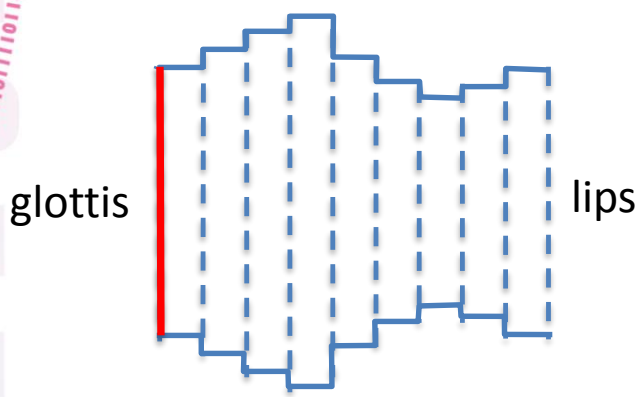


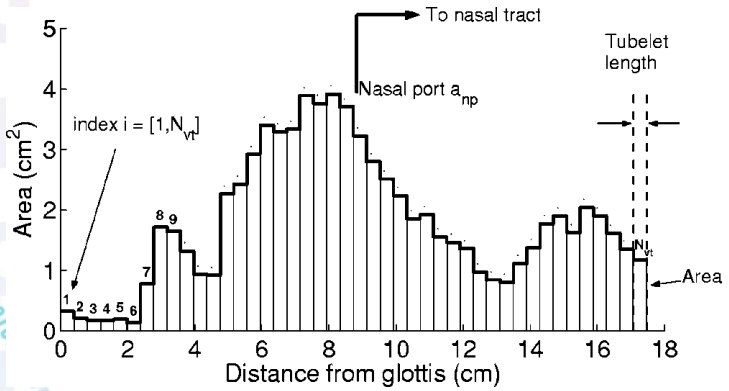


Loria
Laboratoire lorrain de recherche en informatique et ses applications

B) A brief history of articulatory models

No articulatory model: model of area functions





- solution adopted in the early stages of acoustic simulations of the vocal tract
- solution adopted by Story (2005, JASA):
  - $A(i,t) = V(i,t) \prod_{k=1}^{N_c} C_k(i,t) \quad i = [1, N_{vt}]$
  - where $V(i,t)$ is the vowel substrate (implemented with 2 modes of deformation) and $C_k(i,t)$ the consonantal superposition functions (implemented as a negative gaussian) and N_c is the number of active consonantal functions (often 1 and sometimes 2).

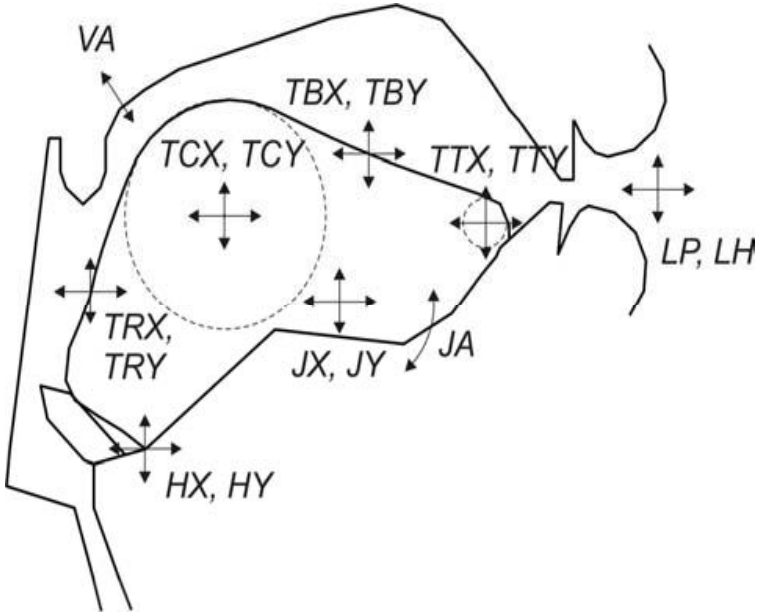
Models based on geometric primitives

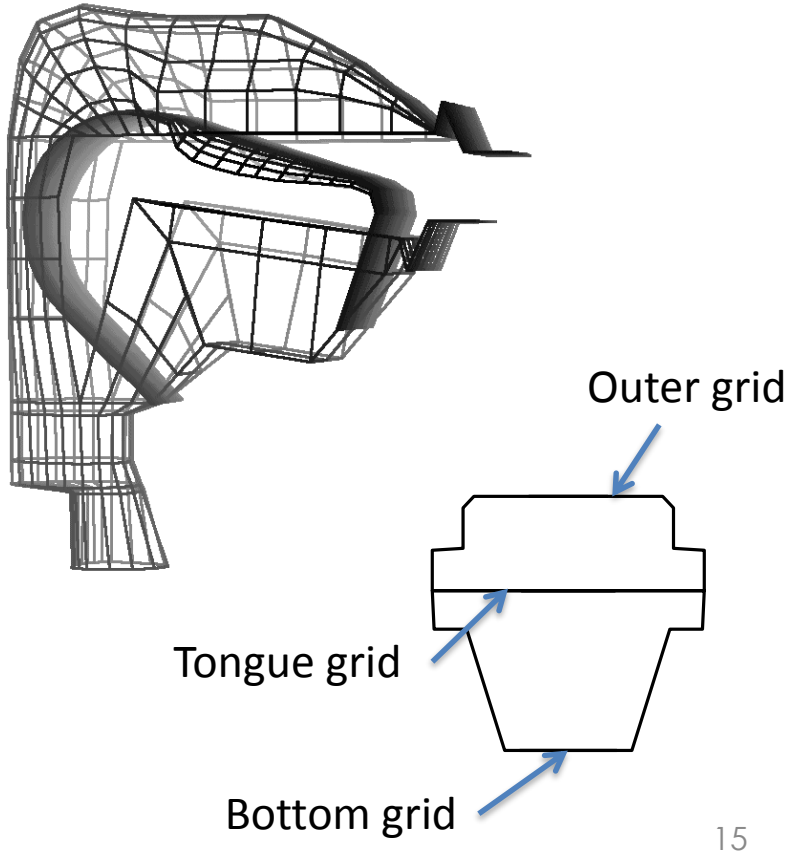
- Geometric model of Coker (1968):
  - A collection of purely geometric primitives (lines given by their extremities, circles by their centers and radius)

  - Probably very difficult from a control point of view.

Two examples of articulatory models based on geometric primitives

The 2D model of Mermelstein (1973)	The 3D model of Birkholz (2003)





Models based on geometric primitives

- Advantages:
  - do not require any articulatory data, or at most a very limited amount only
  - could account for any speaker with some effort.
- Disadvantages:
  - no guarantee that vocal tract (VT) shapes correspond to real shapes
  - many arbitrary parameters to adjust
  - how to move from one shape to another one? How to pilot the model?

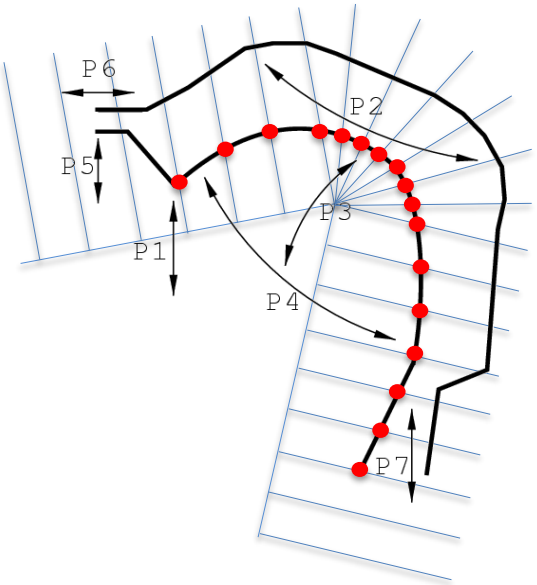
Loria

Models derived from articulatory data

- Construction
  - Exploits a database “as vast” as possible:
    - Between 500 and 1000 images from an X-ray or MR film
    - Around between 50 and 100 3D MR images
  - Relies on a factor analysis method to derive a small number of deformation modes:
    - Generally PCA (Principal Component Analysis) or other related methods ICA (Independent Component Analysis) for instance.
    - Requires an overall strategy to identify which articulators should be modeled (jaw, tongue, lips, larynx, velum) and whether there are dependency relations or not.
- A shape is thus represented by a small number of parameters, each parameter giving the score of one deformation mode.

Loria

Models derived from articulatory data

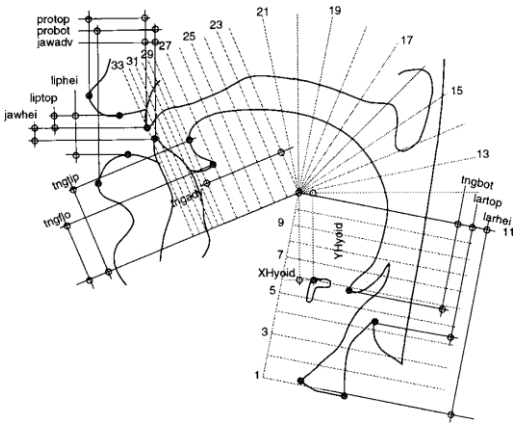


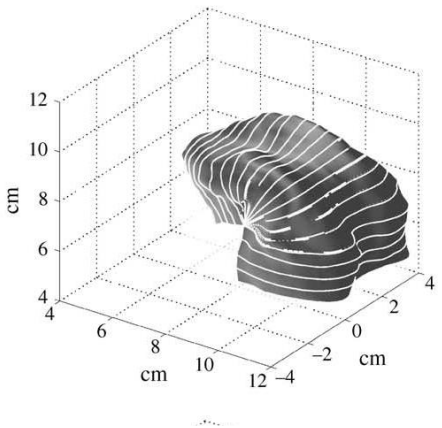
- Model proposed by Maeda (1979) with 7 parameters (jaw, tongue position, tongue shape, apex, lip aperture and protrusion, larynx).
- Semipolar articulatory grid
- Model constructed by guided PCA (choosing a variable which renders a key articulatory gesture, for instance tongue position).

- Advantages:
  - generate realistic vocal tract shapes since deformation modes are derived from true VT shapes
  - interpolation between two vectors (describing a vocal tract shape) is easy.
- Weaknesses:
  - adapted to vowels
  - semipolar articulatory grid which imposes acoustics

Other models derived from articulatory data

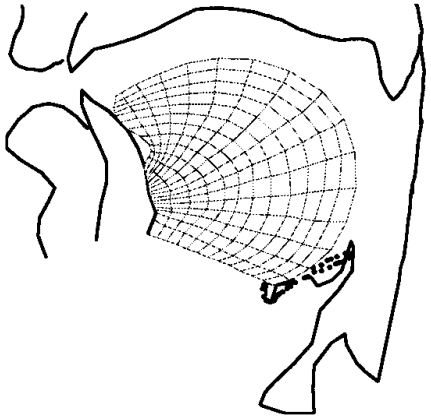
- Beautemp et al. (2001)
  - Variable articulatory grid.
- Models derived from 3D MR images
  - Tongue model Engwall (2000), Badin et al. (2002)
  - The delineation of each tongue is long and is subject to many errors (difference between tongue tissues and others).
  - Require some “human” supervision to get acceptable results.
- A good and bad point:
  The model is “locked” on ONE speaker.





2D Biomechanical models

- First 2D model proposed by Perkell (1974).
- 2D finite-element method (FEM) tongue model (Perrier et al. 2003).
- Mesh designed to account for the anatomical arrangement of the main muscular components.
- 3D tissue quasi incompressibility is implemented as area conservation.



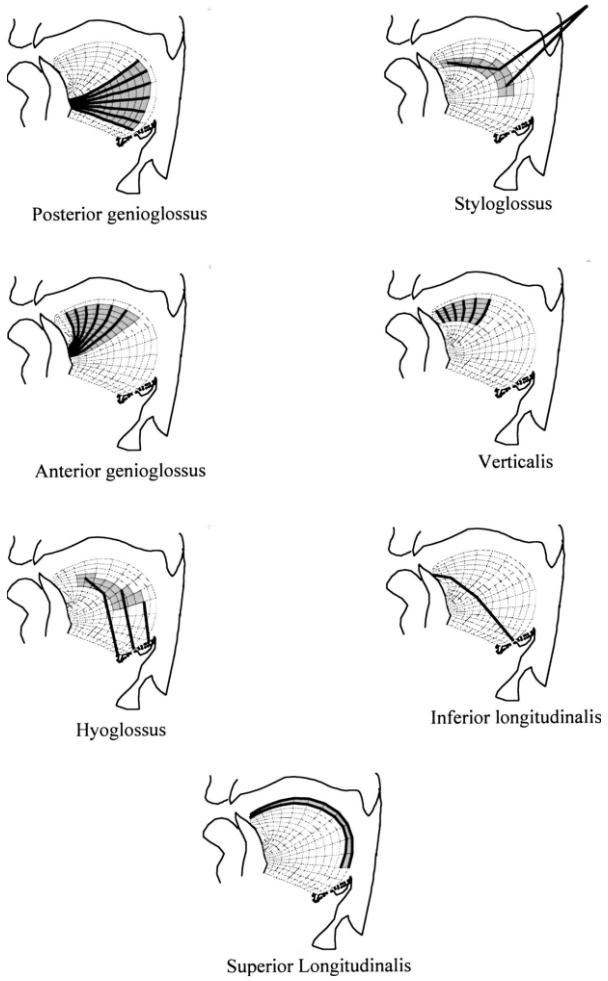
- 221 nodes (intersections of lines in the figure) define192 quadratic elements..
- Elasticity modulus (Young's modulus) giving the relationship between the force applied and the deformation.
- Stiffness parameters adjusted from observations and by adjustments.
- Detection of contacts and then reaction force (penalty method).
- Many ad-hoc adjustments.

Loria

2D Biomechanical models

Representation of the seven muscles taken into account in the Perrier et al. model.

Bold lines represent macrofibers, over which the global muscle force is distributed. The gray-shaded quadrilaterals are selected elements within the FE structure, whose mechanical stiffness increases with muscle activation.



- Consonant and vowel targets defined in terms of muscle activations after a number of trials…

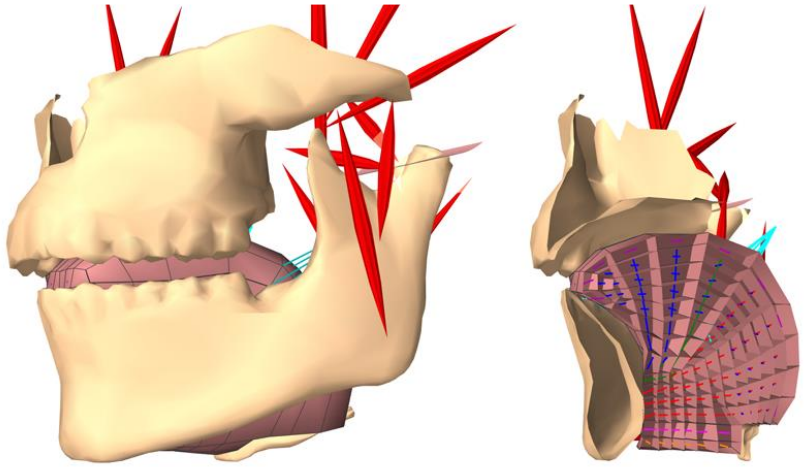
3D biomechanical models

Pros:

- Volume conservation instead of surface.
- No need to use a 2D → 3D transformation.

Contras:

- A larger number of parameters to adjust targets and movements.
- Slow and unexpected effects sometimes.
- Complete VT requires a very large number of muscles and adjustments.



Stavness et al. (2011)

Loria

3D or 2D models?

Advantages:

- A 3D model gives the area function almost directly.
- Complex shapes (for instance /l/) can (or should) be accessible directly.

Disadvantages:

- Requires more data (MRI data) which may not cover all the vocal tract deformations.
- It is not sure that the 3rd dimension improves acoustics. See Ericsdotter ICPhS (2007).

→ an improved 2D model is probably sufficient… in many cases.

Loria

Loria
Laboratoire lorrain de recherche en informatique et ses applications

C) Data for creating an articulatory model

Means of articulatory data acquisition

	EMA	MRI	Ultrasound	X-ray	X-ray microbeam
Whole V.T.	No	Yes	No	yes	No
Tongue imaging	Pellets	Full-length	Full-length	Full-length	Pellets
Tongue root	No	Yes	No	Yes	No
Velum imaging	Yes[2]	Yes	No	Yes	Yes
Time resolution	200 Hz	≥ 30 Hz	30-200 Hz	50 Hz	40-160 Hz
3D	No	Yes	No	No	No
Health hazard	No	No	No	Yes	No/Yes
Natural art.	Affected	Yes[3]	Yes	Yes	Affected
Acoustic noise	Low	High	Acceptable	Low	Acceptable
Head Mvt.	Restricted[4]	Restricted	Restricted[5]	Free	Free
Portable	No	No	Yes	No	No
Inexpensive	No	No	Yes	No	No

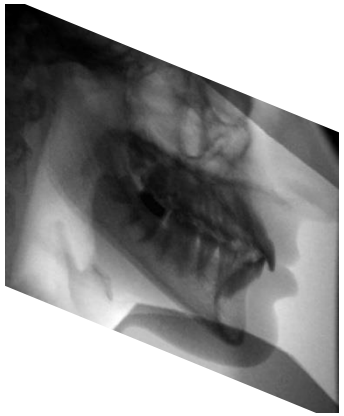
→Only x-ray and now MRI films can provide whole VT shape and dynamics

→Data with other modalities need to be fused to get the whole picture of VT

Loria

Differences between X-ray and MRI

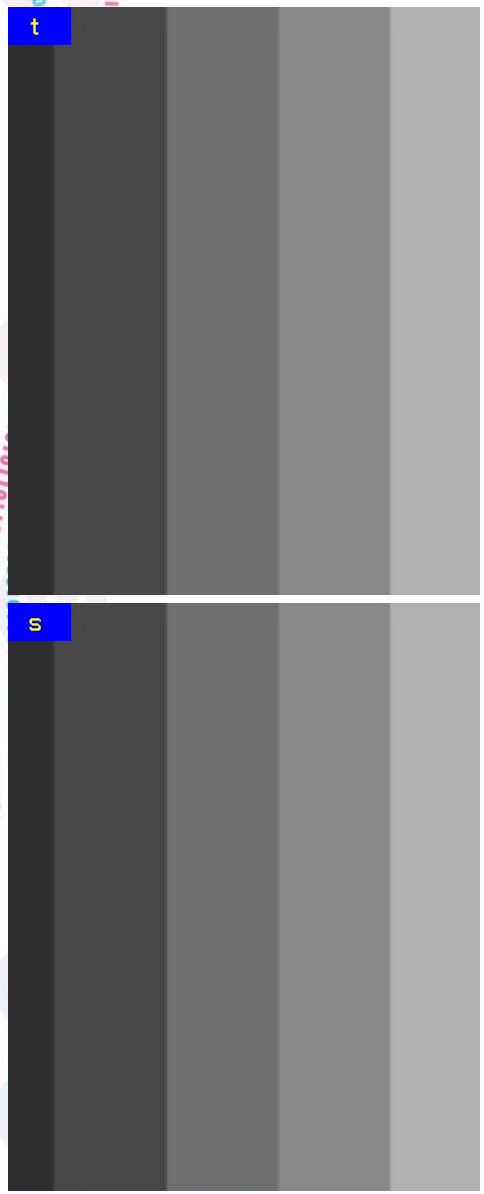
- **X-ray = projection**



- **MRI = a “true” numerical slice**

From http://www.coppit.org/brain/

X-ray films



A rich X-ray database in IPS including a number of speakers and languages.

X-ray imaging thus:
→ organs overlap on images,
→Images/films cannot be used directly.

Preprocessing is thus necessary!

The example shows the original image plus contours of articulators superimposed onto images.

Static or dynamic data?

- At the articulatory level, speech is a collection of overlapping gestures (mandible, tongue, velum…).

→ dynamic data is more natural

  - the sampling frequency should be close to 100 Hz.
  - the 3D geometry is required to get the possibility of checking the acoustics precisely.

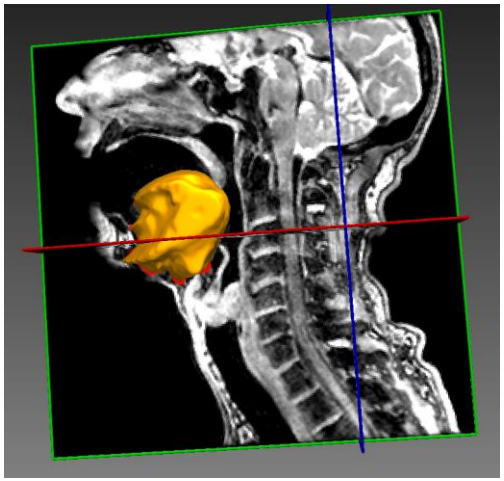
- These two conditions cannot be fulfilled yet and the fallback solution is:
  - 2D dynamic mid-sagittal (or any arbitrary direction) films between 30 and 100 Hz,
  - 3D static images.

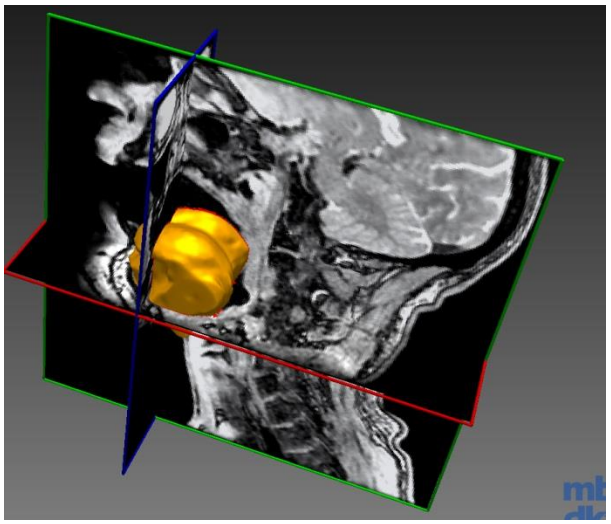
Loria

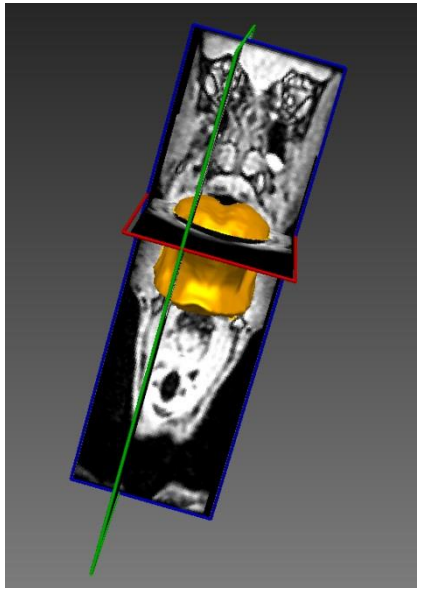
Static MRI images

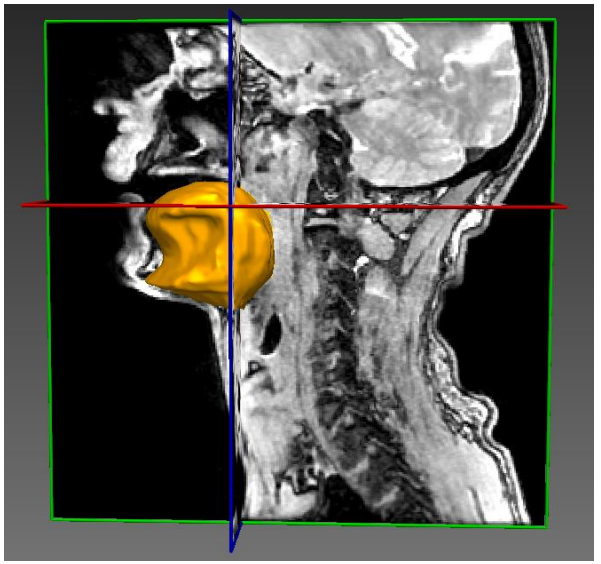
- Static images require the subject to maintain the same articulation (between 5 and 10 seconds approximately).
- Offer a good resolution in the three directions (sagittal, coronal, axial).
- Several strategies for the speaker:
  - Silent and blocked articulation (but not exactly the articulation with phonation)
  - Phonation (now possible for vowels)
- Requires dental cast to get the position of teeth.
- Requires processing to get the vocal tract walls (Itk-Snap, Osiryx, MITK…)

Example with semi-manual delineation of the tongue









Dynamic MR images

- Similar to X-ray films except that the 2D image is one “true” slice (between 3 or 8 mm thick).
- Obstacle: each elementary (line, radius, spiral) acquisition in the k-space of MRI requires 3ms.

→ elaborated acquisition strategy and reconstruction.

→ exploits sparsity of the Fourier MRI coefficients (compressive sensing).

- Narayanan team in USC
- Jens Frahm (Max Planck Institute, Göttingen)

Examples of cineMRI data

- cineMRI from the Signal Analysis and Interpretation Lab (http://sail.usc.edu/SPAN/index.php)

- cineMRI from Institute of Cognitive Neuroscience (University College London - http://www.icn.ucl.ac.uk) Dr. Zarinah Agnew

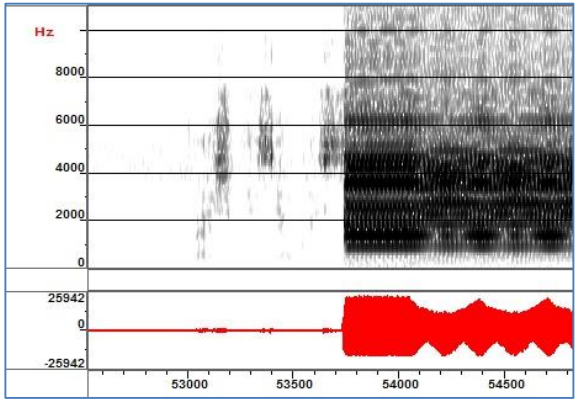
Max Planck Institute Göttingen – Jens Frahm

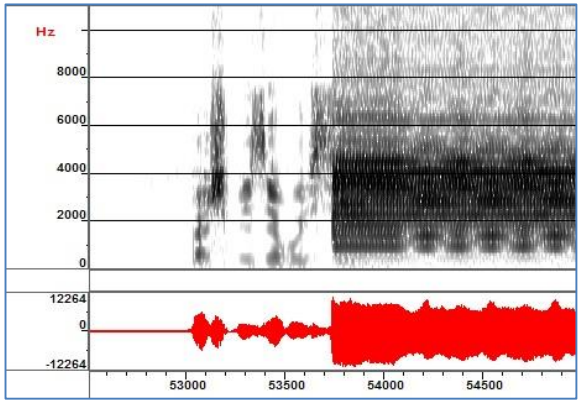
Natural speech (before complete denoising).
128x128 pixels at 55 Hz.

Obstacles to cineMRI

- Strong noise due to the gradient coils
  - → triggers Lombard effect
  - → requires optical microphones
  - → requires denoising:

Microphone environment	
Microphone close to mouth	
Denoised speech (E. Vincent)	
Speech in a quiet environment	





- Supine position

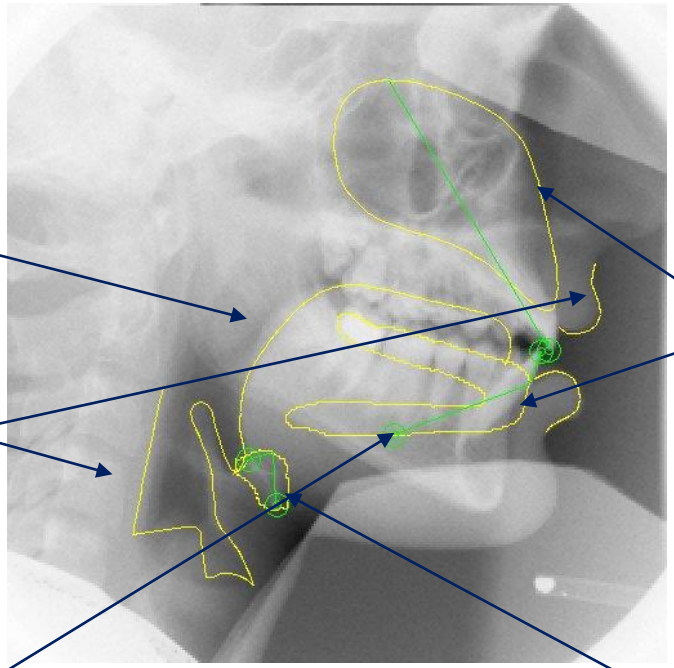
Ideal MRI database

- Static 3D MRI
  - Acquire a static database for several subjects with 100 images (vowels + CV from the list: i e ɛ a ɔ o u y ø œ ɑ̃ ɔ̃ ɛ̃ (13 vowels) p t k f s ʃ m n ʁ l (10 consonants).
  - 3D scan of the dental cast.
    - Delineate shapes (3D meshes).
    - Investigate the recovery of the 3rd dimension.
    - Acoustic validation.
- CineMRI
  - A corpus of sentences covering the phonetics of the target language.
    - Improve noise reduction (real time to remove Lombard effect).
    - Sampling rate > 50 Hz, in several directions.

Loria

Extracting contours of articulators

Graphical annotations of IPS X-ray data by phoneticians from IPS and LORIA by using Xarticul software.



Tongue: drawn by hand (medio-sagittal contour)

Larynx, lips: tracked semi automatically

Landmarks to point specific points: lower and higher incisors, hyoid top and bottom, tongue origin.

Registration regions: head and jaw in order to cancel head movements and jaw movements.
Rigid structures tracked via correlation.

Hyoid bone: tracked via correlation

Loria
Laboratoire lorrain de recherche en informatique et ses applications

D) Model construction

Dimensionality reduction

- Consider the tongue regularly sampled:
  - 100 points x 2 coordinates in 2D
- Or a more compact representation via control points of a B-spline:
  - 20 points x 2 coordinates
- Much bigger problem if 3D surfaces are considered.

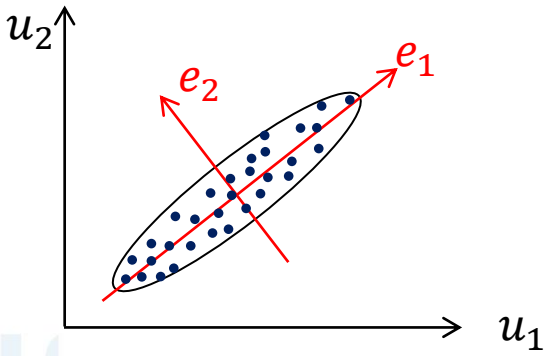
→ dimensionality reduction

Loria

Reduction of dimensionality

x=$\begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_N \end{bmatrix}$ is represented by y=$\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_K \end{bmatrix}$ with $K << N$

- x_n is a tongue point or intersection with the semi polar grid, y_k a component representing a mode of deformation.
- N is the number of points or intersections (50 or more) and K is the number of components (less than 6) in the case of the tongue contour.
- Principal Component Analysis (PCA) is one solution to reduce the dimensionality.



The red base describes data better than the black does. These 2D data could be approximated by projecting them onto e_1.
Idea: search for such a base.

Principal component analysis

$x = x_1 u_1 + x_2\, u_2 + \cdots + x_N\, u_N$ complete vector

$y = y_1 e_1 + y_2\, e_2 + \cdots + y_K\, e_K$ reduced representation, i.e. $K \ll N$

- The compact vector y is given by:

$$y = Tx$$

where $T = \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1N} \\ t_{21} & t_{22} & \cdots & t_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ t_{K1} & t_{K2} & \cdots & t_{KN} \end{bmatrix}$

T maps data from a high dimensional space to a lower dimensional sub-space.

T is obtained by minimizing $\|x - y\|$over the examples used to build the model. T is given by the first eigenvectors of the covariance matrix, those which explains variance of input data at best.

Loria

Other related techniques

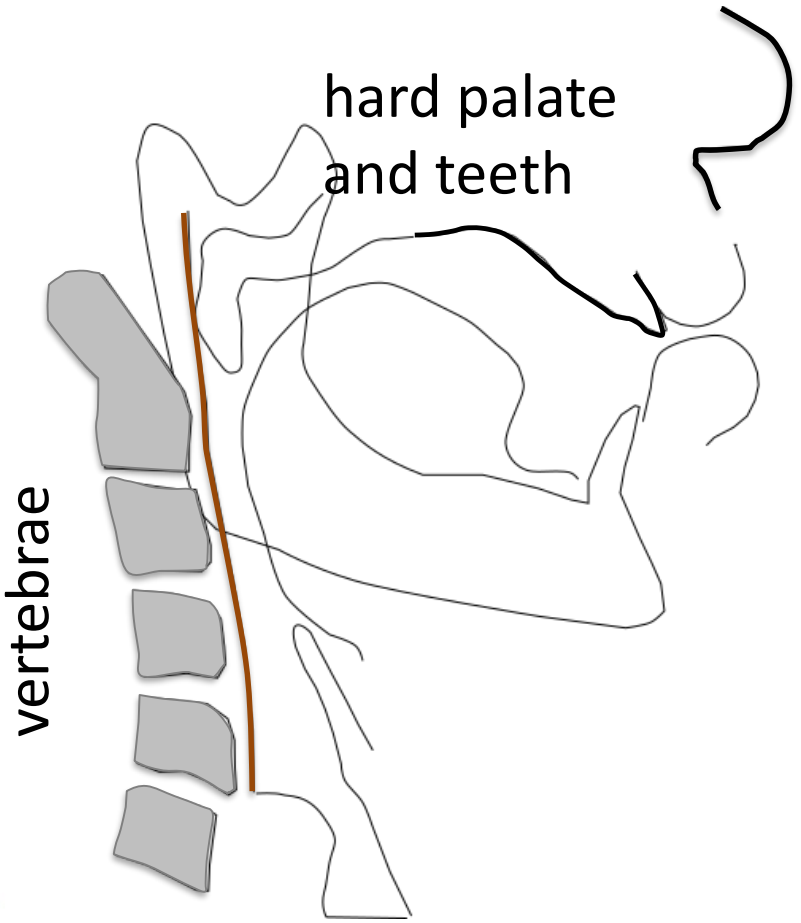
- Guided PCA (used by Maeda 79, Overall 1962):
  - The idea is to choose variables which fit the a priori knowledge (for instance the importance of the front-back movement of the tongue given by one point/measure of the tongue).
  - The principle is to subtract the correlation with this measure and to iterate the process.
- Independent component analysis (ICA):
  - finding deformation modes that are statistically independent.
  - the Central Limit Theorem says that the combination of distributions is all the more Gaussian since the original distributions are not gaussian.
  - the non gaussian character given by a kurtosis far from zero
    - $kurtosis(X) = E(X^4) - 3E(X^2)^2$

Loria

Loria
Laboratoire lorrain de recherche en informatique et ses applications

E) Construction strategy

Links between articulators

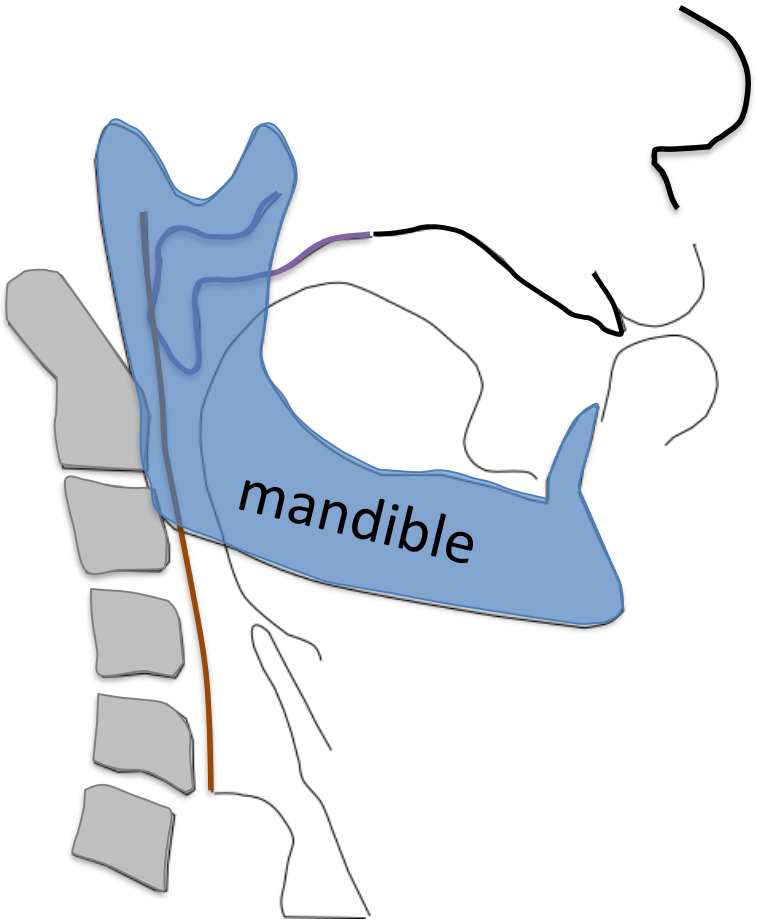


Motionless structures during one acquisition, and assumed to be fixed in the model:

- Hard palate and higher central incisor tooth (not visible directly on MR images)
- Pharyngeal wall: not true but acceptable in a first approach.

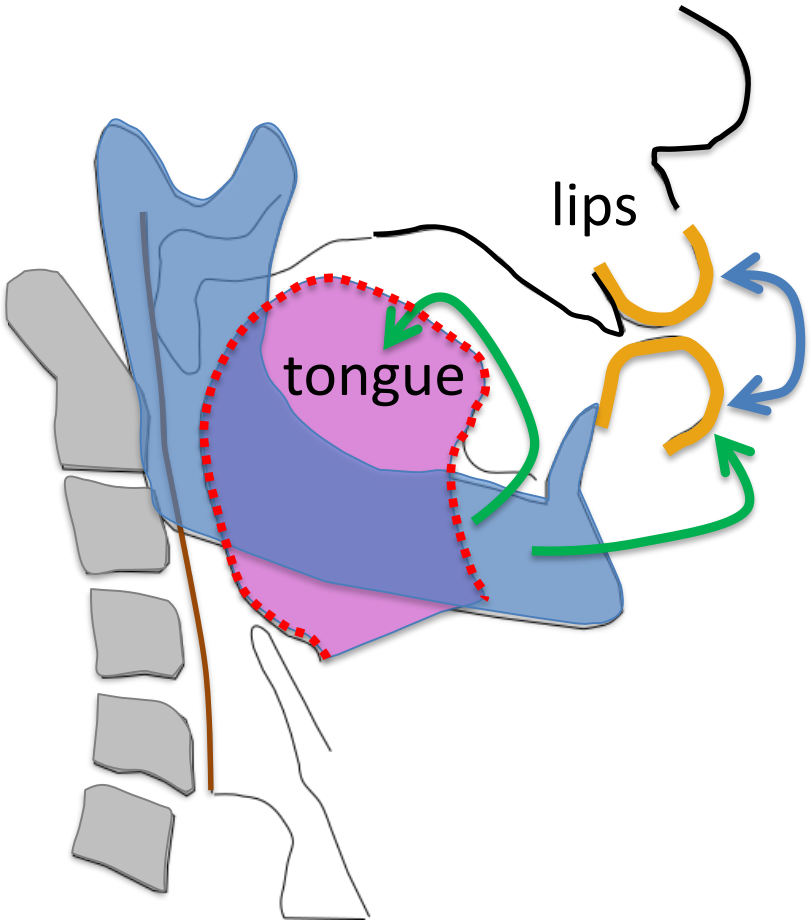
Loria

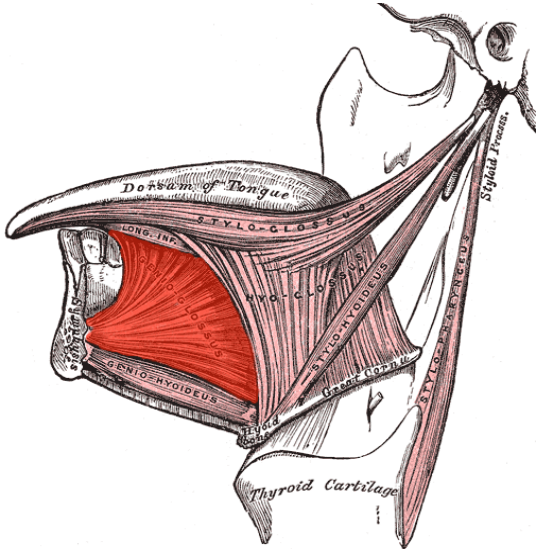
Links between articulators



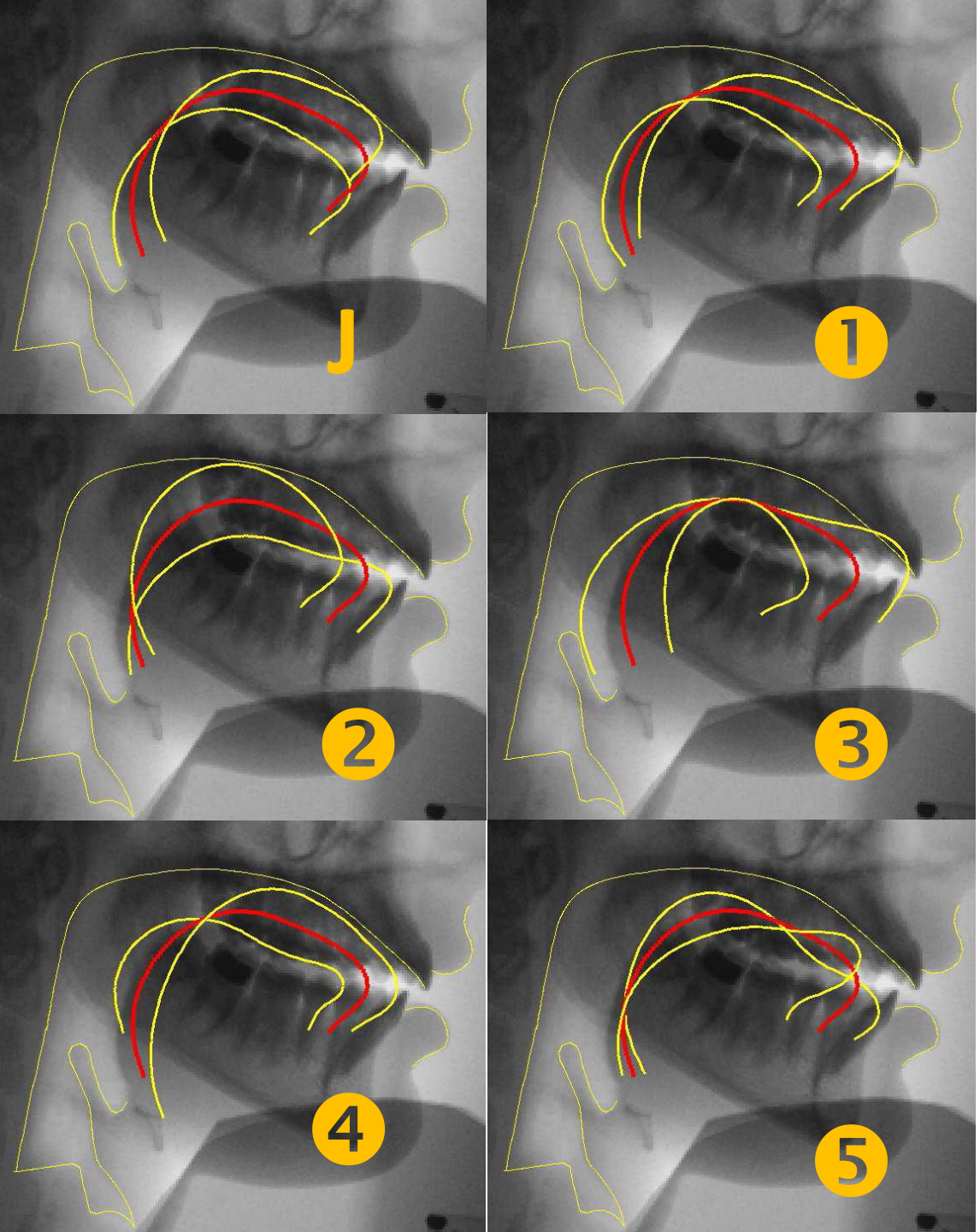
- The mandible is not visible directly and is approximated by the central lower incisors.
- Mandible (rigid) movement given by a 2D shift vector and a rotation (one angle), i.e. 3 parameters in 2D.
- PCA to get 2 parameters only.

Links between articulators



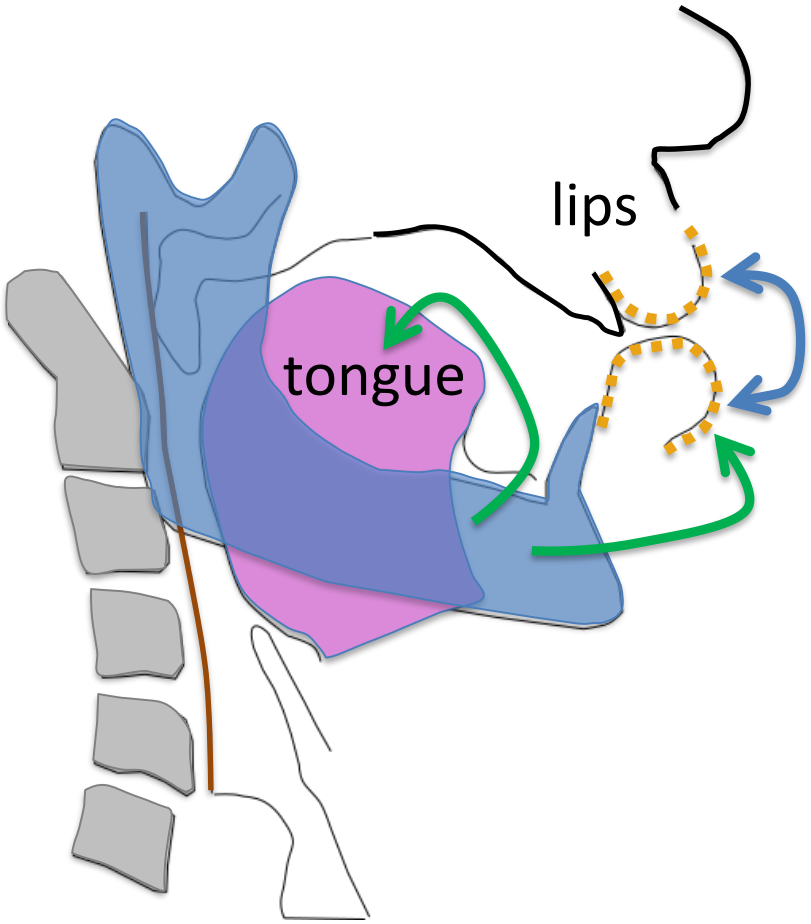


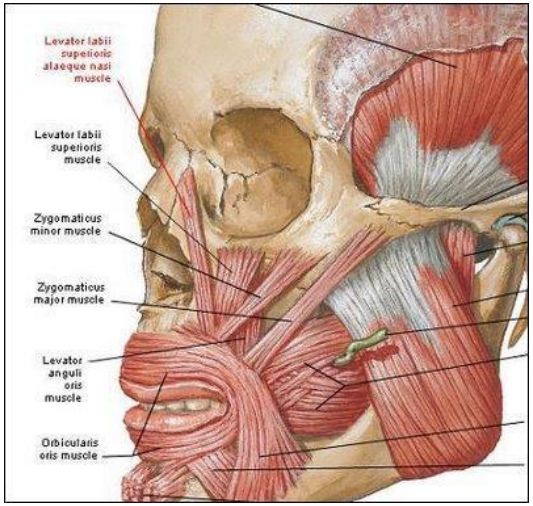
- The movement imposed by the mandible is subtracted from the tongue contour points.
- PCA to get between 6 and 12 parameters.
- More parameters to allow the fine deformations corresponding to consonants.



First linear component of the jaw and first five components of the tongue

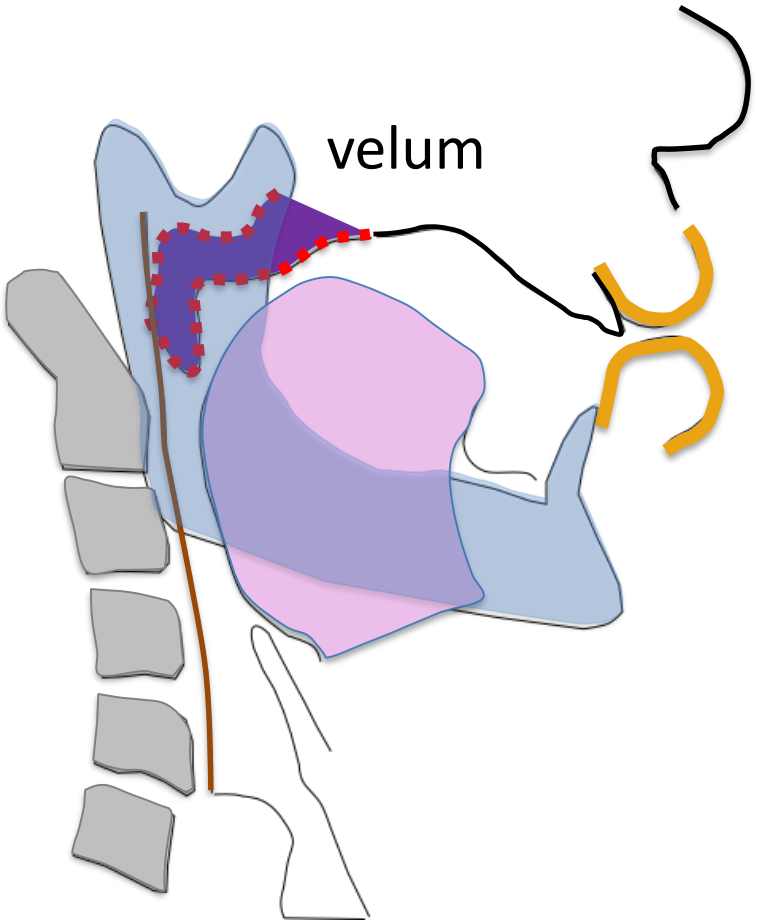
Links between articulators





- The influences of the mandible is more complex since there are muscles coupling both lips
  ➔ subtract correlation
- PCA to get 3 parameters

Links between articulators



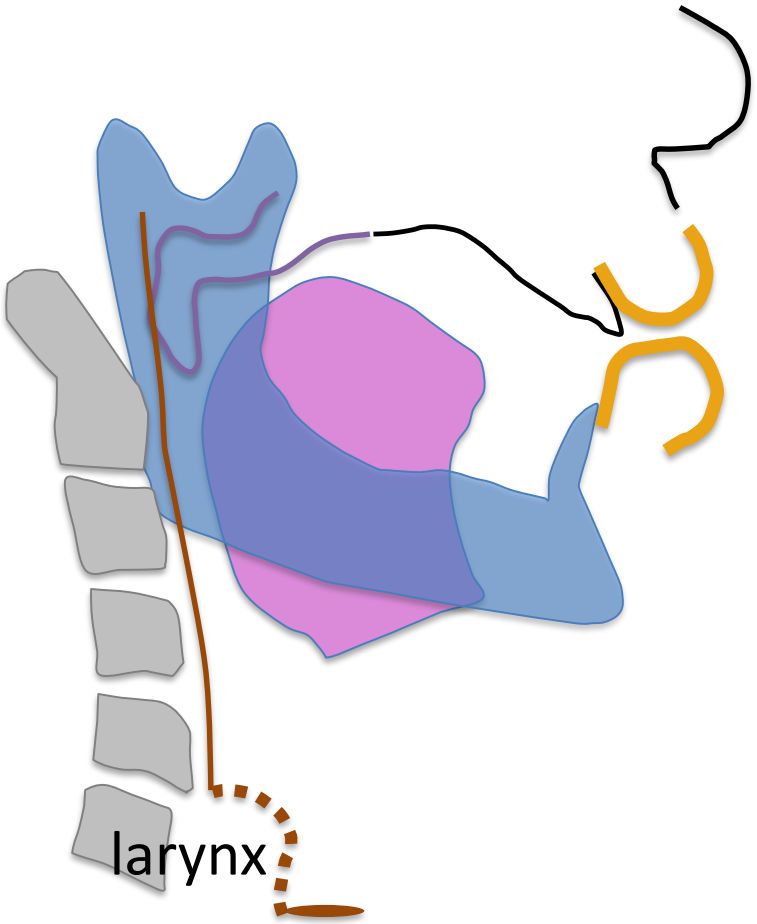
- No big influence of the mandible.
- PCA applied on the contour points.
- The crucial point is the velopharyngeal opening.

→ 3 factors

→ Not ideal because the velum can roll onto itself when contacting the tongue groove.

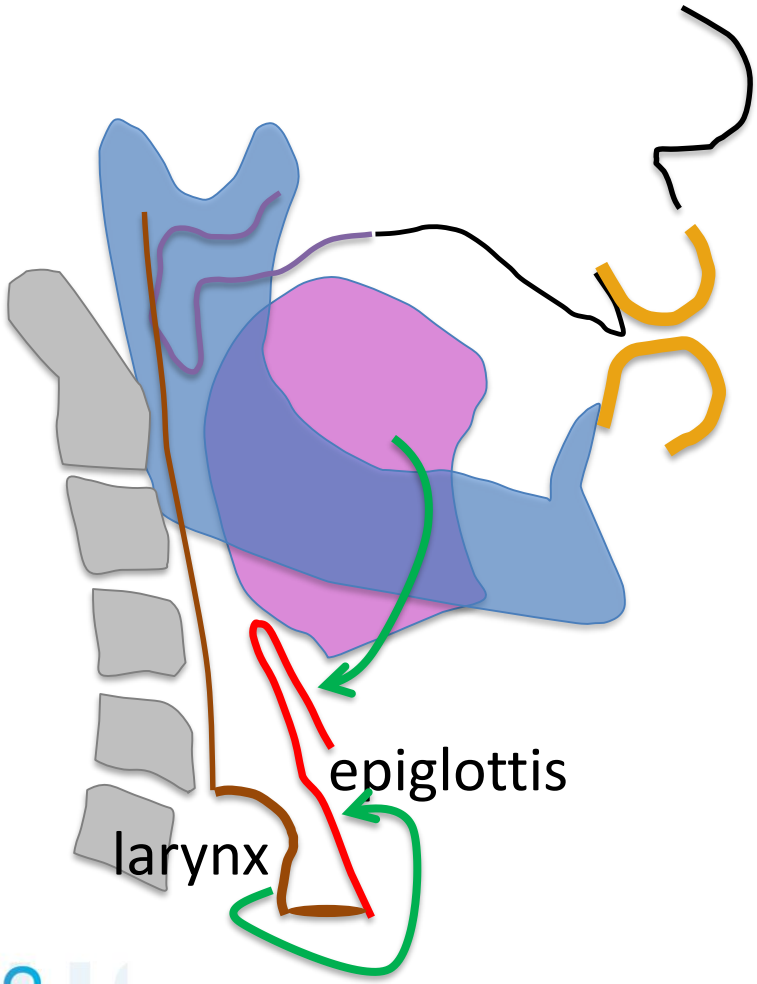
Loria

Links between articulators

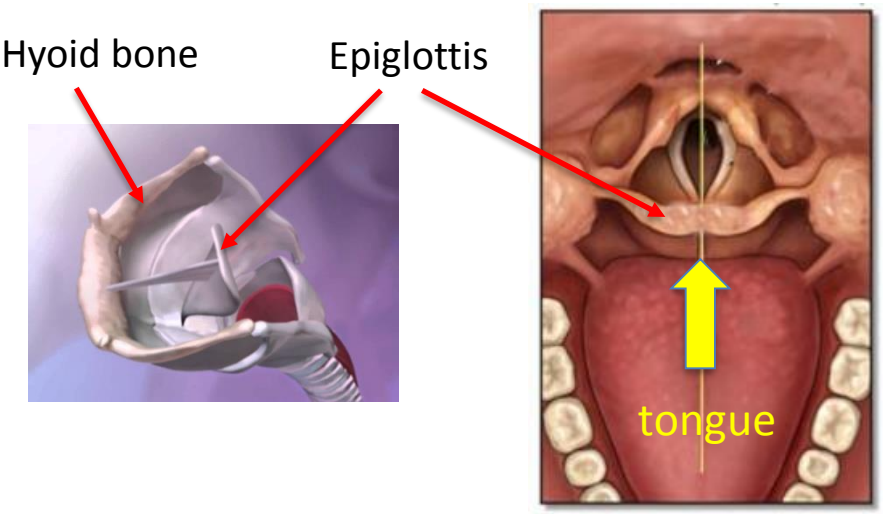


- Larynx considered as independent.
- PCA applied on the contour points
- 2 factors: vertical larynx position and laryngeal vestibule width

Links between articulators



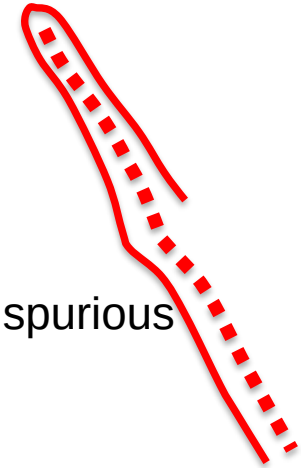
Epiglottis is a cartilage attached to the hyoid bone and thyroid cartilage.



Implementation of the influences in the model:

- The tongue (which pushes the epiglottis)
- The larynx (tightly connected to the hyoid bone).

Model of the epiglottis



- Epiglottis is a cartilage
  - → constant width
  - → only the centerline needs to be modeled
  - → centerline represented as a B-spline
  - → benefit: prevent the errors of delineation to be converted in spurious deformation modes
  - → few deformation modes
- Implementation of the dependencies
  - Via multi-linear regression

$$P_l = jaw_{0,l} B_0 + \sum_{j=0}^{T-1} tg_{j,l} C_j + lx_{0,l} D_0 + E_l$$

  - Where P_l is a control point, $jaw_{0,l}$ is the first factor of the mandible movement, $tg_{j,l}$ the factor of the j[th] tongue deformation mode, $lx_{0,l}$ the factor of the first larynx deformation. B_0, C_j and D_0 are the regression coefficient vectors, and E_l the residue not explained by mandible, tongue and larynx.

Model of the epiglottis

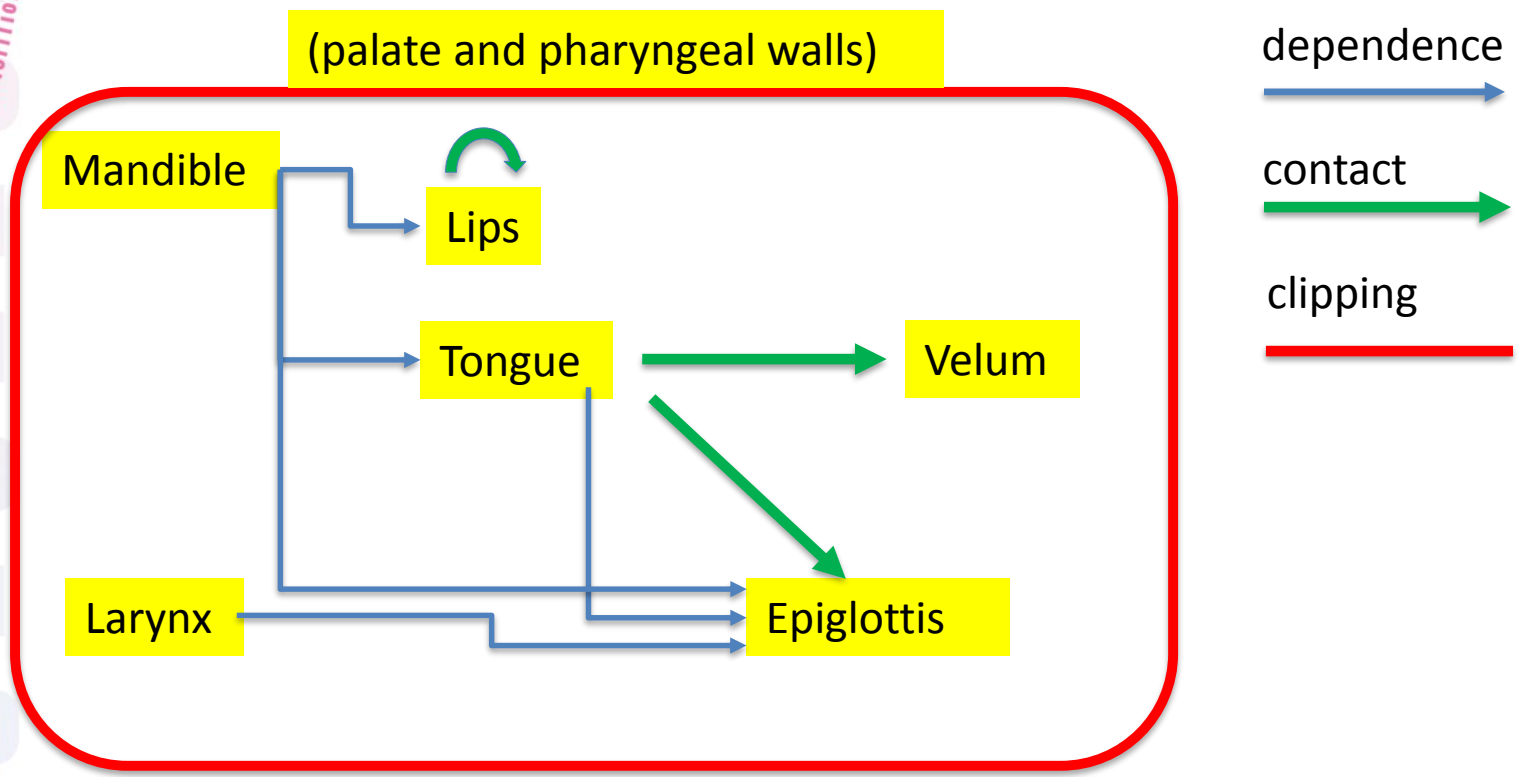
$$\begin{bmatrix} P_{1,1} & P_{1,2} & \dots & P_{1,M} \\ P_{2,1} & P_{2,2} & \dots & P_{2,M} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ P_{N,i} & P_{N,2} & \dots & P_{N,M} \end{bmatrix} = \begin{bmatrix} X_{1,1} & X_{1,2} & \dots & X_{1,K} \\ X_{2,1} & X_{2,2} & \dots & X_{2,K} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ X_{N,1} & X_{N,2} & \dots & X_{N,K} \end{bmatrix} \times \begin{bmatrix} B_{1,1} & B_{1,2} & \dots & B_{1,M} \\ B_{2,1} & B_{2,2} & \dots & B_{2,M} \\ \vdots & \vdots & & \vdots \\ B_{K,1} & B_{K,2} & \dots & B_{K,M} \end{bmatrix} + \begin{bmatrix} E_{1,1} & E_{1,2} & \dots & E_{1,M} \\ E_{2,1} & E_{2,2} & \dots & E_{2,M} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ E_{N,i} & E_{N,2} & \dots & E_{N,M} \end{bmatrix}$$

- P is the NxM matrix of the N observations of the control points. Since each point is a (x,y) vector P is a $2NxM$ matrix.
- X is $2NxK$ the factor matrix of known deformation modes (mandible, tongue and larynx)
- B is the $2NxK$ regression matrix
- And E is the NxM matrix of the residues of the control points nor explained by mandible, tongue and larynx deformations.
- Or in a matrix form: $P = XB + E$
- B is given by $B = (X^t X)^{-1} X^t Y$

→PCA can be applied on matrix E to get intrinsic epiglottis deformation modes.

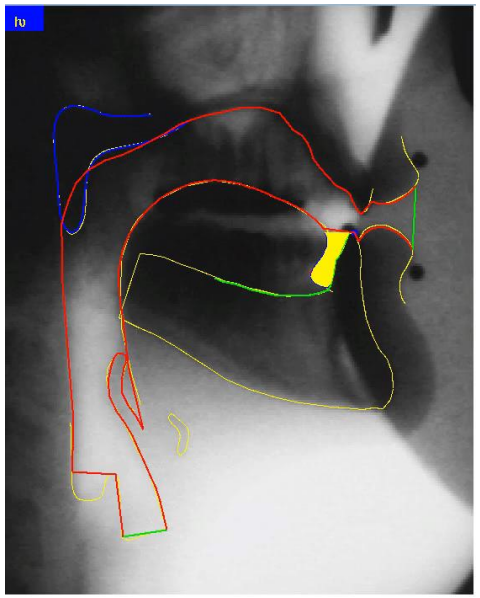
Loria

Overall view of the links

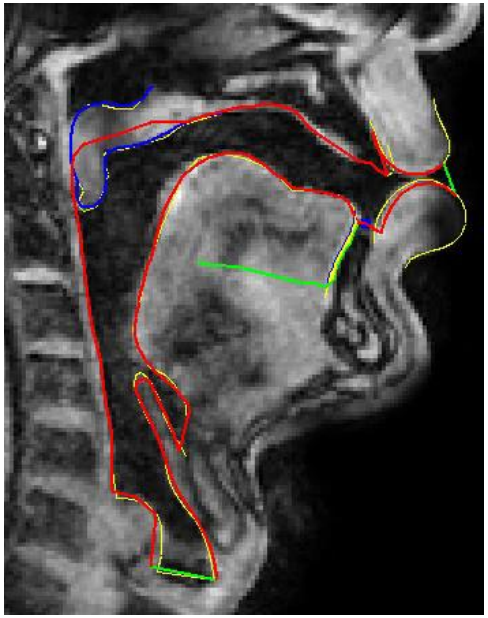


Loria

Videos of the model



One model derived from X-ray film.
The yellow region is the subglottal cavity.
The red curve is the model.

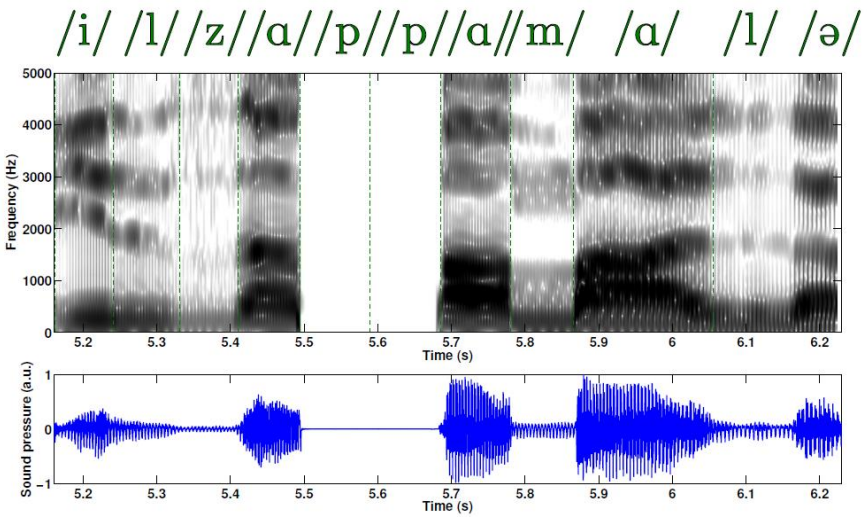


One model derived from MRI static images.
The red curve is the model.

Loria

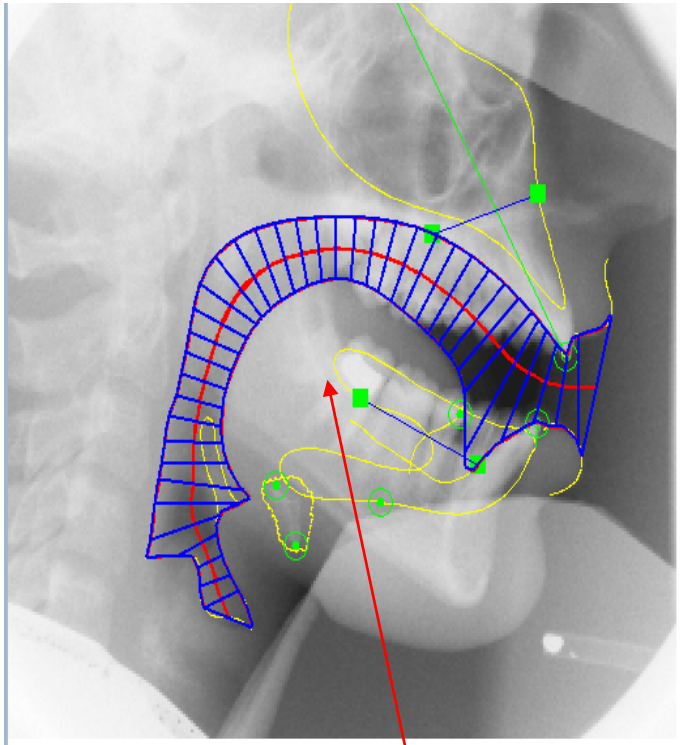
What comes after the model?

- Determination of the area functions at every time point of the synthesis
- Coordination with the source
- Acoustic simulation

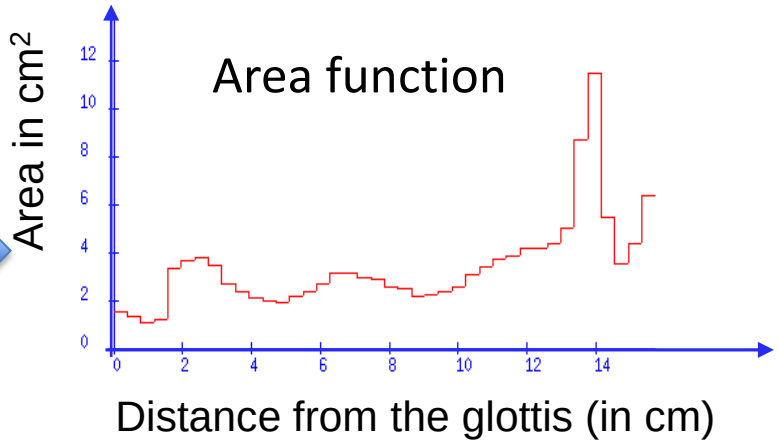


Example: “Il zappe pas mal.”
He is zapping a lot.

Recovering the area function



Automatic decomposition of the vocal tract into a series of tubes.



Transverse area approximated by same α β coefficients proposed by Heinz & Stevens (1965):

$$A(x) = \alpha d(x)^{\beta}$$

Very simple and old but more efficient that more recent approaches.

Loria

Determination of the centerline

The centerline is used to divide the vocal tract into uniform tubes consistent with the propagation of a wave plane.

Two steps:

1. A dynamic programming algorithm that chooses an optimal (with respect to a plane wave propagation) set of segments with one extremity on the exterior wall (~pharyngeal wall and palate) one on the interior wall (~tongue) of the vocal tract,
2. A regularizing algorithm to find a smooth contour equally distant from relevant points on the exterior walls.

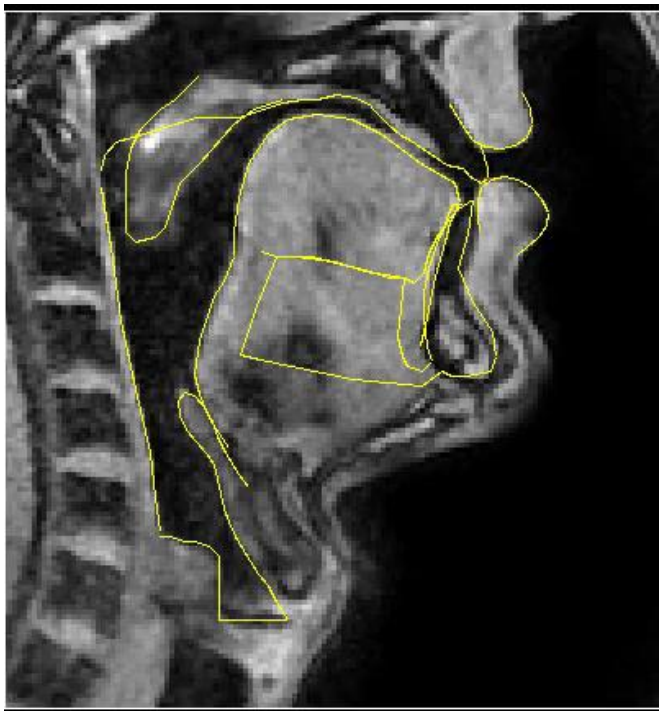
Centerline obtained an X-ray image of a sound /i/.

Advantages:

- Does not use any fixed articulatory grid
- Smoothing taking into account both walls.

Loria

Splitting the vocal tract into tubelets

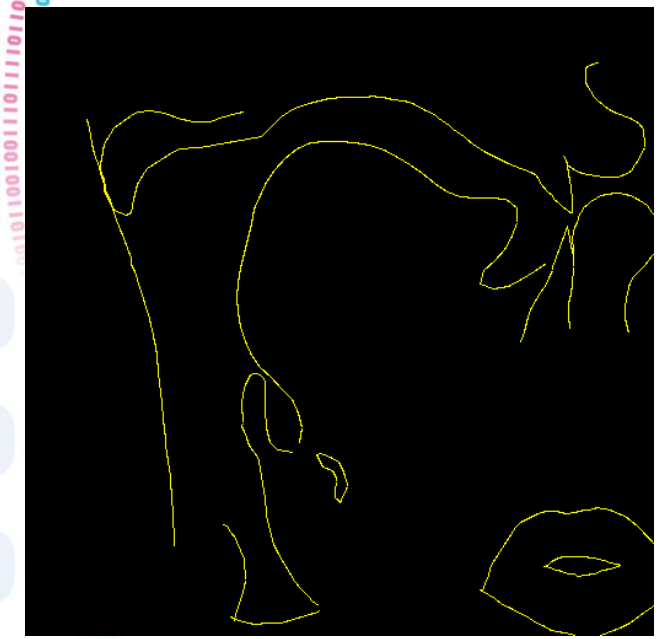


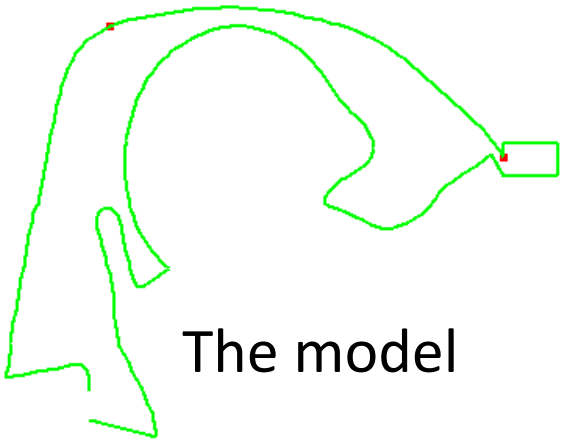
Loria
Laboratoire lorrain de recherche en informatique et ses applications

F) Evaluation and adaptation

E1) Evaluation of the model

- "Historical" data used by S. Maeda (1979) because very few X-ray data are available with delineated contours.
- The tongue contour is often not complete.



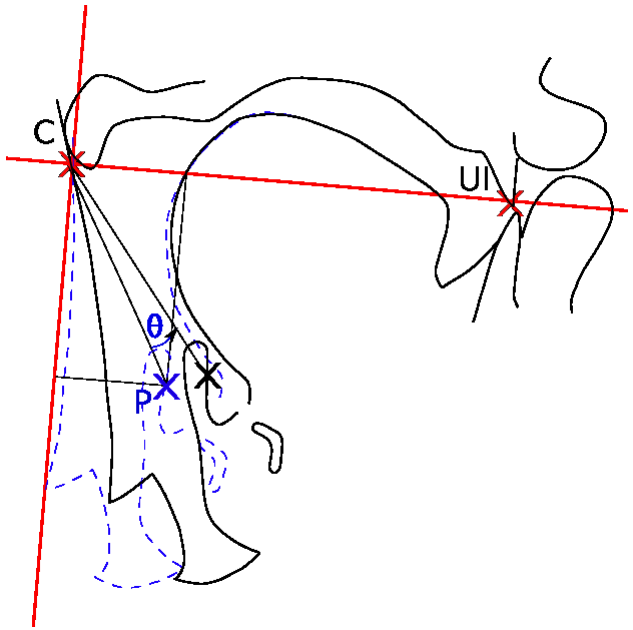


Adaptation of the model

- Two scale factors (one for the mouth direction and one for the pharynx direction).
- One rotation for the mouth but affects the whole head.

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{bmatrix} \alpha_m \cos\phi & \alpha_p \sin\phi \\ \alpha_m \sin\phi & \alpha_p \cos\phi \end{bmatrix} \begin{pmatrix} x - x_{UI} \\ y - y_{UI} \end{pmatrix} + \begin{pmatrix} x_{UI} \\ y_{UI} \end{pmatrix}$$

- A second rotation for the pharynx. Only points below the red line (C,UI) and close to the pharynx are concerned. θ is a function of the distance to (C,UI).



Evaluation of the model

- A good reconstruction for the tongue

# of comp	Error in mm	σ In mm
8	0.428	0.215
7	0.507	0.251
6	0.550	0.257
5	0.668	0.298
4	1.188	0.473

Loria

E2) Models for consonants

Objective:

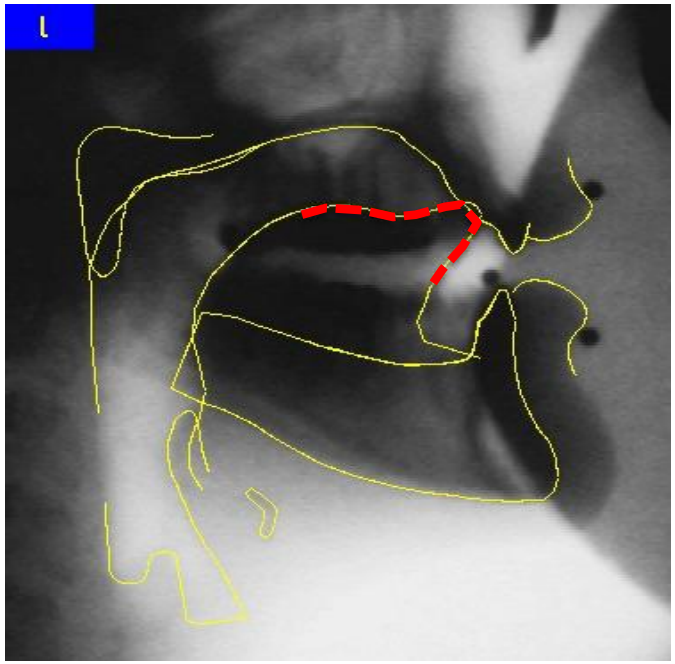
- The model should give a small geometrical error especially at the constriction

But:

- More images corresponding to vowels than consonants.
- The tongue is no longer independent from the palate and teeth because of contacts.

Hence:

- Change the weight of consonants
- Edit tongue contours to add virtual targets

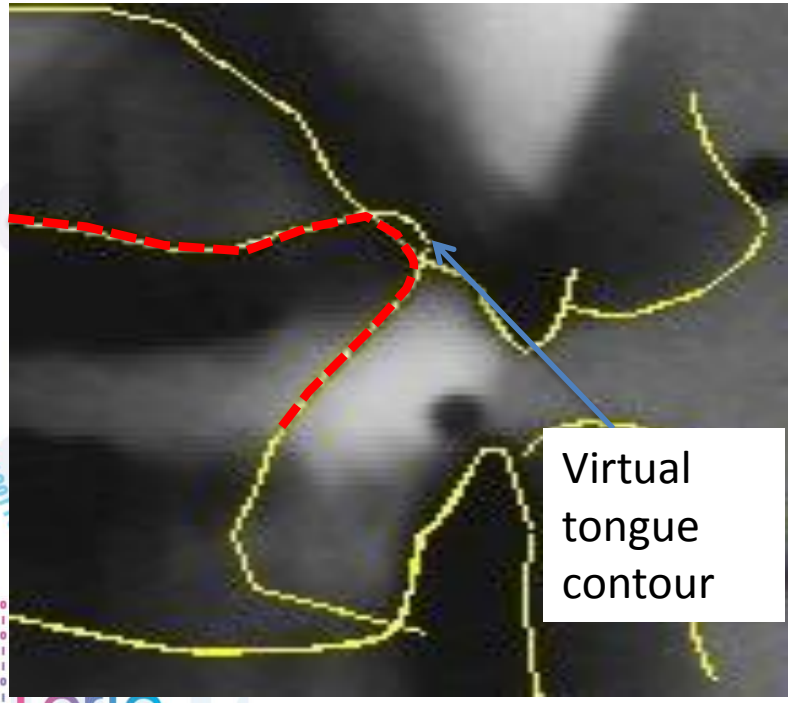


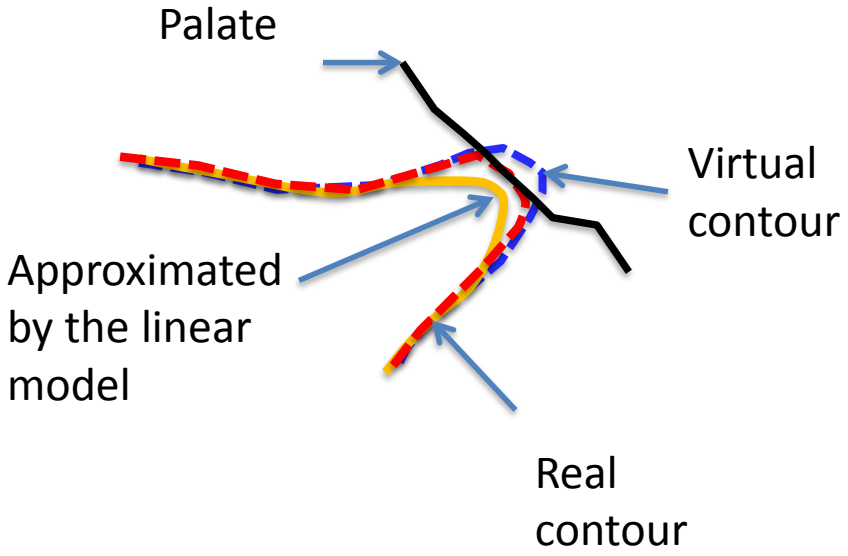
Loria

Virtual targets

- The tongue is target a point beyond the palate (which thus cannot be reached by the true tongue)

→ in some sense remove the contact effect from the deformation modes learnt from images.





Optimization of the contour in the constriction region

- In the standard linear model the coefficients can be calculated by projecting the tongue contour onto the base vectors (the linear modes).
- With a model that should approximate consonants shapes:
  - It is no longer possible because the virtual contour is not available in films from other speakers.
  - the constriction region plays a critical acoustic role and should be approximated as a priority.

→ the coefficients are optimized (Powell)

→ the cost function gives more importance to tongue points close to the palate (or any contacting region)

→ the model tongue contour is clipped (by the palate contour) before optimization.

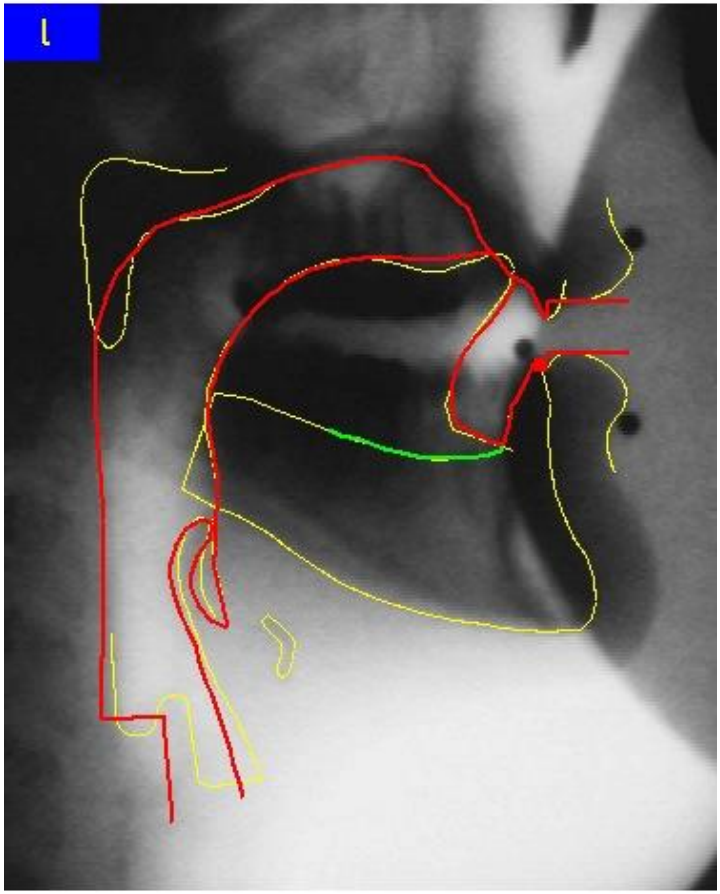
Loria

Results

# of factors	Global deviation in mm (σ in mm)	Deviation at the contriction (σ in mm)
12	0.307 (0.226)	0.205 (0.146)
8	0.366 (0.347)	0.236 (0.239)
6	0.830 (0.599)	0.567 (0.575)

Deviations (and their standard deviations) over the whole contour and in the constriction region.

The film (Vaxelaire) has 1015 images corresponding to small sentences.
More components are necessary for consonants than vowels to guarantee a good geometrical fitting.



Loria

Conclusions

- Articulatory models:
  - Representing the vocal tract in a realistic manner without incorporating a biomechanical approach.
  - Speaker variability is a big issue which can now be addressed because collecting MRI is easier.
  - Link between models derived from static and dynamic data?
- Future:
  - Full 3D models (many technical issues).
  - Binding deformation modes learnt from factor analysis to biomechanical models to make them easily usable.

Loria