High spatiotemporal resolution cineMRI films of the vocal tract using Compressed Sensing for acquiring articulatory data

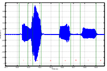
Benjamin Elie[1,2], Yves Laprie[1], Pierre-André Vuissoz[2], Freddy Odille[2]

[1]LORIA, INRIA/CNRS/Université de Lorraine
Nancy, France
benjamin.elie@inria.fr
www.loria.fr/~belie/
[2]IADI, INSERM/CHU-Nancy,CIC-IT
Nancy, France

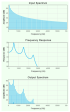
July 20, 2016

Context : Articulatory synthesis

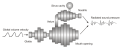
Classification of techniques for speech synthesis



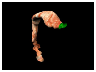
1)



2)



3)



Signal-based approach

Physics-based approach

Concatenative | Source-filter | Simplified models | Finite elements

- Speech synthesis based on physical/acoustical models
- Continuous time-domain, word/phrase level utterances
- Simulation of acoustic and articulatory phenomena

1) `http://www.phon.ucl.ac.uk/`

2) `http://www.vocaltractlab.de/`

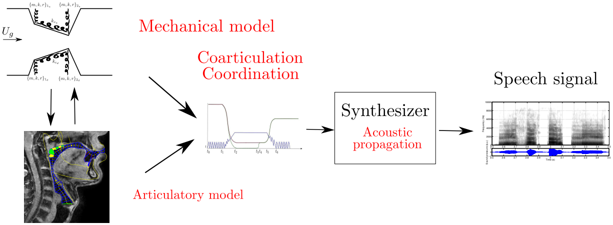
3) `http://www.magic.ubc.ca/`

Principle

Speech synthesis (utterances), **complete** and **realistic**, based on purely acoustical model

Example of an articulatory synthesizer

Phonatory source



Vocal tract deformation

- Realistic acoustics
- Articulatory control

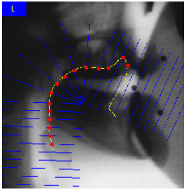
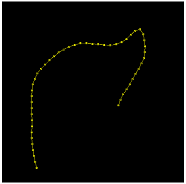
Applications: Medicine, audiovisual, language learning, text-to-speech...

Articulatory data

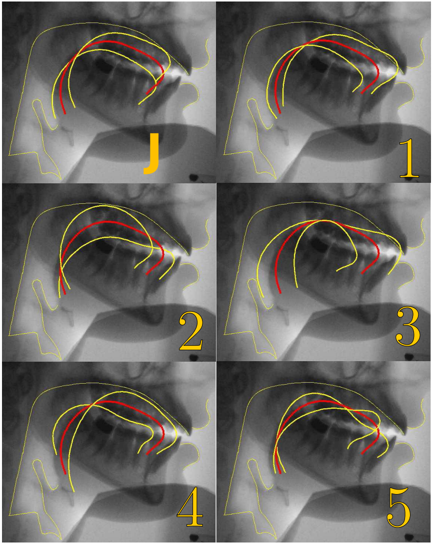
Making the articulatory model

- Large database
- Factorial analysis to reduce the number of components (PCA)
- Geometry of the vocal tract reduced to a few number of parameters

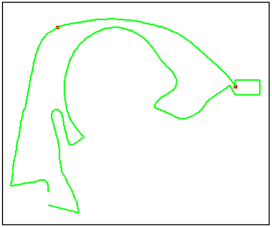
Which data ?


Tongue modes

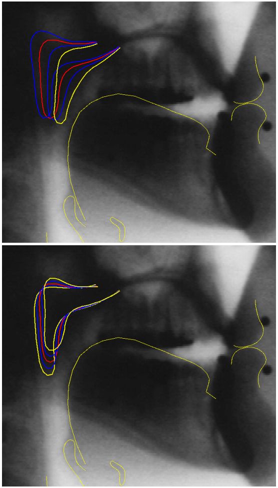


Fist mandible mode
and
5 first tongue modes

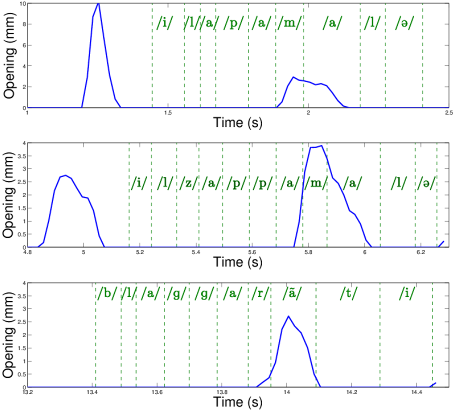


Complete model

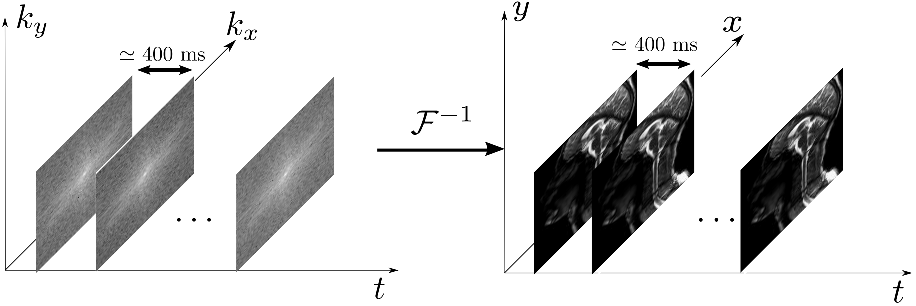
Velum modes



VPO for a few French utterances
(Laprie and Elie, ICPhS, 2015)



Acquisitions by MRI techniques: principles

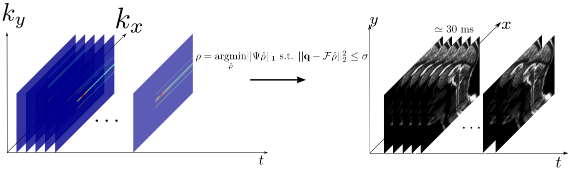


Reconstruction of midsagittal slices

- Full *k*-space sampling
  $\rightarrow$ bad temporal resolution

Sparse reconstruction (*Compressed Sensing*)

Using the sparsity for better temporal resolution



Compressed Sensing : definition

- $\rho \in \mathbb{C}^n$ is the set images to be recovered
- Ψ is the sparse transform, so that $\Psi\rho$ is K-sparse, with $K < n$
- $\mathbf{q} \in \mathbb{C}^m$, with $n > m > K$, is the observation vector (the subsampled version of the k-spaces $\mathcal{F}\rho$)
- $\Phi \in \mathbb{R}^{m \times n}$ is a CS encoding matrix that contains only 0 and 1

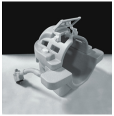
Then, in the presence of noise, and a tolerance ϵ, ρ is the solution of the convex problem

$$\rho = \underset{\hat{\rho}}{\text{argmin}} ||\Psi\hat{\rho}||_1 \quad \text{s.t.} \quad ||\Phi\mathcal{F}\hat{\rho} - \mathbf{q}||_2^2 \leq \epsilon$$

Acceleration techniques in MRI

Multi-measurement vector compressed sensing

Antenna is a $l = 16$ multi-coil receiver $\rightarrow$ 16 versions of $\mathbf{q}$
Using the fact that non-zero coefficients share the same locations



$$\mathbf{X} = \underset{\hat{\mathbf{X}}}{\operatorname{argmin}} ||\Psi\hat{\mathbf{X}}||_{1,2} \quad \text{s.t.} \quad ||\Phi\mathcal{F}\hat{\mathbf{X}} - \mathbf{Q}||_{2,2} \leq \epsilon,$$

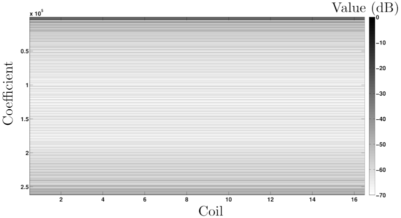
where $\mathbf{X} \in \mathbb{C}^{n \times l}$ is the l versions of the images to be recovered, and $||\mathbf{X}||_{1,2} = \sum_{i=1}^{n} ||\mathbf{X}_i||_2$

Sparse transform

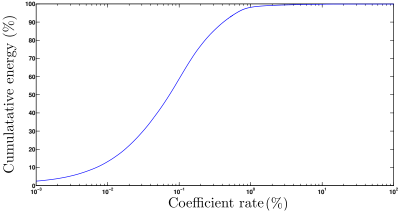
- $x - f$ space : Temporal Fourier transform of the image space
- $w - f$ space : Temporal Fourier transform of the wavelet transform of the image space

Sparsity

Joint sparsity: energy of the coefficients for each coil

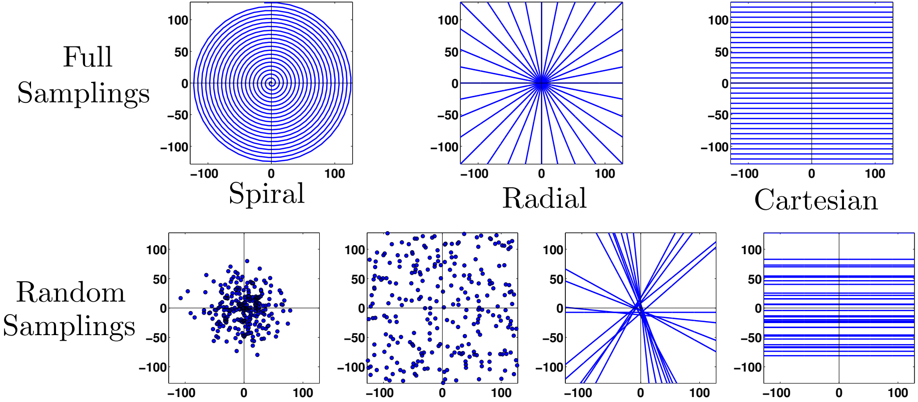


Row sparsity



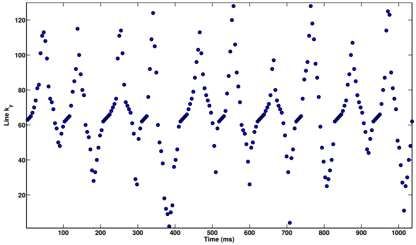
Sampling trajectory

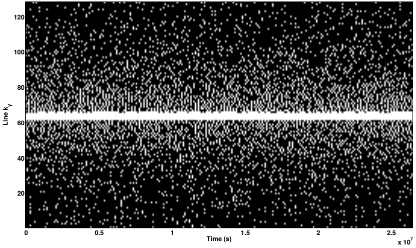
For one image, several possibilities



Sampling trajectory used in speech MRI

Pseudorandom Cartesian: an alternative to be used with CS and homodyne reconstruction



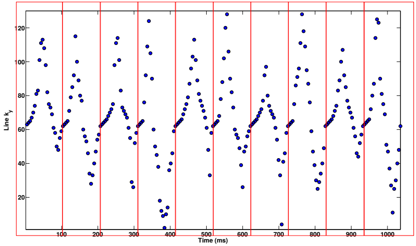


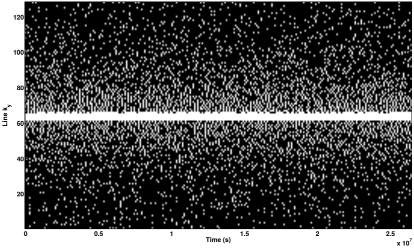
Suitable probability density

- full sampling of the central lines
- *pdf* $\propto 1/r^2$
- Partial phase line encoding for partial Fourier reconstruction

Sampling trajectory used in speech MRI

Pseudorandom Cartesian: an alternative to be used with CS and homodyne reconstruction



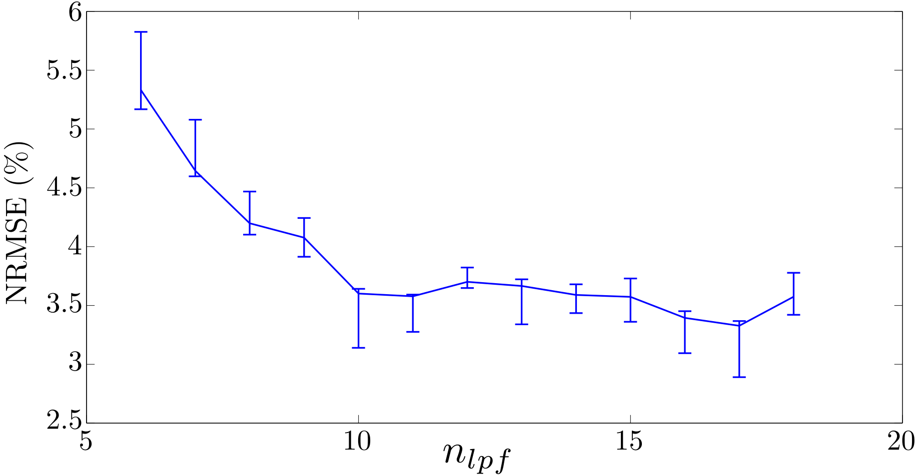


Suitable probability density

- full sampling of the central lines
- $pdf \propto 1/r^2$
- Partial phase line encoding for partial Fourier reconstruction

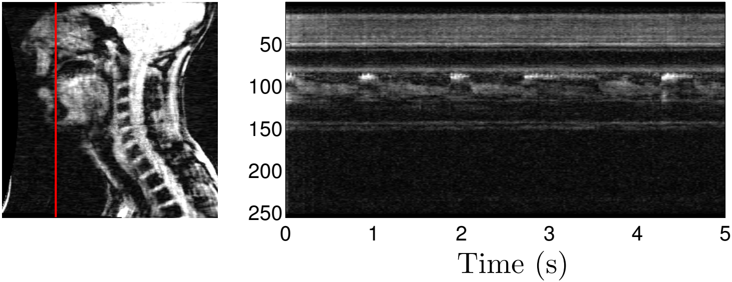
Validation

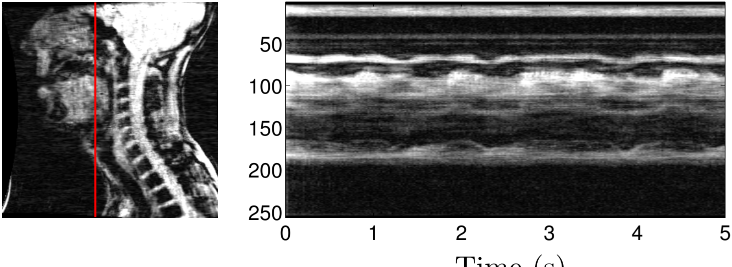
Reconstruction error from numerical phantoms



Results, fast acquisition

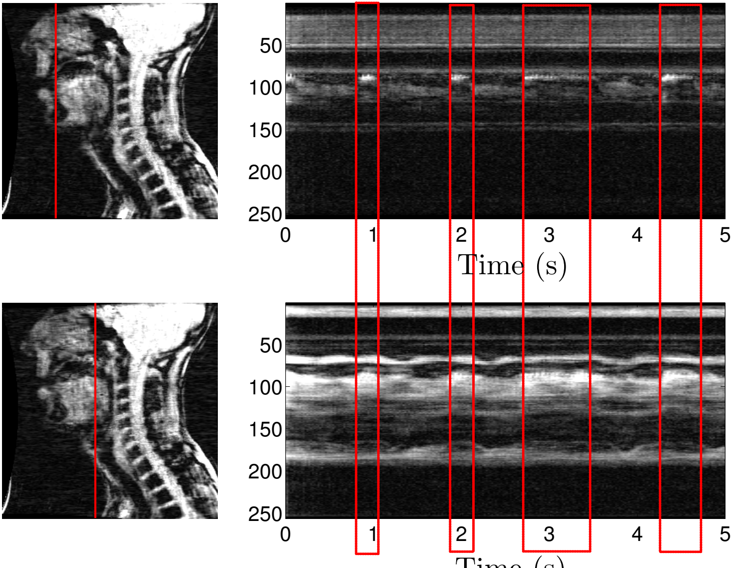
Alveolar trill, /ɑrɑ/, 48 fps, 1×1 mm, GE 3T Signa HDxt





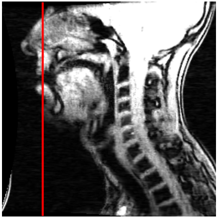
Results, fast acquisition

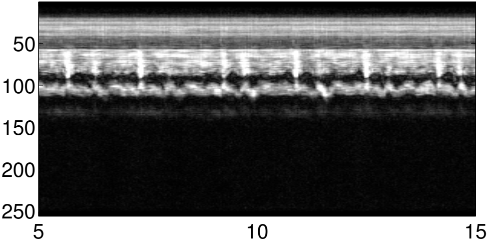
Alveolar trill, /ara/, 48 fps, 1×1 mm, GE 3T Signa HDxt

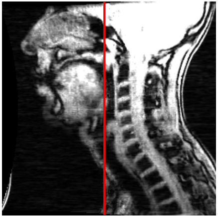


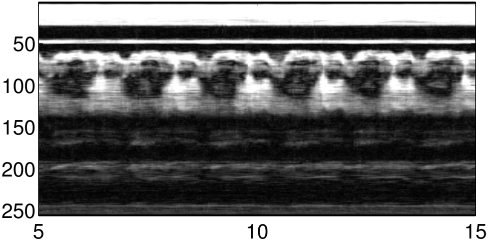
Results: moderate acquisition

"J'ai pigé la phrase" /ʒe.pi.ʒe.la.fʁɑ.zə/, 29 fps, 1×1 mm, GE 3T Signa HDxt



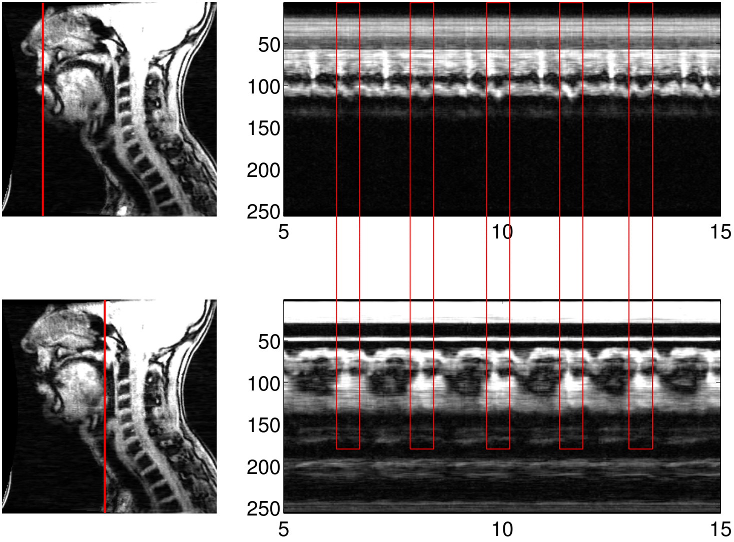






Results: moderate acquisition

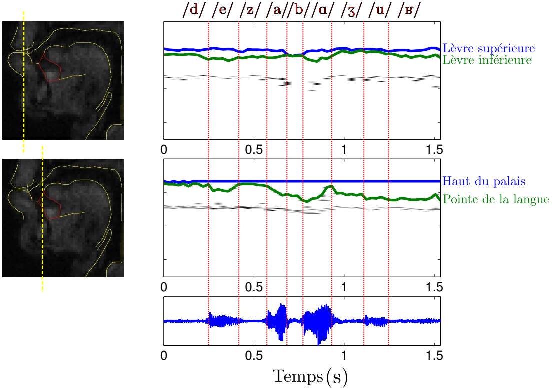
"J'ai pigé la phrase" /ʒe.pi.ʒe.la.fʁɑ.zə/, 29 fps, 1×1 mm, GE 3T Signa HDxt



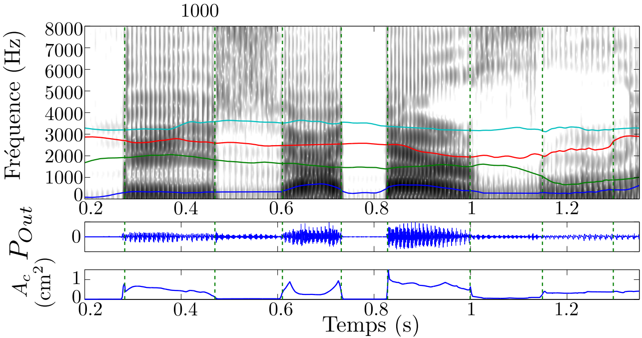
Contours

Video with contours of articulators:

- "Des abat-jours" (/dezɑbaʒuʁ/), 36.5 Hz, 2×2 mm



Acoustic synthesis



Acoustic synthesis

Conclusions

Speech MRI

- Method for visualization of articulatory movements of natural speech
- Good spatiotemporal resolution
- Choice of the trade-off speed/image quality
- **More acquisitions planned for the next future (ANR ArtSpeech)**

Extracting the articulatory parameters

- Time-tracking the contours of the articulators
- Acquisition of the time evolution of the VT deformations for building an articulatory model of the VT
- Acoustic synthesis reproducing the acoustic features of natural speech

Further works

- 3D+t compressed sensing
- Towards a 3D articulatory model