Deducing language from non-language

Björn Lindblom

Workshop on *Phonological systems and complex adaptive systems*, LYON, France, 4-6 July, 2005.

Where do phonetic units come from?
(in development and evolution)

- Carré 2004
- Clements 2003
- Flemming 2005
- Goldstein 2003
- Lacerda 2003
- Nowak & Krakauer 1999
- Oudeyer 2003
- Stevens 2003
- Studdert-Kennedy 2002
- Zuidema & de Boer 2005

A selective walk through the history of phonetics

From ’targets’ to ’gestures’

- Notion of ‘target’
- Speech perception & sensory systems
- ‘Dynamic specification’
- ‘Phonetic gestures’
- The ‘particulate principle’

The notion of ’target’

Phonetic units specified as *static* articulations

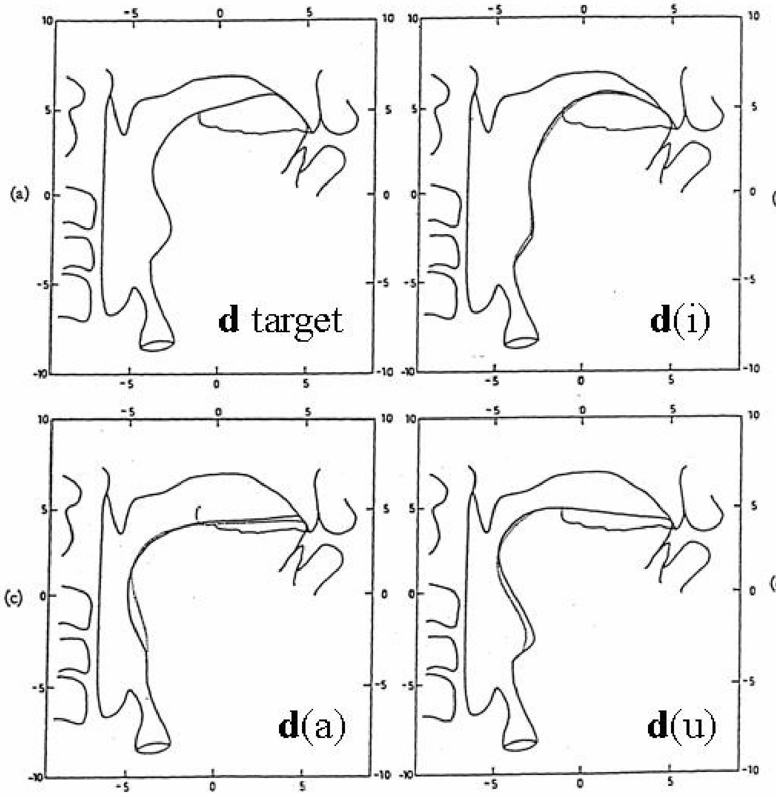
’*Stellungslaute*’

Three early studies

Stevens & House 1963
Lindblom 1963
Öhman 1967

Target-based account of VCV coarticulation
(Öhman 1967)

single target underlies the variants



Main observations

Strong isomorphism

phonetic categories

&

articulatory processes.

Units

as timeless **static**

articulatory ‘targets’.

Implications

If the talker controls what the listener wants, target theories of speech production would seem to imply:

Implications

If the talker controls what the listener wants, target theories of speech production would seem to imply:

- Speech perception is basically a matter of recovering **static targets**.

Implications

If the talker controls what the listener wants, target theories of speech production would seem to imply:

- Speech perception is basically a matter of recovering **static targets**.
- This implication seems to be at odds with what we know from **sensory physiology**.

Implications

If the talker controls what the listener wants, target theories of speech production would seem to imply:

- Speech perception is basically a matter of recovering **static targets**.
- This implication seems to be at odds with what we know from **sensory physiology**.
- Visual and auditory systems are more sensitive to changing stimulus arrays than to purely static ones.

Implications

If the talker controls what the listener wants, target theories of speech production would seem to imply:

- Speech perception is basically a matter of recovering **static targets**.
- This implication seems to be at odds with what we know from **sensory physiology**.
- Visual and auditory systems are more sensitive to changing stimulus arrays than to purely static ones.
- If **perception likes change**, why assume production control in terms of **static targets**?

Moving-edge detectors

(Lettvin et al 1959)

”The frog .... is not concerned with the detail of stationary parts of the environment of the world around him.

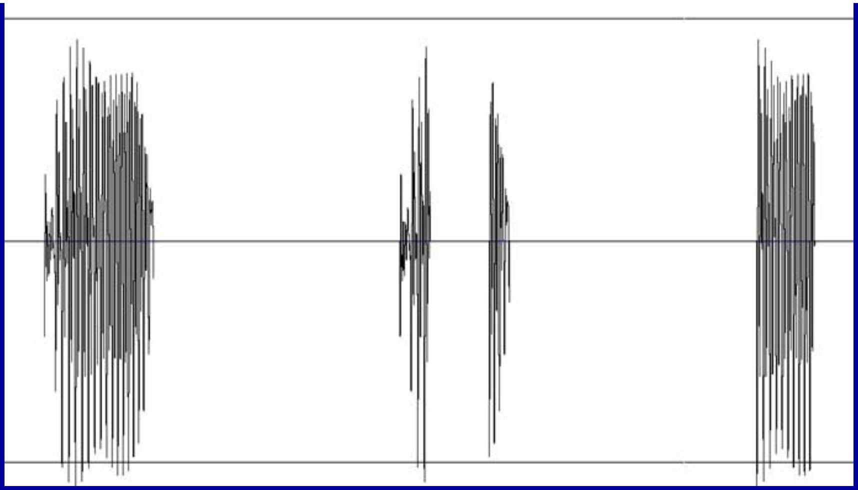
He will starve to death surrounded by food if it is not moving.

Lettvin J T et al (1959): “What the frog's eyes tells the frog's brain”

The ’silent center’ paradigm
Winifred Strange

Syllables presented to listeners in 3 different ways:

Full syllable Silent nucleus Center only.



Results

- Listeners were able to identify vowels with high accuracy although the center portions of CVC stimuli had been removed.

- Vowel perception is possible also in ‘silent-center’ syllables that lack information on the alleged ‘target’ but include an initial stop plosion and surrounding formant transitions.

Vowels as gestures

(Strange 1989)

"… vowels are conceived of as characteristic gestures having intrinsic timing parameters (Fowler, 1980). These dynamic articulatory events give rise to an acoustic pattern in which the changing spectrotemporal configuration provides sufficient information for the unambiguous identification of the intended vowels."

Gesturalist frameworks

- Motor Theory (Liberman&Mattingly)

- Direct realism (Fowler)
- Articulatory phonology (Browman&Goldstein)

Some gesturalist assumptions

- What we perceive is the articulatory activity of the VT.
- The building blocks of that process are the phonetic gestures.
- Phonetic gestures are dynamically specified.
- They are implemented as coordinative structures with intrinsic timing properties.
- They are the basic units of speech, the primitives of phonetic theory

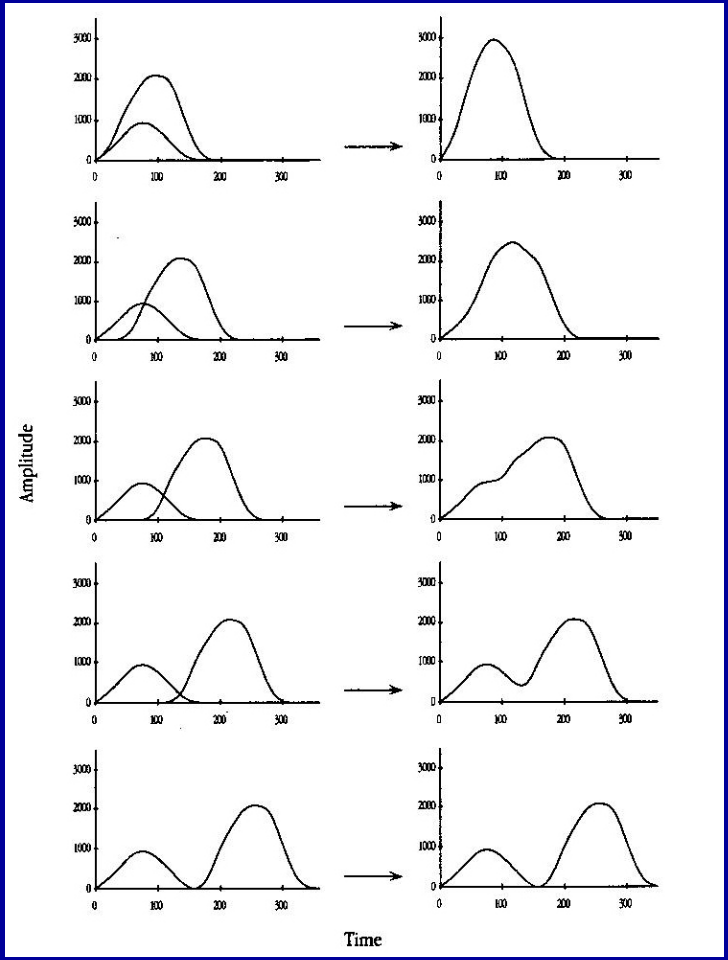
Gestural ’blending’

(Munhall & Löfqvist 1992)

Fast

Input gestures

Slow



Output movements

Some doubts & objections

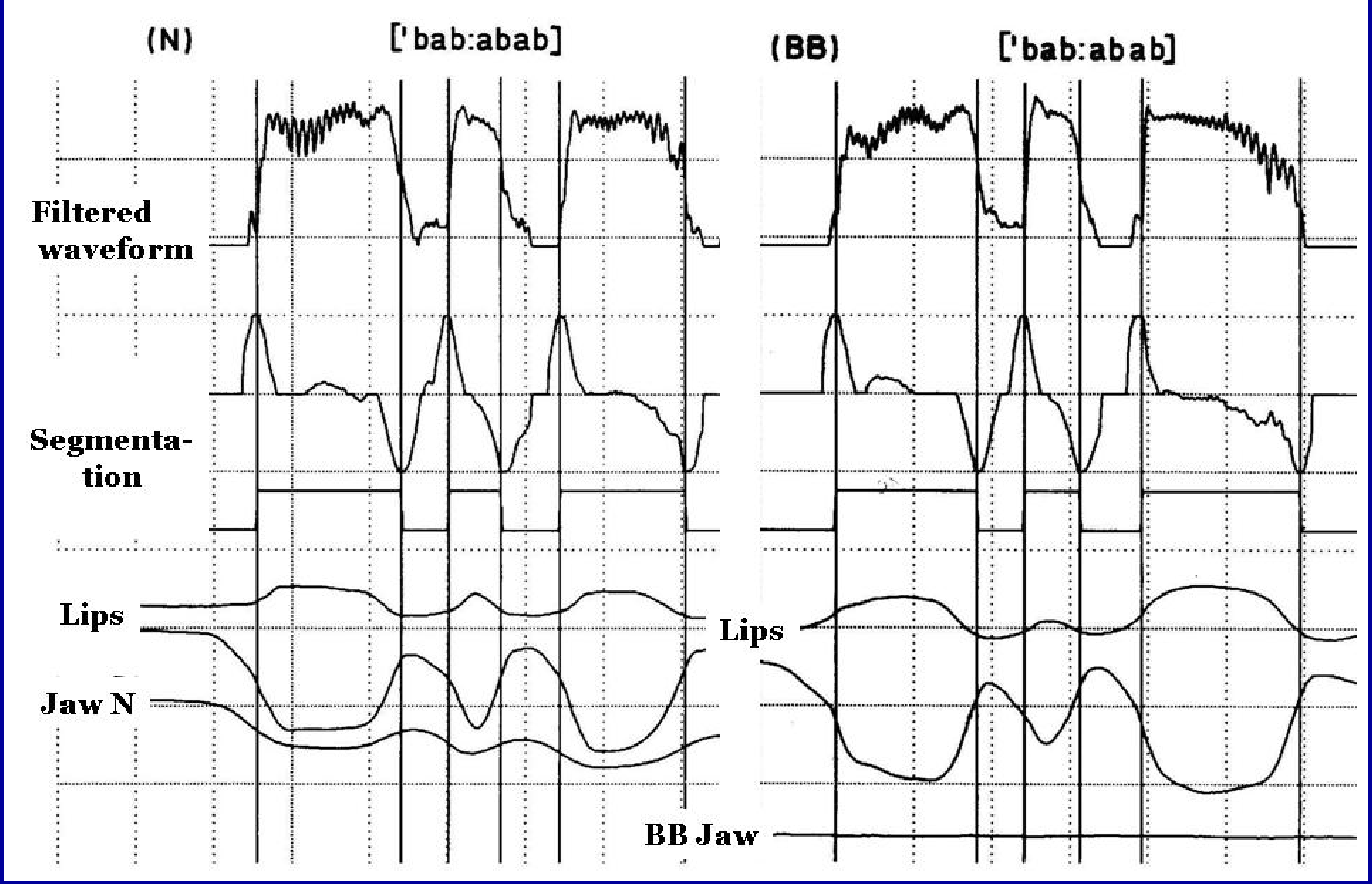
- Sign & speech eminently “gestural”;
- The notion of ‘gesture’ acknowledges the crucial role played by ‘signal dynamics’ (spectro-temporal variations) for speech perception;
- But the gestural account misrepresents the nature of the underlying control commands;
- It does so by failing to consider the inevitable contribution to ‘dynamics’ of system response characteristics (physiological & mechanical).
- In short, it lacks parsimony using gesture (read: movement) to explain movement.

A crucial choice

- It turns out that the choice between targets and gestures has serious consequences in the pursuit of answers to “Where does phonetic structure come from?”

Compensatory articulation:

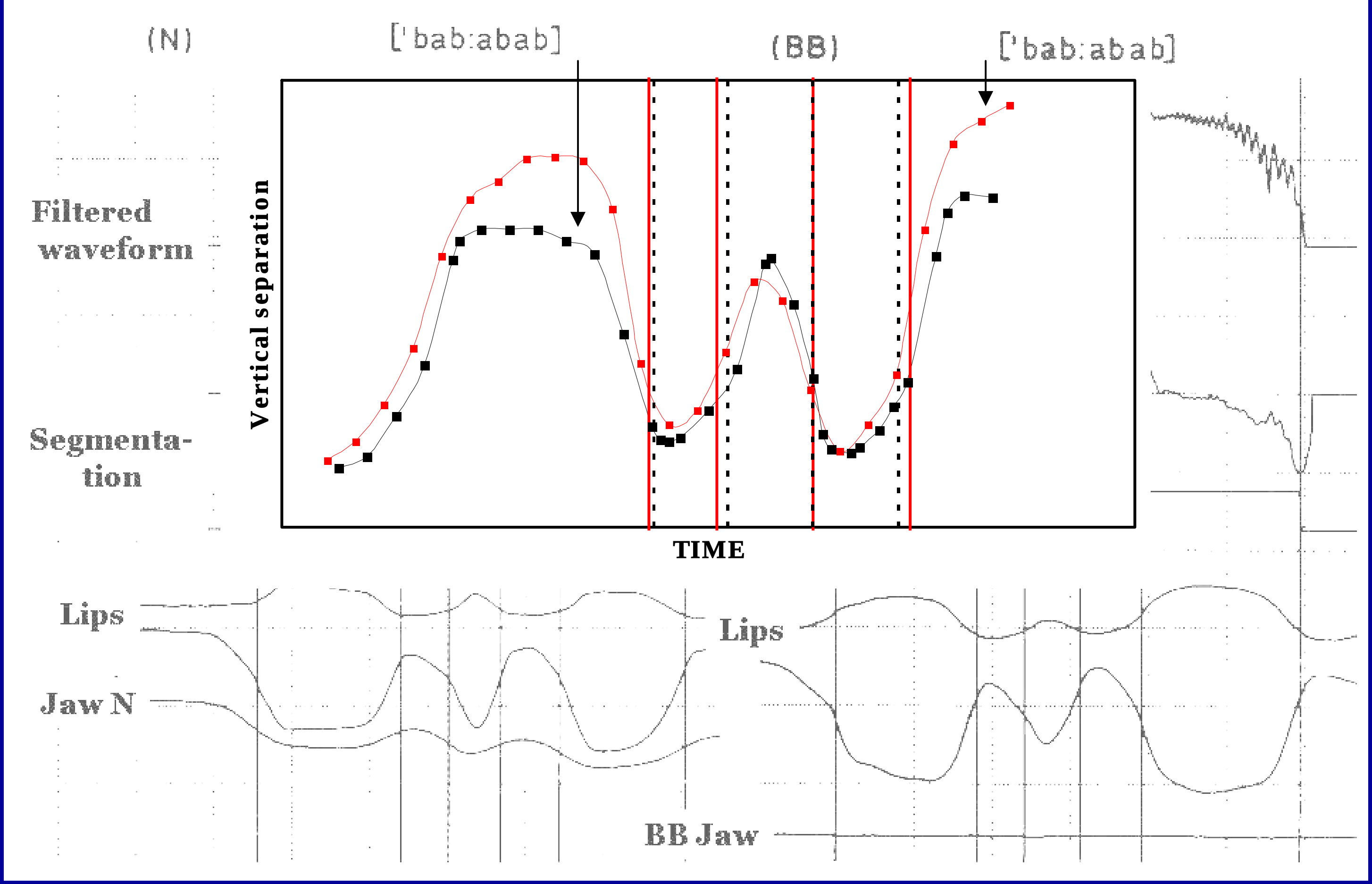
Labial closure



Lindblom et al (1987): ”The concept of target and speech timing”, in Channon R & Shockey L (eds): *In honor of Ilse Lehiste*, Foris publications.

Compensatory articulation:

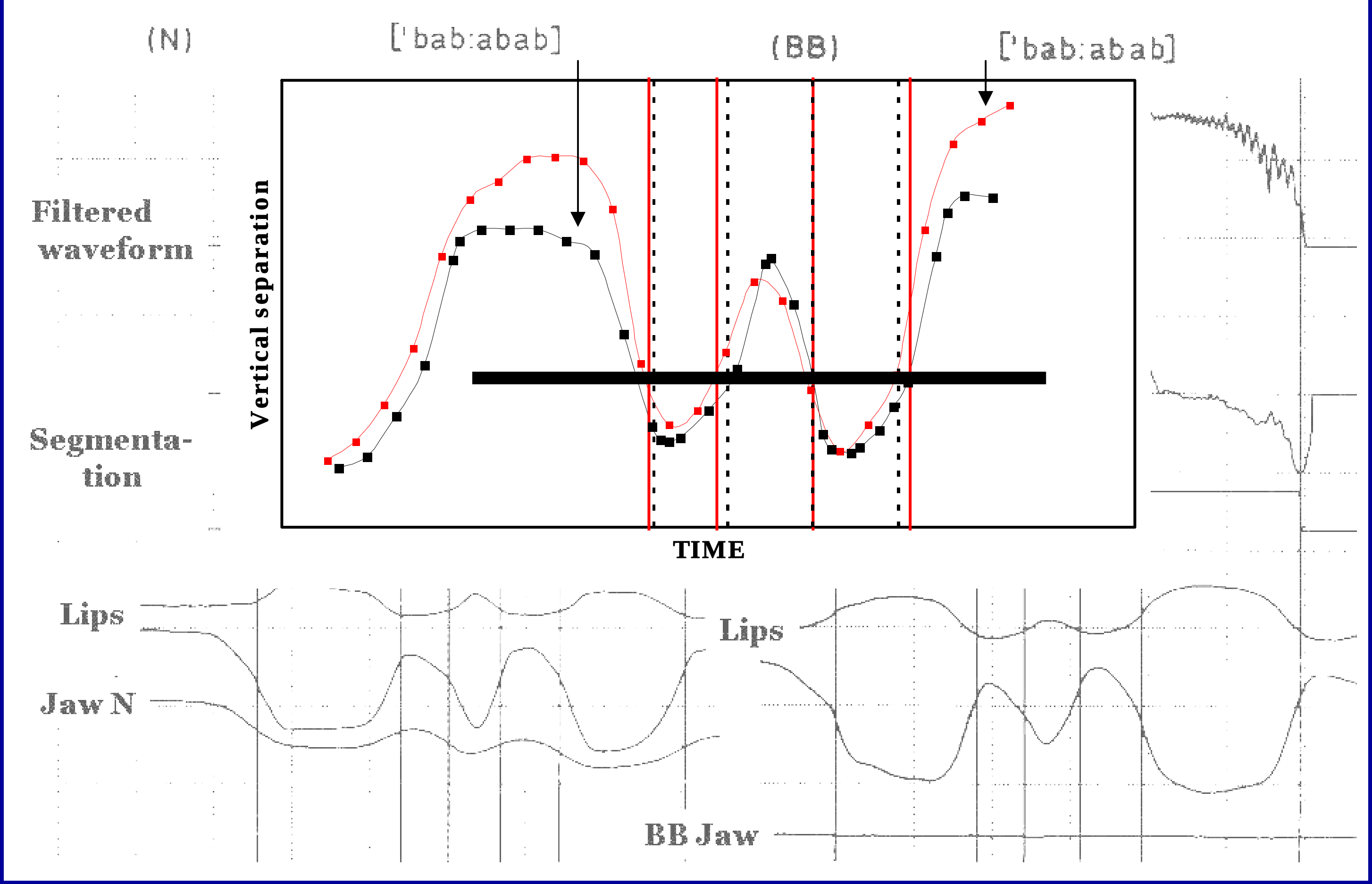
Labial closure



Lindblom et al (1987): ”The concept of target and speech timing”, in Channon R & Shockey L (eds): *In honor of Ilse Lehiste*, Foris publications.

Compensatory articulation:

Labial closure



Lindblom et al (1987): ”The concept of target and speech timing”, in Channon R & Shockey L (eds): *In honor of Ilse Lehiste*, Foris publications.

Interpretation

Labial closure has two meanings:
1. Closing of the lips; 2. State of closed lips.

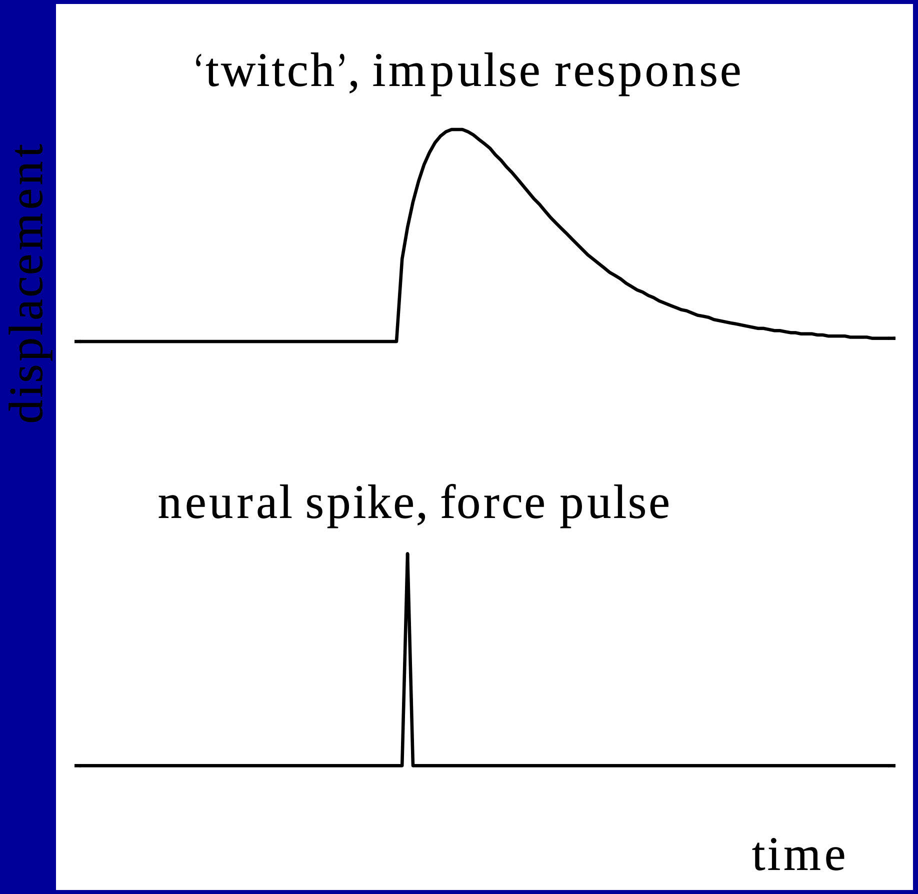
Q: What remains invariant across conditions?

A: Reaching the state of closed lips (i.e., not the 'gesture' but the static target with its dynamic (aerodynamic & acoustic consequences)

Primary: The goal (minimally the static 'spatial target');

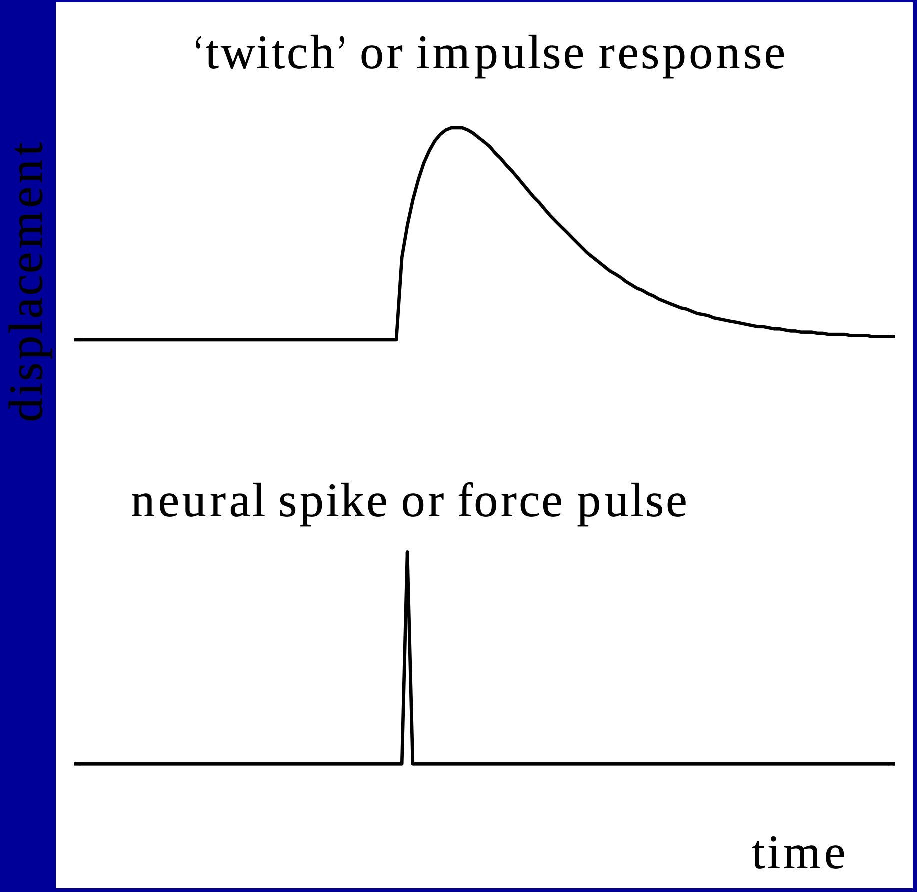
Secondary: How to get there (the 'gesture').

Dynamic response characteristics

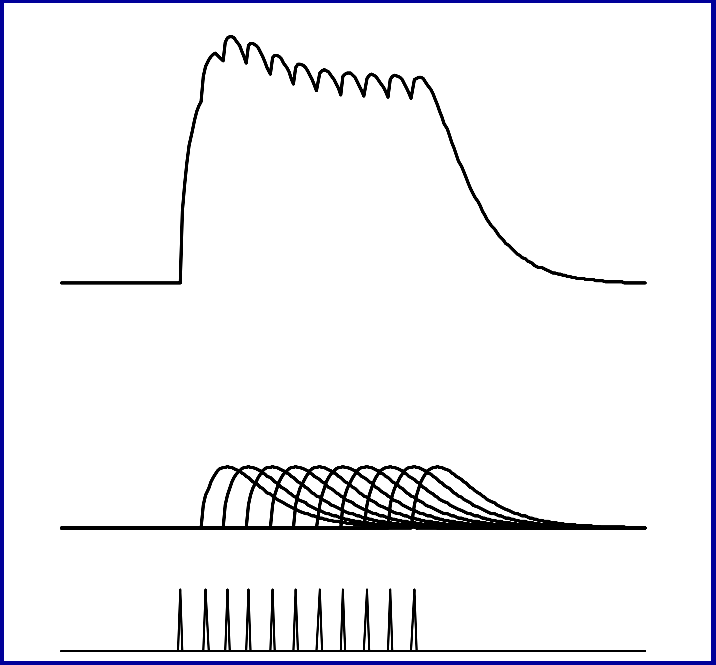


Dynamic response characteristics

physiological and mechanical



summed response



spike train, sampled force

Arm movements

- Numerous reports indicate that normal subjects typically perform a movement from A to B (say a reaching task) with great consistency.
- The trajectory is smooth and comes close to a straight path.
- There is an infinite number of paths that such a movement could follow in principle.
- Q: How does it do that?
- A: Movement is conceptualized in terms of

  spatial targeting &

  general motor mechanisms for trajectory formation.

Clues from work on non-speech movement

- Equilibrium Point Hypothesis

  Target information is coded in terms of **muscle lengths**

  Trajectory shape in terms of **stiffness**

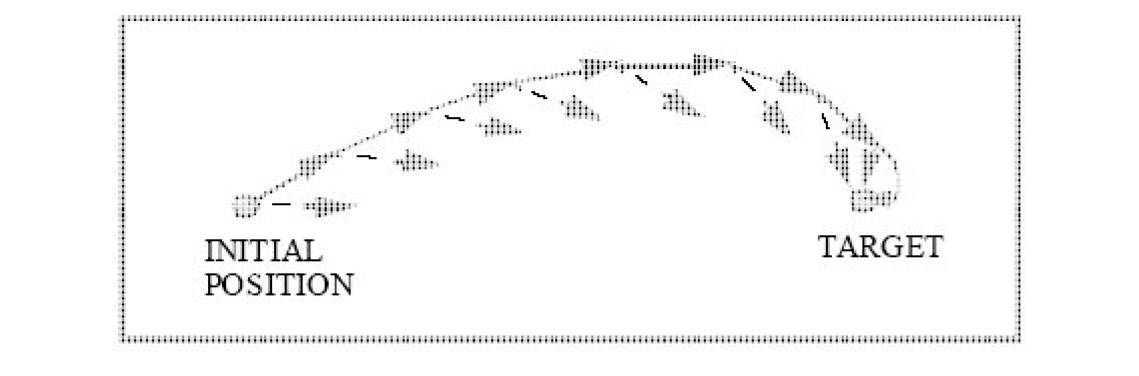
- Optimization models

  Smooth trajectories from optimization criteria

  energetics (minimum jerk, minimum work, ...)

  precision (minimum variance)

- DIVA model



Present claim

The key to answering “Where does phonetic structure come from?” lies in

- According a significant role to the general motor mechanism handling movement paths in non-speech and speech in driving phonetic recombination;
- Viewing the control signals (targets) as the elements to be recombined.

A toy model of phonetic learning

Search space & ambient input

Definition of ‘phonetic pattern’

A movement made within a fixed time frame

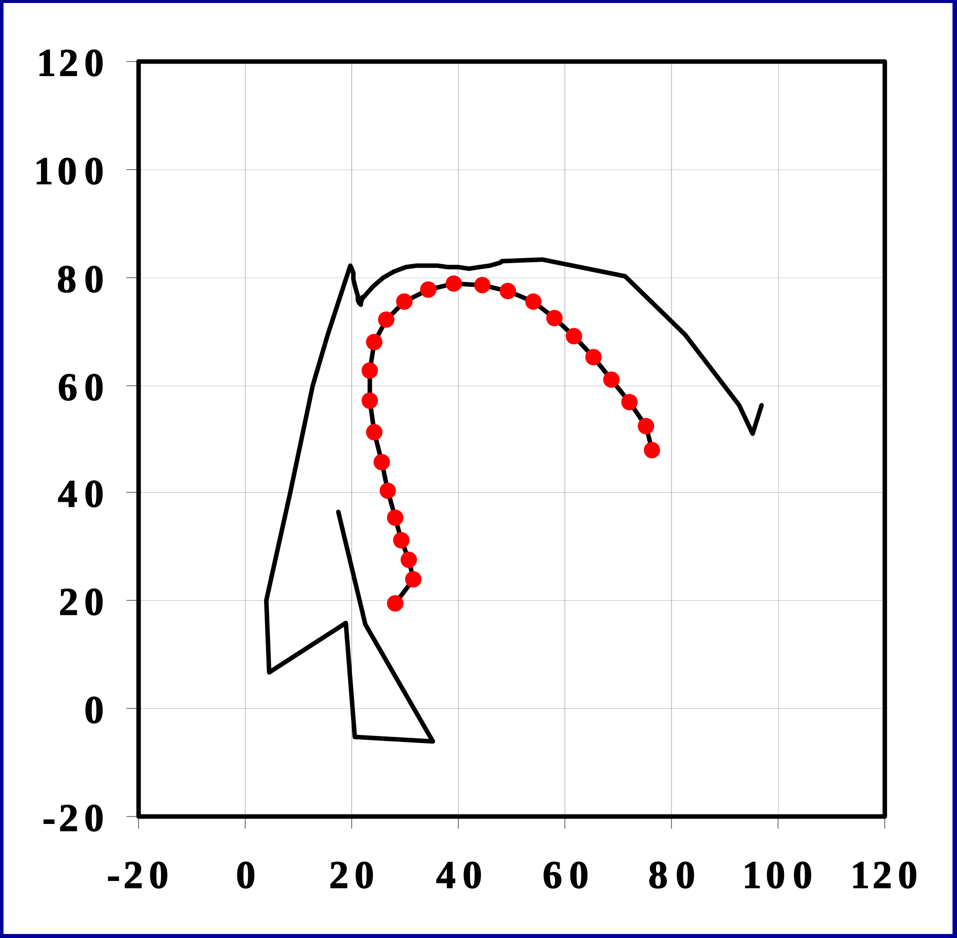
&

from a constricted to a more open articulation

*10 constricted * 21 open configurations = 210 patterns*

Source of data

- X-ray data (SU data base, single subject)
- Speech samples (Swedish)
- Tracings (ca 500 images)
- PCA analysis (numerical specification of articulatory profiles)



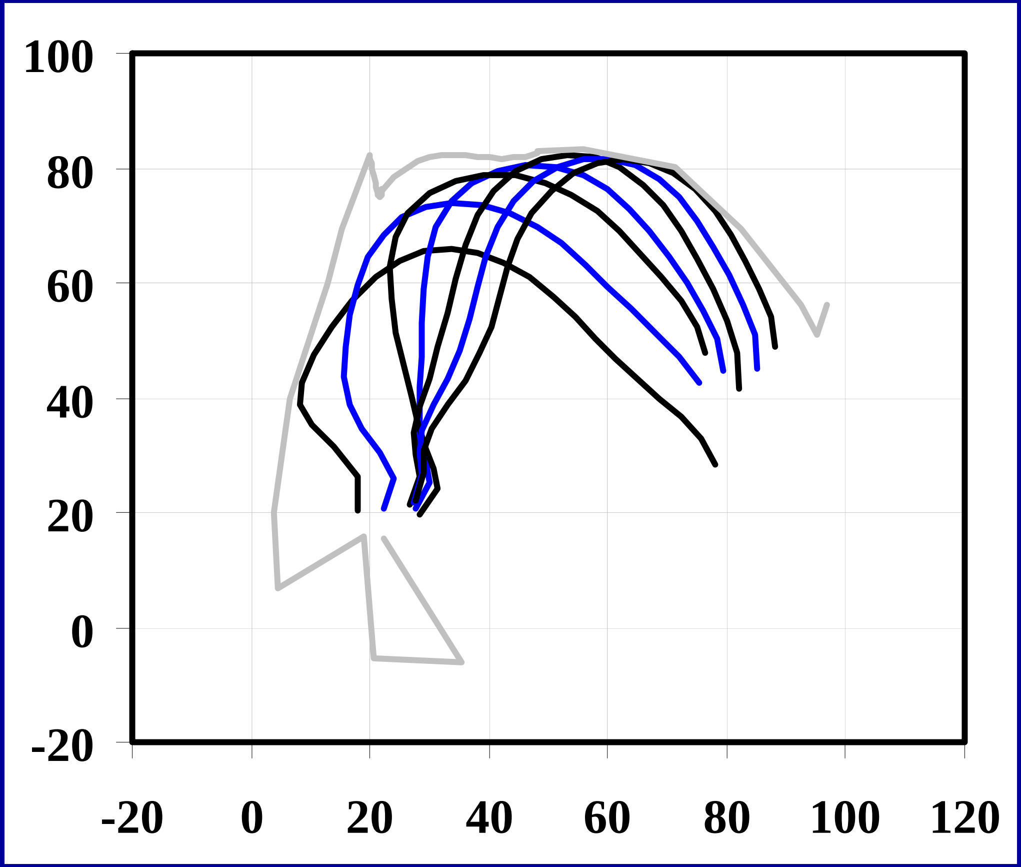
Arbitrary tongue shapes

linear combinations of a small number of Principal Components

$$= \left[\sum_{i=1}^{4} w_i PC_i(x)\right] + c(x)$$

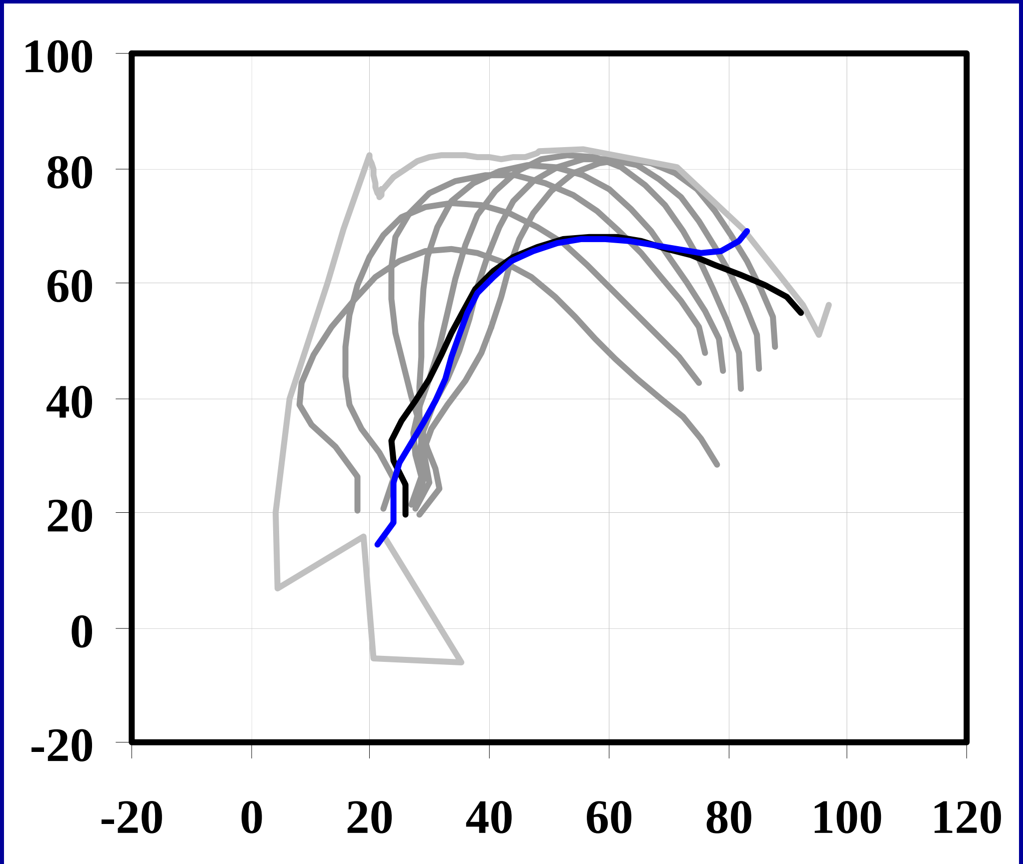
Constrictions

Dorsal articulations



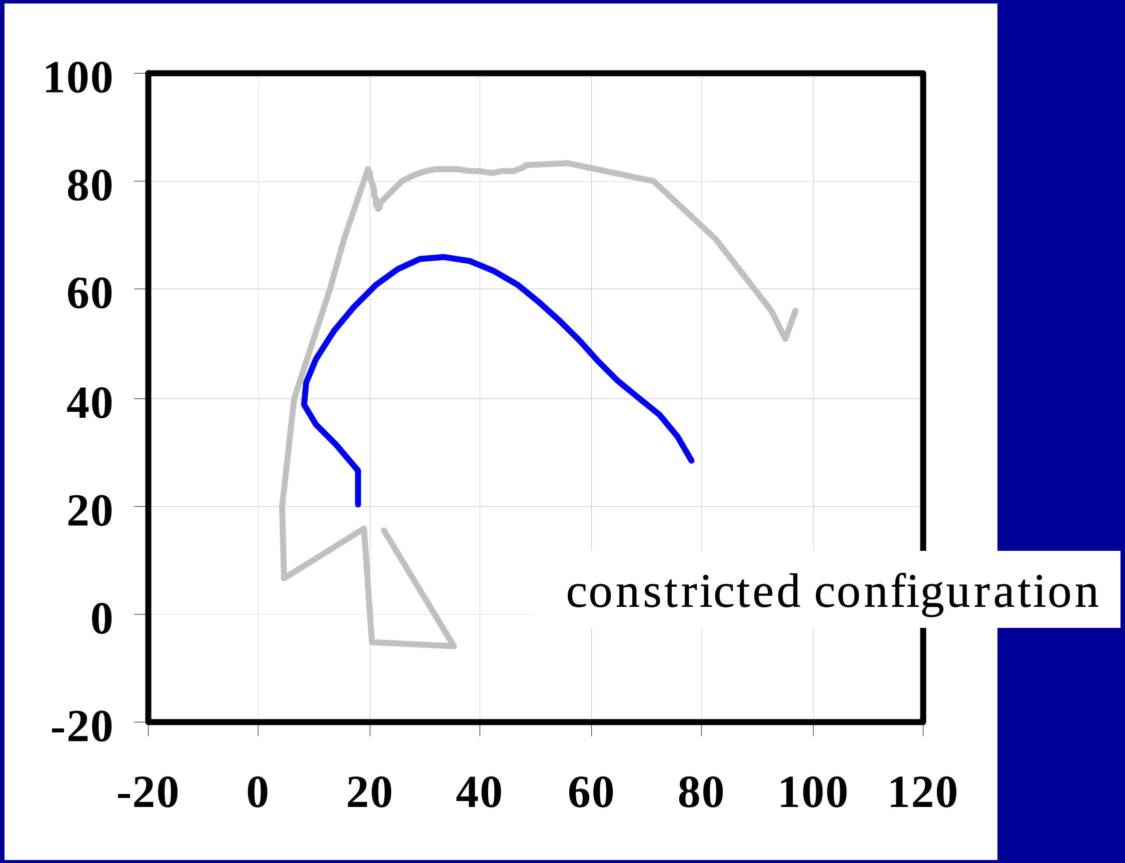
Constrictions

Dental & retroflex shapes



Open configurations

are derived by interpolation

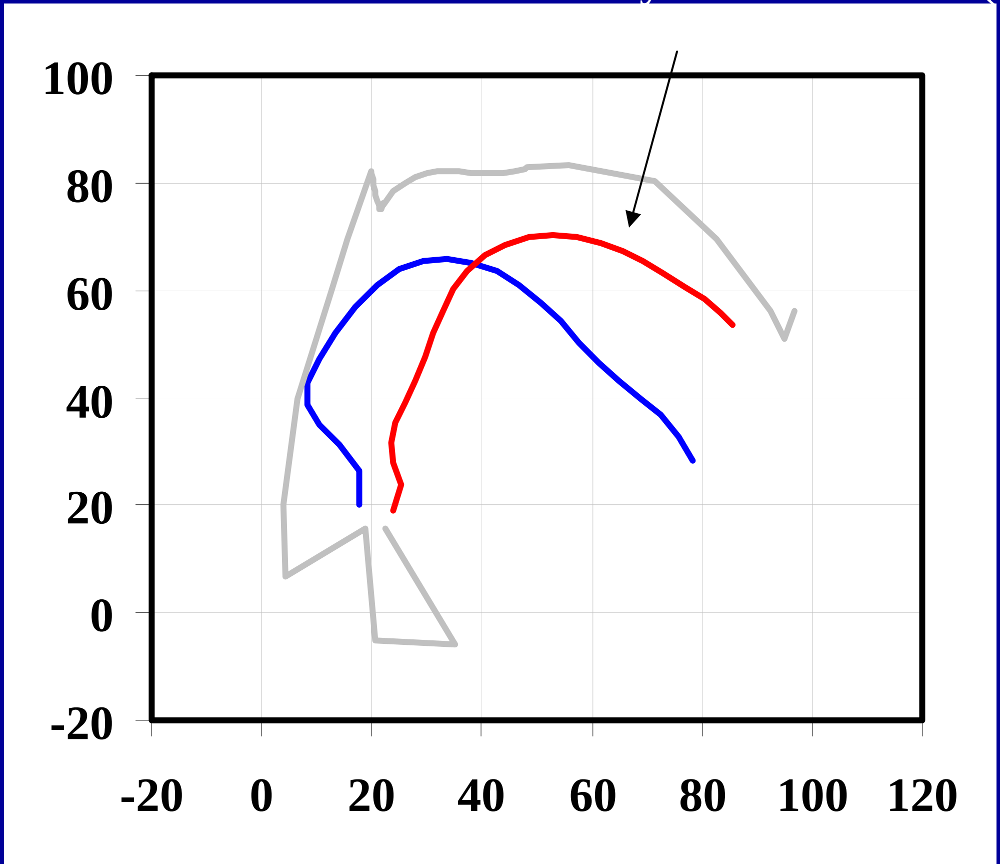


constricted configuration

Open configurations

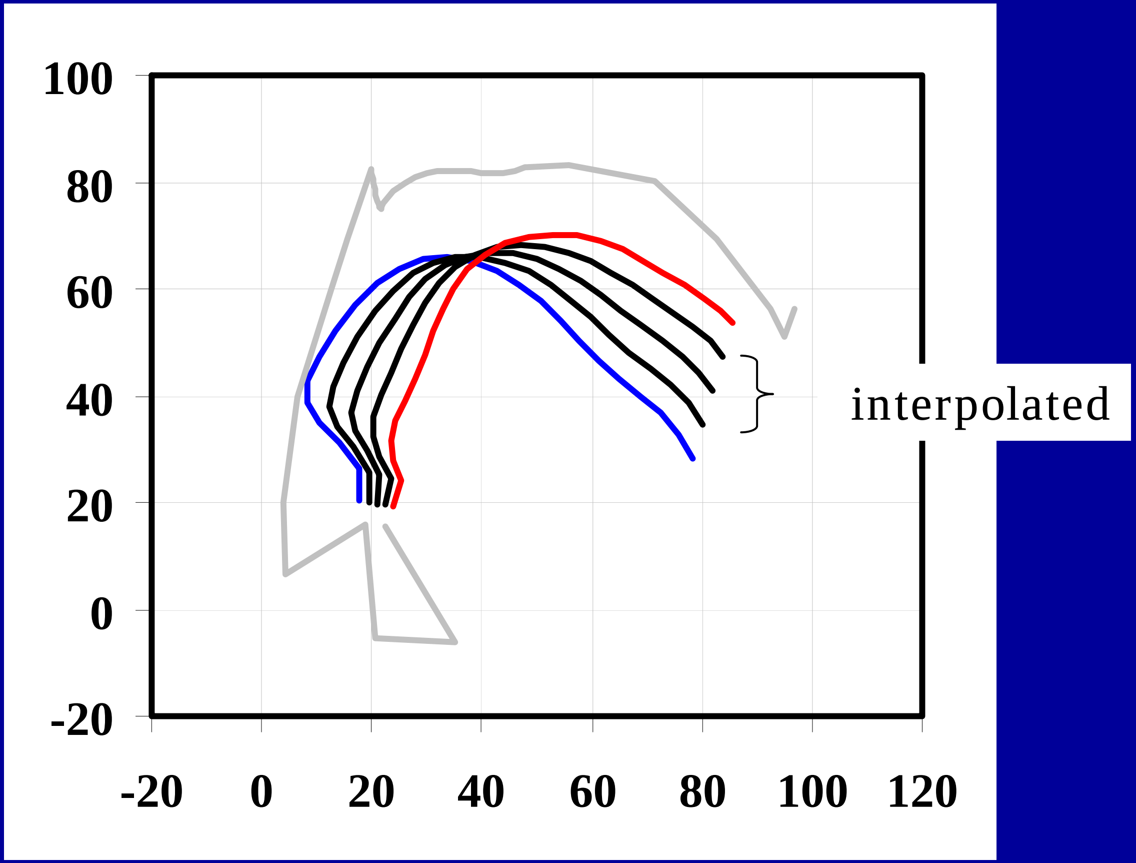
are derived by interpolation

maximally unconstricted (=average) contour



Open configurations

are derived by interpolation

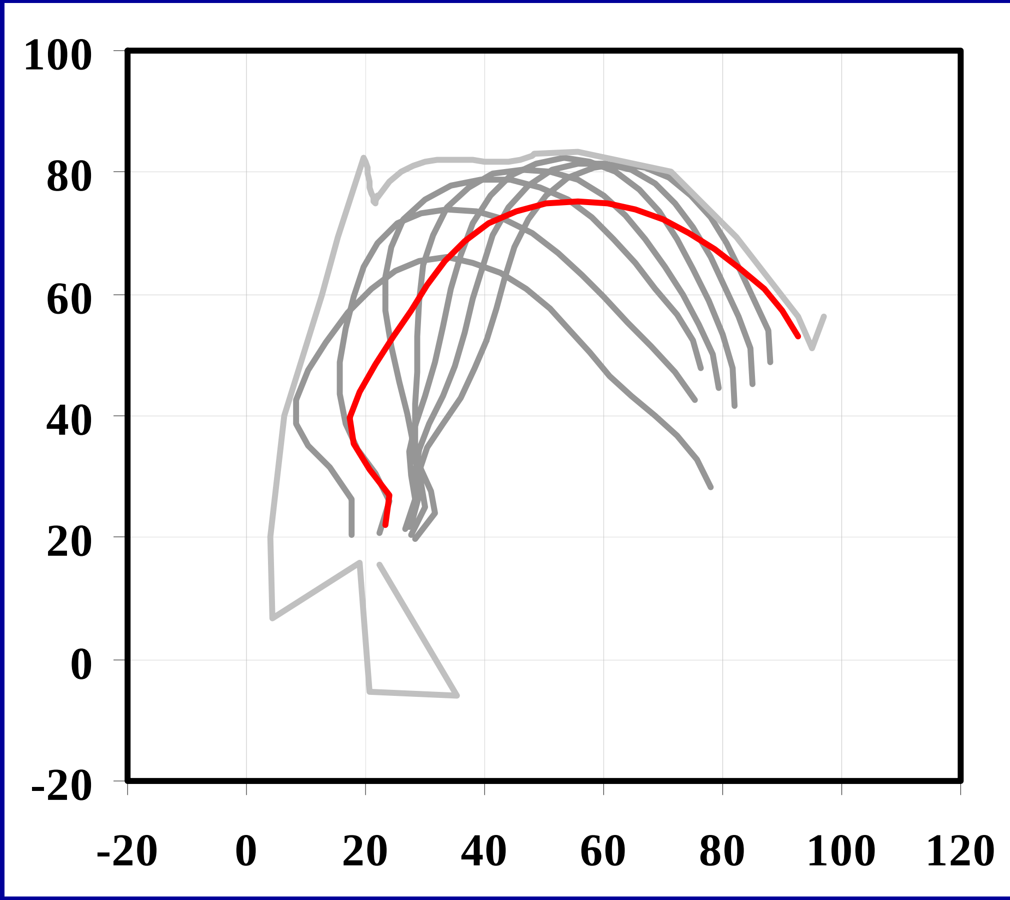


Quantifying ‘articulatory cost’

a simplified biomechanical account

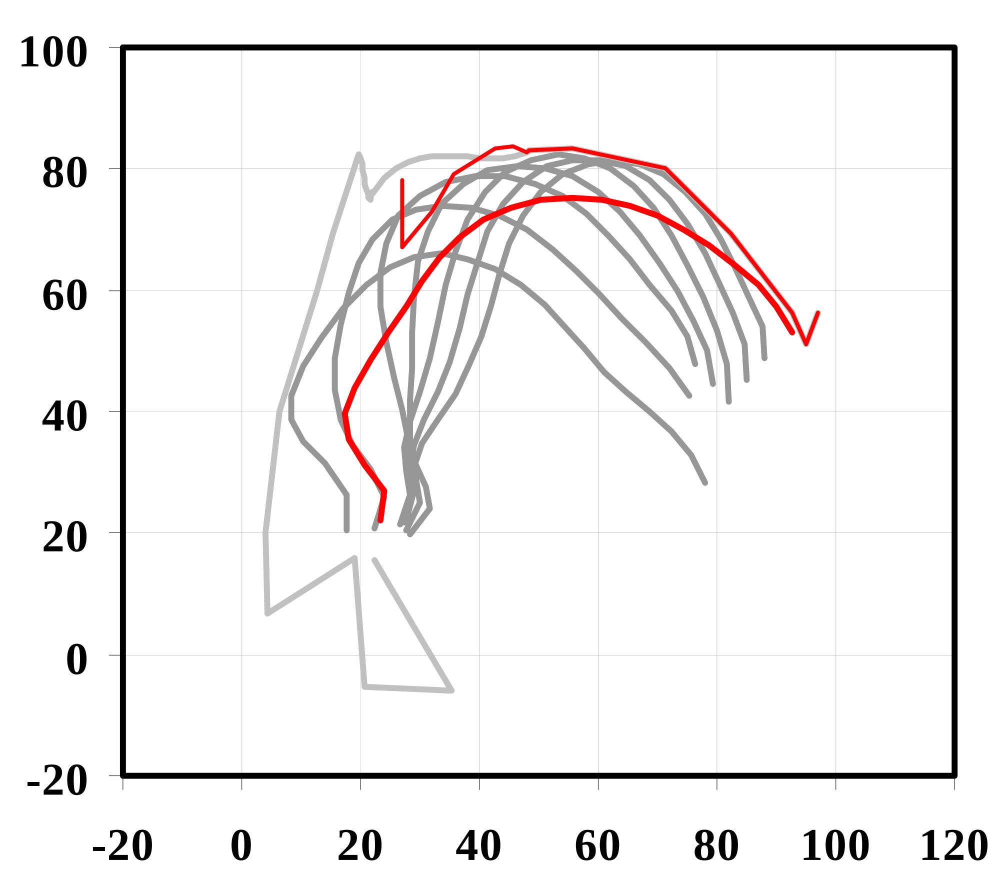
Rest position

- The position during quiet breathing ('habitual rest' in odontology);
- Jaw raised, tongue fronted, breathing through nose.

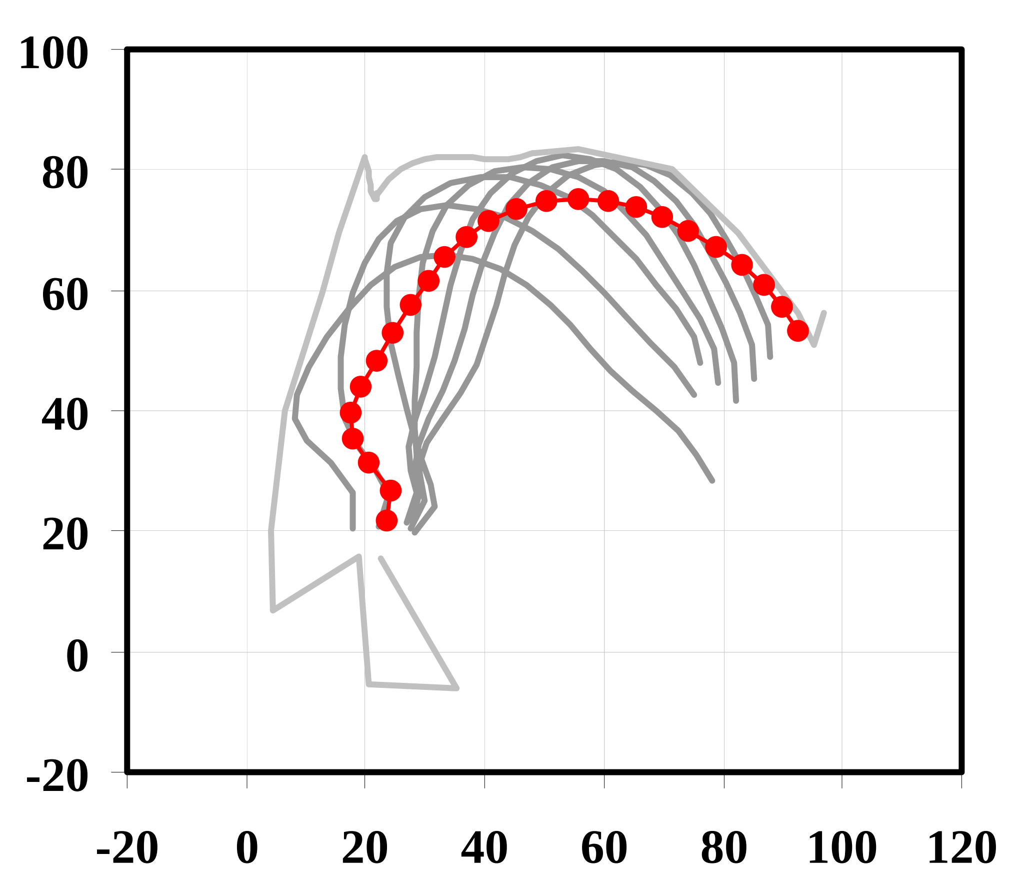


Rest position

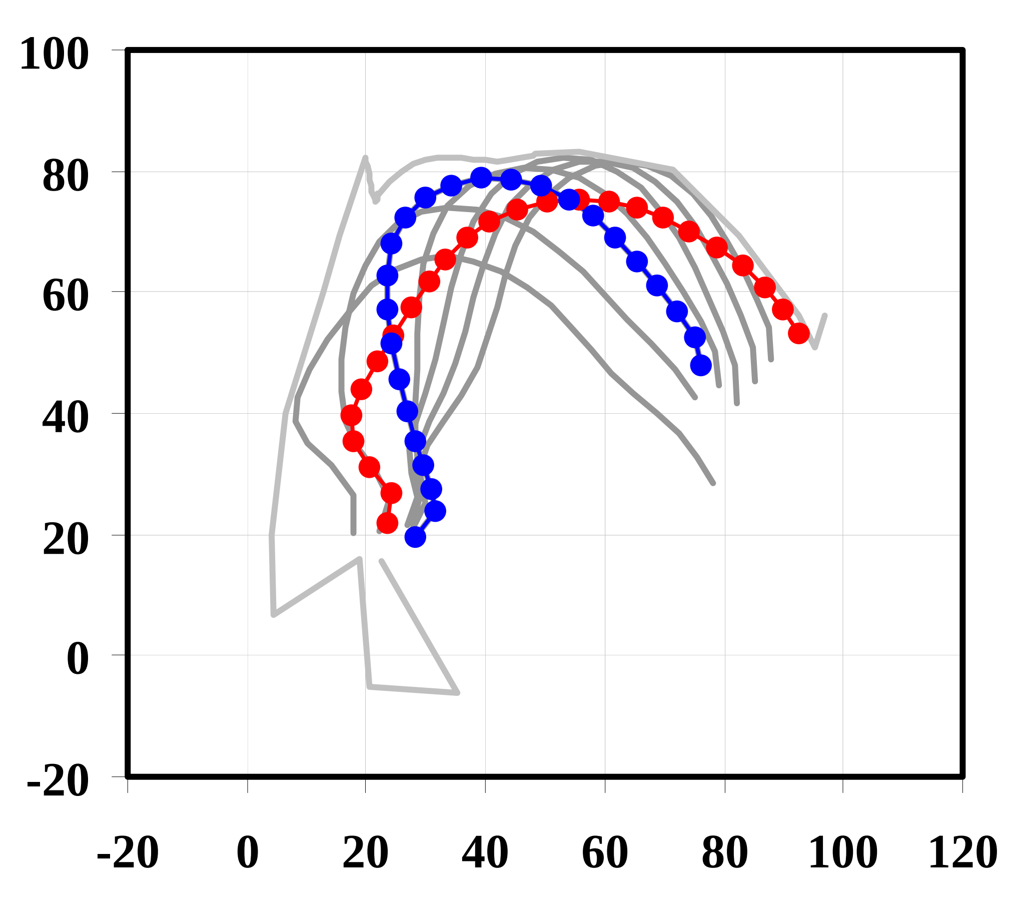
- The position during quiet breathing ('habitual rest' in odontology);
- Jaw raised, mouth closed, breathing through nose, tongue fronted.



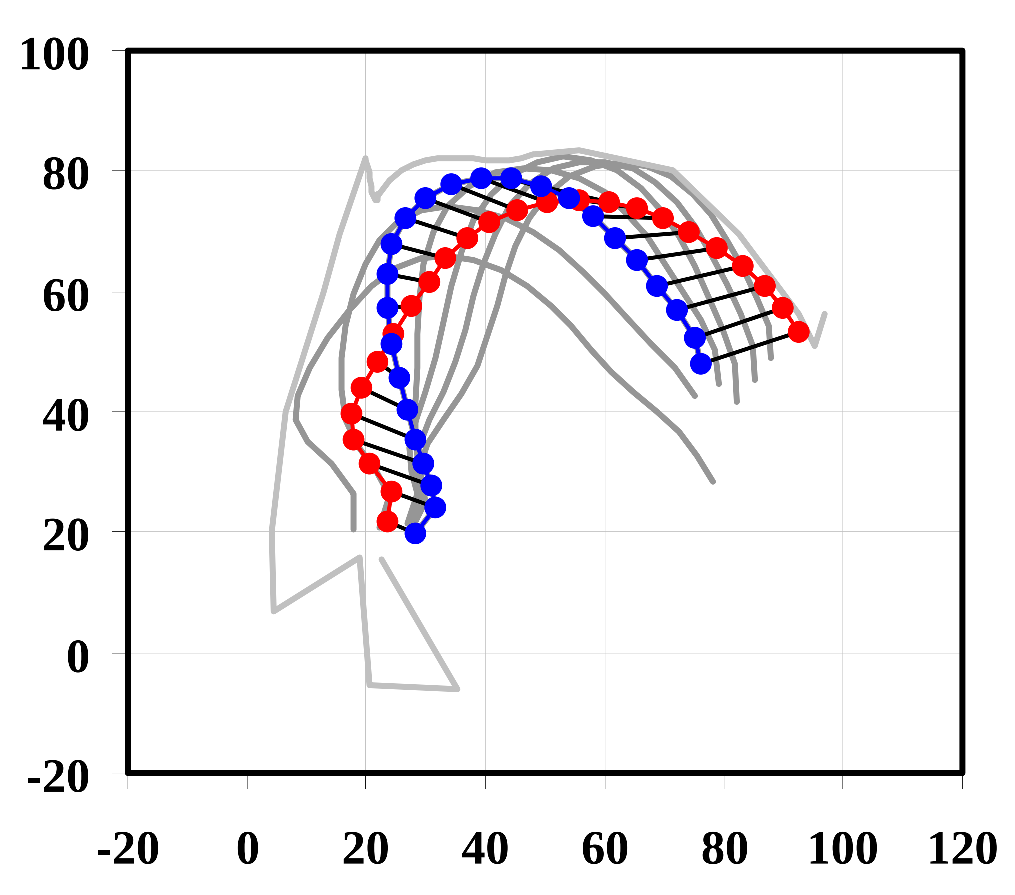
Deviation from rest



Deviation from rest



Deviation from rest

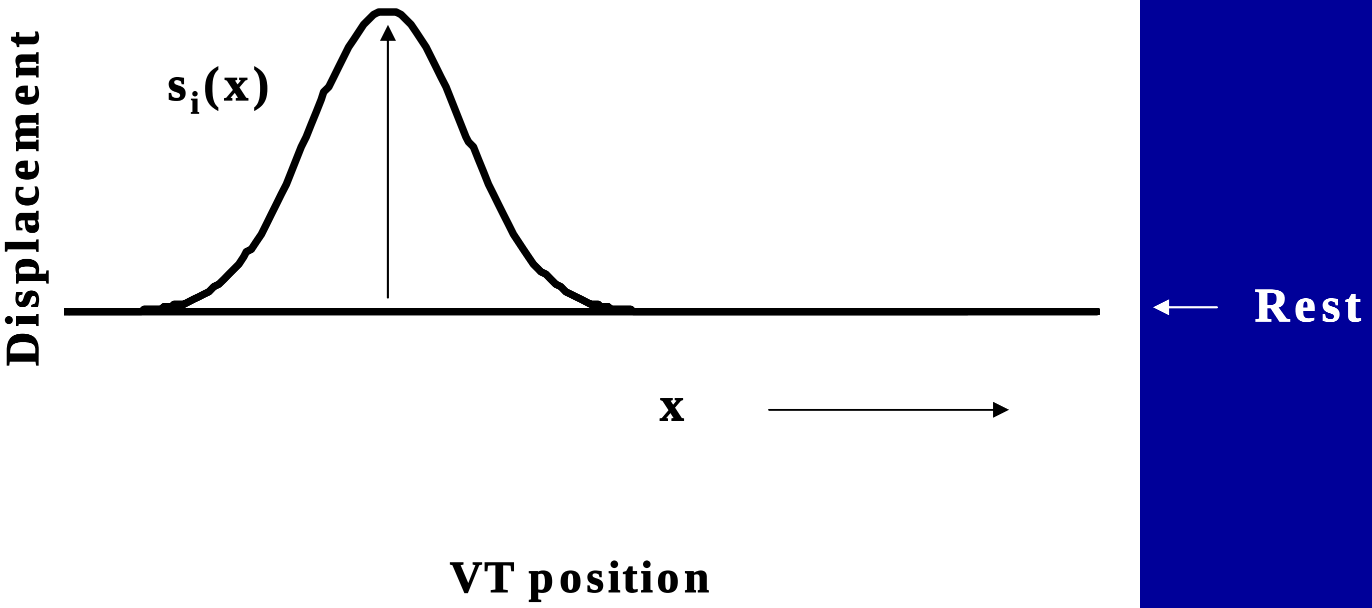


Rms distance =

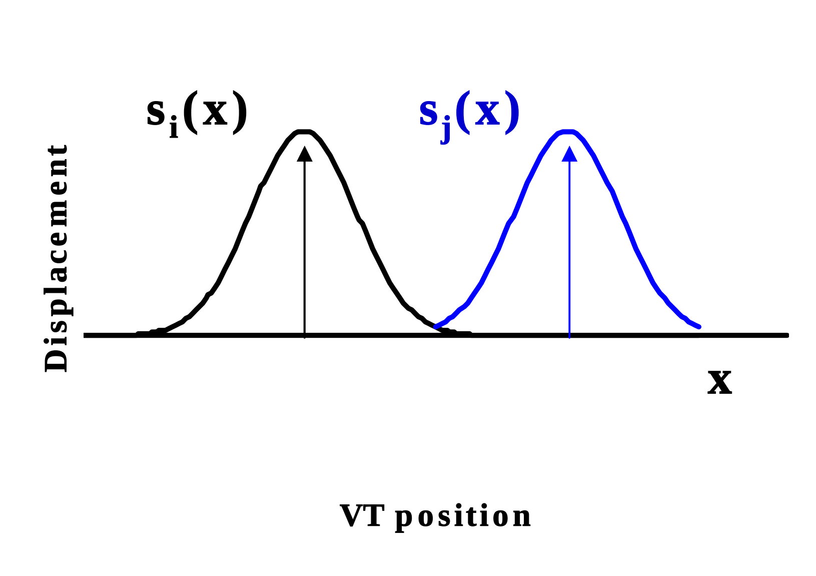
$$\sqrt{\left(\sum_{1}^{25}[a(x)-b(x)]^2\right)/25}$$

Single event

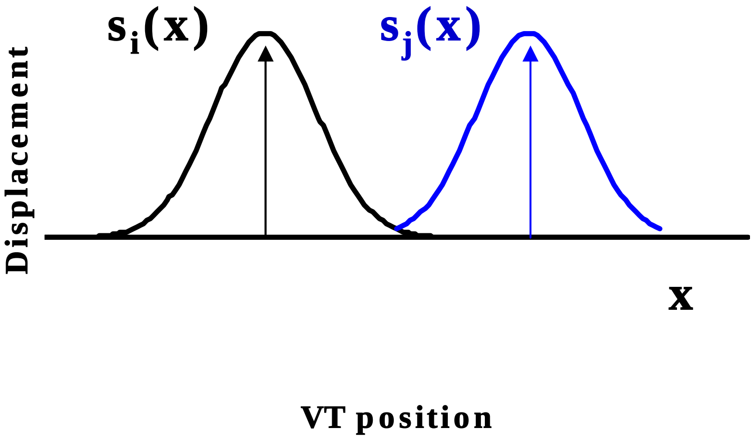
A-cost = Dist[rest(x) - s_i(x)]

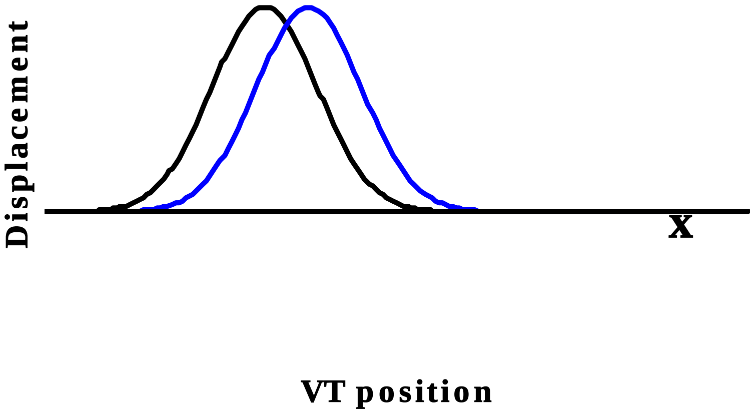


Multiple events

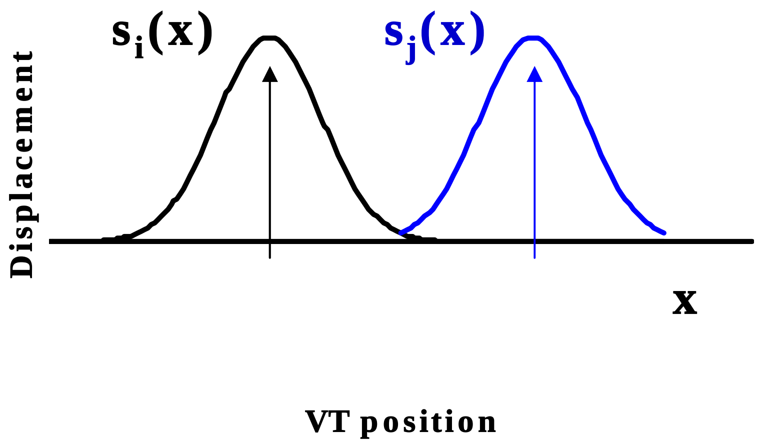


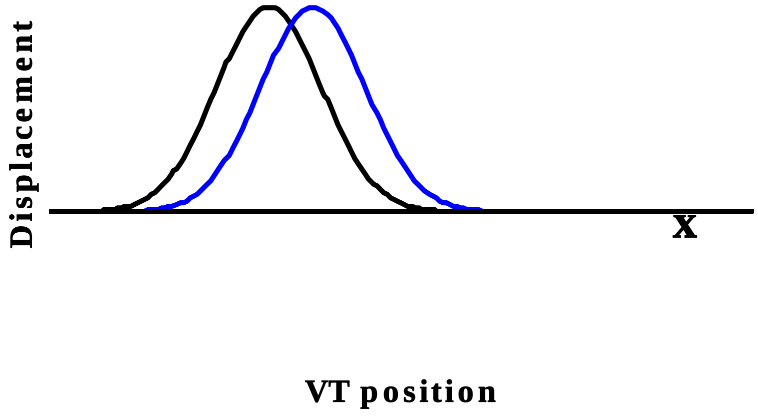
Multiple events

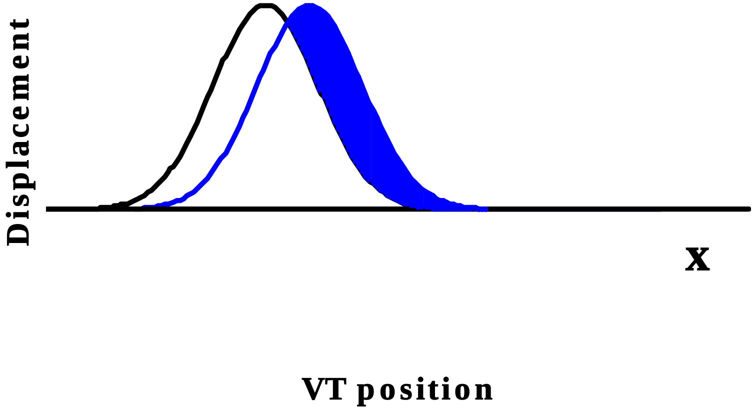




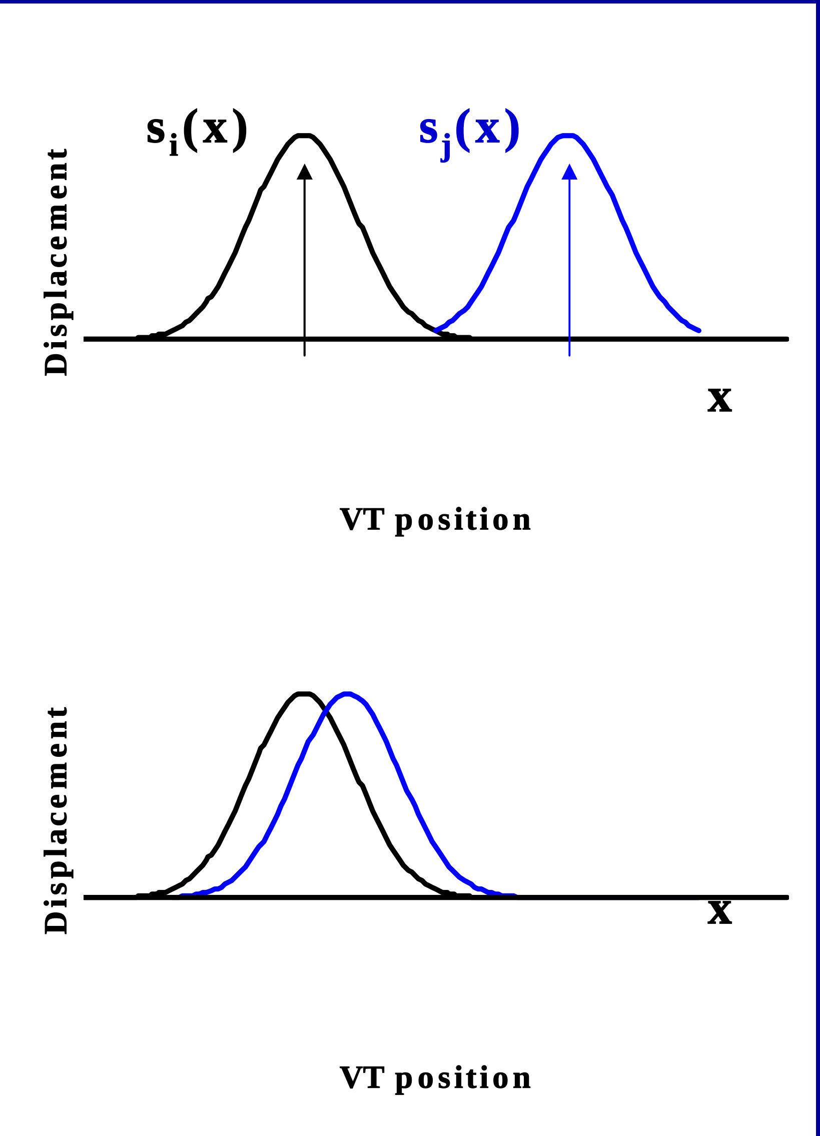
Multiple events

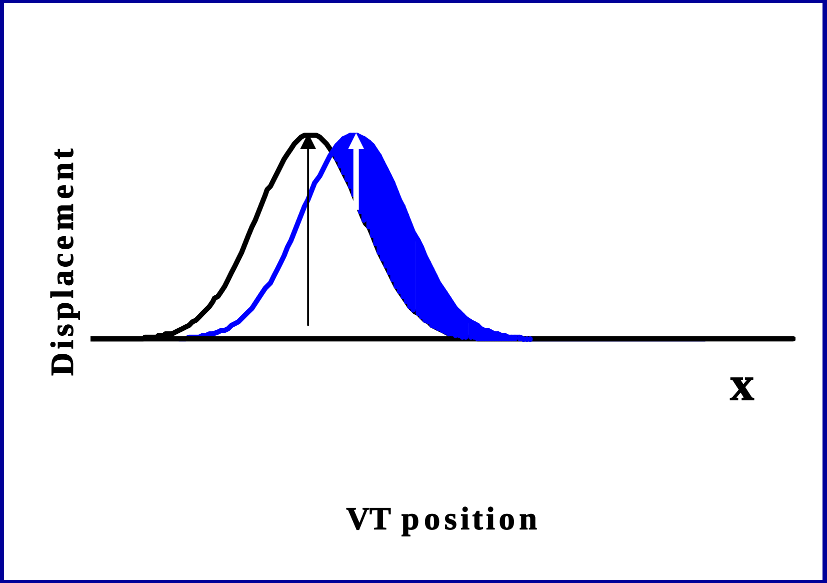




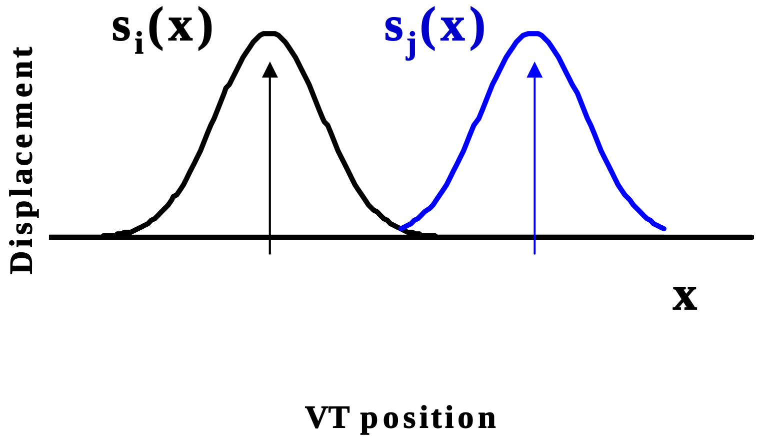


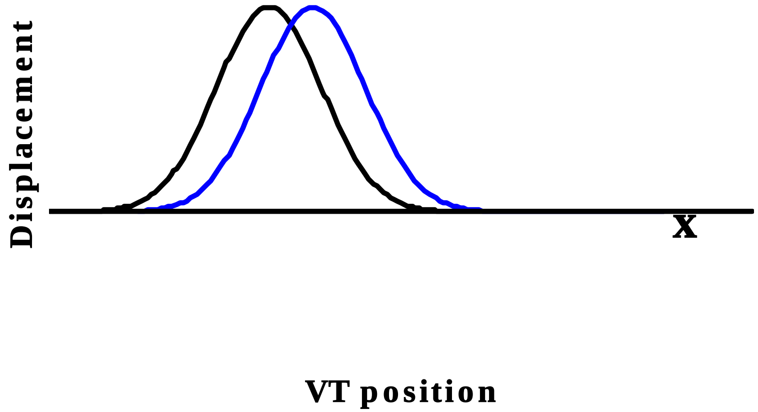
Multiple events





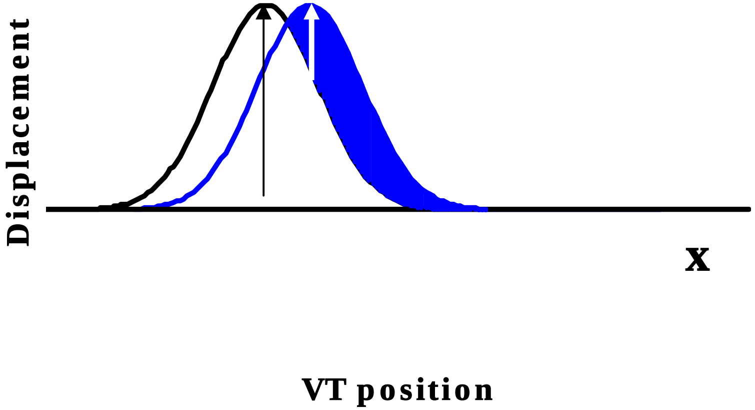
Multiple events





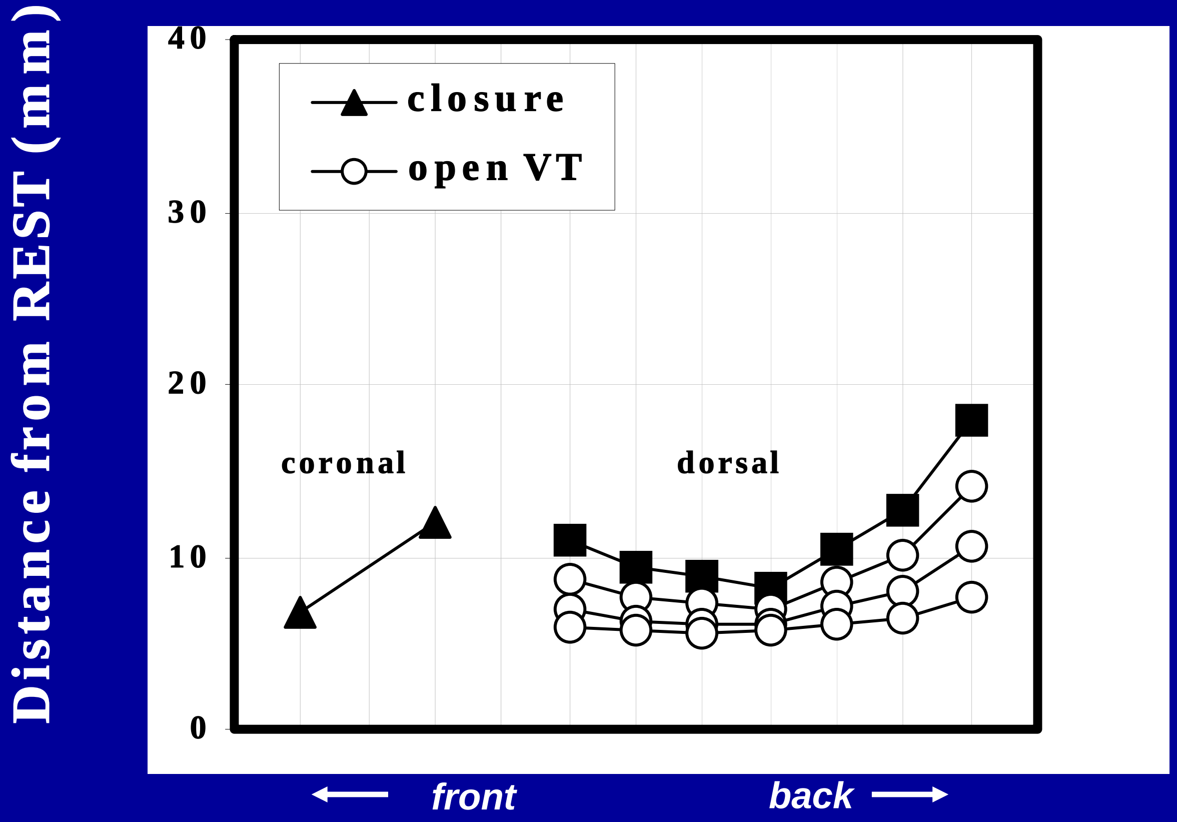
$$A\text{-}cost =$$

$$Dist[\ rest(x) - s_i(x)\ ] + Dist[\ s_j(x) - s_i(x)\ ]$$



Articulatory ‘costs’

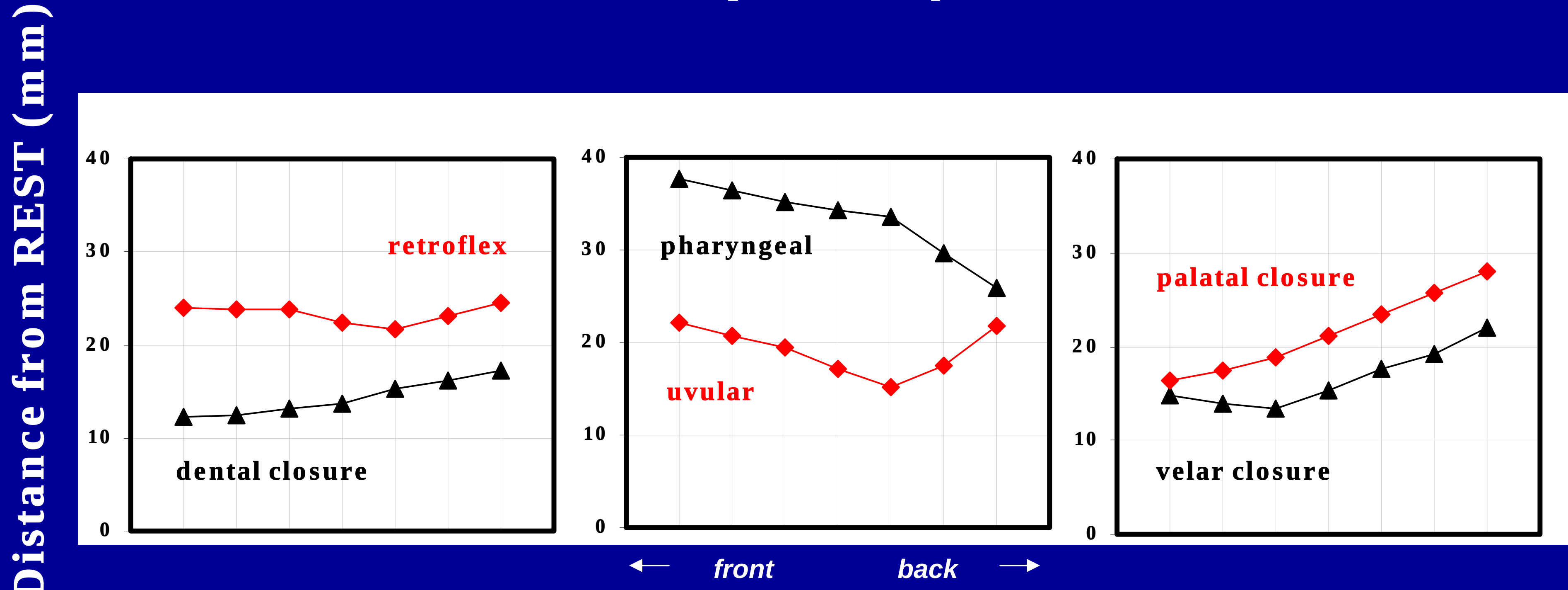
for closures & open gestures separately



Place of articulation of closures & open VT:s

Articulatory ‘costs’

for closed-open VT sequences



Place of articulation of open VT

1. Articulatory ‘costs’ vary as a function of place.
2. Front onsets more ’costly’ when followed by back contexts and conversely.

Comparing two ‘learning’ strategies

1. Convert the A-costs into probabilities

$$p_{ij} = \frac{A_{ij}}{\sum_{i=1}^{10} \sum_{j=1}^{21} A_{ij}}$$

Aij = A-cost for movement from closure *i* to open VT *j*

p_{ij} = probability of spontaneous production or imitation of ij^{th} phonetic pattern.

2. Define equations that simulate ‘learning’ by recalibrating these probabilities

Gestural control

Treating each trajectory as a whole

1. Learning equation I:

$$p_{ij} + k_{ij} * \varepsilon = 1$$

t

2. Solve for k_{ij}, the *amount of practice* needed to reach the criterion.

$$k_{ij} = (1 - p_{ij}) / \varepsilon$$

Endpoint control & GPM

learning, not the gesture, only its 'least action' representation

1. Learning equation II:

$$p_{ij} + k_{ij} * (\varepsilon + r) = 1$$

where

re-use factor →

$$\boxed{r} = (w(i) + z(j))$$

w(i) & *z(j)* reflect the degree of previous activation along the i^{th} & j^{th} dimensions

2. Solve for k_{ij}, the *amount of practice* needed.

$$k_{ij} = (1 - p_{ij}) / (\varepsilon + r)$$

How the models differ

Gestural control

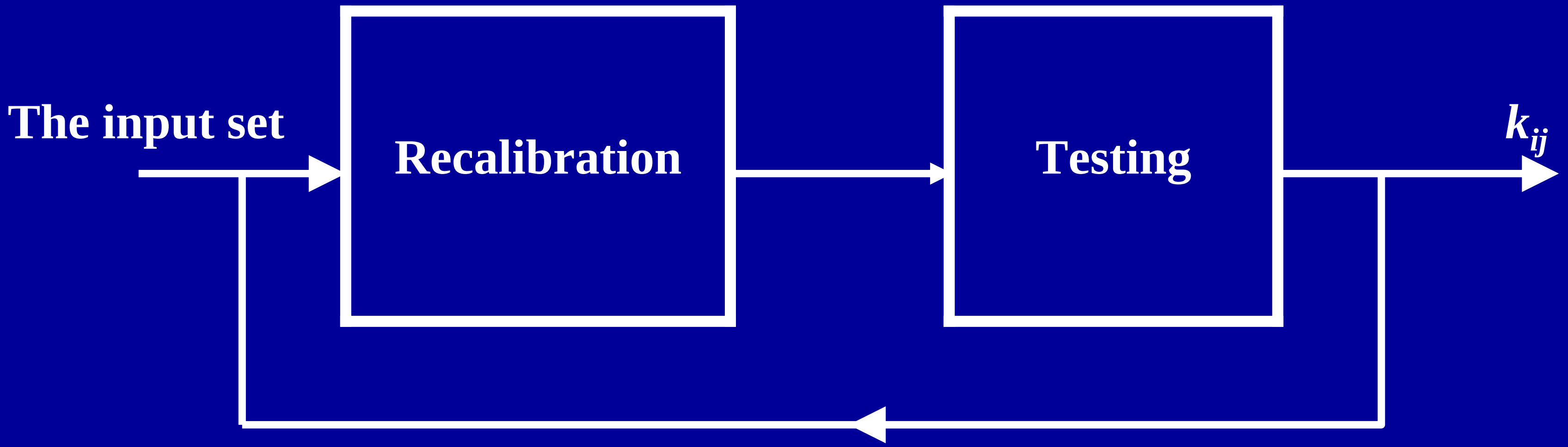
$$k_{ij} = (1 - p_{ij}) / \varepsilon$$

Store the whole thing!

Target control

$$k_{ij} = (1 - p_{ij}) / (\varepsilon + \enclose{circle}{r})$$

Store minimally necessary info!

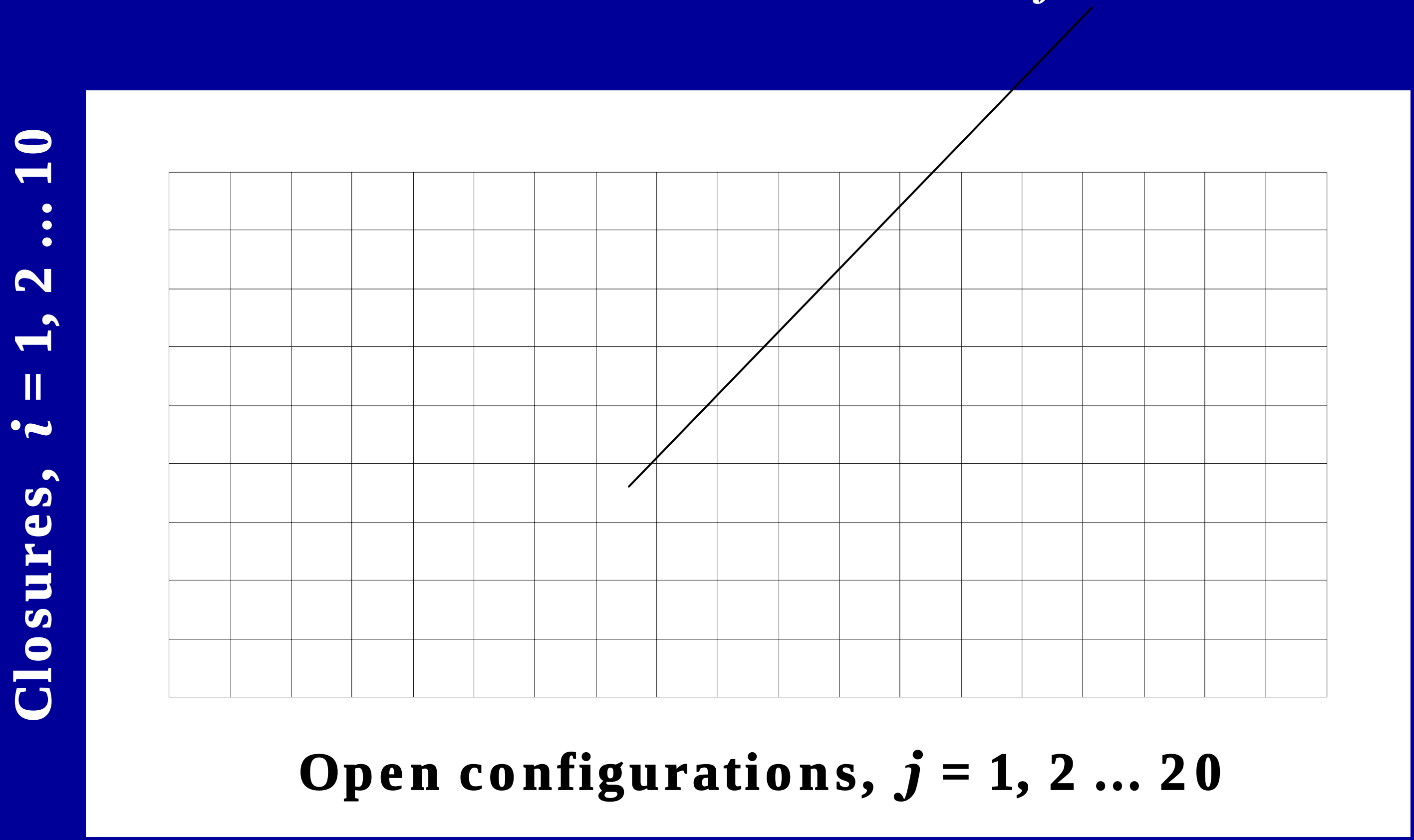


- Let $k = 0$;
- Present input system;
- Update k: $k = k + 1$;
- Compute Learning Equation;
- Output k when criterion =1;
- Repeat from 2;
- Stop when criterion =1 for all input items.

Recalibration matrix

Each cell contains the current value of the learning equation:

$$p_{ij} + k*\varepsilon$$



Closures, i = 1, 2 … 10

Open configurations, j = 1, 2 … 20

Arrays specifying current activation status

Closure array *w(i)*, *i* = 1, 2, … 10

0	0	1	0	…	…	1	0	1	0

Open VT array *z(j)*, *i* = 1, 2, … 20

0	0	0	1	1	0	…	…	…	0	1	0	1	1	0

Comparing degrees of re-use

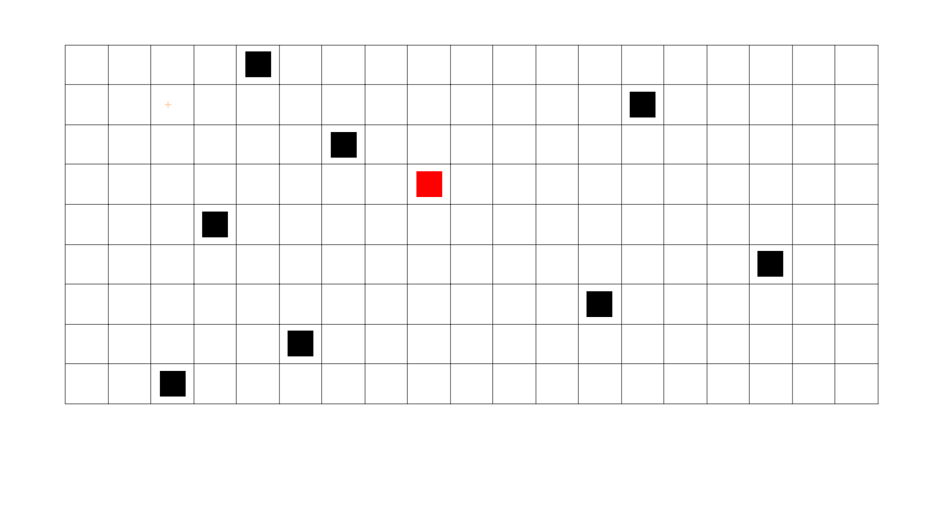
(a) *r=0*

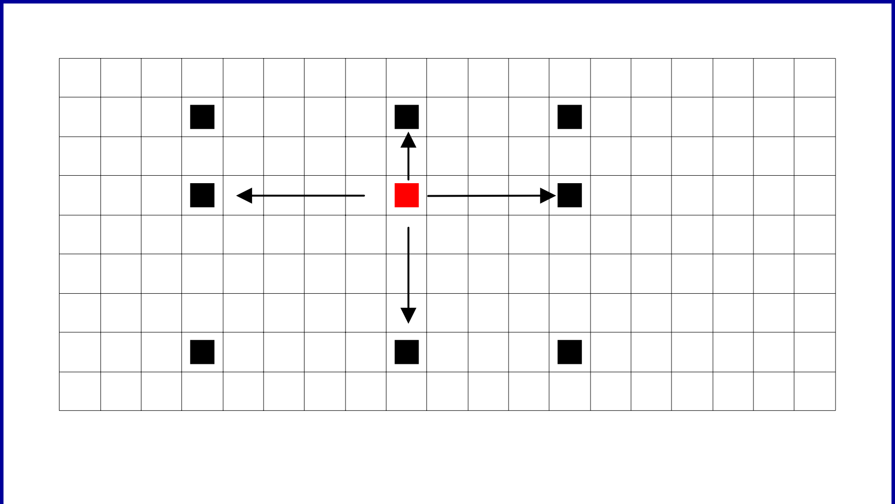
Minimum re-use

(b) *r=4*

Maximum re-use

3-by-3 system





Learning of red pattern (identical for **a** & **b**)

1. The gestural model makes no difference between **a** and **b**
2. According to the target model, **case b** should be learned faster than **case a** because of the re-use factor.
3. In **a** the learning time for the red pattern is determined solely by its articulatory 'learnability'.
4. In **b** the learning time also depends on systemic factors (e.g., "feature economy".

How much faster because of re-use?

- Compute:
  Ratio between ‘gesture’ practice and ‘target’ practice:

$$\frac{(1-p_{ij})(\varepsilon+r)}{\varepsilon(1-p_{ij})}=\frac{(\varepsilon+r)}{\varepsilon}$$

- Conclusion:

  Learning speed is directly proportional to re-use factor, *r*. *Hence* re-use promotes learning.

Phonetic systems

- Distinctiveness & auditory realism (Diehl et al 2003)
- Articulatory factors (Maddieson 1984:16)

  /i ẽ a̤ o̰ uˤ/

- Size Principle (Lindblom & Maddieson)

- Ohala's predicted 7-consonant system (chairman's comments ICPhS 1979):

  ɗ, kʼ, ts, ɬ, m, r, ʡ

"Rather than maximum differentiation of the entities in the consonant space, we seem to find something approximating the principle which would be characterized as "**maximum utilization of the available distinctive features**". This has the result that many of the consonants are in fact, perceptually quite close – differing by a minimum, not a maximum number of distinctive features."

Summary of present proposals and claims

1. Adult speech movements:

   * Spatial targets play a primary role in speech motor control;

   * Transitions between segments are determined by targets, the GPM and constraints set by syllabic and prosodic factors;

2. Speech development: End-state of phonetic learning

   * Target representations and the GPM are found by behavior driven by a ‘least action’ criterion;

3. Origin of phonetic structure:

   * Re-use of discrete elements is significantly promoted by the Target-GPM organization:

     GPM = a key mechanism behind recombination

     Targets = the units to be recombined