
# Acoustic-to-articulatory Inversion Using An Episodic Memory

INRIA

**Sébastien Demange**

**Research engineer, INRIA**

**France**

Nancy-Université
Université Nancy 2

**Slim Ouni**

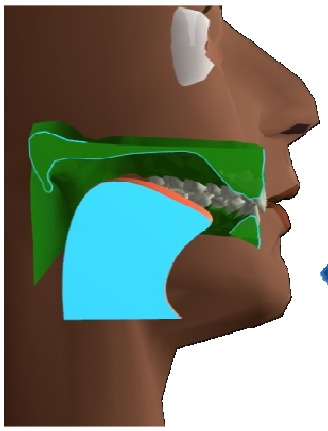
**Associate professor, University of Nancy2**

**France**

# OUTLINE

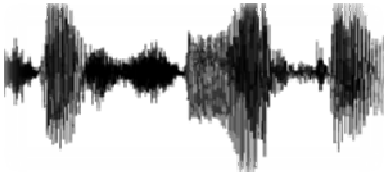
1. Introduction
   A. Acoustic-to-articulatory inversion
   B. Episodic modelling
   C. A new memory for the inversion (G-Mem)
2. Generative episodic memory (G-Mem)
   A. Building the G-Mem
   B. Inversion
3. Comparative evaluation
   A. Compared approaches
   B. Corpora
   C. Experiment set-up
   D. Quality measures
   E. Results
4. Conclusion and future work

# 1.A. ACOUSTIC-TO-ARTICULATORY INVERSION



**Speech production**
1 articulatory configuration➔ 1 sound

**Acoustic-to-articulatory inversion**
1 sound➔ many articulatory configurations



ICASSP 2011

## Goal

- Infer the articulatory movements of a speaker from the speech signal
- Difficult problem because many solutions are possible

## Articulatory dynamics is an important cue

Articulatory dynamics can solve (at least partially) the non-uniqueness of the solution as it accounts for

- The coarticulation effect
- The physical properties of the articulators (mass, velocity, degree of freedom,…)
- The speaker's articulatory strategy (expected to be more or less consistent)

# 1.B. EPISODIC MODELLING

## Biological basements

- Encoding and retrieval of past events (episodes)
- Evidences that we use our episodic memory during speech processing

## Computational models for speech processing

- Template based **speech recognition**
- Speech **synthesis** by unit selection

## Episodes

- Realizations of lexical units : phones, diphones, syllables, words…
- Description
  - time ordered sequences of acoustic observations
  - contextual information

Results expressed as **concatenations of episodes** which best explain the input with regard to some criteria

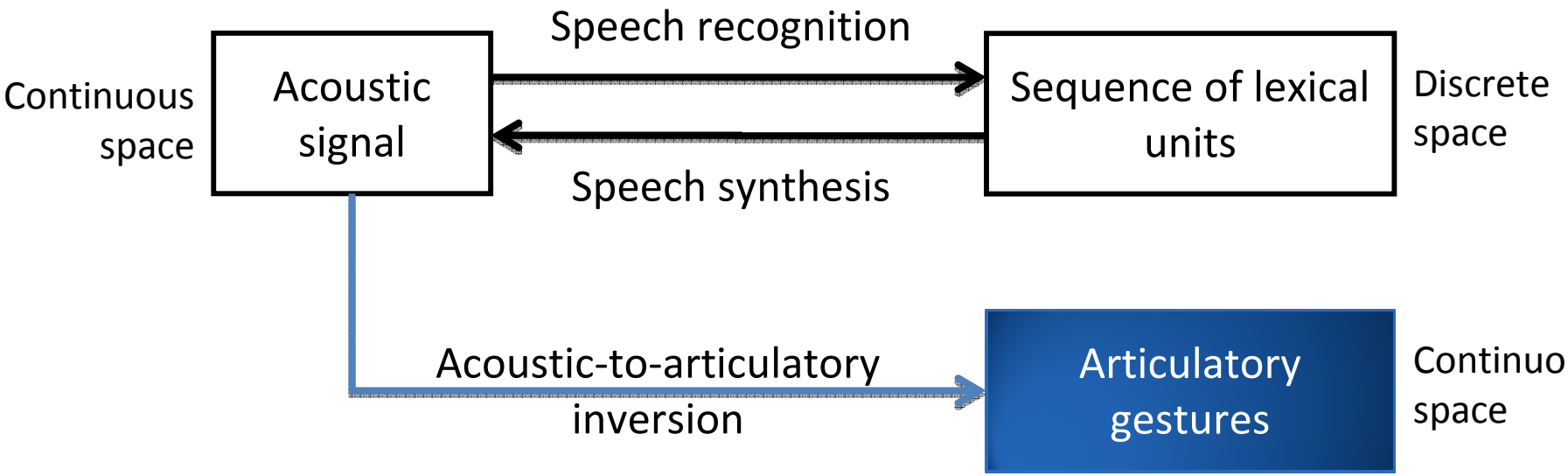
⇒ **Concatenative memories (C-Mem)**

# 1.C. A NEW MEMORY FOR THE INVERSION (1)

## An episodic memory is an attractive model because…

- We do not need to formulate any assumption about the mapping function
- The articulatory dynamics are preserved and encoded and can be easily retrieved during the inversion

## However the usual concatenative memory needs generalization capabilities because….



# 1.C. A NEW MEMORY FOR THE INVERSION (2)

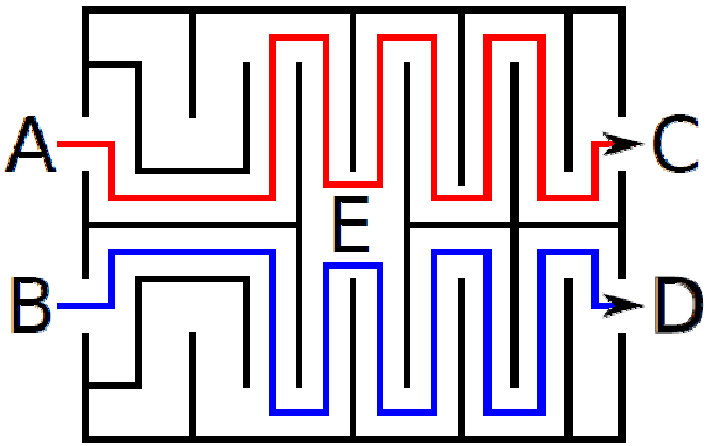
## Proposed solution

- Find local articulatory similarities between episodes
- Exploit these similarities to switch from an episode to another during the inversion

## Combining episodes allows the memory to produce unseen articulatory gestures

→ **Generative memory (G-Mem)**

- Example:




- 2 solutions after few trials
  - A → E → C
  - B → E → D
- Local similarity: E
- 2 new solutions can be deduced without any new trial
  - A → E → D
  - B → E → C

# 2.A. BUILDING THE G-MEM (1)

## Episodes

- Sequences of synchronized acoustic and articulatory observations
- Lexical unit : phone

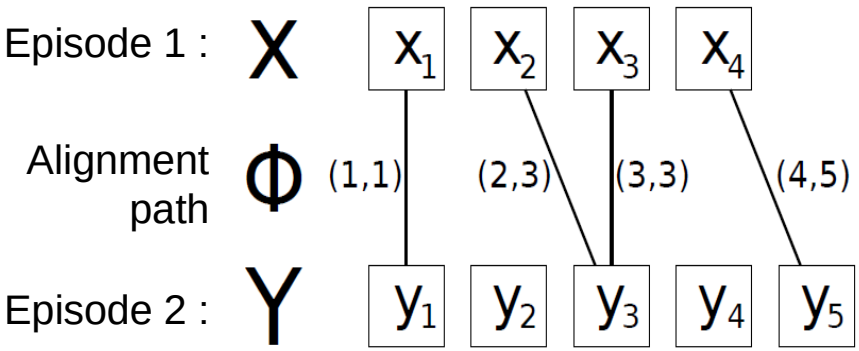
## Local similarities

- Similar articulatory configurations which occur at similar instants during different realizations of a given phone



**Temporal similarities**

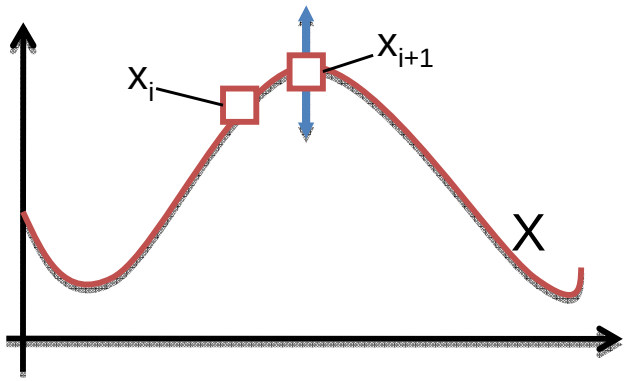
DTW + Itakura constraints




**Spatial similarities**

Articulatory target interval (ATI)

$$ATI_{x_i} = [x_{i+1} - \delta, x_{i+1} + \delta]$$



# 2.A. BUILDING THE G-MEM (2)

Let X and Y be two different articulatory realizations of a particular phone

Create a transition from any $x_i$ to any $y_j$ if

1. $D(X,Y) \leq \Delta$

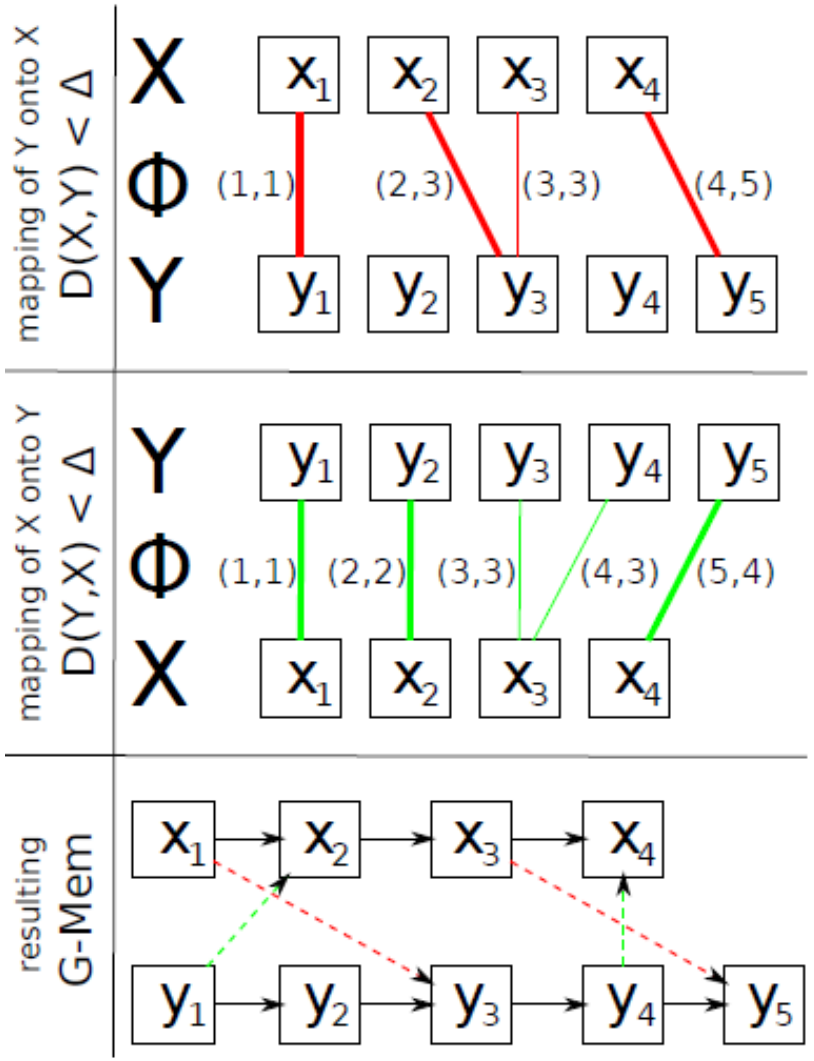
   Prevents from combining dissimilar episodes

2. $\Phi_{y,i+1} = j$

   Accounts for the **temporal similarity**

3. $y_j \in ATI_{x_i}$

   Accounts for the **spatial similarity**
   Ensures locally the articulatory naturalness and physical validity

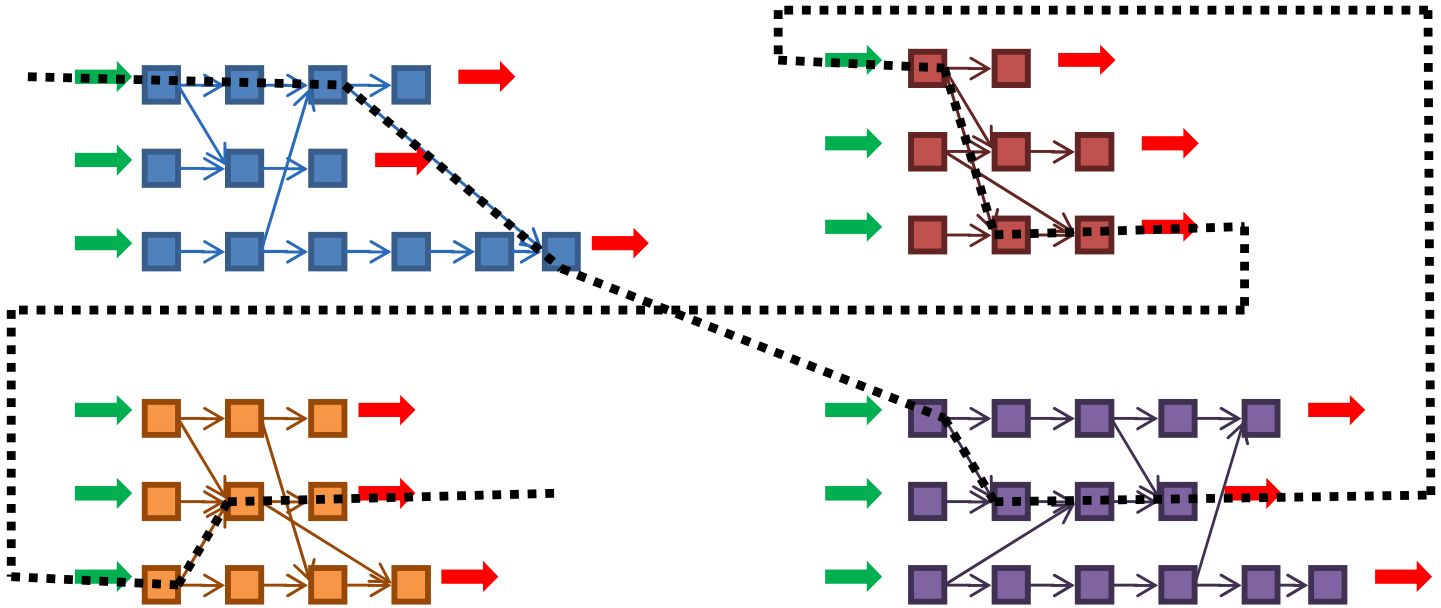


# 2.B. INVERSION

## G-Mem = oriented graph

- **Nodes** : acoustic/articulatory observations
- **Edges** : allowed articulatory transitions

A possible decoding path



- Decoding path based on the acoustic matching between the input signal and the acoustic component of each node
- Articulatory gesture derived from the articulatory component of each node part of the winning path

# 3.A. COMPARED APPROACHES

## Episodic memories

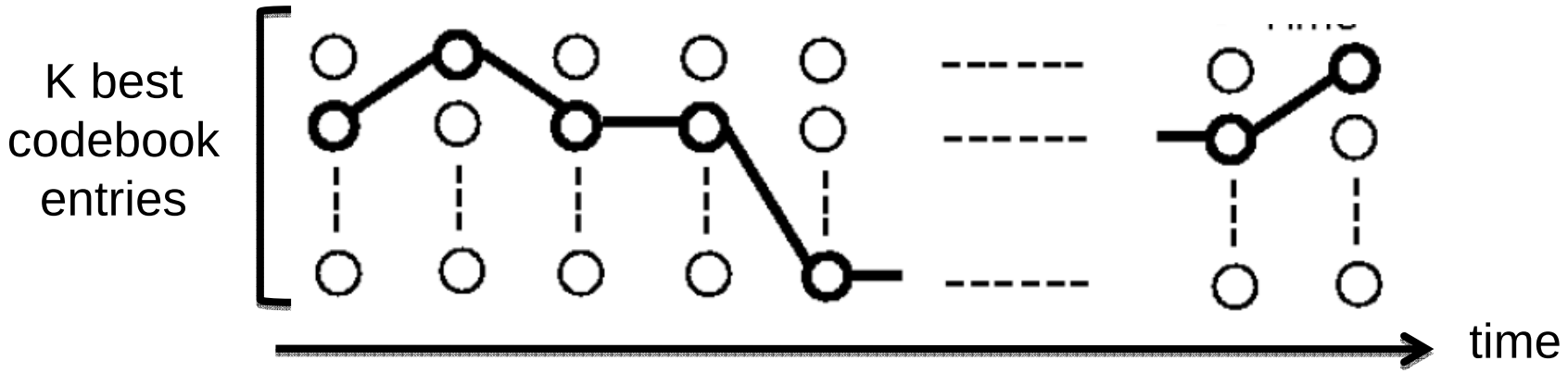
- **C-Mem** : concatenative episodic memory
- **G-Mem** : generative episodic memory

## Codebook [Suzuki et al. 98]

- For each acoustic test frame select the K best matching codebook entries
- Find the path which minimizes the weighted sum of the

**acoustic distances** and **articulatory constraints**

$$\min_{(c(t),\, x(t))} \sum_t \{ D_s(c_s(t),\ c(t)) + w \cdot D_x(x(t-1),\ x(t)) \}$$



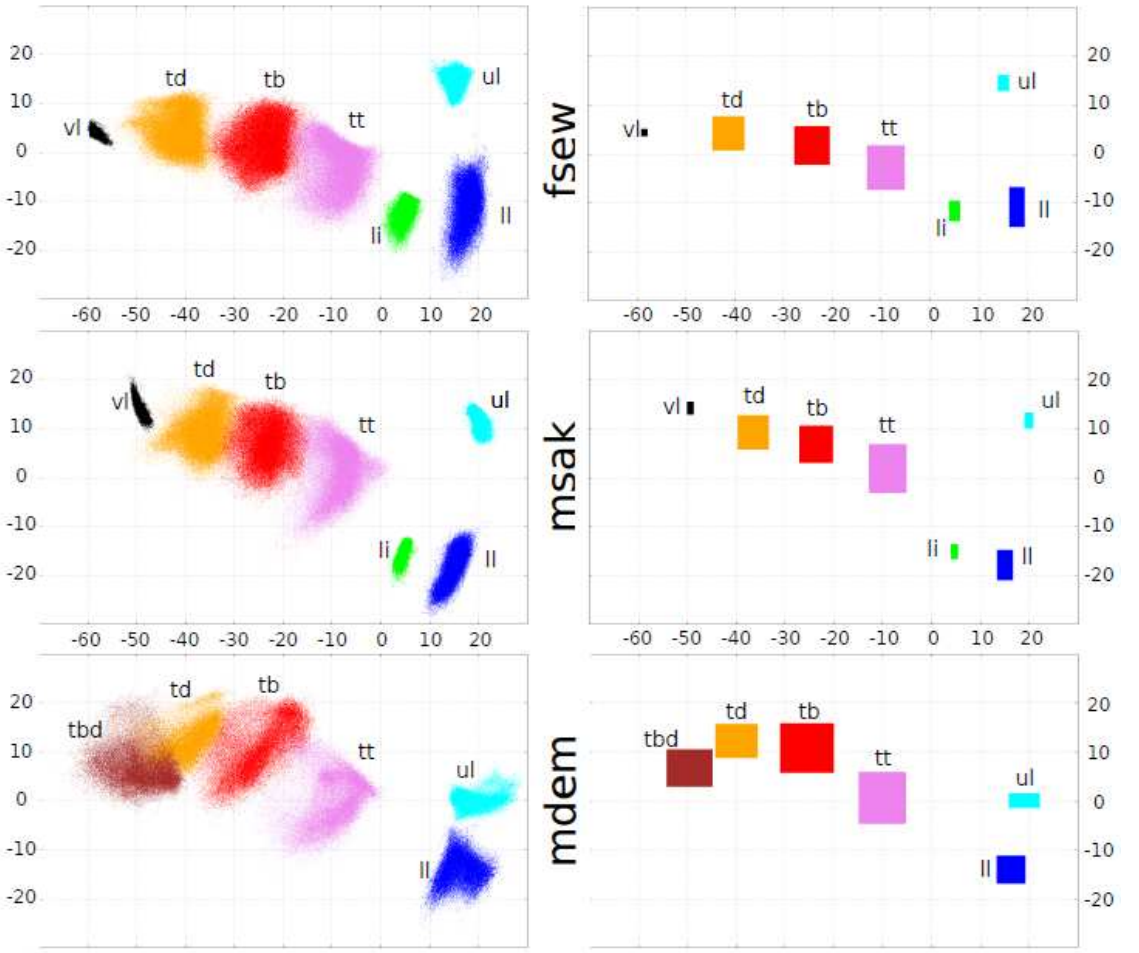
# 3.B. CORPORA

## MOCHA (fsew – msak)

- 2 speakers (female and male)
- Language: English
- 7 coils

## mdem

- 1 speaker (male)
- Language: French
- 6 coils




| Corpora | Sets | Durations | Sentences | Phones |
|---|---|---|---|---|
| *fsew* | train | 16 min 35 sec | 368 | 11179 |
| | dev | 1 min 57 sec | 46 | 1324 |
| | test | 2 min 5 sec | 46 | 1457 |
| *msak* | train | 13 min 59 sec | 368 | 11179 |
| | dev | 1 min 41 sec | 46 | 1324 |
| | test | 1 min 45 sec | 46 | 1457 |
| *mdem* | train | 8 min 24 sec | 319 | 6355 |
| | dev | 1 min 2 sec | 40 | 817 |
| | test | 1 min 3 sec | 40 | 814 |

# 3.C. EXPERIMENTAL SET-UP

## Data processing

- **Acoustic**
  - Removal of the silences at the sentence boundaries
  - Extraction of 12 static PLP coefficients for all 25ms speech frames shifted by 10ms
- **Articulatory**
  - Low pass filtering to remove aquisition noise (20 Hz)
  - High pass filtering of the mean articulatory positions per sentence to remove speaker adaptation effect

## Experiment design

- Consider inversion for each coil along the up/down and front/back directions as independent problems
- Creation of dedicated memory/codebook for each problems with optimized parameters
- Use of the Euclidean distance in both the acoustic and articulatory spaces

# 3.D. QUALITY MEASURES

## Root Mean Square Error

Quantifies the error between the reference and estimated articulatory trajectories

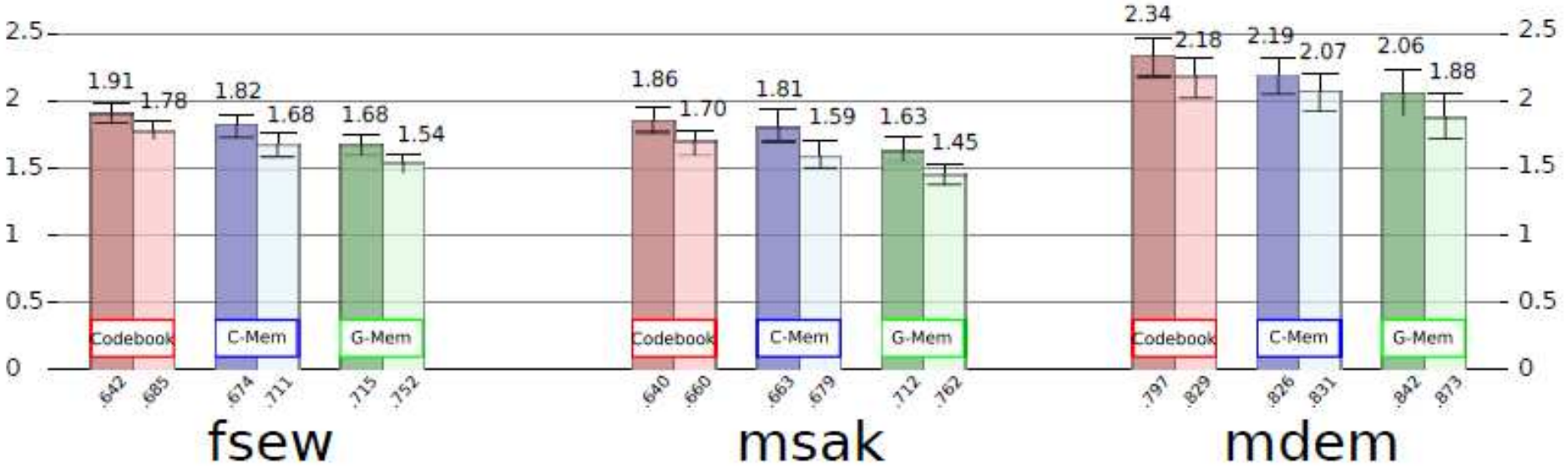
$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(f(x_i) - y_i)^2}$$

## Pearson's correlation

Quantifies the similarity and the synchrony between the reference and estimated articulatory trajectories

$$Cor = \frac{\sum_{i=1}^{N}(f(x_i) - \overline{f(x)}).(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(f(x_i) - \overline{f(x)})^2 . \sum_{i=1}^{N}(y_i - \overline{y})^2}}$$

# 3.E. RESULTS (1)




- Best results obtained with the G-Mem on all experiments w.r.t. the RMSE and the Pearson's correlation

- Probabilities of improvement:
- G-Mem → codebook: 100% [10;15]%
- G-Mem → C-Mem: 100 % [5;10]%

ICASSP 2011

# 3.E. RESULTS (2)

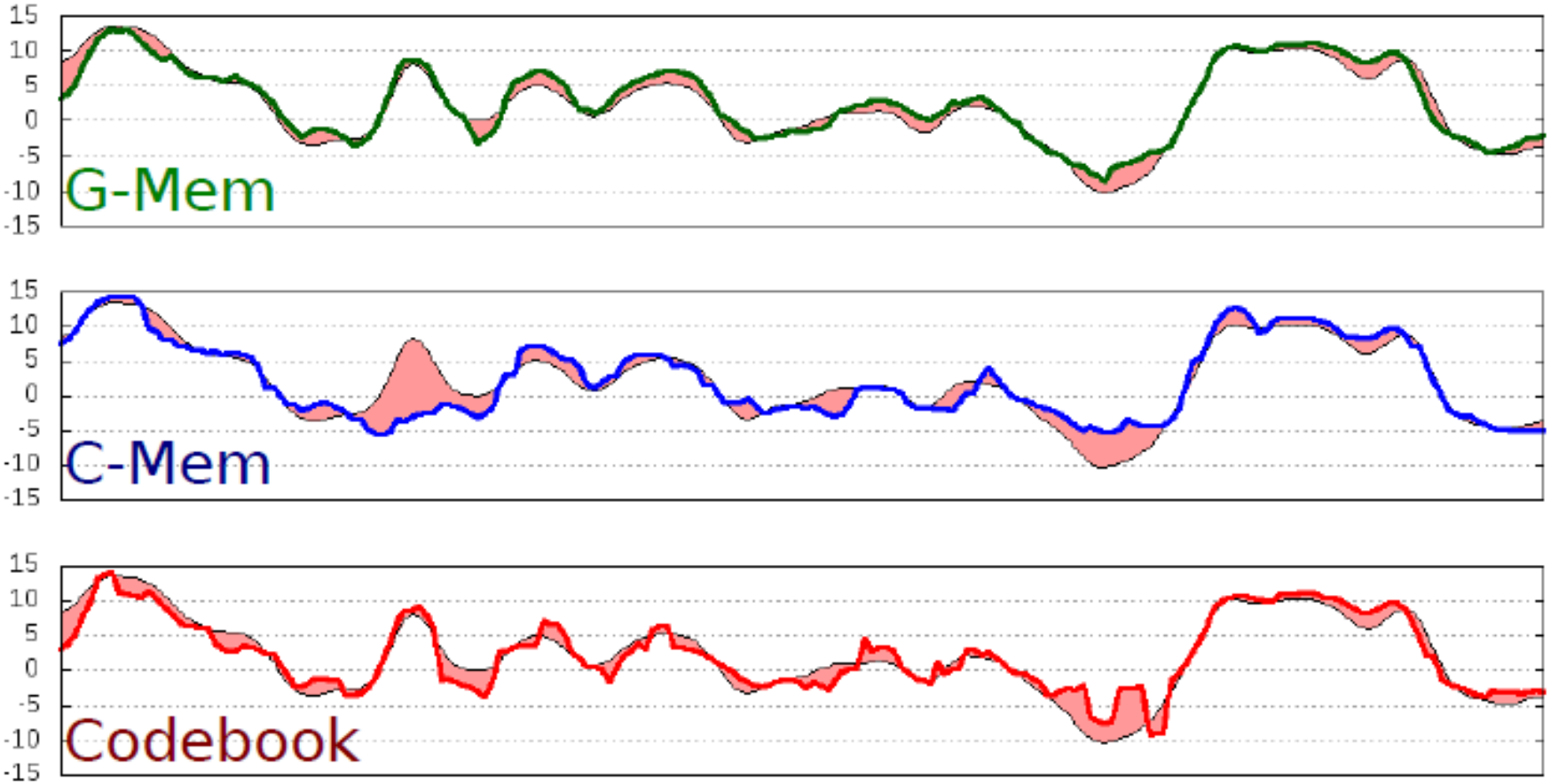
## A visual assessment

- French sentence

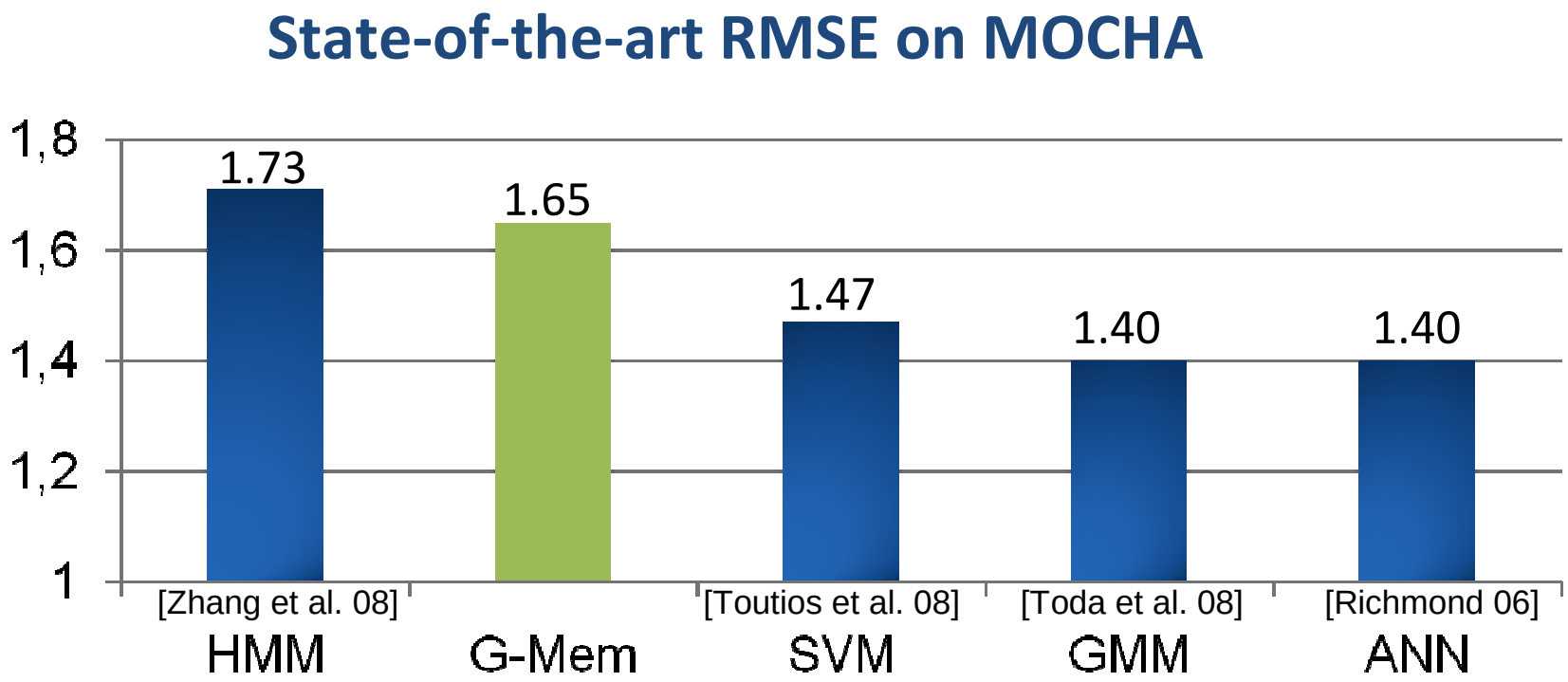
« Juste quelques extrémités de branches gelées »

(Just few frozen extremities of tree branches)

- Tongue tip up/down movement



# 3.E. RESULTS (3)




- Reported RMSE comprise between 1.4 and 1.73 mm
- But articulatory data acquisition error is about 0.4 mm

# 4. CONCLUSION AND FUTURE WORK

## We have proposed a generative episodic memory

- The model **does not require any assumption about the mapping function** but rather relies on synchronized articulatory/acoustic observations
- The **articulatory dynamics is structurally embeded within the memory** and helps to solve the non-uniqueness of the solution
- The proposed **memory is able to produce unseen articulatory gestures** and thus can generalize over the data

## Future work

- Use a more **robust acoustic distance** (e.g. kernel based distances)
- Take into account **correlations between the articulators**
- Use **articulatory gesture based segmentations** instead of phonetic segmentations
- A **local linear regression** could provide further improvements

# THANKS FOR YOUR ATTENTION