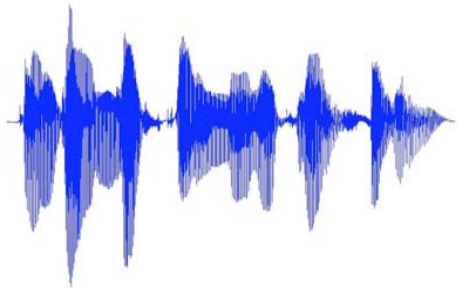
National Technical University of Athens, Greece
LORIA/INRIA, France

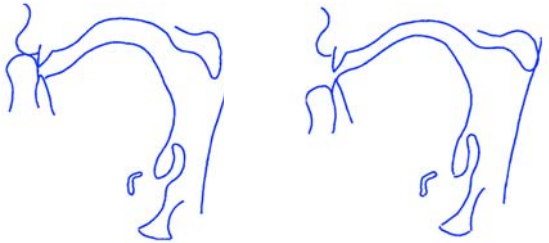
Inversion from Audiovisual Speech to Articulatory Information by Exploiting Multimodal Data

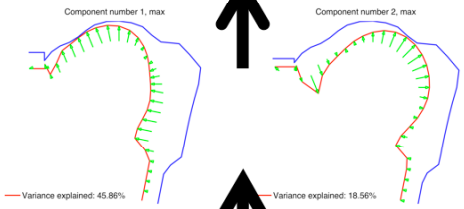
N. Katsamanis, T. Roussos, P. Maragos,
M. Aron, M.-O. Berger

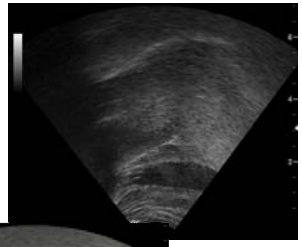
Computer Vision, Speech Communication and Signal Processing Group
http://cvsp.cs.ntua.gr



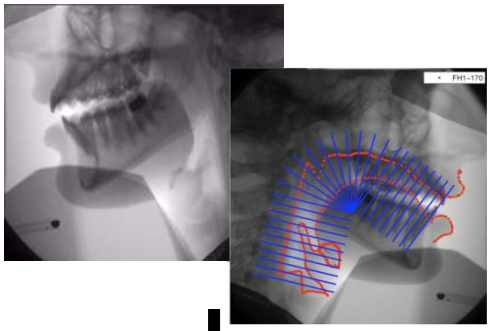
Audiovisual Speech Inversion







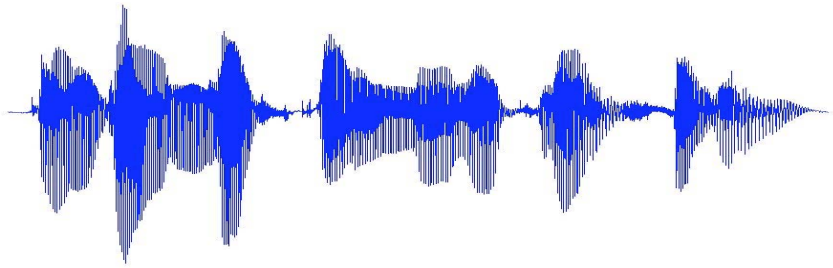
Articulatory Parameter Extraction



Articulatory Model Training

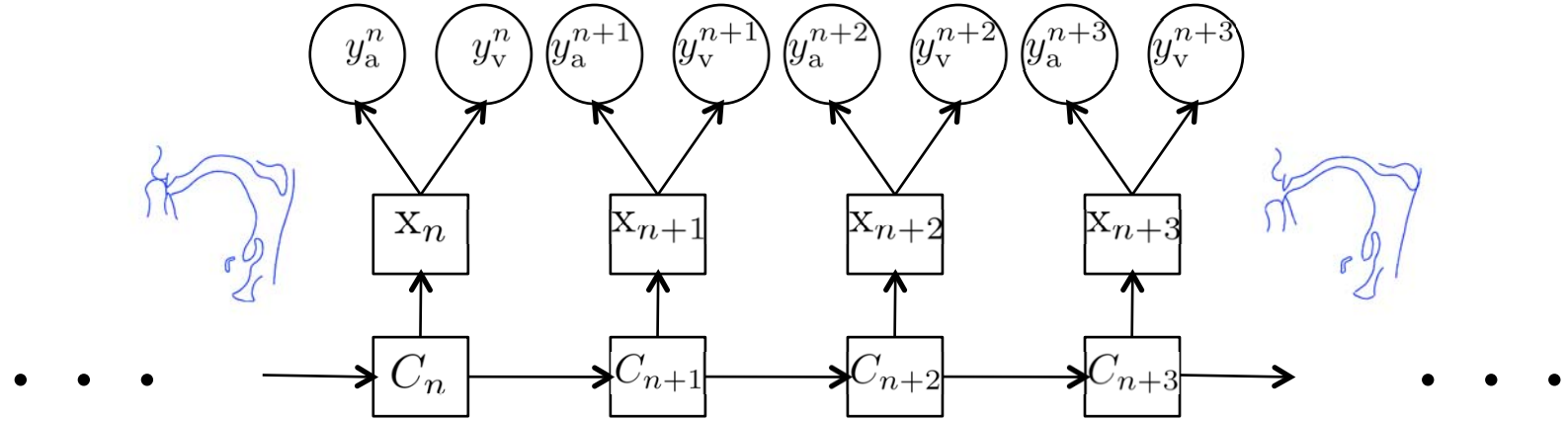
Audiovisual Speech Inversion

$\mathbf{y_a}$



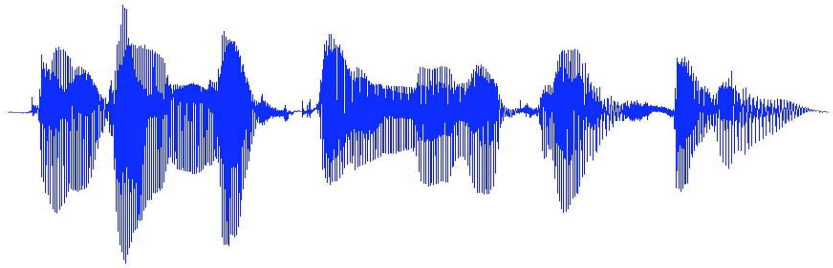
$\mathbf{y_v}$

$\mathbf{x}$



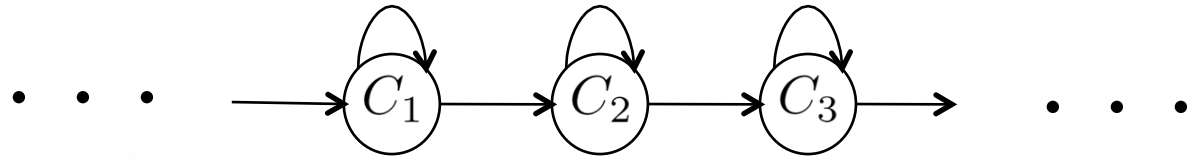
Audiovisual Speech Inversion

$\mathbf{y_a}$

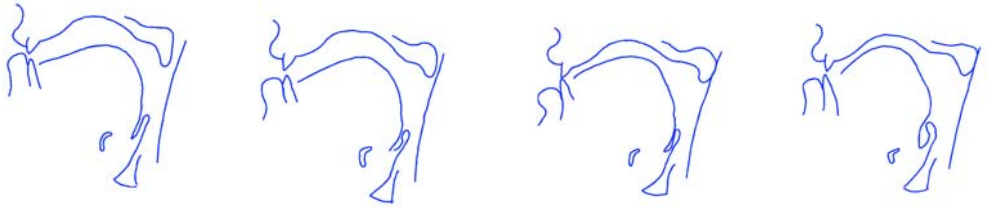


$\mathbf{y_v}$

phoneme

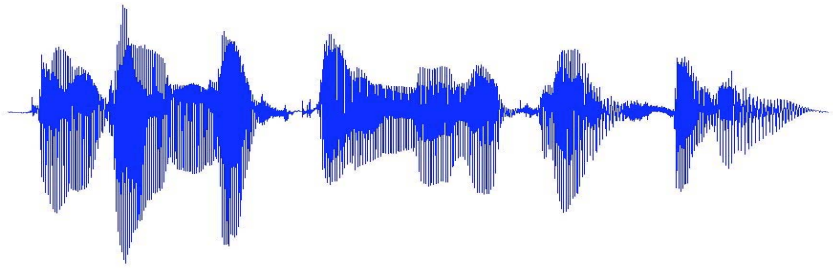


$\hat{\mathbf{x}}$



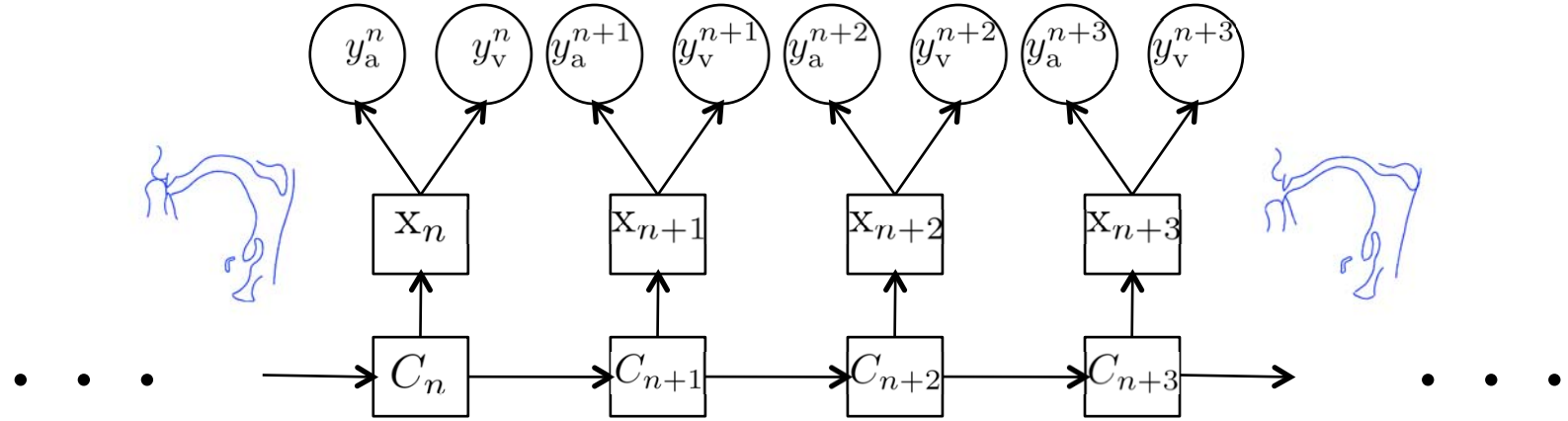
Audiovisual Speech Inversion

$\mathbf{y_a}$



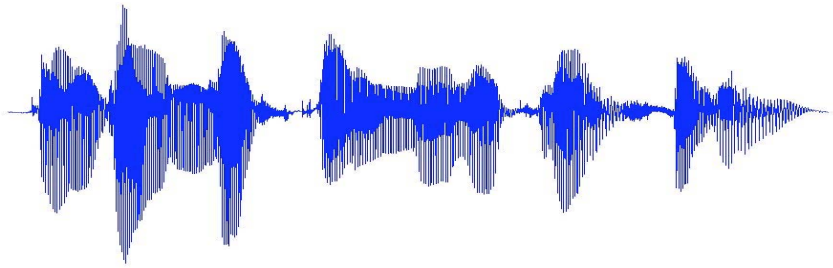
$\mathbf{y_v}$

$\mathbf{x}$



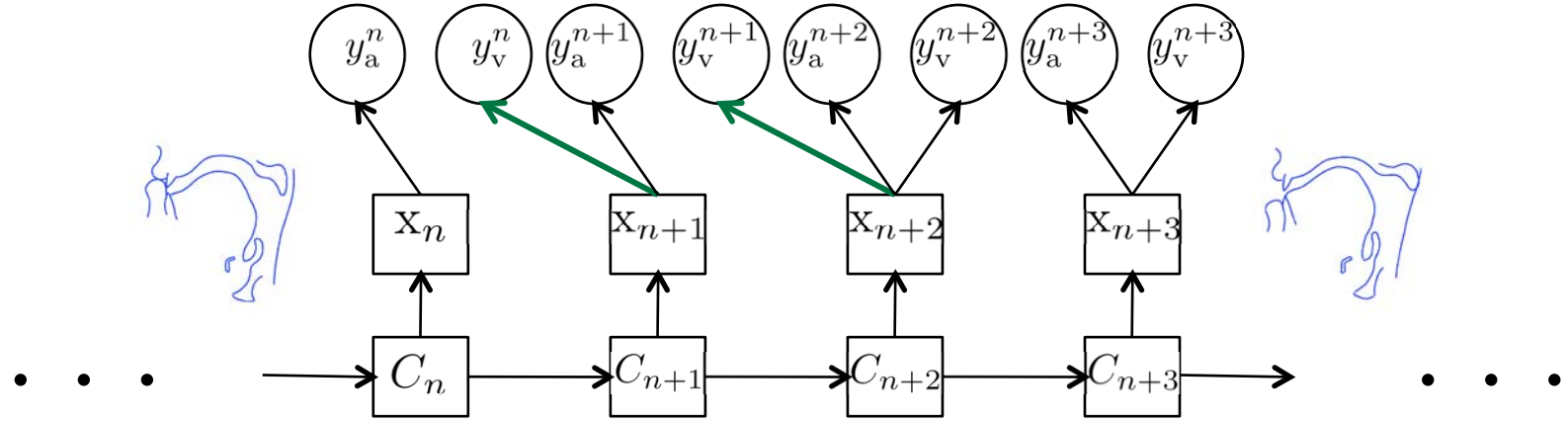
Audiovisual Speech Inversion

$\mathbf{y_a}$

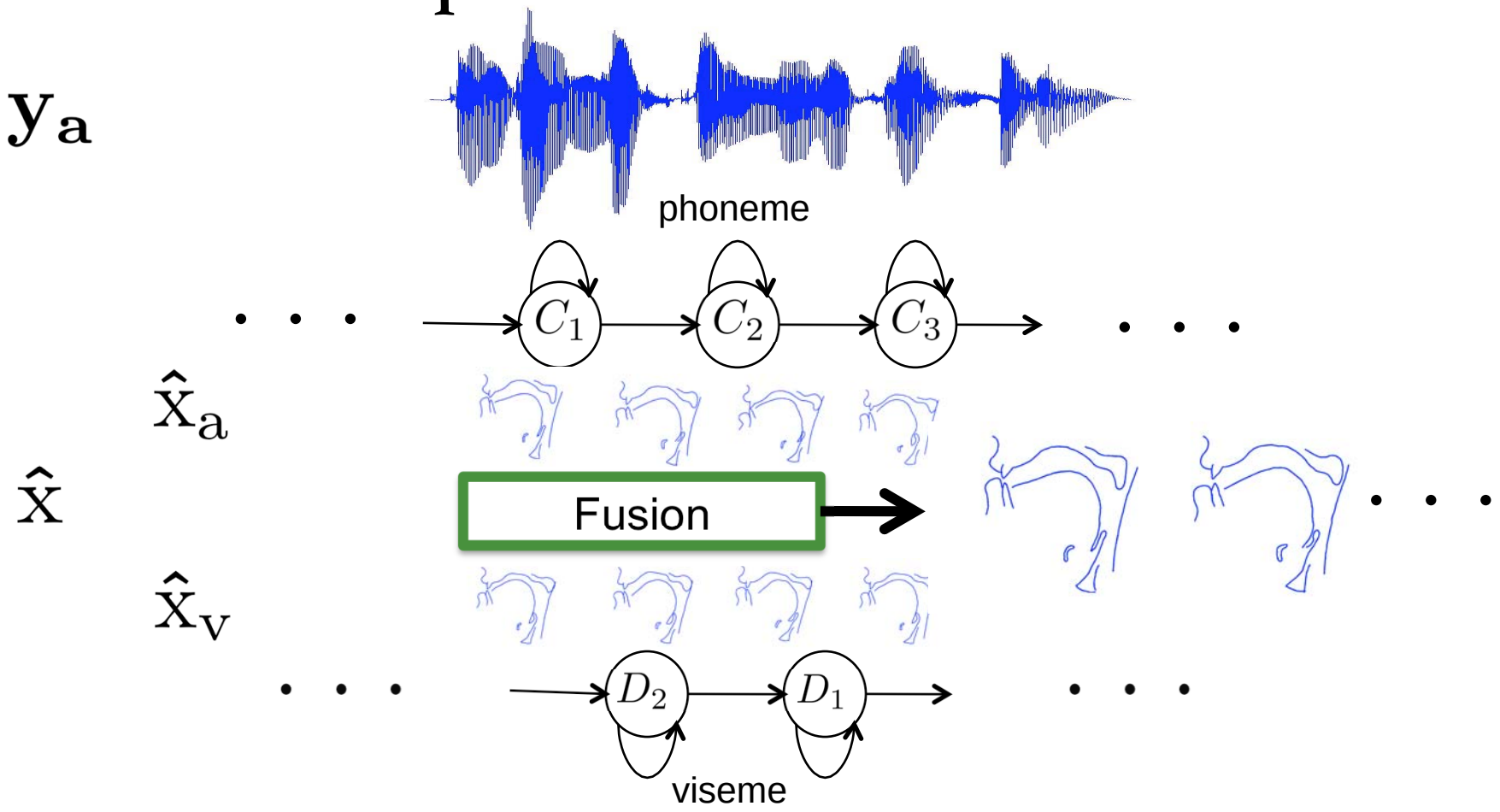


$\mathbf{y_v}$

$\mathbf{x}$



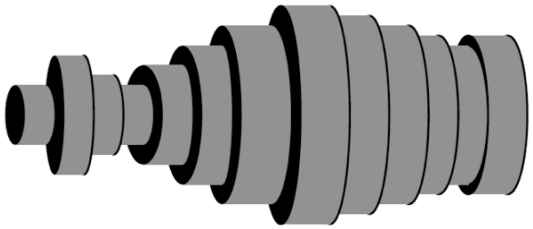
Audiovisual Speech Inversion



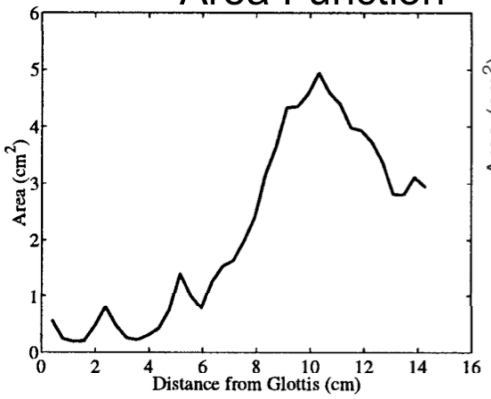
Vocal Tract Representation

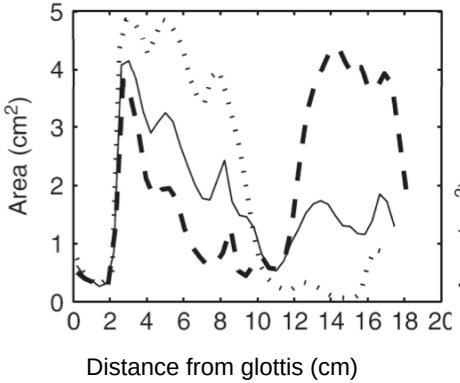
Area Function Model

Tube model

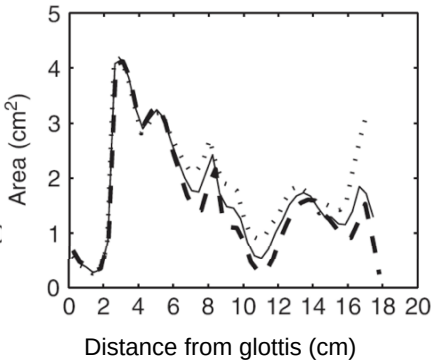


Area Function

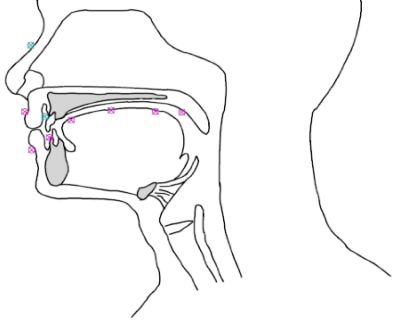


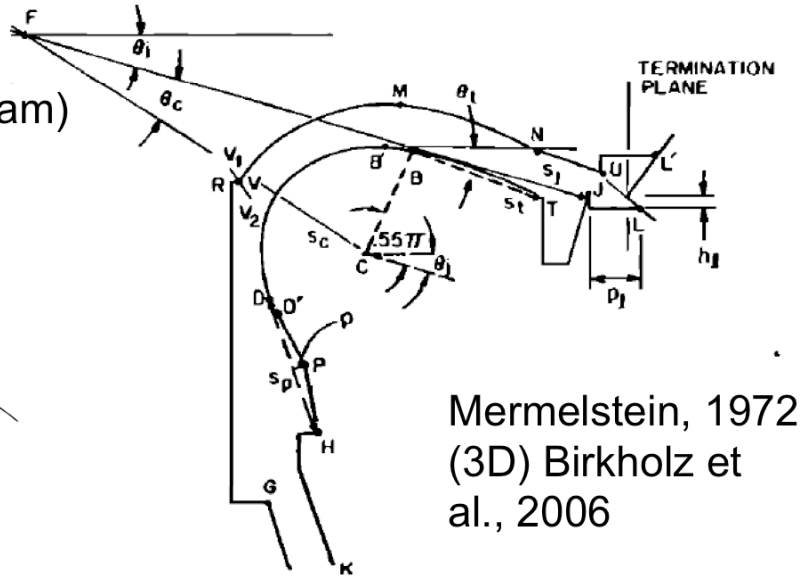


Mokhtari et al. 2007

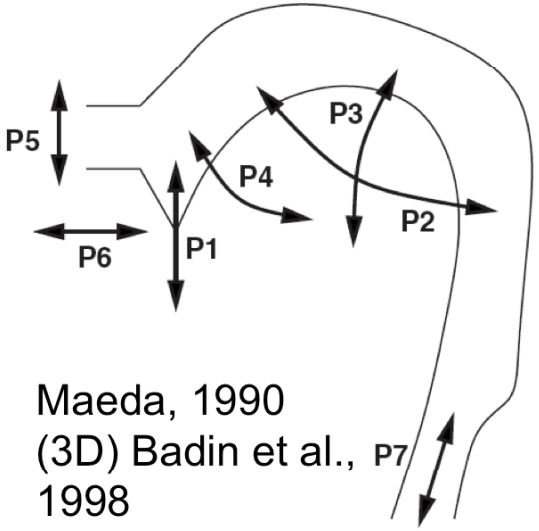


Points on articulators
(EMA, X-ray Microbeam)



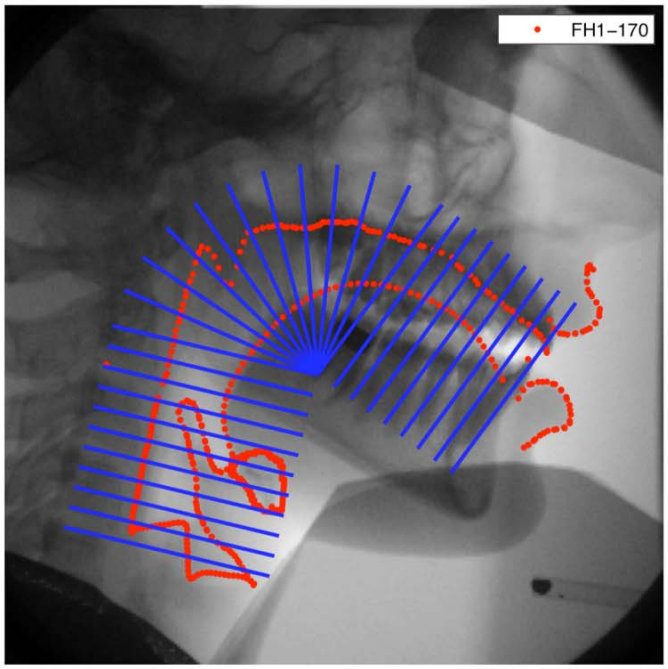


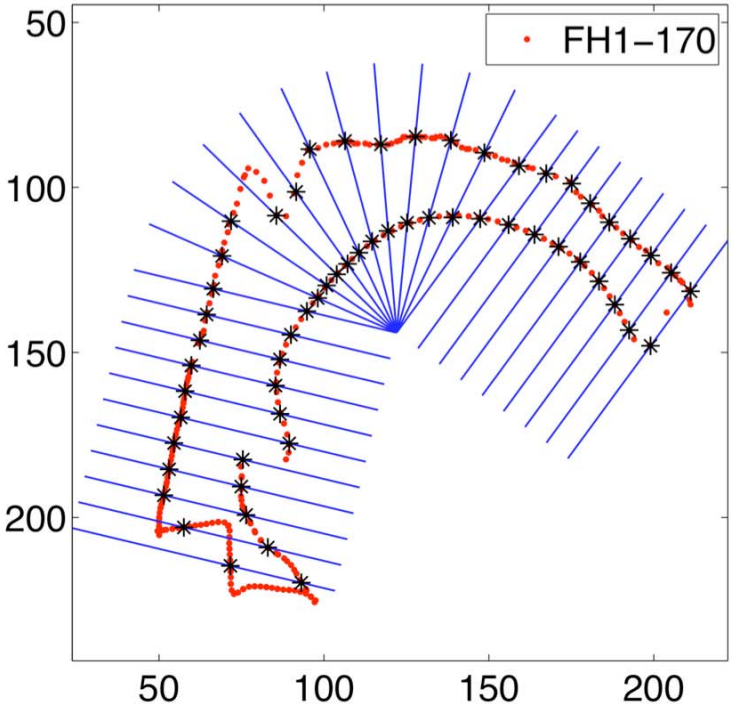
Mermelstein, 1972
(3D) Birkholz et al., 2006



Maeda, 1990
(3D) Badin et al., 1998

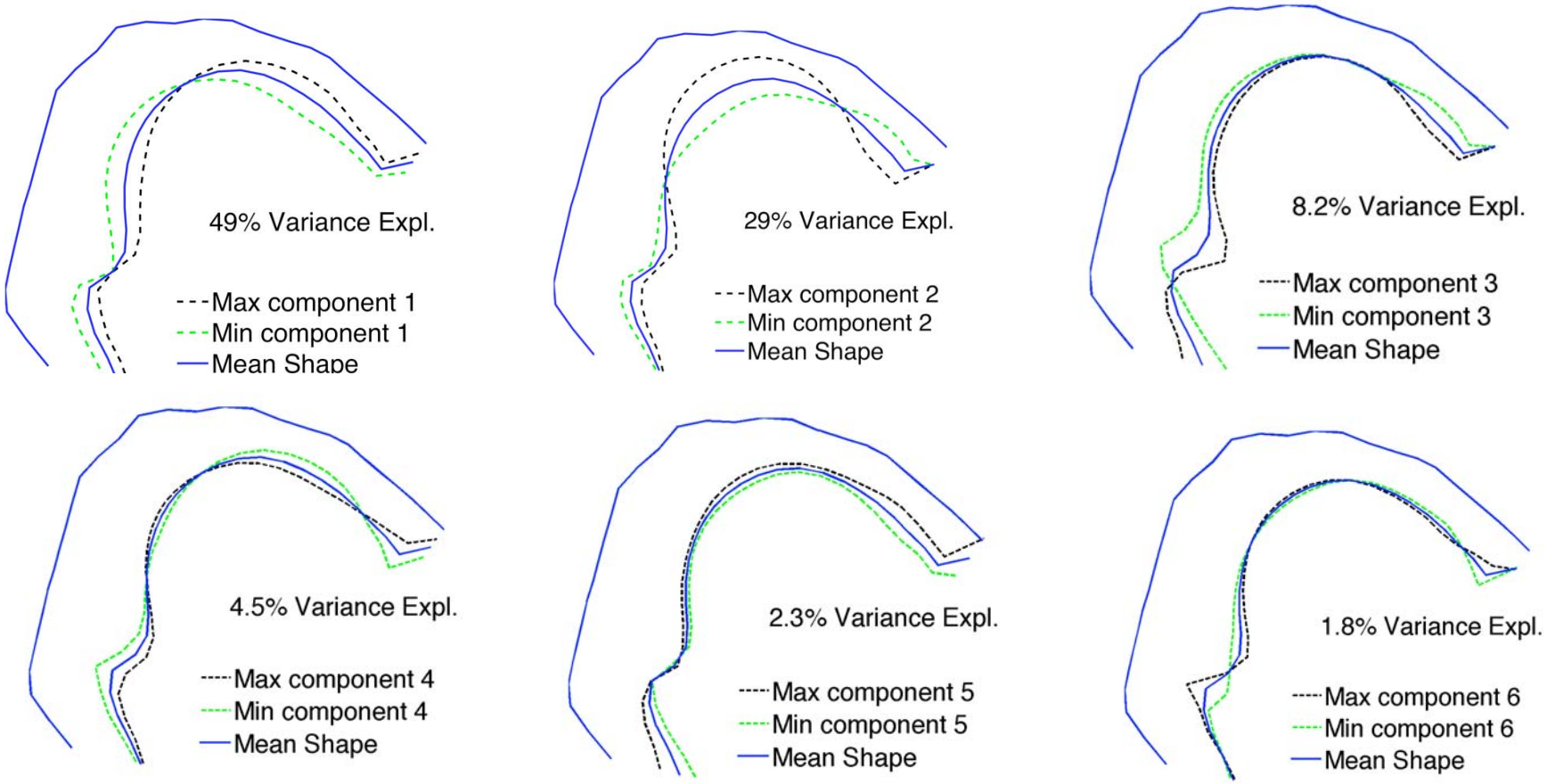
Articulatory Model Building (Xrays and Grid)



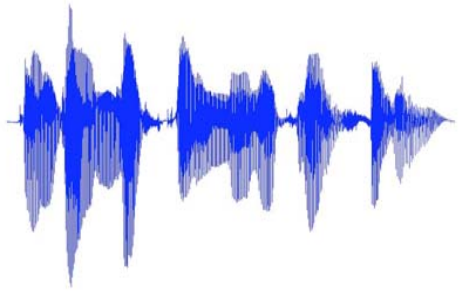


- 3-part semipolar grid (30 gridlines in total)
- approx. 700 VT contours (IPS data)
- Vocal tract semi-automatically annotated (IPS & LORIA)

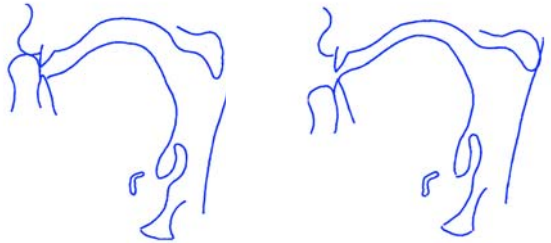
Articulatory Model Building (Principal Component Analysis)

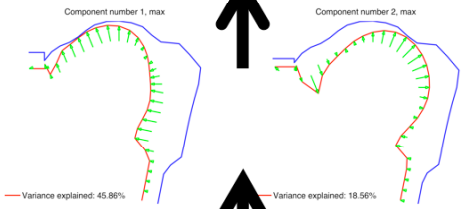


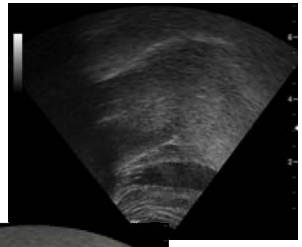
- 96% of the variance is explained by the 6-parameter model



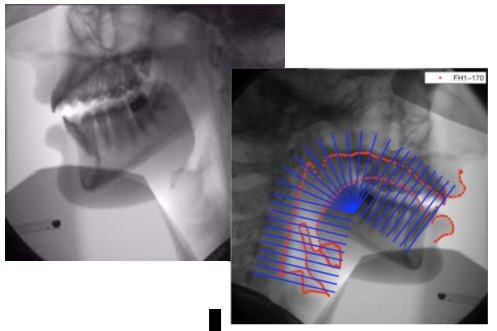
Audiovisual Speech Inversion





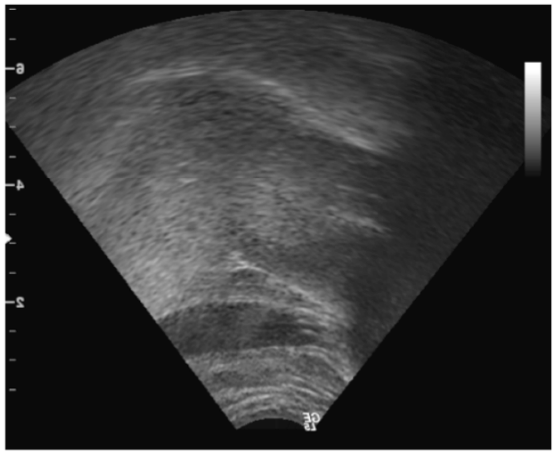


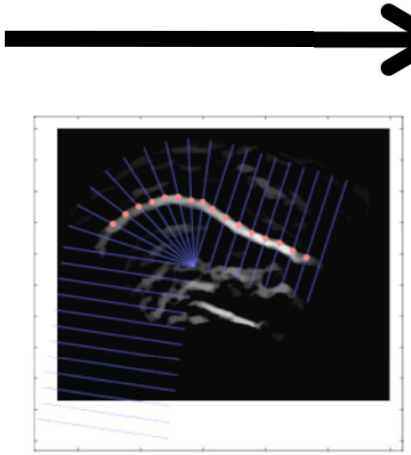
Articulatory Parameter Extraction

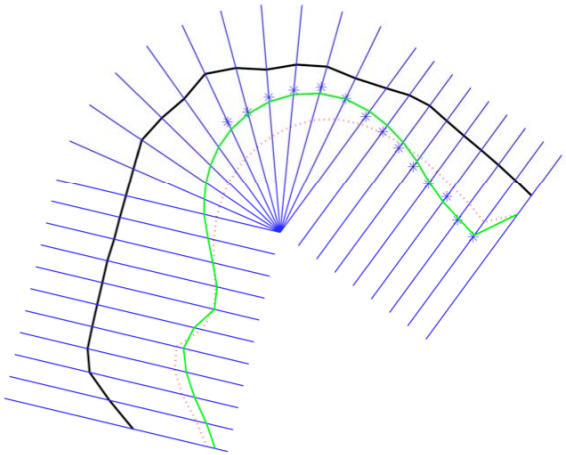


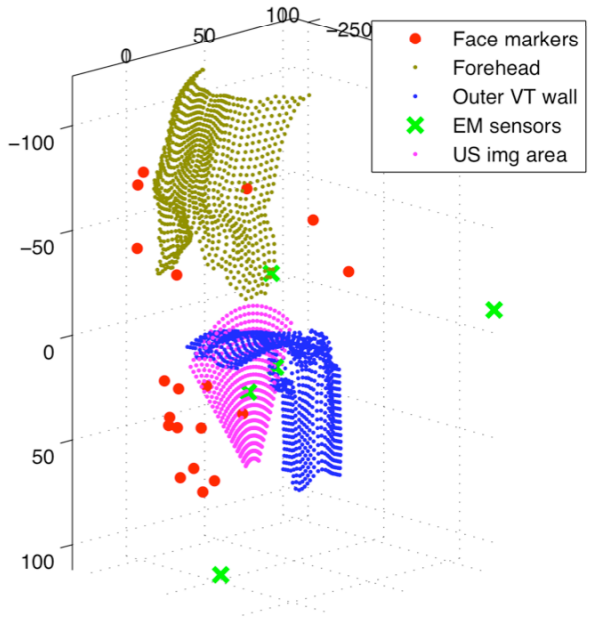
Articulatory Model Training

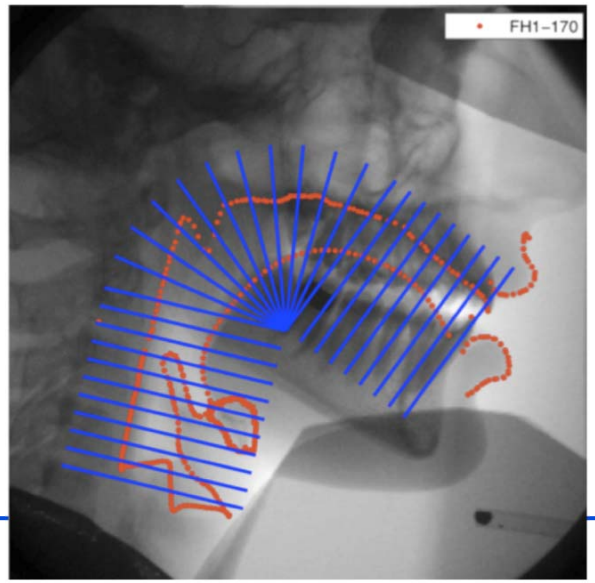
Articulatory Parameter Extraction



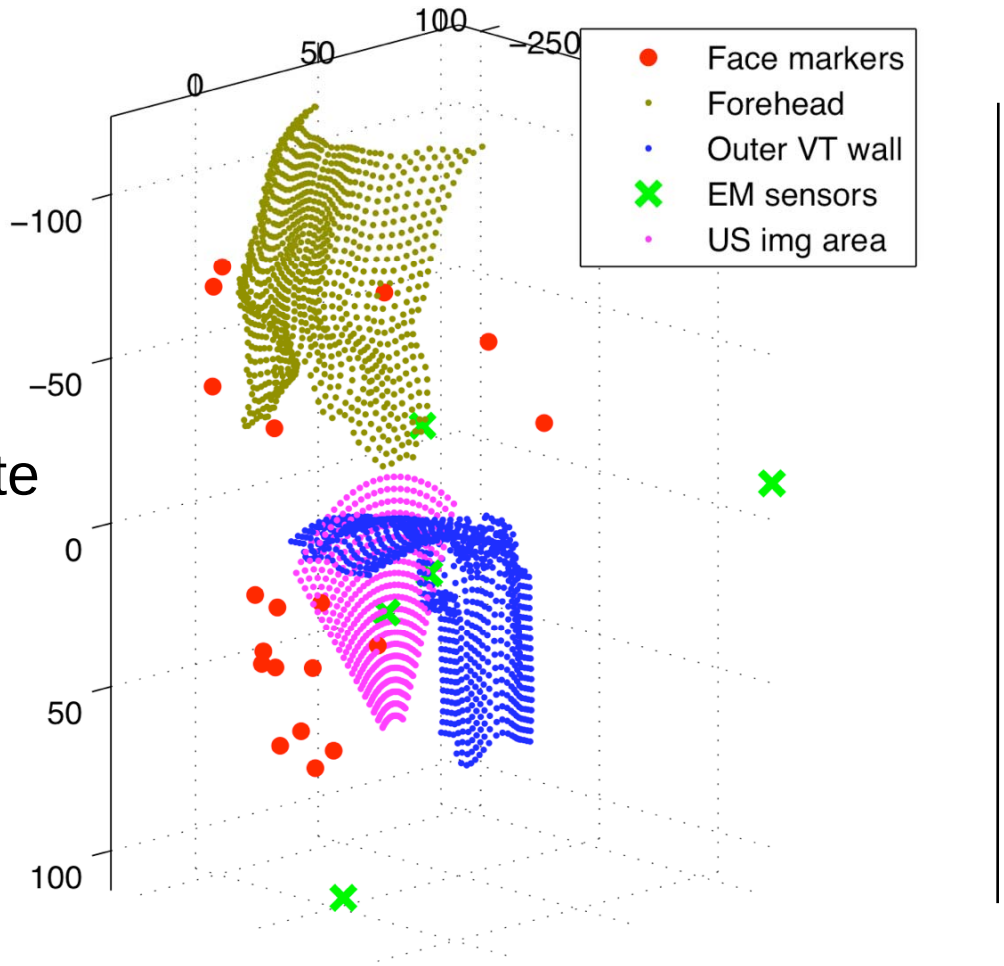




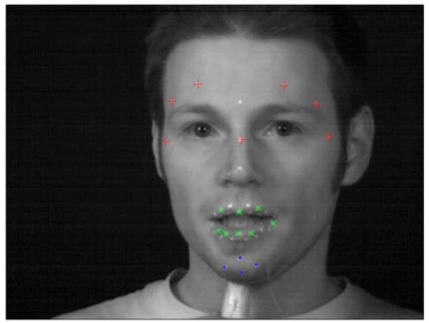




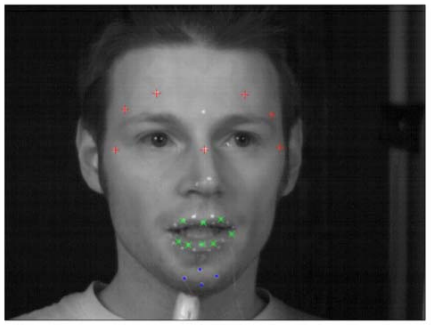
Articulatory Parameter Extraction
Registration of Multimodal Data



Head Coordinate System

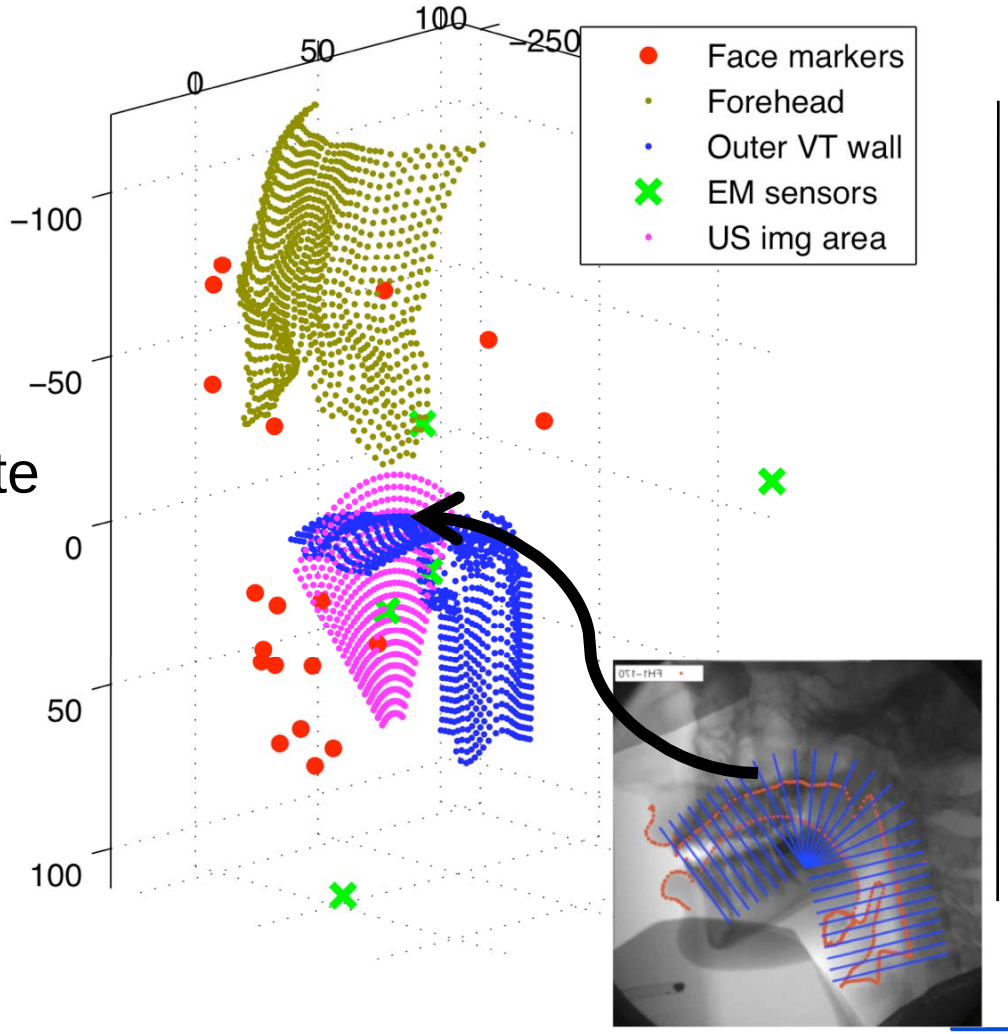


CAM 1

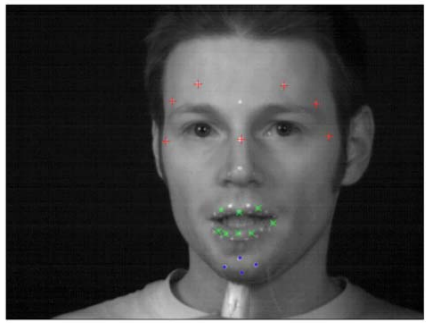


CAM 2

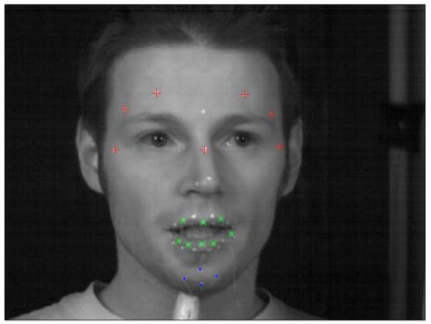
Articulatory Parameter Extraction
Registration of Multimodal Data



Head Coordinate System

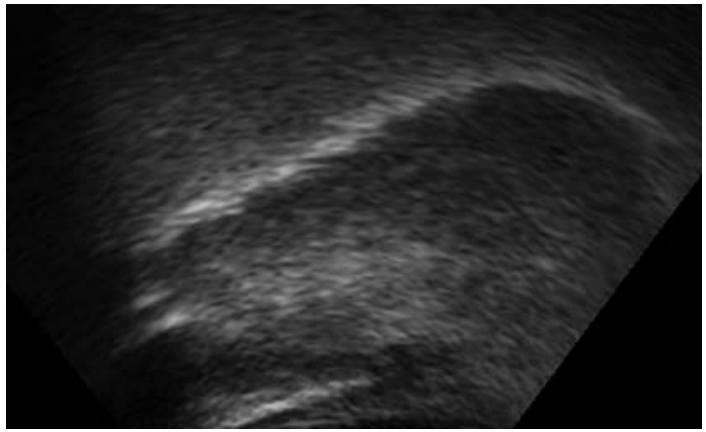


CAM 1



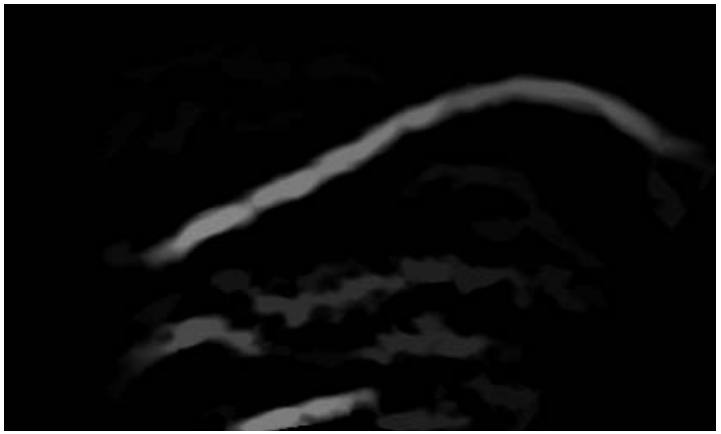
CAM 2

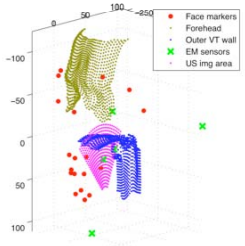
Articulatory Parameter Extraction
Ultrasound Tongue Tracking

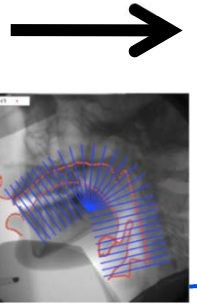


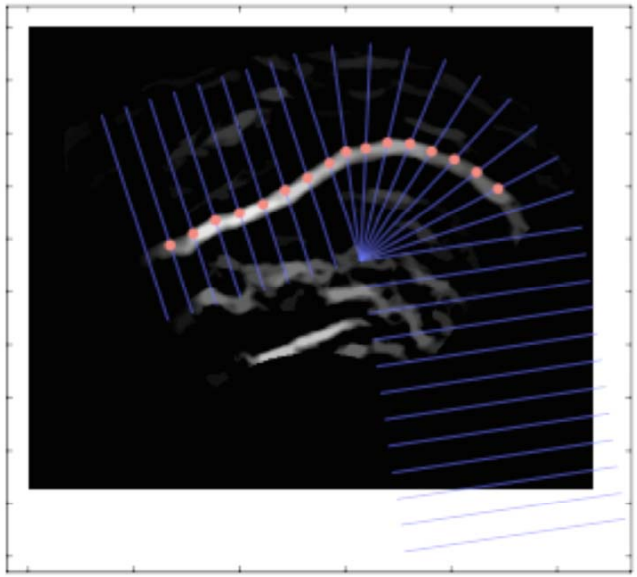
Denoising

Aron et al. 2008





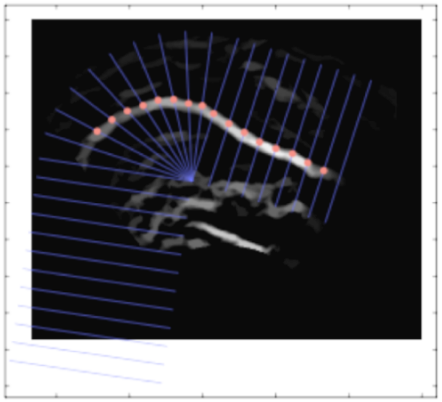


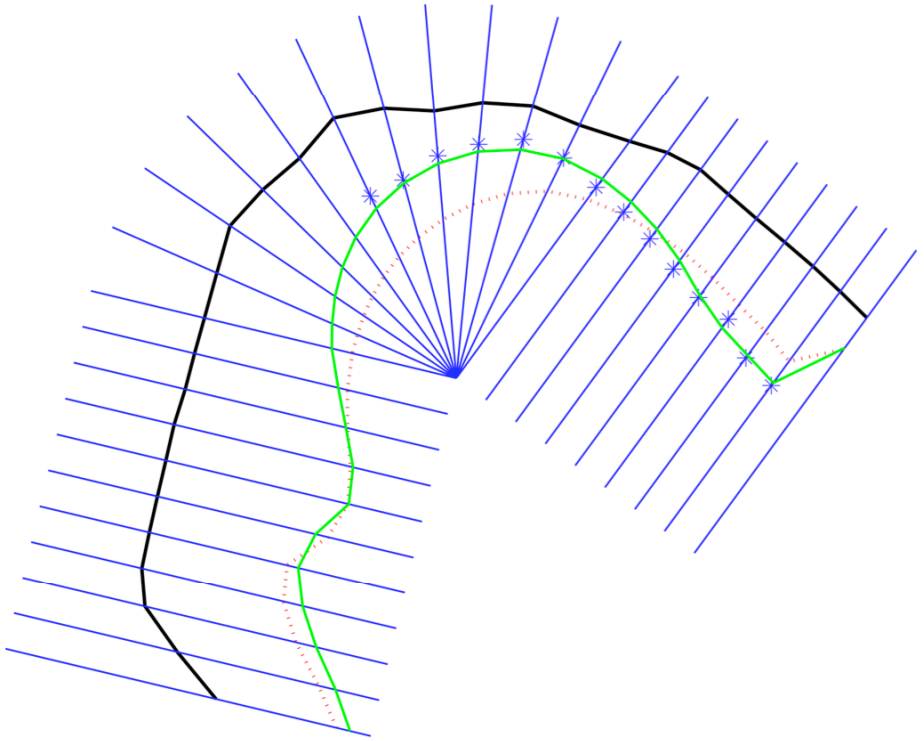


Detection of maximum image intensity on each gridline

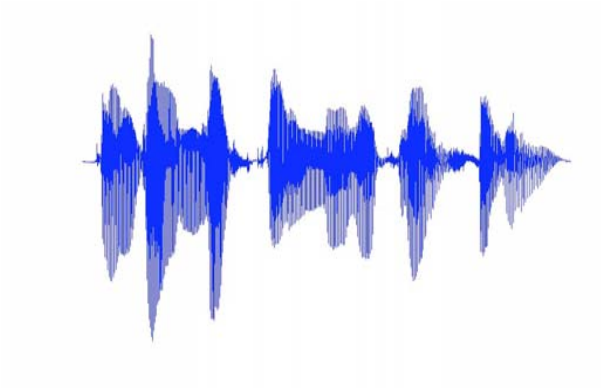
Articulatory Parameter Extraction
Model Fitting

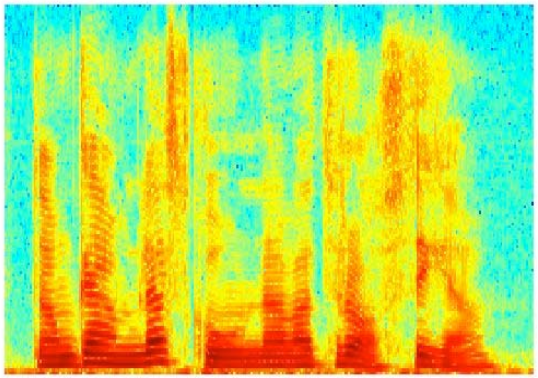
6 parameters are specified for each US frame



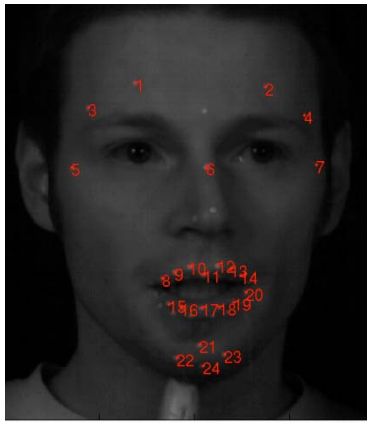


Audiovisual Speech Representation

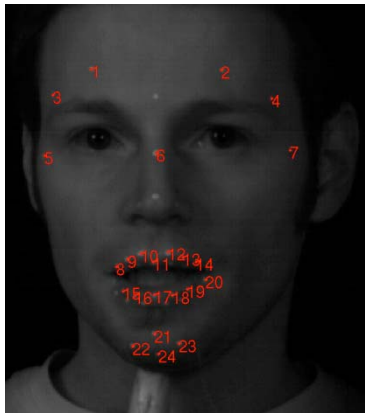




MFCCs



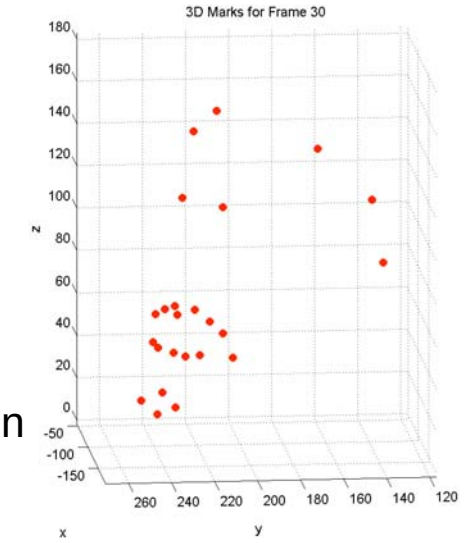
CAM 1

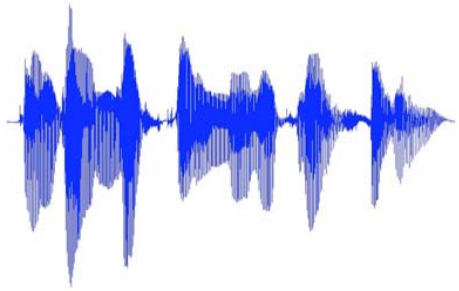


CAM 2

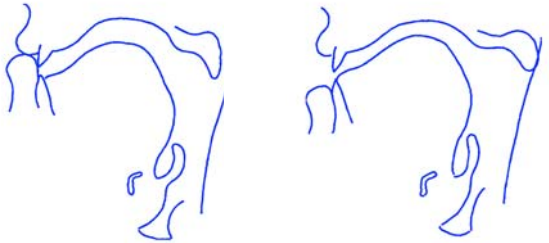
optical flow

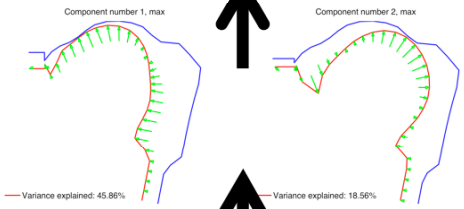
head movement compensation

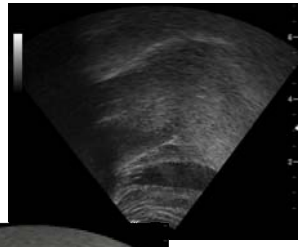




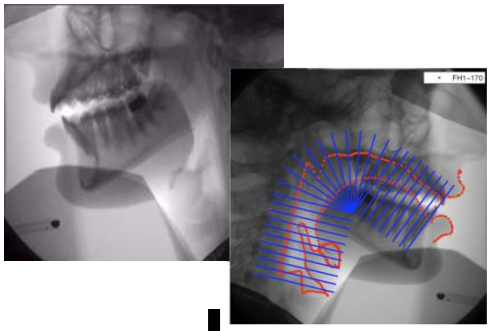
Audiovisual Speech Inversion





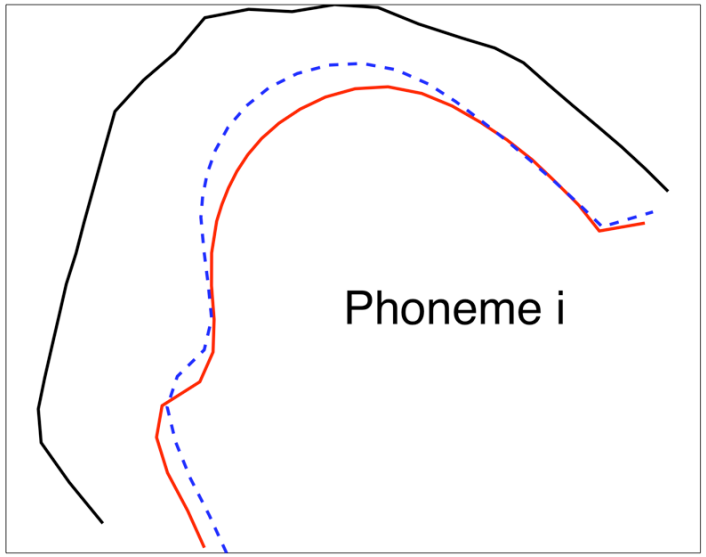


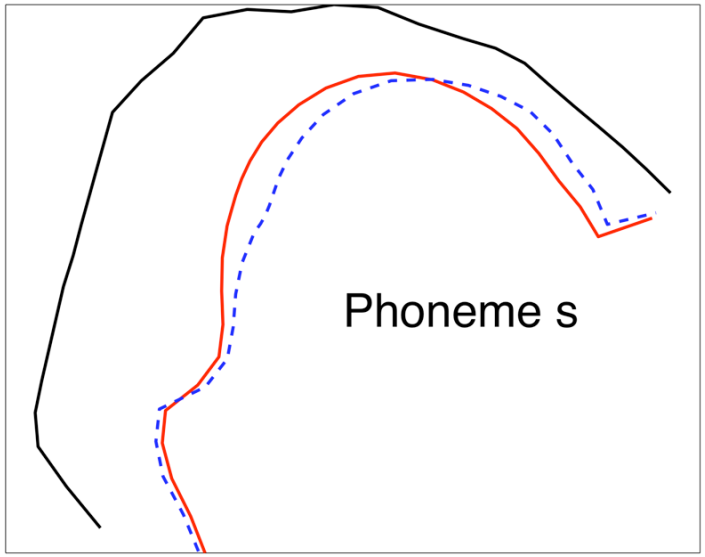
Articulatory Parameter Extraction



Articulatory Model Training

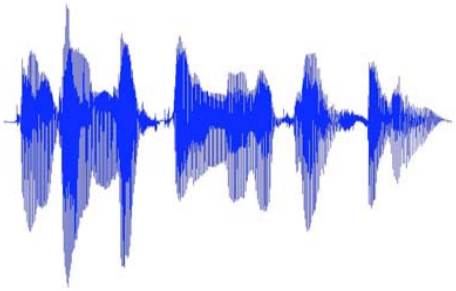
Inversion Experiments and Results



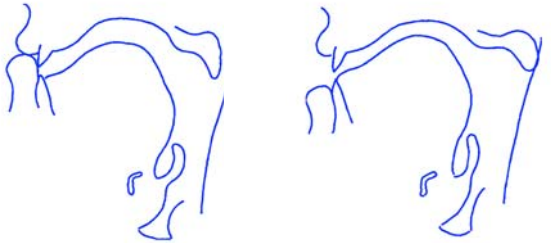


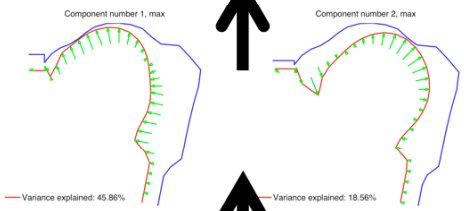
- Reference shapes are in dashed lines
- Corpus: VCVs, VVs and a number of phonetically balanced French sentences (38 phonemes)
- 6 minutes of recordings (US+EM+SV+A)
- 10% testing data
- US @ 25Hz, SV @ 120Hz, EM @ 40Hz, A @ 44kHz

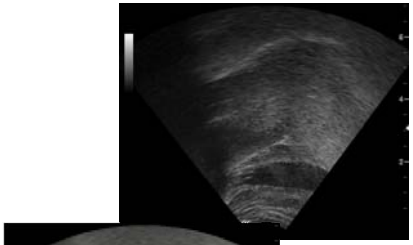
Conclusions



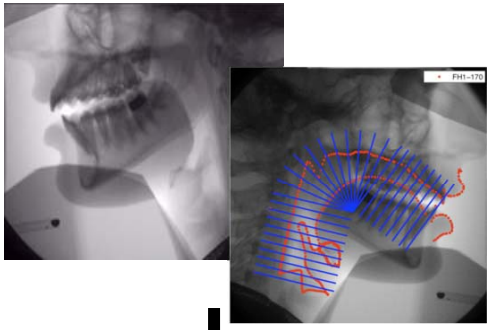
Audiovisual Speech Inversion



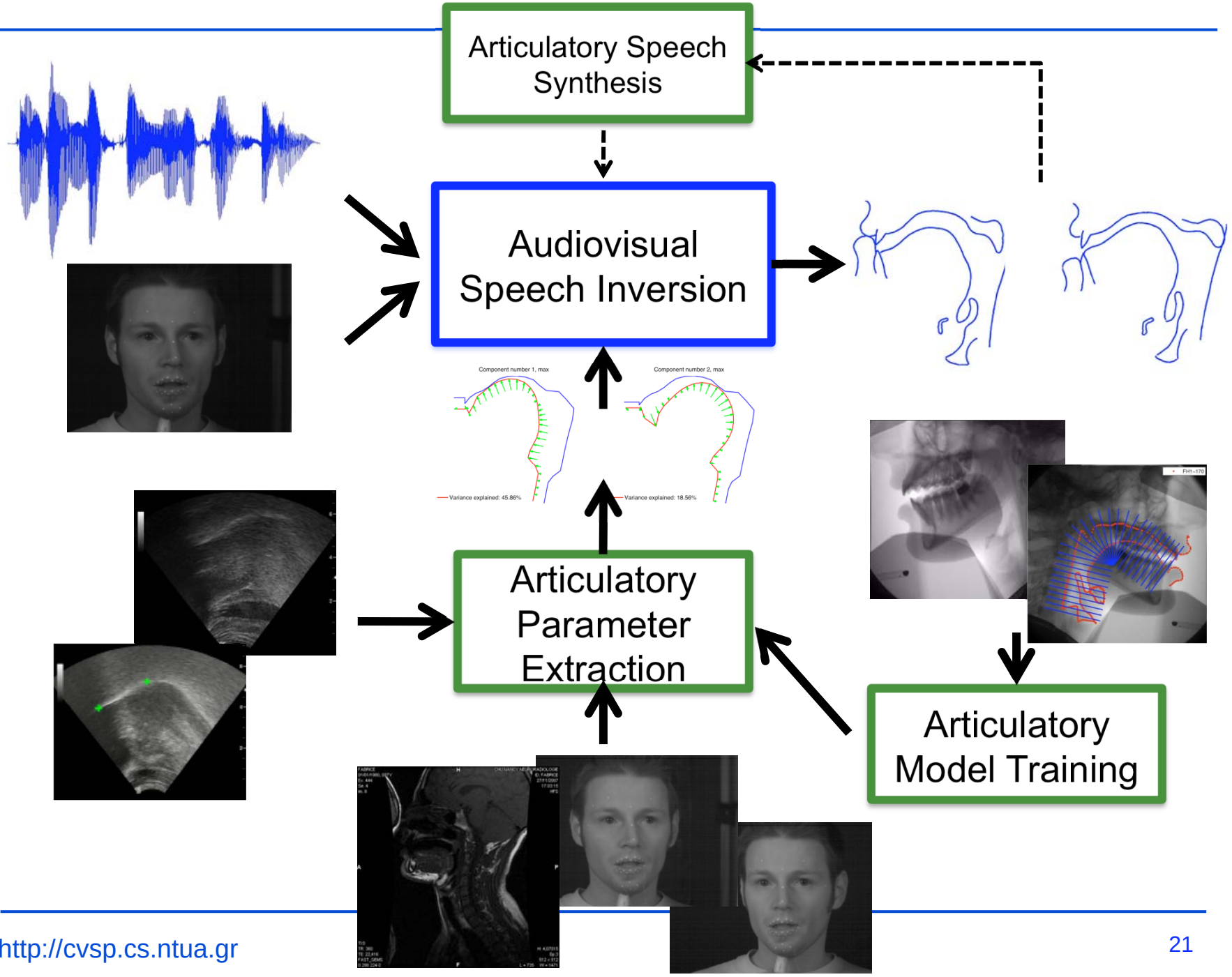


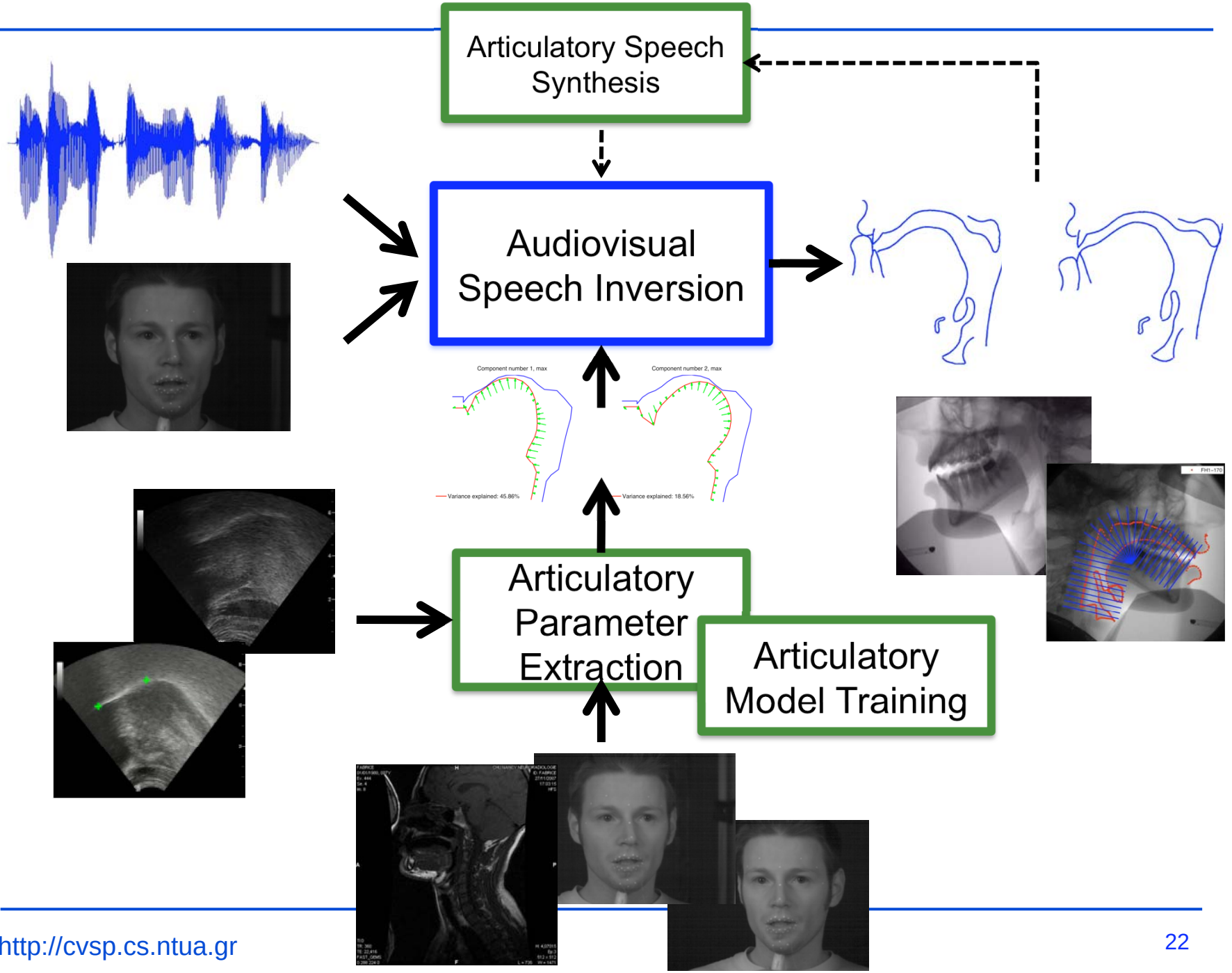


Articulatory Parameter Extraction



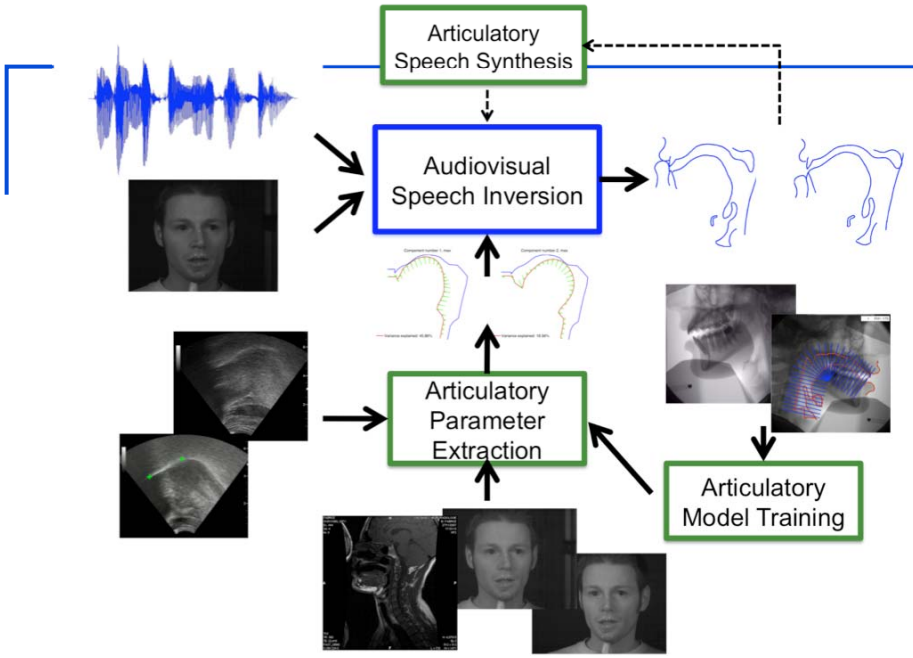
Articulatory Model Training

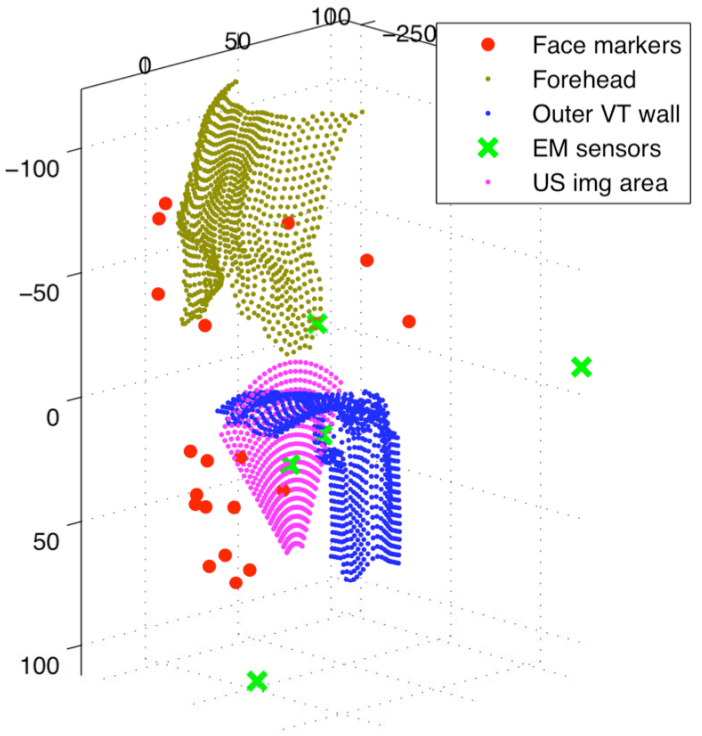

Articulatory Speech Synthesis
Audiovisual Speech Inversion
Component number 1, max
Component number 2, max
Variance explained: 45.86%
Variance explained: 18.56%
Articulatory Parameter Extraction
Articulatory Model Training


Articulatory Speech Synthesis
Audiovisual Speech Inversion
Component number 1, max
Component number 2, max
Variance explained: 45.86%
Variance explained: 18.56%
Articulatory Parameter Extraction
Articulatory Model Training

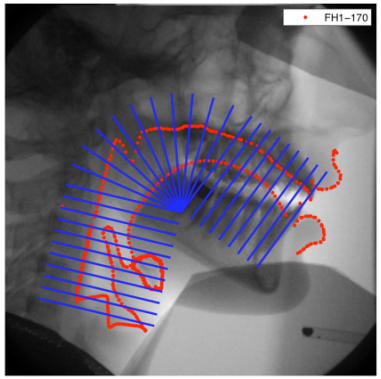
Acknowledgements

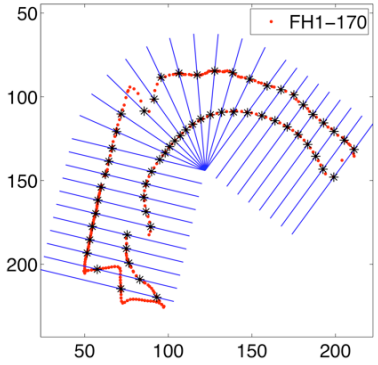
- We would like to thank
  - Yves Laprie and Erwan Kerrien at LORIA for their help in the acquisition and distribution of the articulatory data
  - Shinji Maeda from ENST, Jean Schoentgen from ULB and all other ASPI participants for many fruitful discussions
  - Fabrice Hirsch and Rudolph Sock for enduring the data acquisition process

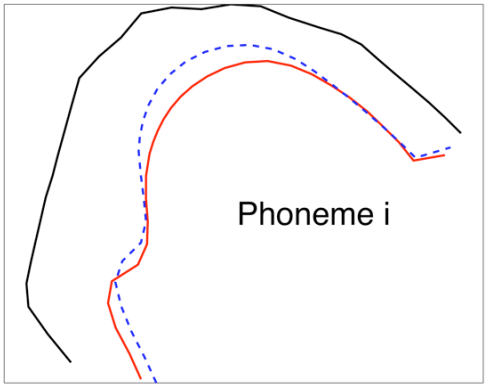

Articulatory
Speech Synthesis
Audiovisual
Speech Inversion
Articulatory
Parameter
Extraction
Articulatory
Model Training


Face markers
Forehead
Outer VT wall
EM sensors
US img area
100
50
0
−250
−100
−50
0
50
100

Thank you !


FH1−170


FH1−170
50
100
150
200
50
100
150
200


Phoneme i