

# **A model of optimal speech production planning integrating dynamical constraints to achieve appropriate articulatory timing**

**Ralf Winkler<sup>1,2,3</sup>, Liang Ma<sup>1,4</sup> & Pascal Perrier<sup>1</sup>**

<sup>1</sup>DPC/GIPSA-lab, Grenoble, France

<sup>2</sup>ZAS/Phonetik, Berlin, Germany

<sup>3</sup>Technische Universität, Berlin, Germany

<sup>4</sup>Zhejiang University, China

# Outline

- Introduction
- The model: GEPPETO
  - Control model
  - Physiological tongue model and acoustics
- Planning (Internal models)
  - Motor commands -> acoustics
  - Motor commands -> force
  - Sequence planning (Optimization)
- Experiment with improved planning
- Summary & Conclusions

# Introduction

- production of speech gestures could be based on an optimal planning in the central nervous system
- planning would use internal representations of the speech production apparatus (Guenther et al. (1998), Bailly, (1997)) to determine the motor command patterns that:
  - allow for reaching speech goals
  - with minimum of effort

# Introduction

- GEPPETO (the speech production model presented in this talk) has been designed within this general theoretical framework
- In the GEPPETO model, speech planning already incorporates linguistic constraints
- It will be shown how dynamical (global force) constraints can be taken into account during speech planning
- Hypotheses:
  - For slow speaking rates or low accuracy requirements, a low level of force can be used
  - fast speaking rates or great accuracy requires a strong level of force

# The speech production model: GEPPETO<sup>1</sup>

- Speech goals
  - Target regions in the acoustic space (inspired from Keating, 1988, see also Guenther, Hampson & Johnson, 1998)
- Vocal tract: tongue model & acoustics
  - 2D biomechanical tongue model (Payan & Perrier, 1997; Perrier, Payan, Zandipour & Perkell, 2003)
  - Harmonic model of the vocal tract (Badin & Fant, 1984)
- Model of motor control
  - $\lambda$  model (Feldman, 1966, 1986)
- The context: planning and sequence generation
  - Static/dynamic internal model(s)
  - Optimization procedure
  - Movement 's execution

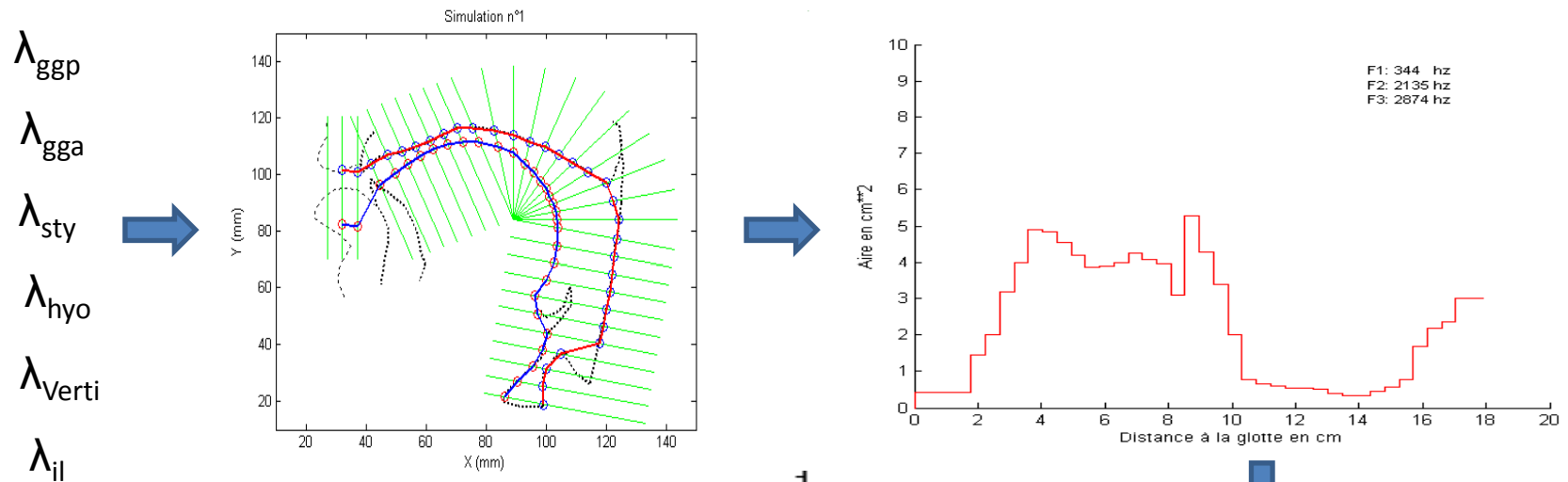
<sup>1</sup>holds for "GEstures shaped by the Physics and by a PErceptually oriented Targets Optimization

# Biomechanical tongue model (2D)

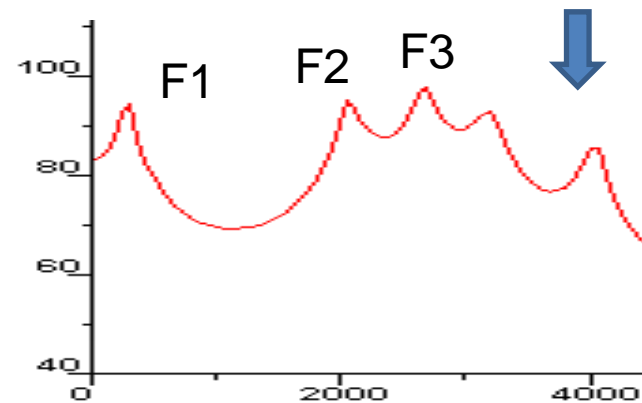
- Driven by constant rate of shift between motor targets
- It is a simplified model
- But it generates realistic articulatory and acoustic patterns:
  - Realistic velocity profiles (Payan & Perrier, 1997)
  - Main degrees of freedom of tongue shaping in the mid-sagittal plane (front raising and back raising) (Perrier et al., 2000)
  - Articulatory loop patterns in V/k/V sequences (Perrier et al., 2003)
- Some limitations:
  - Tongue tip raising is difficult → alveolar consonants non realistic
  - Obviously no shaping in the coronal plane

# Tongue model (overview)

- Sagittal view → Area Function (Perrier et al., 1991)



- Area Function → Formants



# Sequence planning

- Sequence planning (in our context) means finding an **optimal** path in motor space that ensures target reaching in a specified period of time!
- (more technical): (nonlinear) optimization of the  $\lambda$ -distance between motor commands constrained by
  - not leaving the perceptual target regions (static forward internal model)
  - not leaving a specified force range (dynamic forward internal model)

# Sequence planning: internal models and optimization

- Data to learn the two forward internal models
  - Uniform sampling of the motor control space
  - 6 muscles (GGP, GGA, STY, HYO, IL, VERT)
  - 8800 simulations (+ ca. 5000 simulations afterwards)
  - RBF network; learning based on 8293 simulations (50% train, 50% test), 400 radial basis functions
- Optimization; accomplished by sequential quadratic programming (SQP) for constrained function minimization

# Static forward (internal) model

motor control space

$\lambda_{\text{ggp}}$

$\lambda_{\text{gga}}$

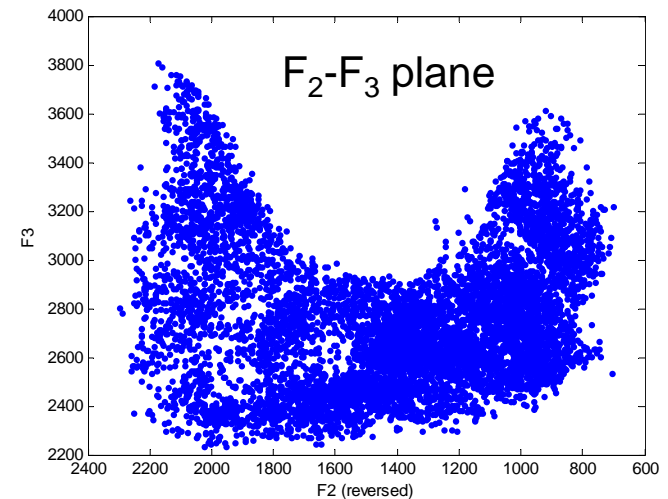
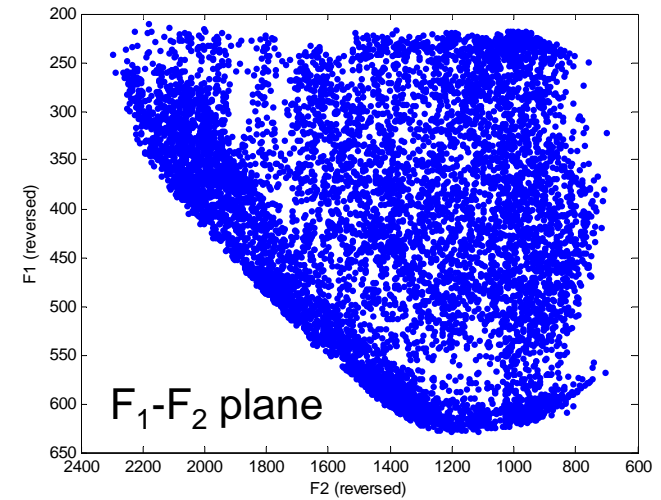
$\lambda_{\text{sty}}$

$\lambda_{\text{hyo}}$

$\lambda_{\text{verti}}$

$\lambda_{\text{il}}$

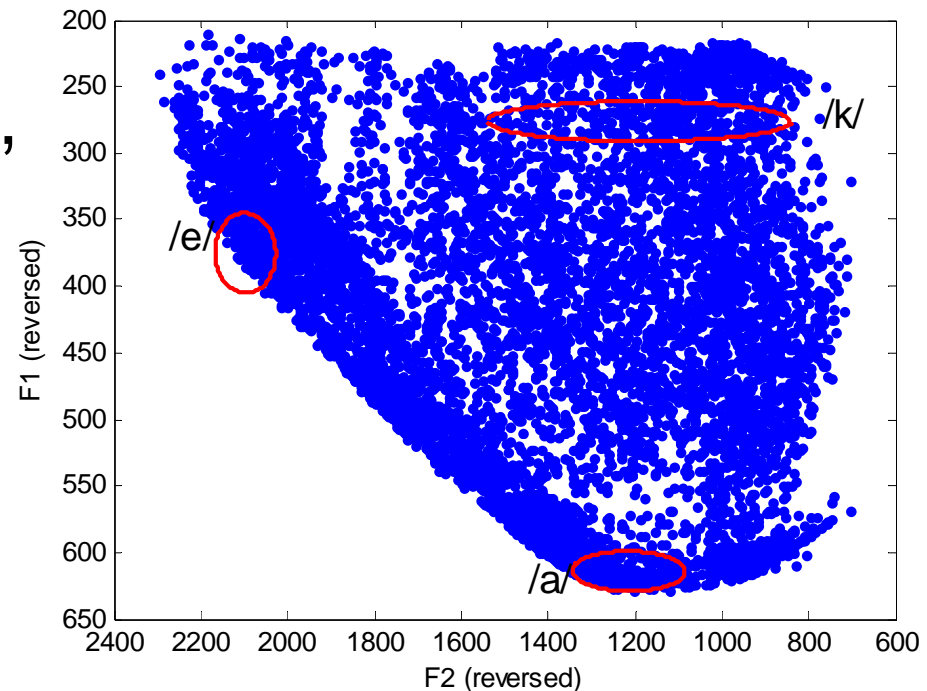
via RBF net



# Static forward (internal) model: speech goals

- Speech goals are target regions
- Ellipsoids in the ( $F_1$ ,  $F_2$ ,  $F_3$ ) space
- defined by:
  - average  $F_1$ ,  $F_2$ ,  $F_3$ ,
  - $\sigma F_1$ ,  $\sigma F_2$ ,  $\sigma F_3$

(for French,  
Calliope, 1989)



# Dynamical forward (internal) model

motor control space

$\lambda_{ggp}$

$\lambda_{gga}$

$\lambda_{sty}$

$\lambda_{hyo}$

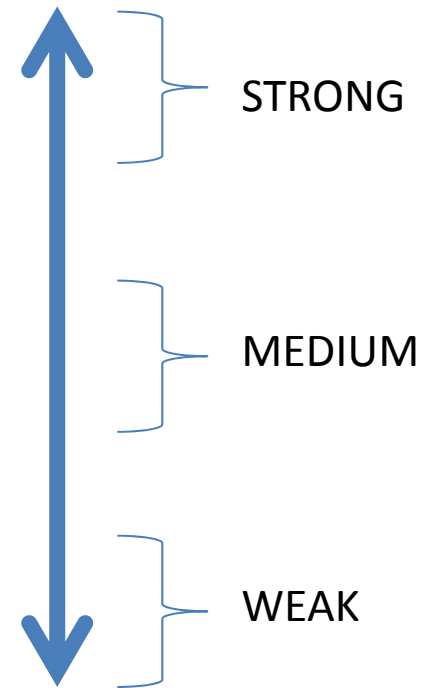
$\lambda_{verti}$

$\lambda_{il}$



Global force level

(scalar)  
global  
force  
value



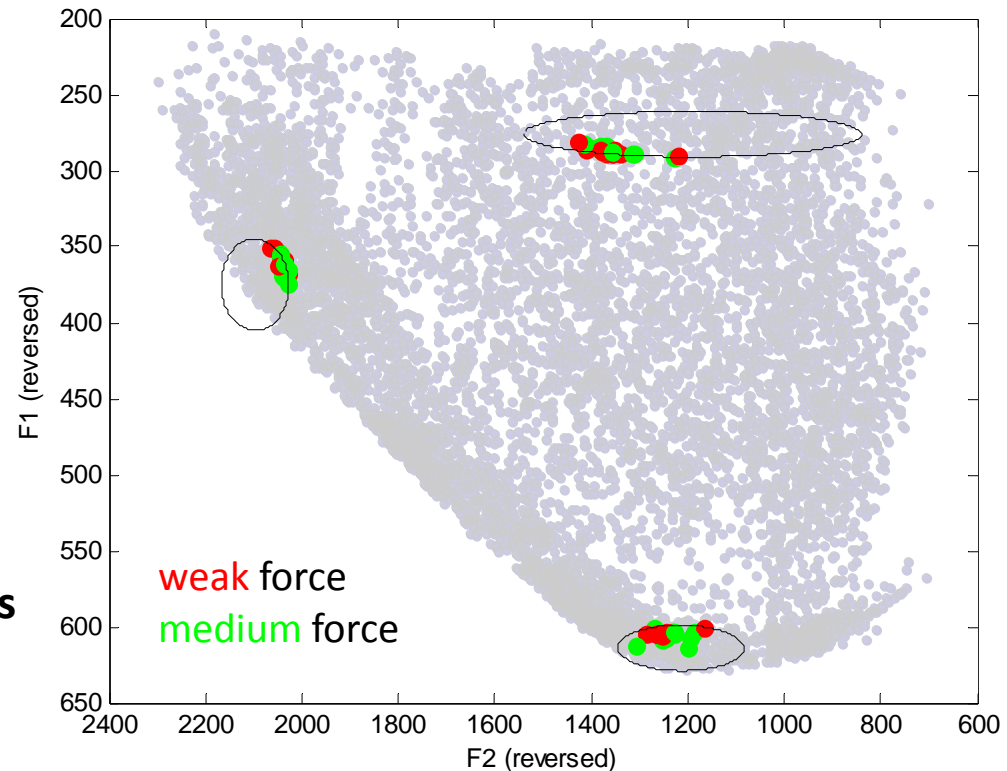
# Sequence generation

- Sequences were generated with optimal motor commands for given speech segments (resulting from optimization) subjected to the constraint:
  - weak (global) force
  - medium (global) force
- /ake/ -sequences were generated differing in the speech tempo:
  - Slow speech rate (transition: 50 ms, hold: 100 ms)
  - Normal speech rate (transition: 24 ms, hold: 60 ms)
  - High speech rate (transition: 20 ms, hold: 50 ms)

# Results: optimal $\lambda$ -commands

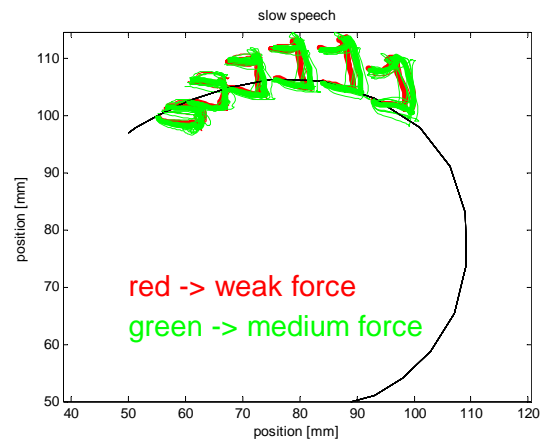
- Optimal solutions (in planning) differ depending from the start point of optimization

Results of several single optimizations show that the spectral patterns associated with the optimal  $\lambda$ -commands does not depend on the force actively generated in the tongue!

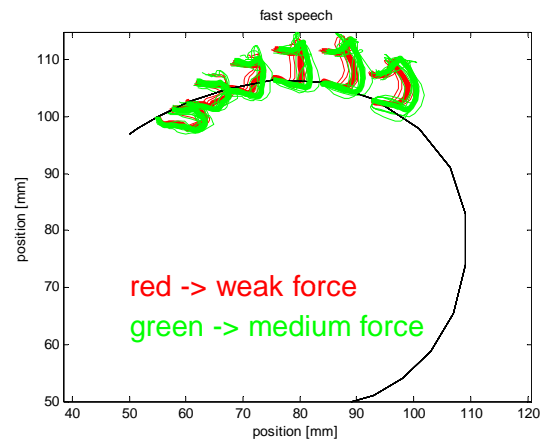


# Results: /ake/

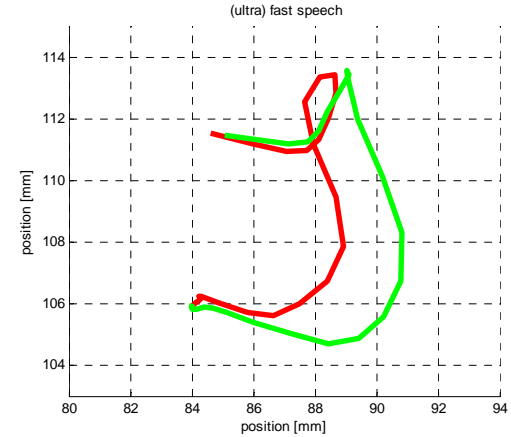
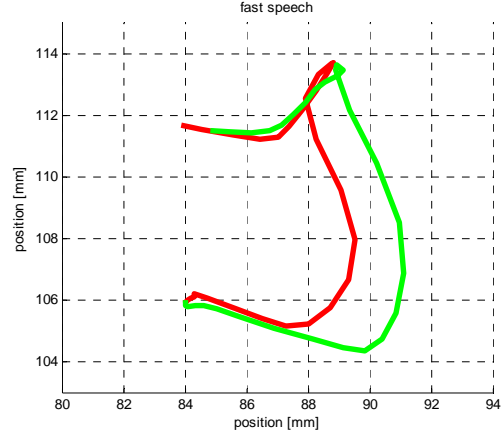
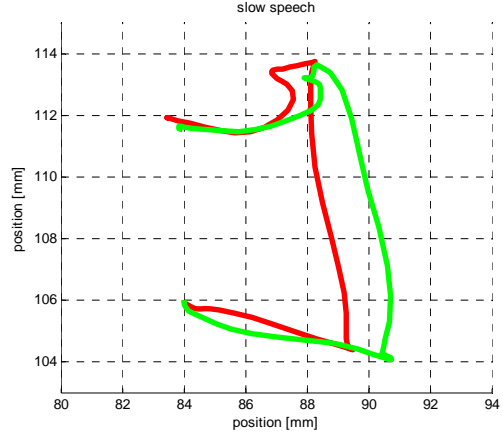
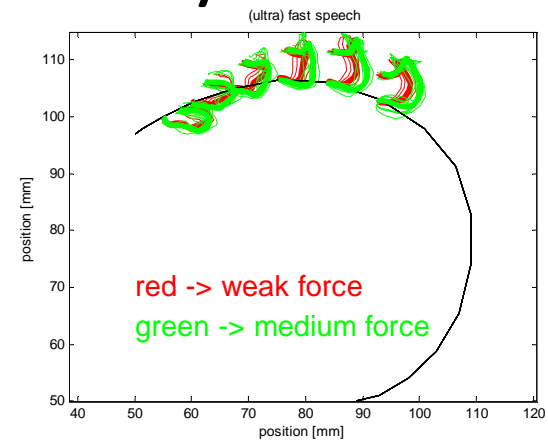
## Slow



## fast

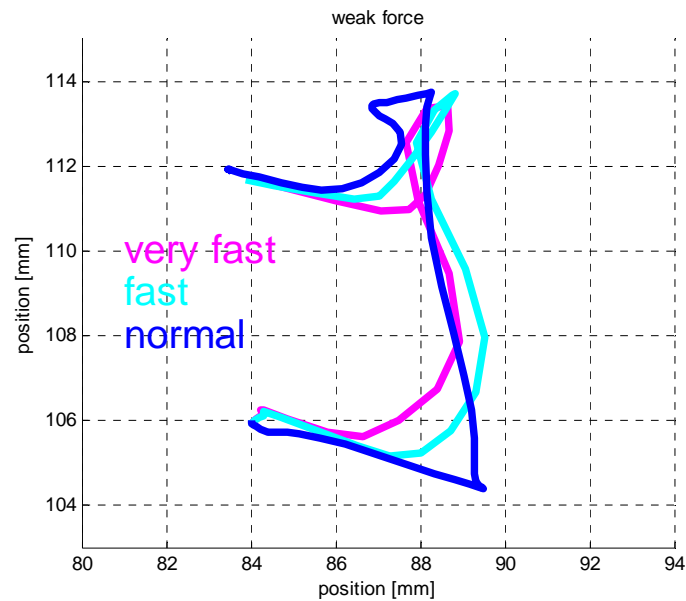


## very fast

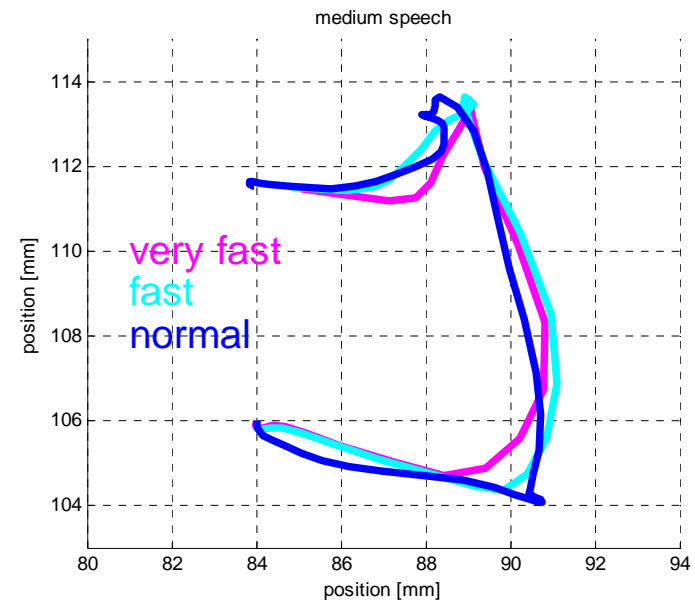


# Results: /ake/

## weak force



## medium force



Target undershoot (left graph) results from fast speech rate produced with “insufficient” force

# Summary & Conclusions

- The sequence planning in our framework led to:
  - a remarkable influence of speech rate on tongue trajectories if force is small
  - a only small effect of speech rate on target position if force is sufficient
- Incorporating dynamical constraints into sequence planning (additionally to perceptive constraints) resulted in tongue trajectories coherent with our hypotheses:
  - For slow speaking rates or low accuracy requirements, a low level of force can be used
  - fast speaking rates or great accuracy requires a sufficient level of force

thanks for listening!