Departement d'Etudes Cognitives - Ecole Normale Supérieure
Ecole des Hautes Etudes en Sciences Sociales
Centre Nationale de la Recherche Scientifique

Modeling early language acquistion

Emmanuel Dupoux

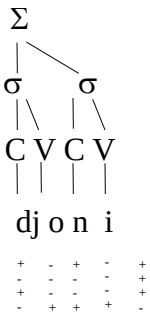

What we say to babies...
Well, Johnny, you've spilled your milk for the last time! I'll not tolerate that behavior any longer!..
What they hear
dawn

Underlying forms

/djoni/

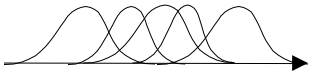
Phonological processes
- assimilation
- epenthesis
- deletion

Surface forms



Phonetic processes
- coarticulation

Articulatory plan



Articulation & transmission
- vocal tract shape
- overshoot/undershoot
- noise, reverb, etc

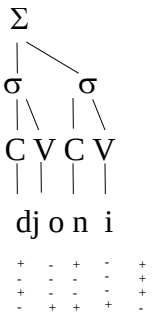
Acoustic signal

Underlying forms

/djoni/

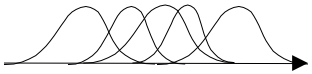
Phonological processes
- assimilation
- epenthesis
- deletion

Surface forms



Phonetic processes
- coarticulation

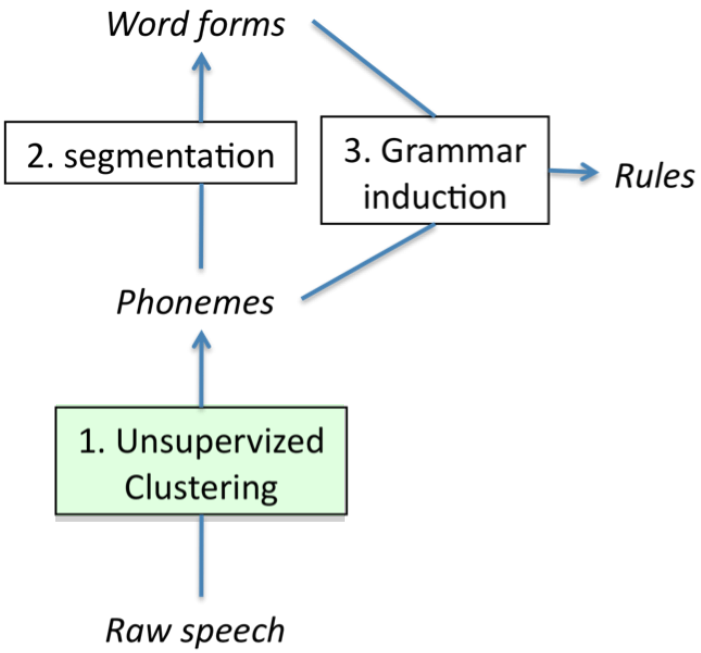
Articulatory plan



Articulation & transmission
- vocal tract shape
- overshoot/undershoot
- noise, reverb, etc

Acoustic signal

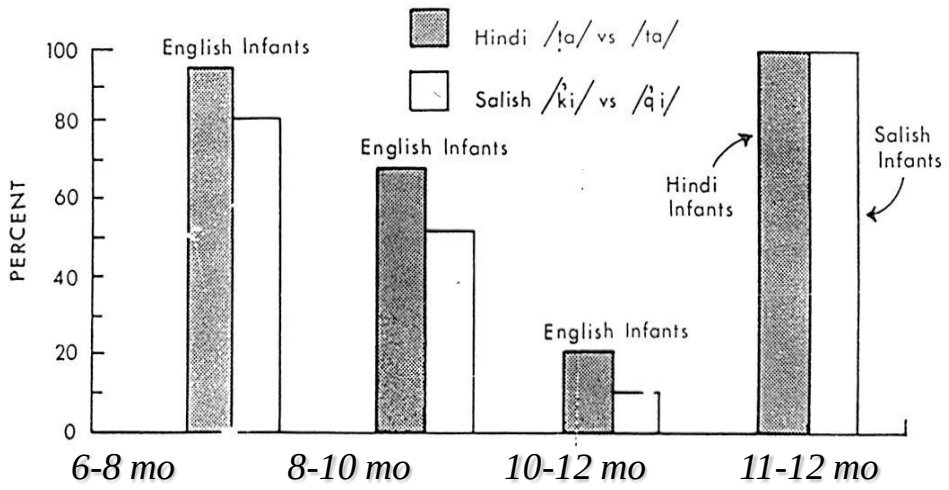
Standard scenario
the sequential bootstrapping scenario



→Modular

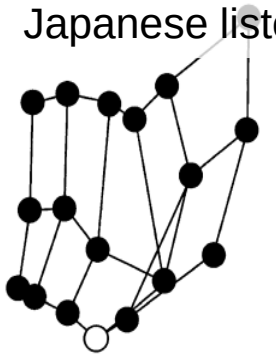
→ Bottom-up

Step 1: When?

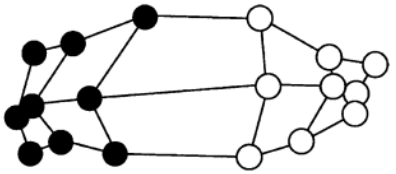


Werker & Tees (1984)

Japanese listeners

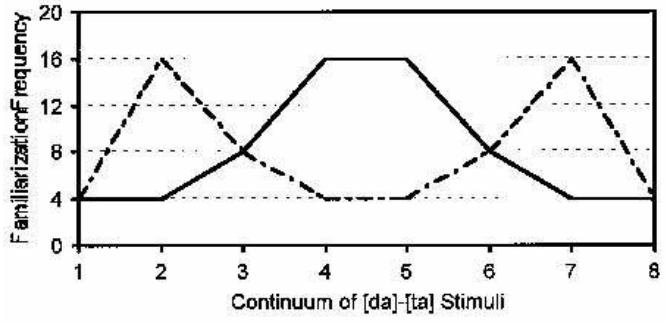


American listeners

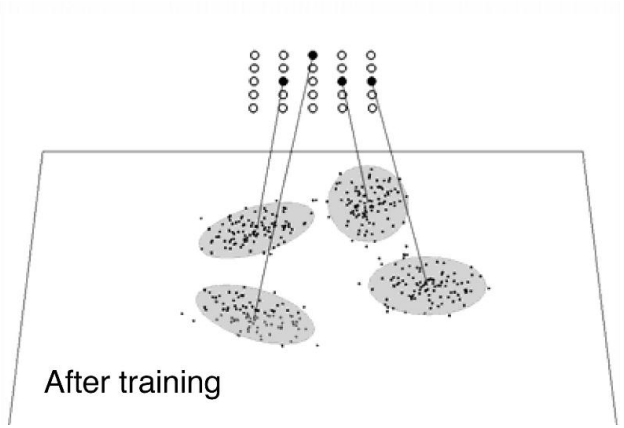


Iverson & Kuhl (2003)

Step 1: How?



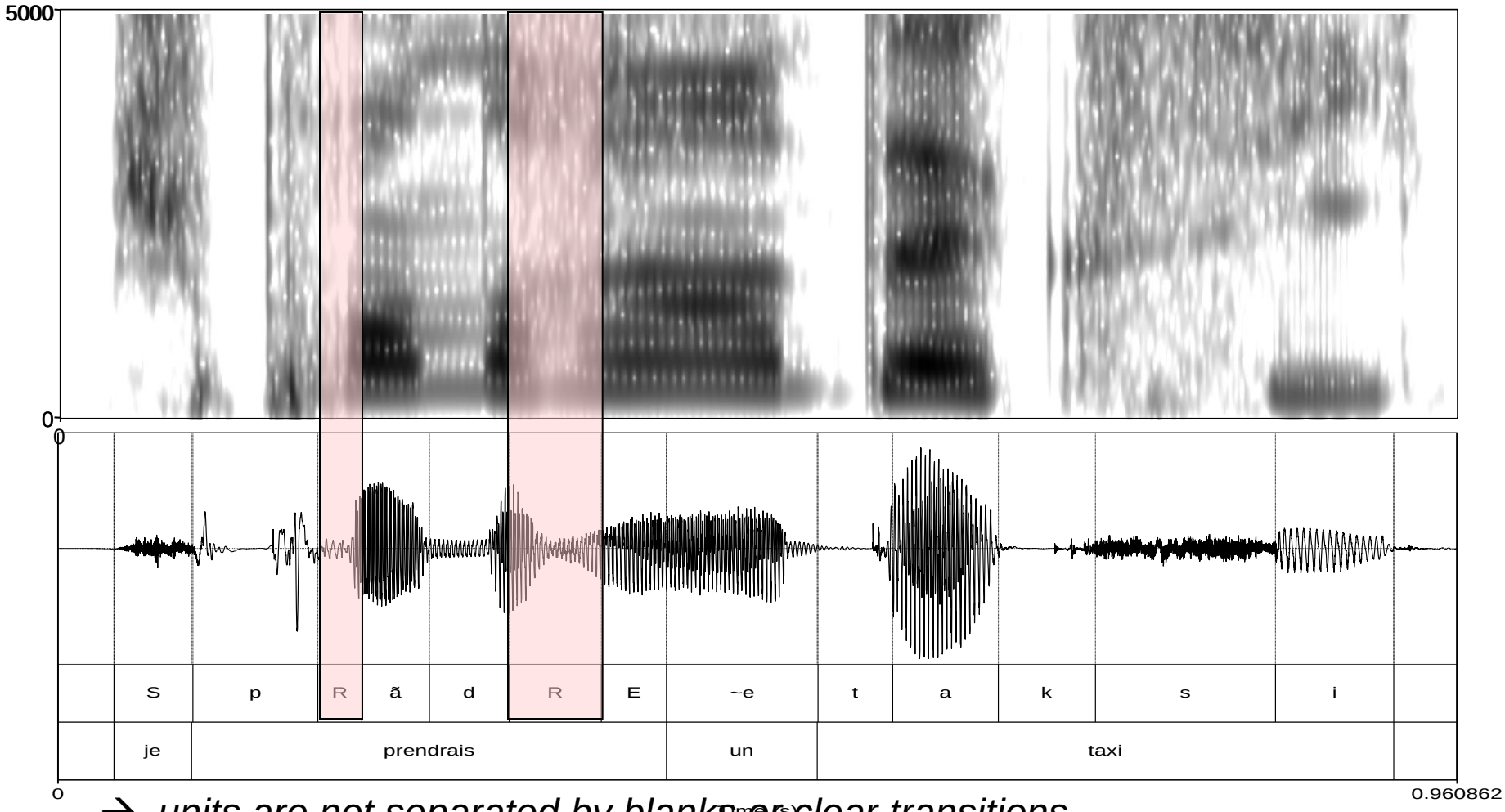
Maye and Gerken (2002)



Vallabha, et al (2007), Gauthier, Shi & Xu (2007)

Problem: not tested on raw unsegmented speech

Studies used parameters extracted by hand: vowel duration, peak F1, F2, etc

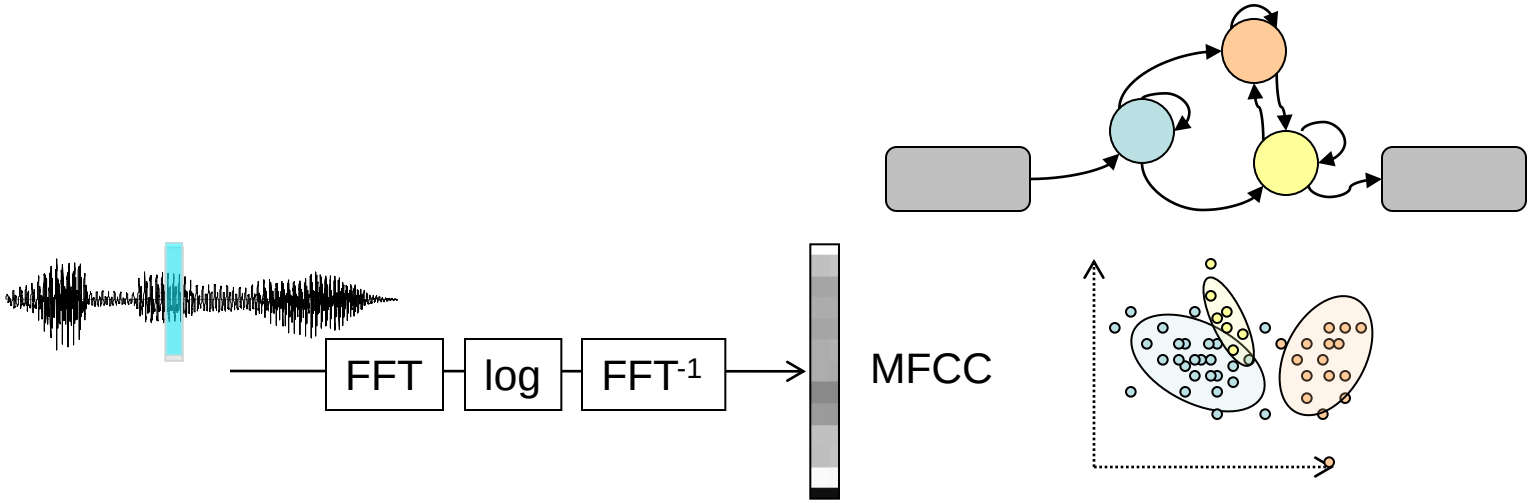


→ *units are not separated by blanks or clear transitions*

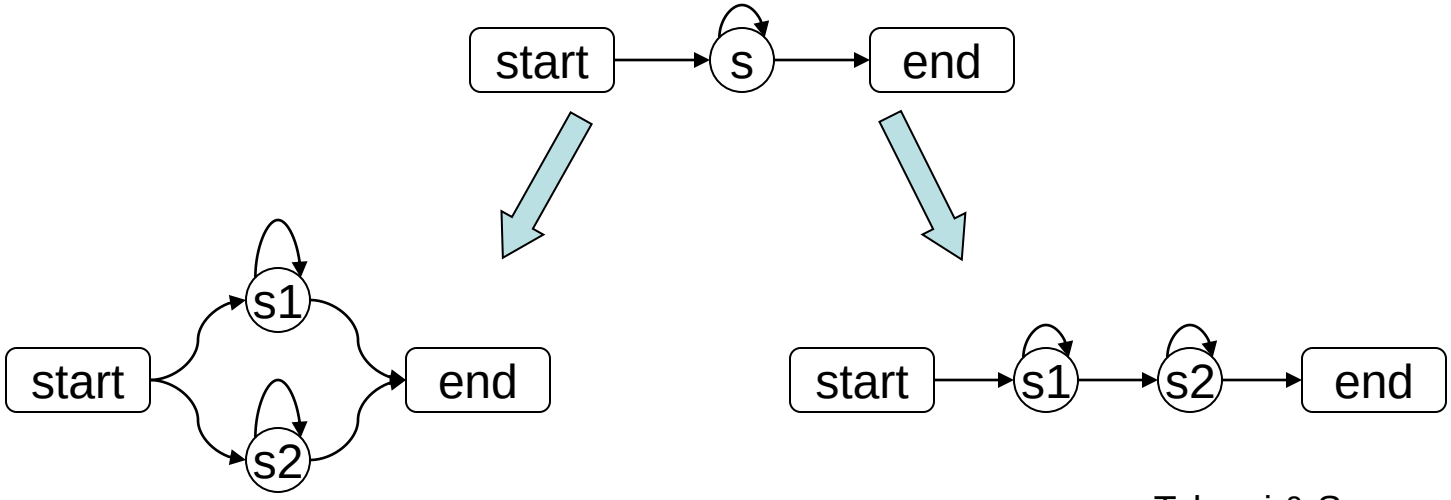
→ *Segmentation varies across languages*

→ *Segmentation is part of the learning problem*

Hidden Markov Models

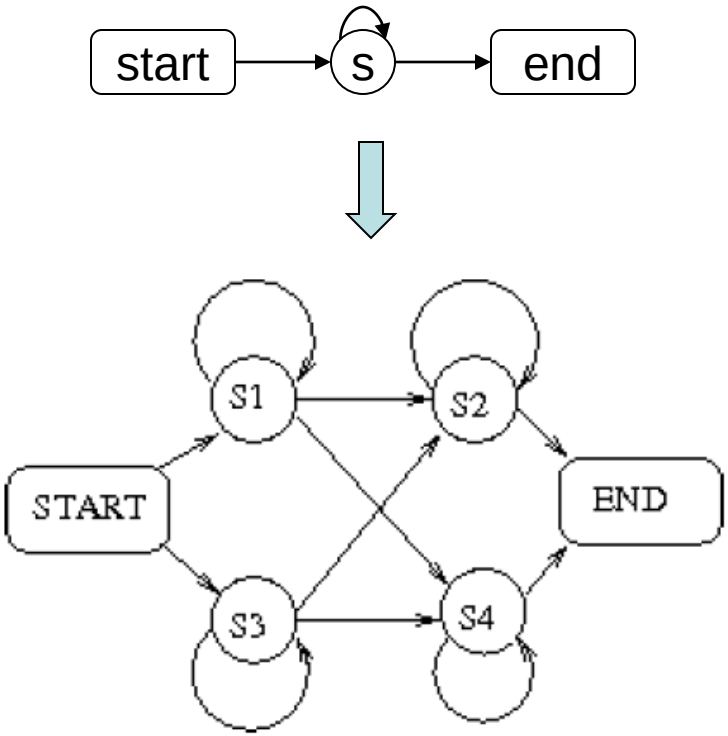


Successive State Splitting



Takami & Sagayama, 1991

Optimized State Splitting



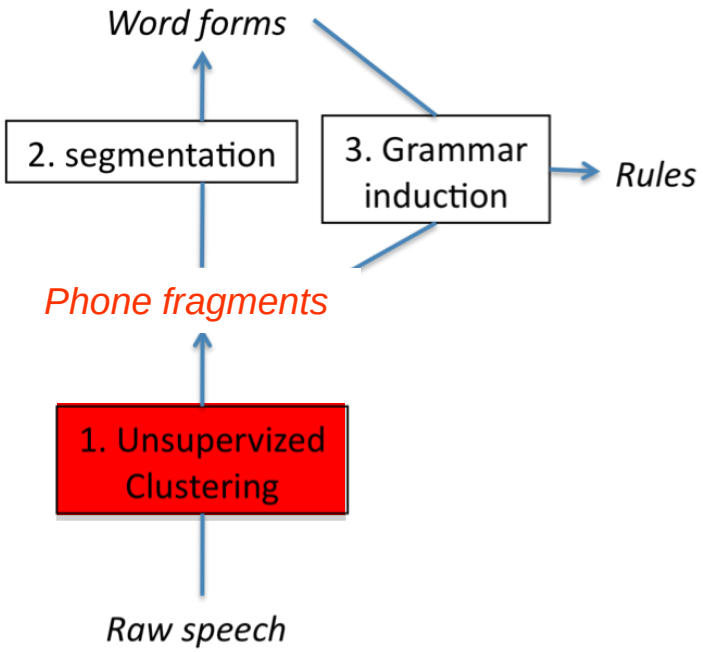
CSJ= 40 hours of spontaneous speech

State seq	Allophones
11,28,32	[V]-**t**+[e|a|o]
15,17,2	[g|k]-**[u|o]**+[⋆]
3,17,2	[k|t|g|d]-**a**+[k|t|g|d]
31,5,13,5	[V]-**[s|sj|sy]**+[V]
17,2,31,11	[g|t|k|d]-**[a|o]**+[t|k]
3,30,22,34	[⋆]-**a**
6 24 8 15 22	[⋆]-o
22 35 11 28 32	[N|i|u|o]-**[t|d]**+[e|o|i]
4 17 24 2 31	[s|sy|z]-**o**+[t|d], [t|d]-**o**+[s|sy|z]

reasonably accurate for recognition, BUT:
Problem 1. The units are *too small*: subphonemic units,acoustic events (30ms)
Problem 2. The units are *too specific* (context-dependent)

Varadarajan, Khudanpur,Dupoux, (2008)

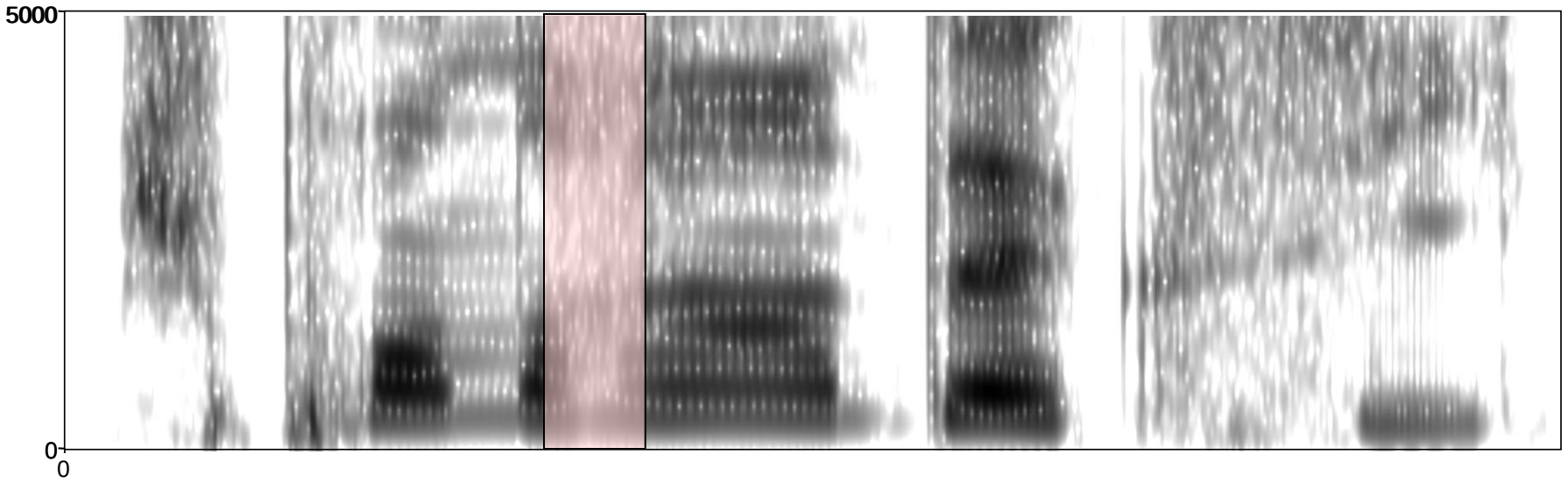
Problem #1: how bad?



→ Segmentation algorithms assume entire phonemes, not phone fragments

→Grammar induction algorithms assume entire phonemes and correct segmentation

Problem #1: why?



→Phonemes are not discrete states, but continuous trajectories

→Diagnostic: MFCC (or spectral features) are too LOCAL

→*Potential solution: replace MFCC by higher order representations (trajectory space)*

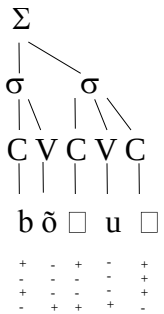
Problem #2: why?

Underlying forms

/bõjur/

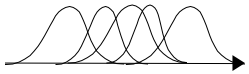
Phonological processes
- assimilation
- epenthesis
- deletion

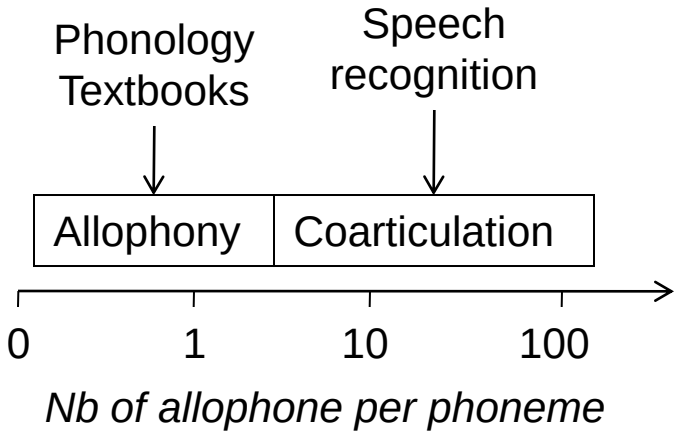
Surface forms



Phonetic processes
- coarticulation

Articulatory plan





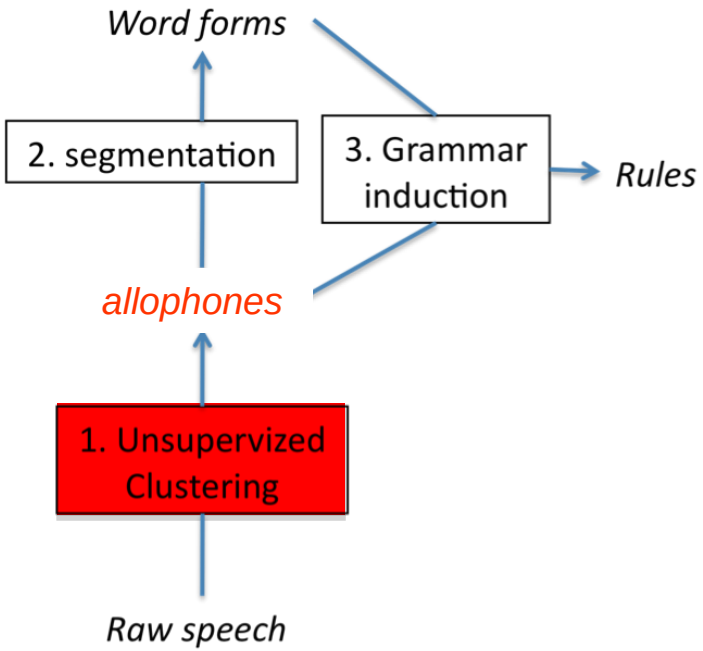
/kanar/ “duck”

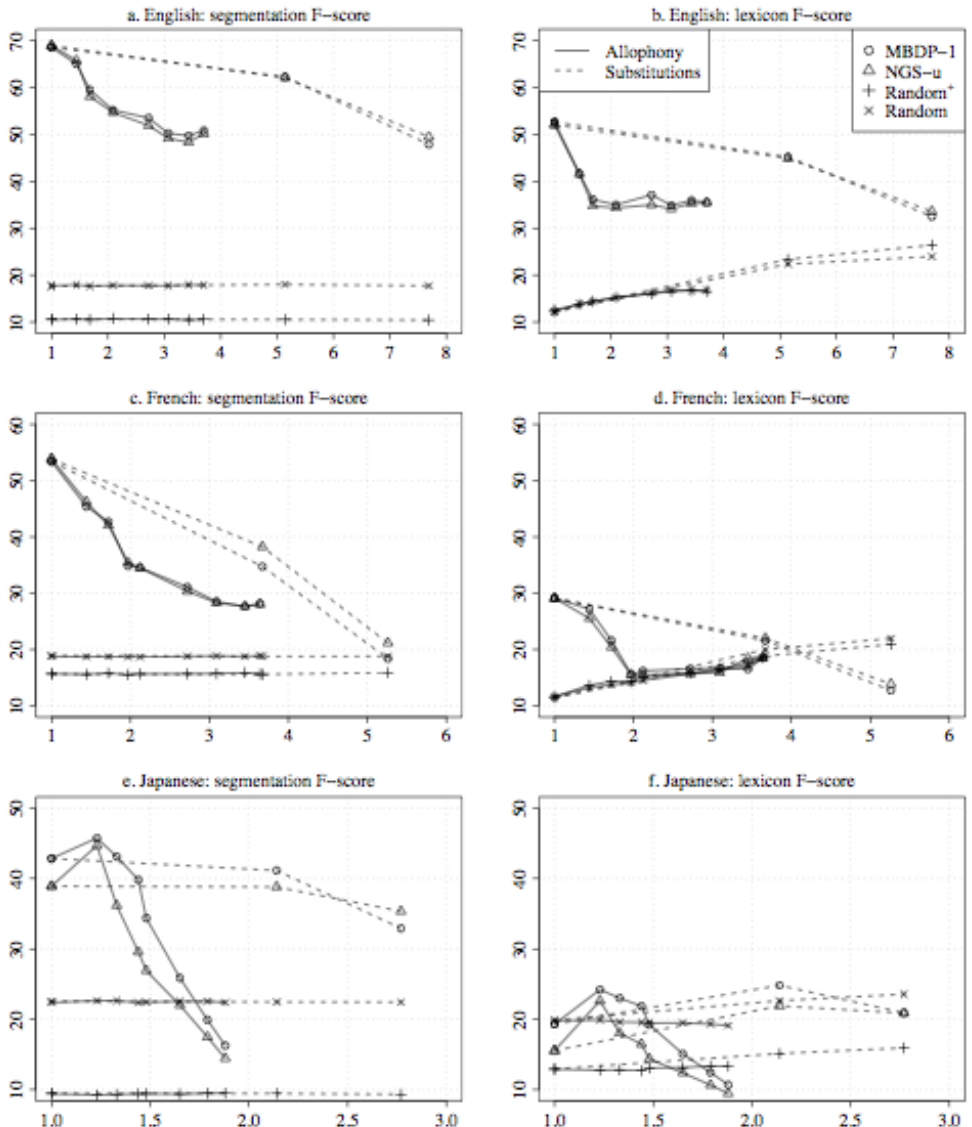
[kana**X**on]
“yellow duck”

[kana**X**flota~]
“floating duck”

ɬˆlɛˀvə̆n
ɬˆlɛvə̆n
ɬˆɬɛvə̆n
əˈlɛˀvə̆n
ɬˆlɛvə̆n
əlɛvə̆n
əˈɬɛvn
ɬɛbm
ˆˈlɛvn
ɛvə̆n
ɬɛˇˀvə̆n
ɛvɱ
ɬɛᵊvə̆n
ˈlɛvɱ

Problem #2: how bad?

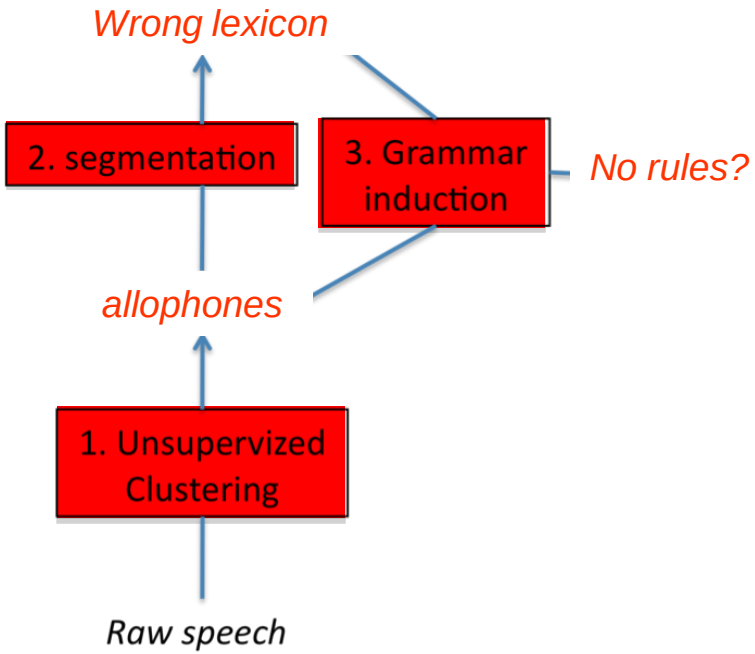


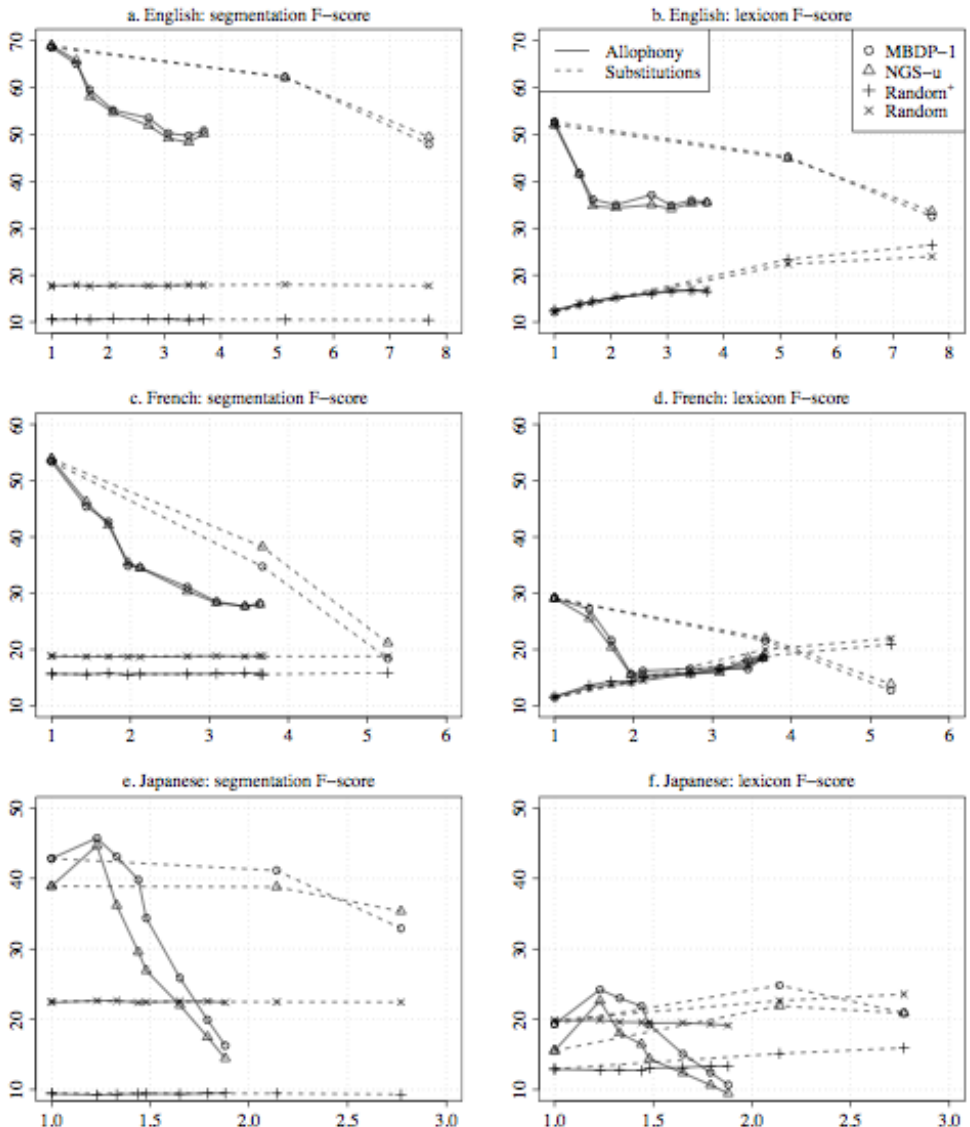


→ Segmentation algorithms assume entire phonemes, not allophones

→Grammar induction algorithms assume entire phonemes and correct segmentation

Problem #2: how bad?





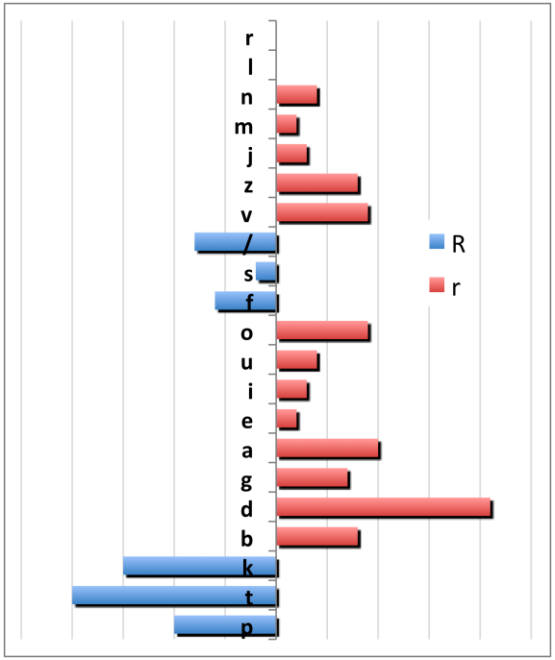
→ Segmentation algorithms assume entire phonemes, not allophones

→Grammar induction algorithms assume entire phonemes and correct segmentation

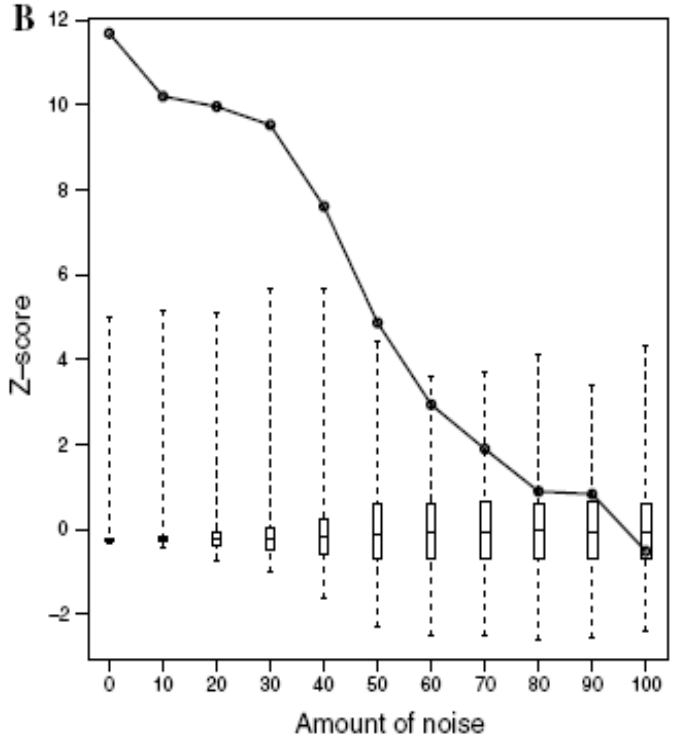
- Three ideas to solve problem #2:
  - Idea #1: complementary distributions
  - Idea #2: linguistic constraints
  - Idea #3: top down information

- Methodology:
  - Phonemically transcribed database
  - Artificial generation of N context-dependant allophones/phoneme

How to reduce the number of allophones? Idea #1: complementary distributions



Simple idea:
→allophonic pairs are in complementary distributions (occur in disjoint contexts)
→ use a measure of divergence in distributions



Pseudo French

KL distance (Kullback-Leibler)

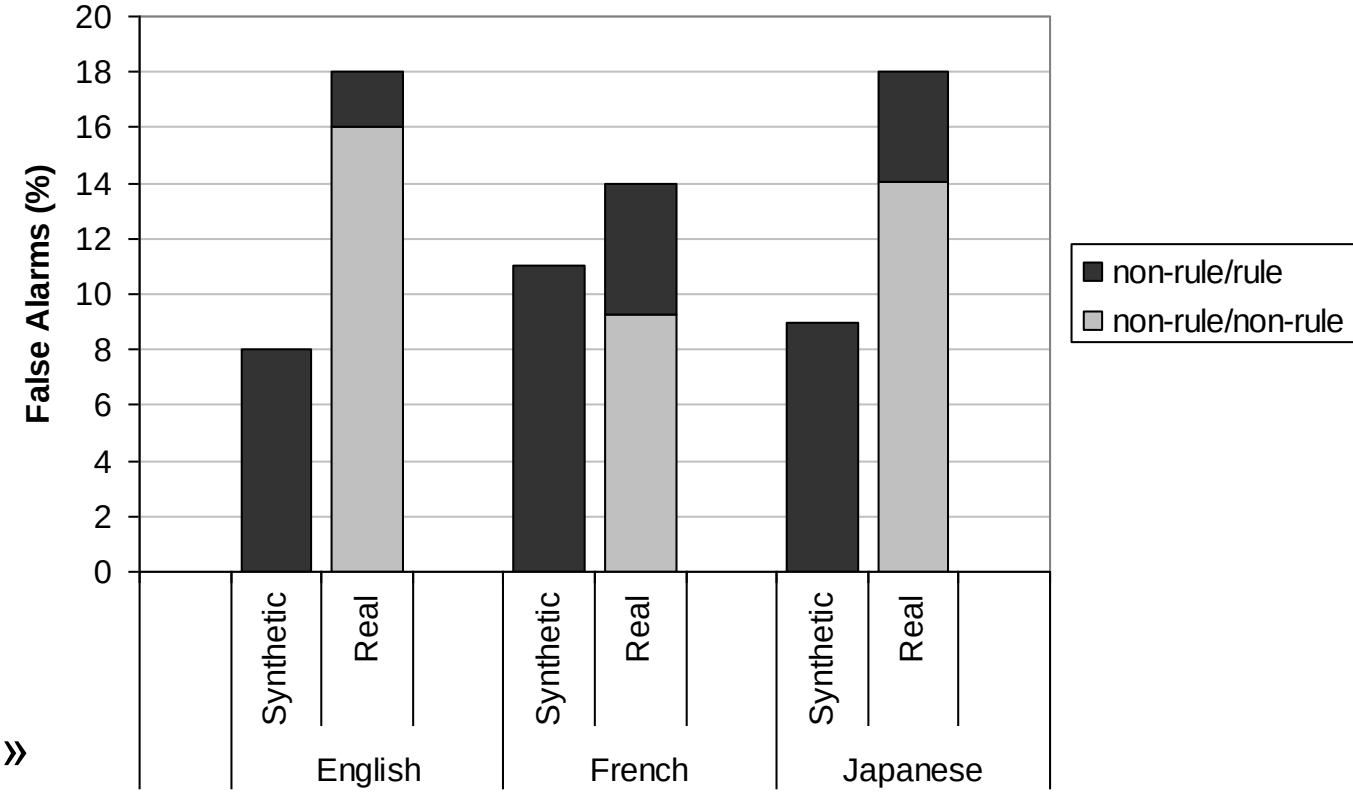
$$m_{KL}(\phi_1, \phi_2) = \sum_c \left(P_1 \ \log\left(\frac{P_1}{P_2}\right) + P_2 \ \log\left(\frac{P_2}{P_1}\right) \right)$$

$$P_1 = P(c|\phi_1), P_2 = \tilde{P}(c|\hat{\phi}_2)$$

Peperkamp, Le Calvez, Nadal, Dupoux, *Cognition*, (2006)

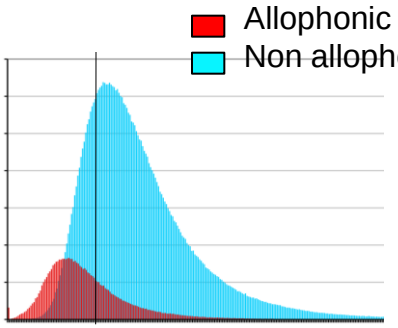
Problem: Effect of phonotactics

- Corpora:
  - Real language (With phonotactics)
  - Synthetic corpus with the same phonemes (no phonotactics)
- Rules:
  - One « allophonic » rule
  - target=most frequent phoneme
  - Context=half of the phonemes

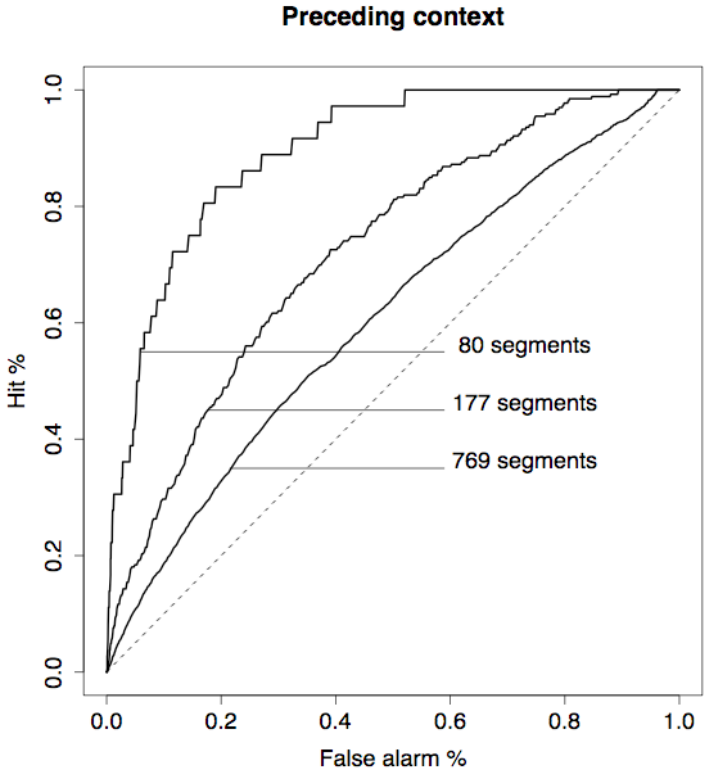


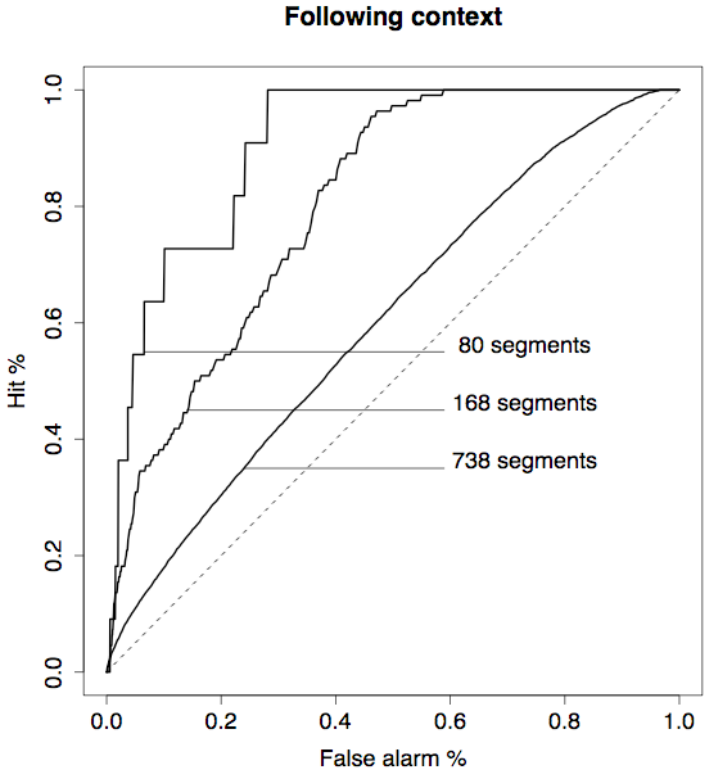
- Test:
  - Z-score threshold of 1
  - Nb of false alarms

Le Calvez (2008)



Problem: effect of nb of allophone





Corpus CSJ (40 hours of spontaneous Japanese)

80 segments: 1 allophone / phoneme
177 segments: 3.4 allophones / phoneme
738 segments: 18 allophones / phoneme

Martin, Peperkamp, Dupoux (submitted)

Limits of KL

- Good:
  - Robust wrt partial application/noisy input
  - Robust wrt rule interaction (not shown)

- Bad:
  - Phonotactics degrades performance

    e.g. in a CV language, every C is in comp distrib with every V
  - Nb of allophones has a catastrophic effect
    - Confusions between allophones (problem of shared contexts). e.g.

      p1 → a1 / _C

      p2 → a2 / _C

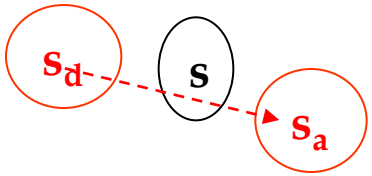
      Then, two other rules are also compatible with the data:

      p1 → a2 / _C

      p2 → a1 / _C

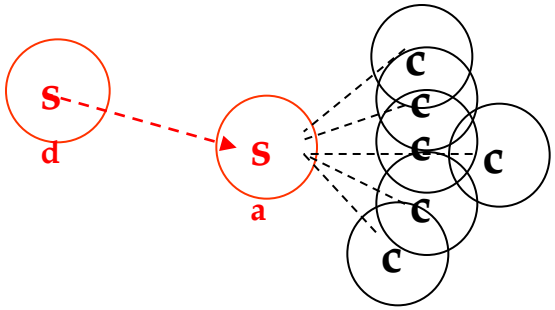
Idea #2. The linguistic/articulatory filters

- allophonic rules generally involve minimal changes



phonetic filter 1: No intermediate segment *allowed between the default segment and allophone*

- allophonic rules are generally assimilatory



phonetic filter 2:*Distance between allophone and context smaller than between default and context*

Peperkamp, Le Calvez, Nadal, Dupoux, *Cognition*, (2006)

Implementation of the filters

- Segments are defined according to phonetic features

Numerical scale along 6 dimensions

Place	from 0 (bilabial) to 8 (uvular)
Sonority	from 0 (voiceless stops) to 12 (low vowels)
Voicing	0 or 1
Nasality	0 or 1
Rounding	0 or 1
Length	0 (simple) or 1 (long vowels and geminates)

filter 1:

$$\exists s \ [\forall i \in \{1, \ldots, 5\}, v_i(s_\mathrm{a}) \leqslant v_i(s) \leqslant v_i(s_\mathrm{d}) \text{ or } v_i(s_\mathrm{d}) \leqslant v_i(s) \leqslant v_i(s_\mathrm{a})]$$

filter 2:

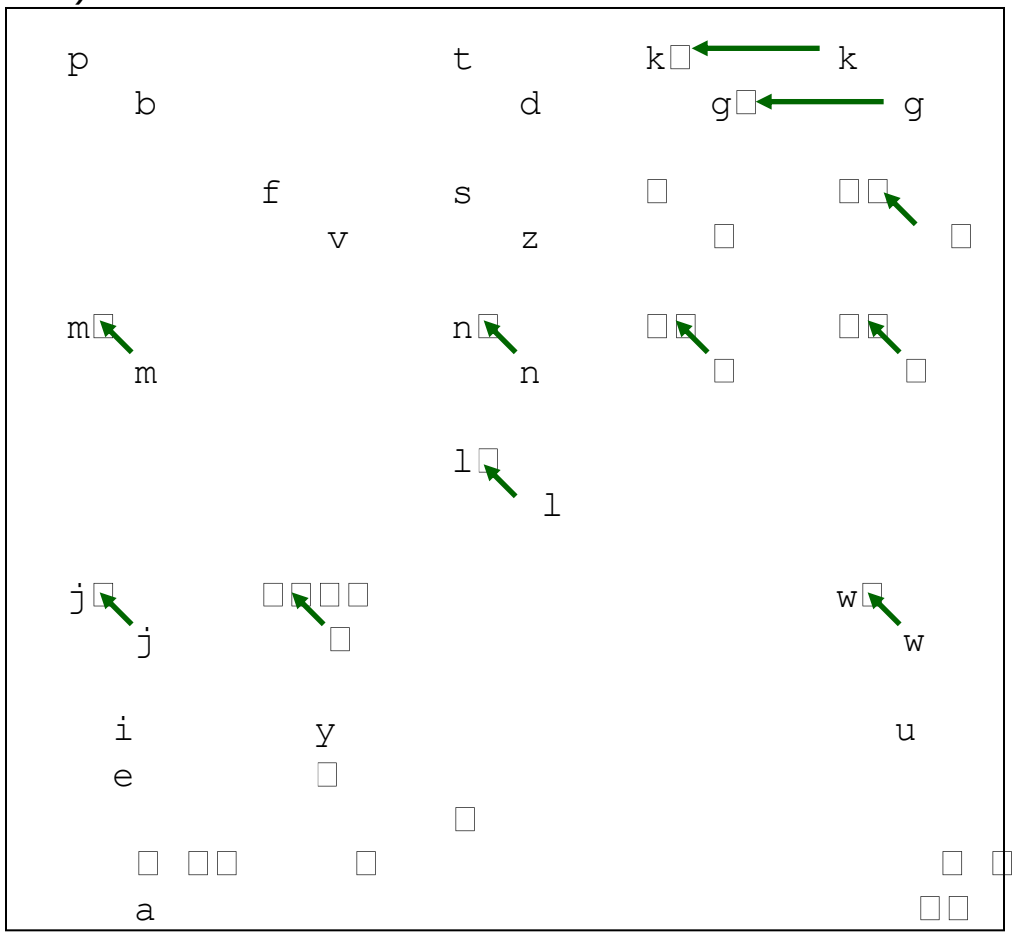
$$\exists i \in \{1, \ldots, 5\}, \left| \sum_{s \in C[s_\mathrm{a}]} (v_i(s_\mathrm{a}) - v_i(s)) \right| > \left| \sum_{s \in C[s_\mathrm{a}]} (v_i(s_\mathrm{d}) - v_i(s)) \right|$$

Tests on French

French Child-directed speech corpus (CHILDES) semi-phonetically transcribed :

≈ 500.000 segments.

- 11 allophonic rules implemented: Palatalisation of /k,g/ (2 rules) and Sonorant Devoicing (9 rules).



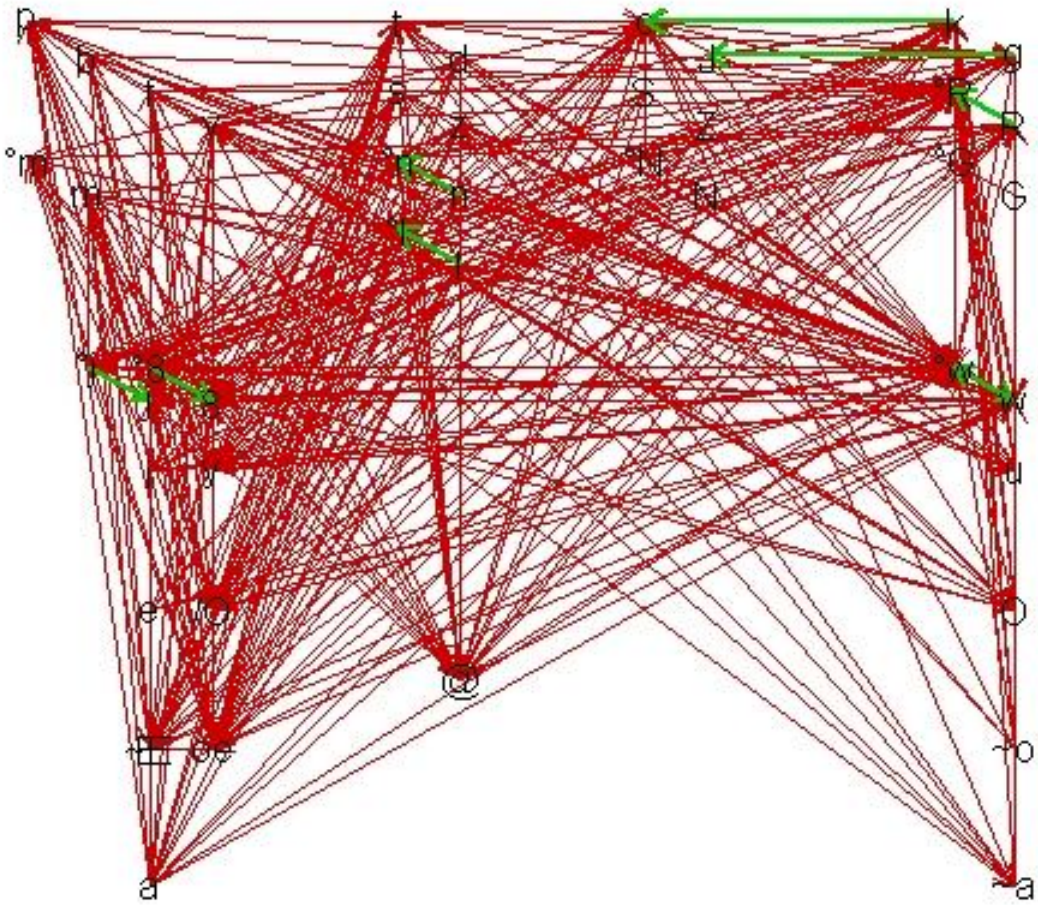
45 segments

2970 pairs of segments

3 contexts (left, right, both)

French

Results: 8 hits out of 11 rules...but 424 False alarms.



→ Hits

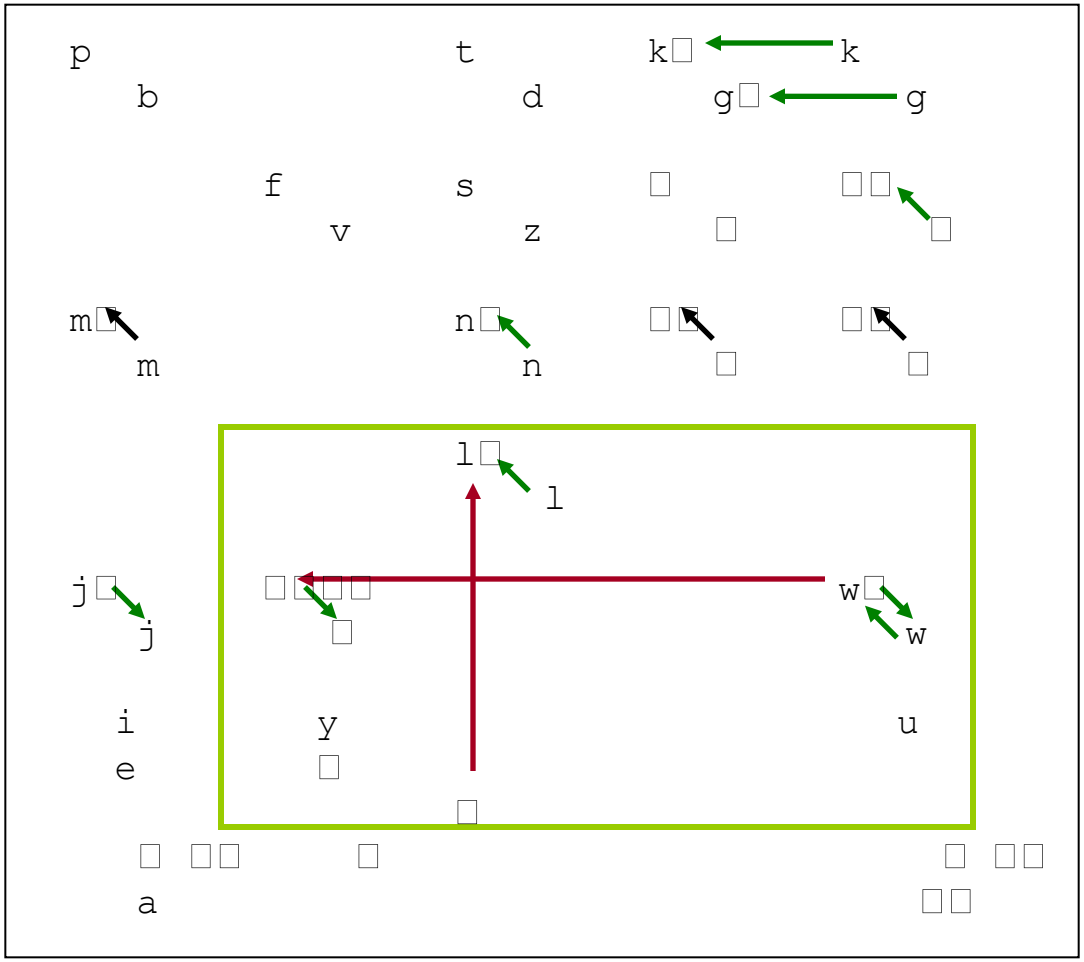
→ False Alarms

=> discard spurious pairs with phonetic filters

French

Before : 8 hits, 424 false alarms.

After: 8 hits, 2 false alarms.



→ Hits
→ False Alarms
→ Misses

Limits of the linguistic/articulatory filters

- How many allophones?
  - The French examples: 11 allophones (i.e. 0.3 per phoneme)
  - State of the art speech recognition:1500 context dependant allophones (i.e., 35 allophones per phoneme)
- Where do features come from?
  - Features are mostly articulatory. Could 12 month olds have access to them?
  - Could we replace linguistic features by acoustic features?

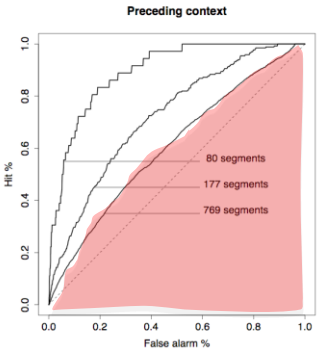
Work in progress

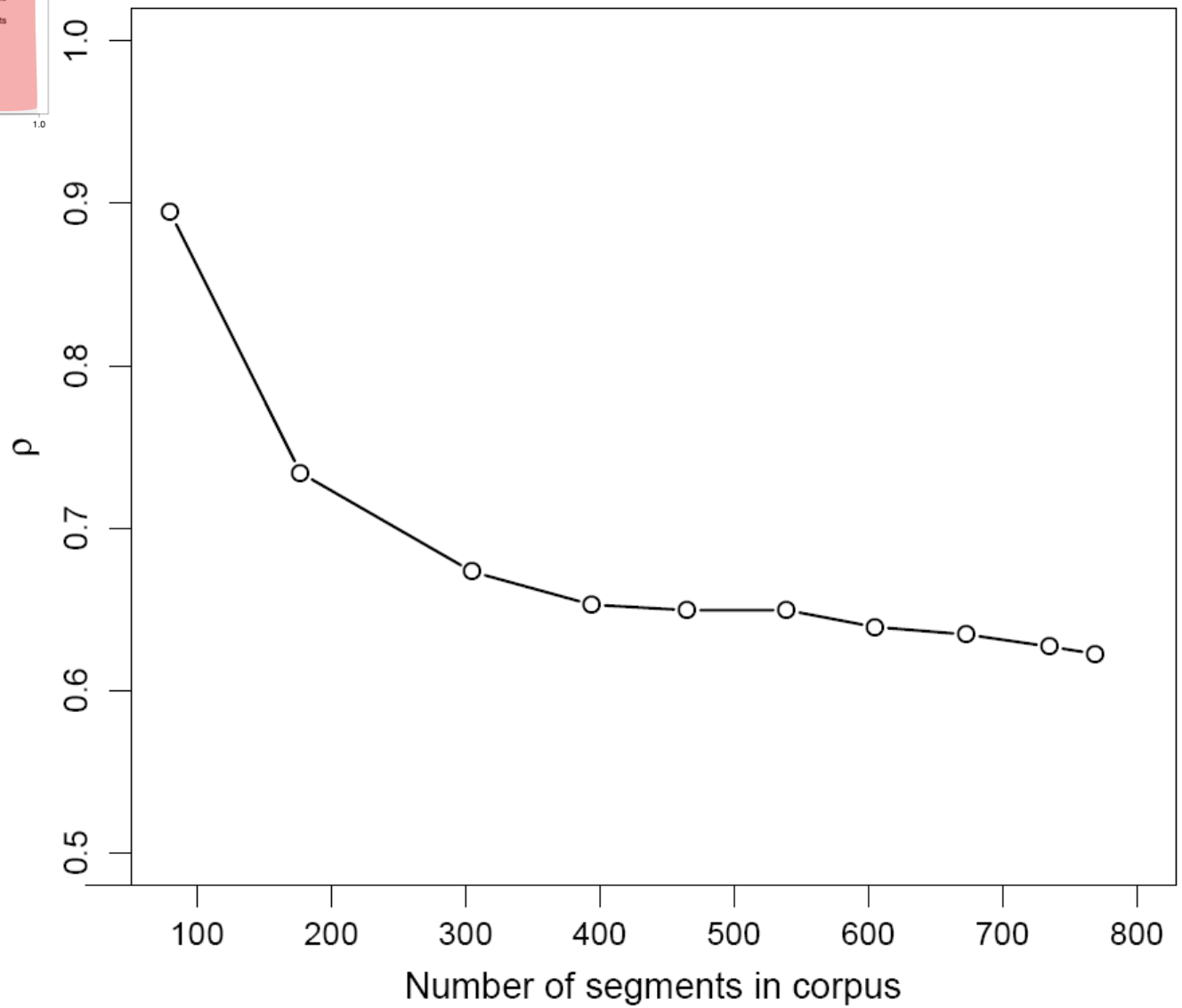
- Use a real speech corpus
  - Replace the linguistic transcriptions with unsupervized labelling
  - Replace the linguistic filters with an acoustic distance or production/coarticulation model in continuous parameter space

Idea #3: use top-down information

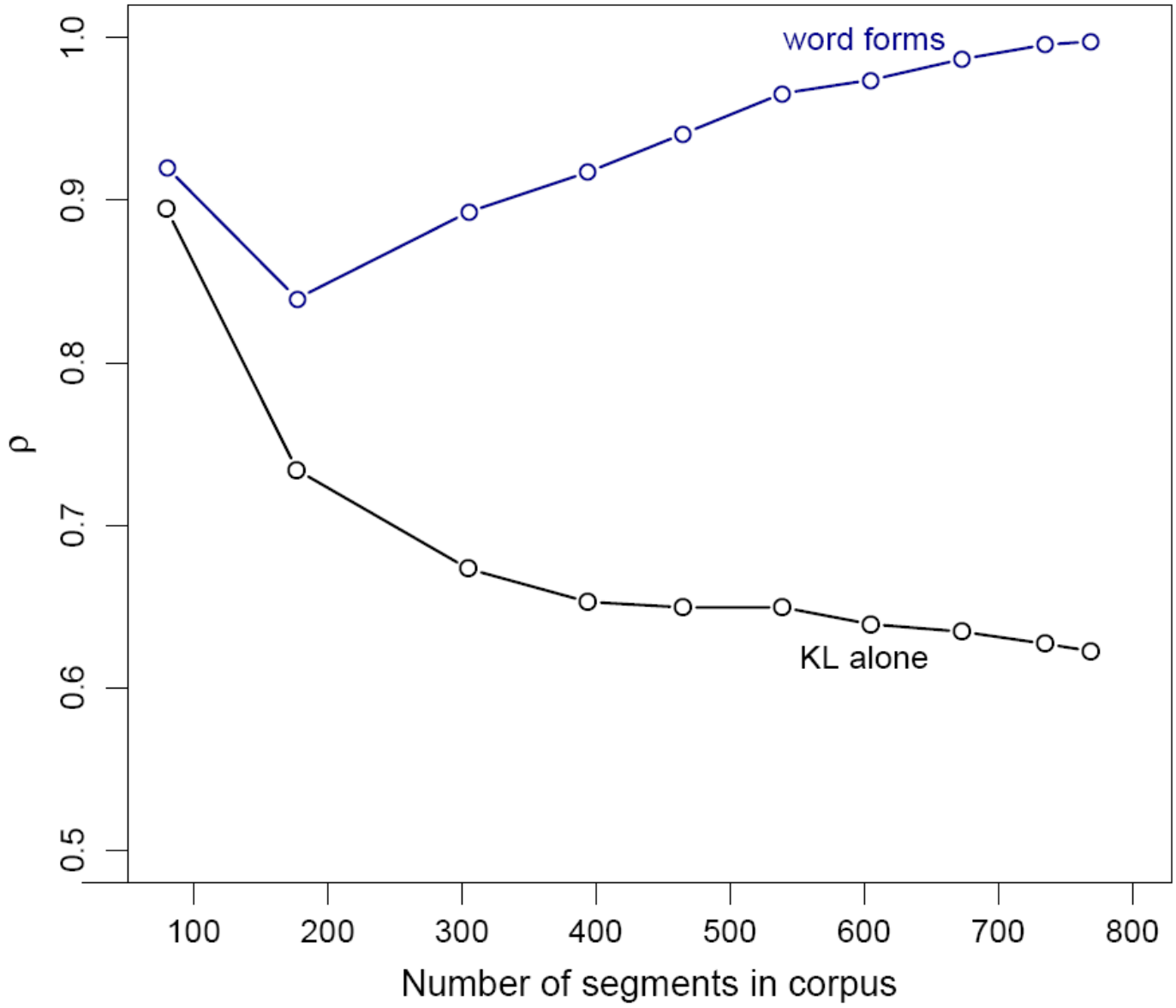
- Intuition:
  - for sufficiently long words, the probability that two different words happen to be identical except for their final segments is very low (example from English: *African* – *affricate*)
  - hence, pairs of word forms differing only in their final segment are likely phonetic variants of the same word

- Word filter on the input:
  - for a given pair of segments {s_1, s_2}, compute KL only if the corpus contains a pair of word forms {Xs_1, Xs_2} where X is a string of segments





Martin, Peperkamp & Dupoux (in preparation)



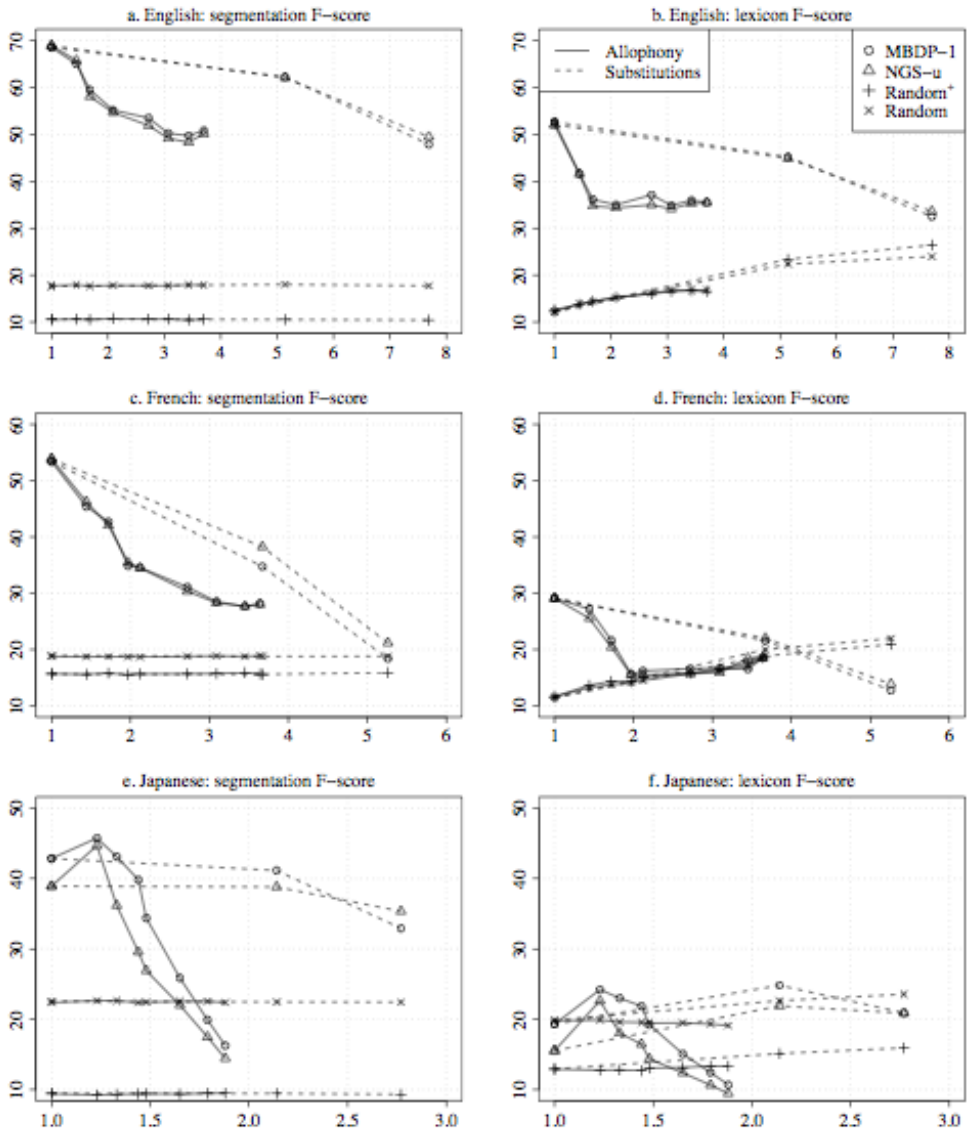
Martin, Peperkamp & Dupoux (in preparation)

But isn’t is cheating?

- Problem:
  - 12-month-olds do not have a large word form lexicon
  - *They probably could not construct one if they have 1500 allophones*

But isn't is cheating?

MBDP-1 Brent (1999).
NGS-u Venkataraman (2001)

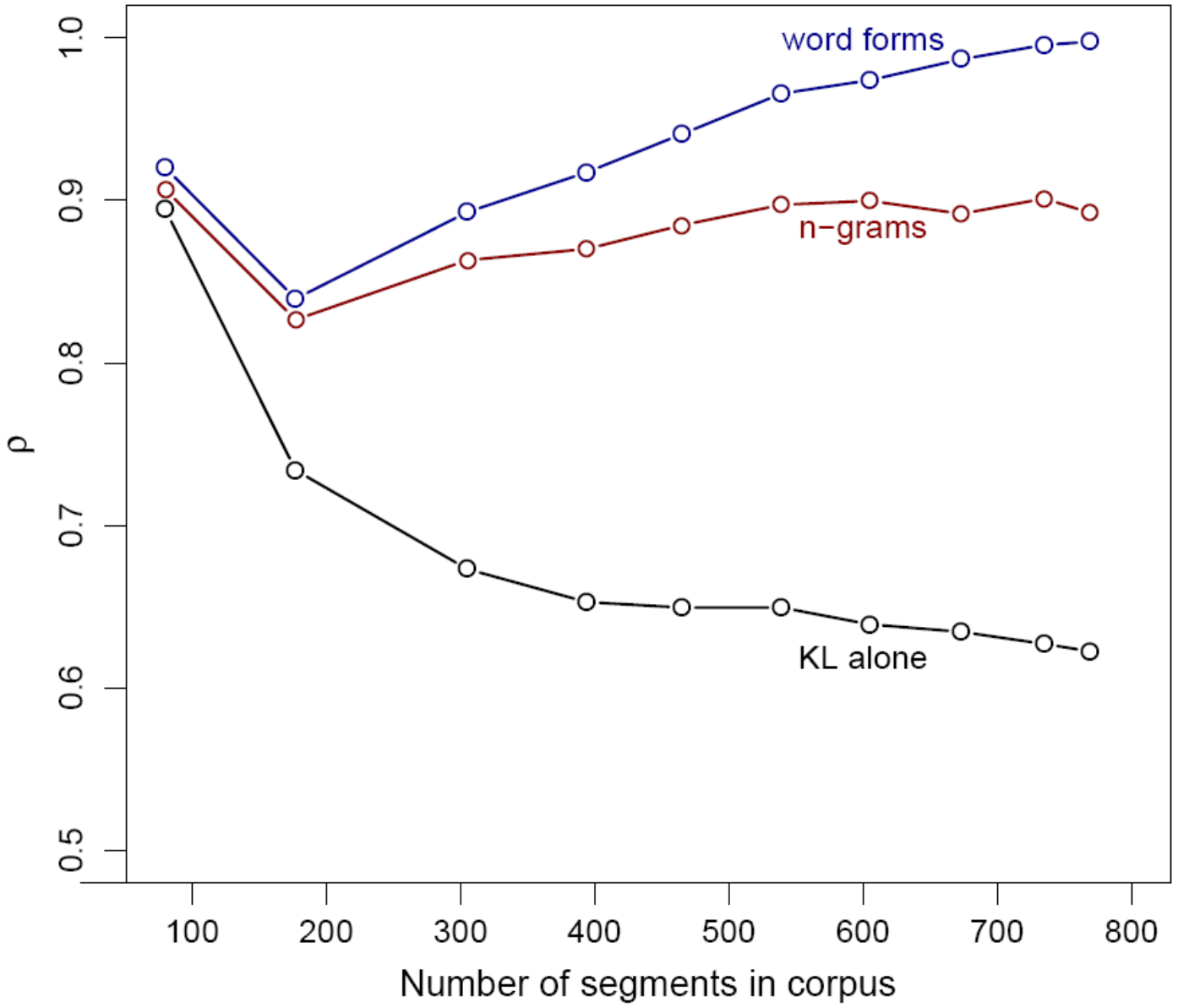


→ *Word segmentation performance degrades with nb of allophones*

Boruta et al, submitted

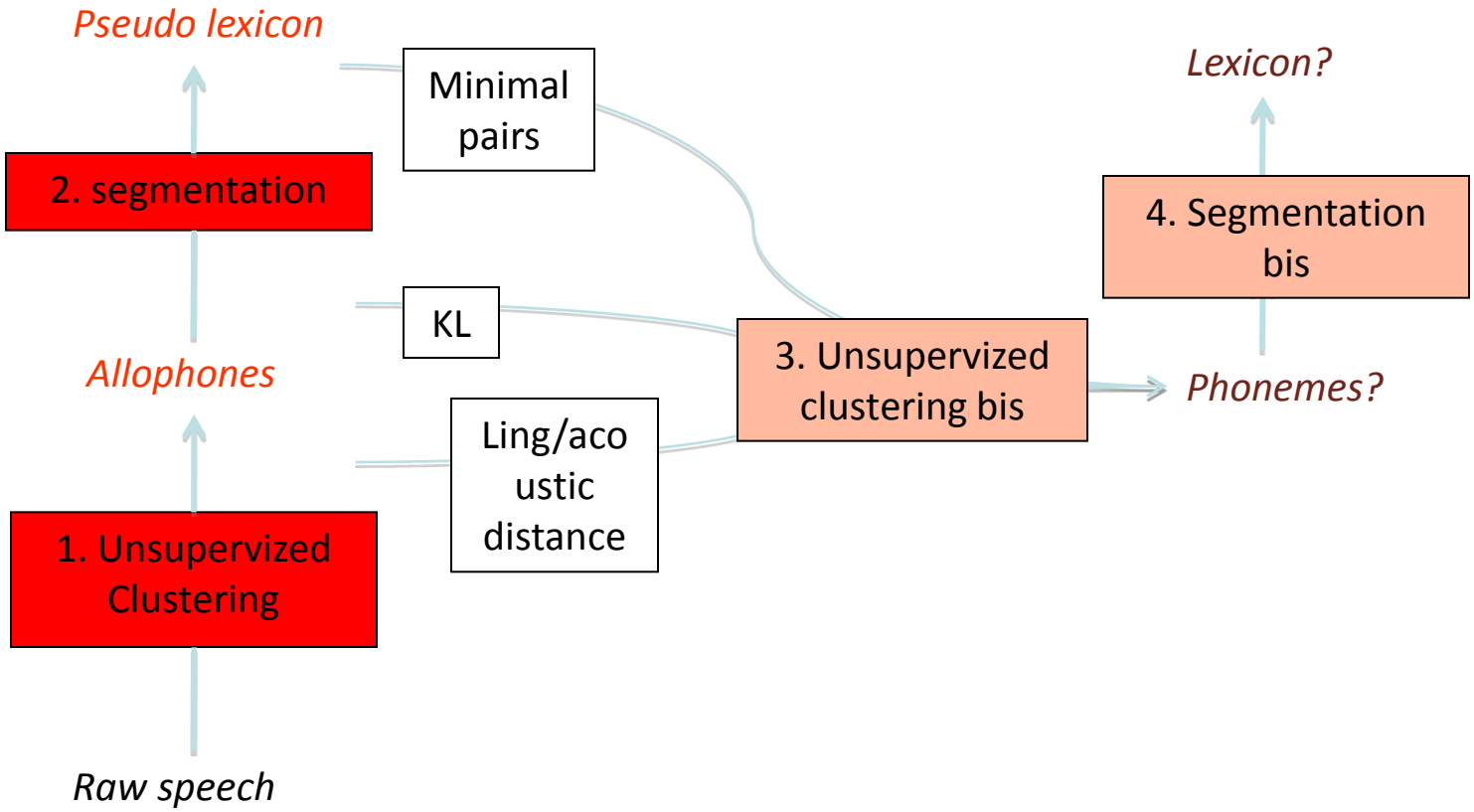
- Solution: use *approximate* word segmentation
  - for a given pair of segments {s_1, s_2}, compute KL only if the corpus contains a pair of frequent *n*-grams {Xs_1, Xs_2} where *X* is a string of segments of length n-1.
    - *n*=7
    - frequency cut-off: 10%

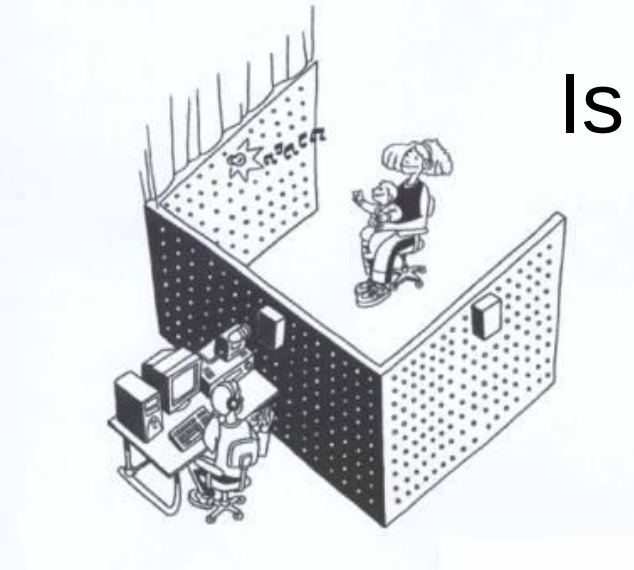
      (a 7-gram is frequent if its frequency rank is within the top 10%)



Martin, Peperkamp & Dupoux (in preparation)

The REVISED Sequential Bootstrapping Scenario

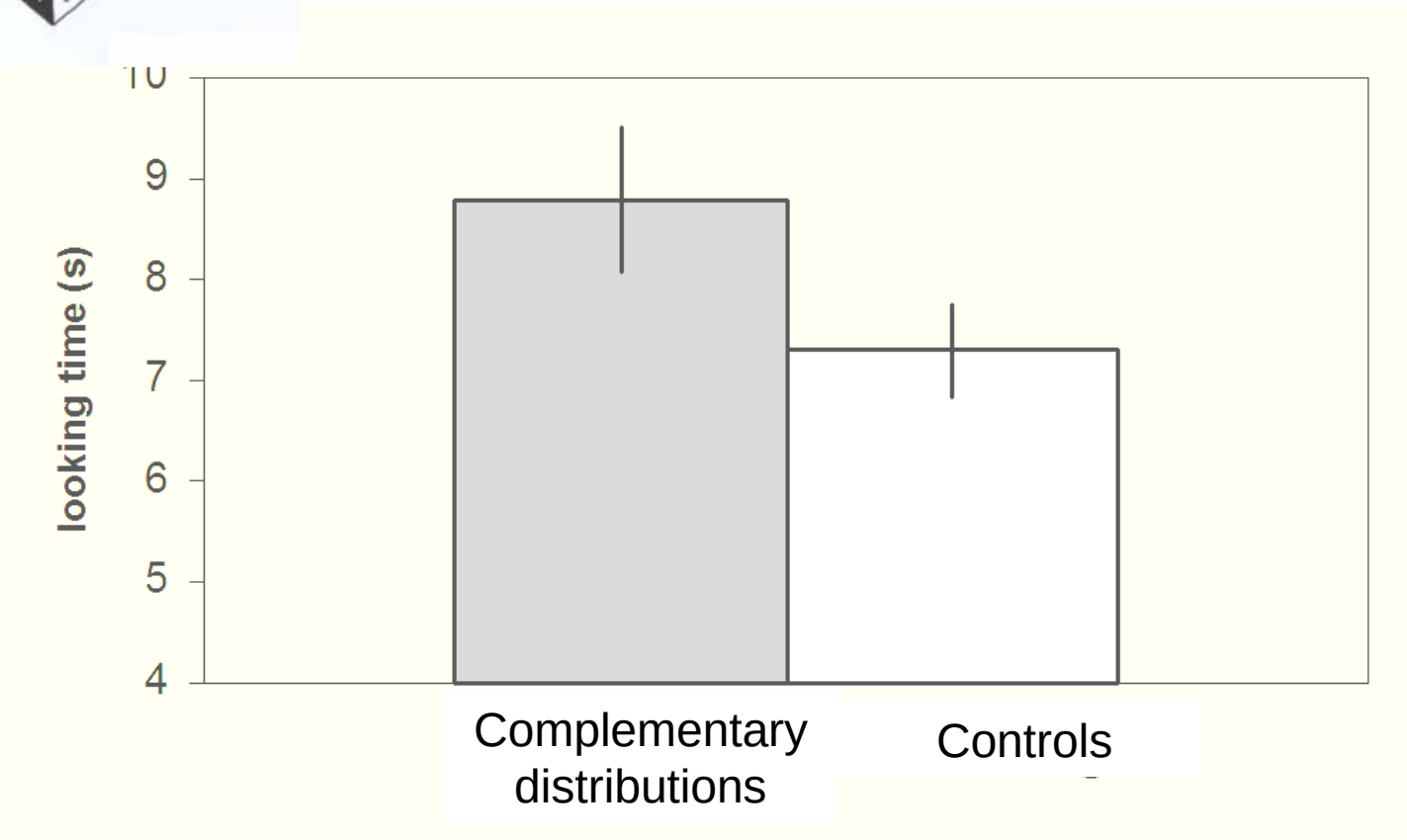




Is this psychologically plausible?

Complementary distributions

*



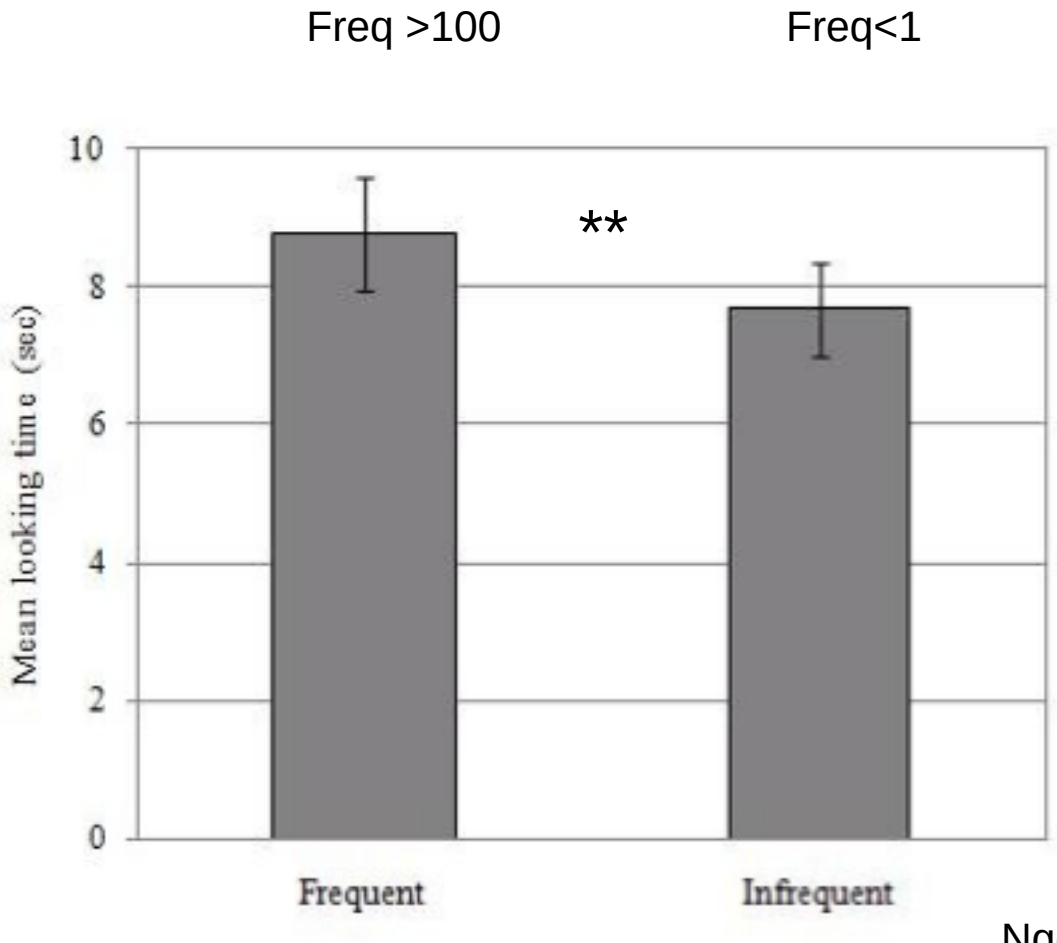
12 month old American infants

White, Peperkamp, Kirk & Morgan (2008)

n-gram pseudo-lexicon

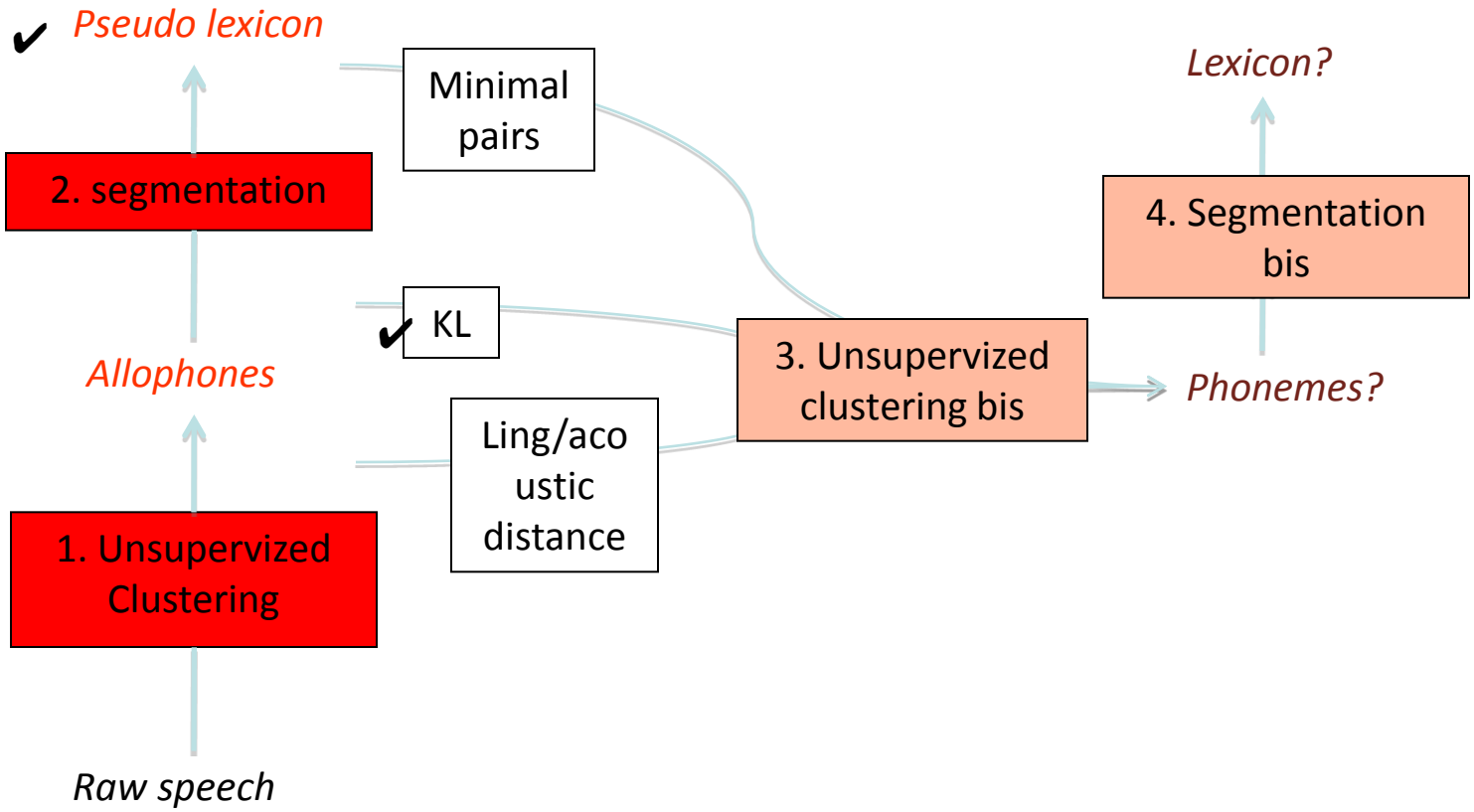
- Bisyllables (4-5 phonemes)
- C initial, Syllabic structure matched pairwise
- Possible but not real words
- Matched mean diphone frequency

•16 French 11-month-olds

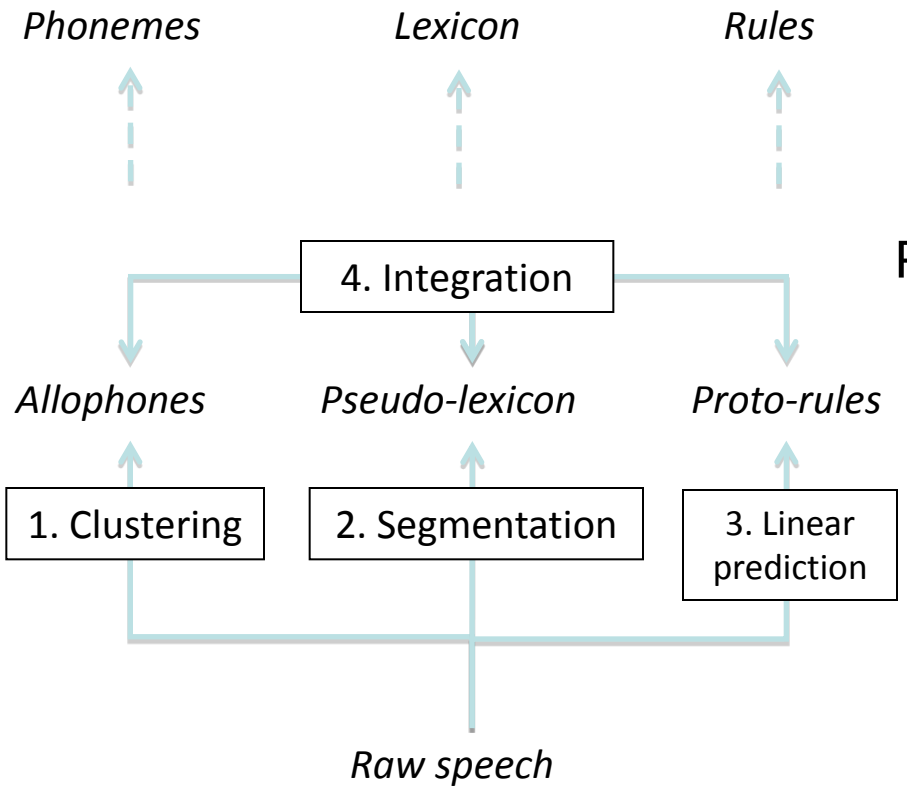


Ngon, in prep

The REVISED Sequential Bootstrapping Scenario

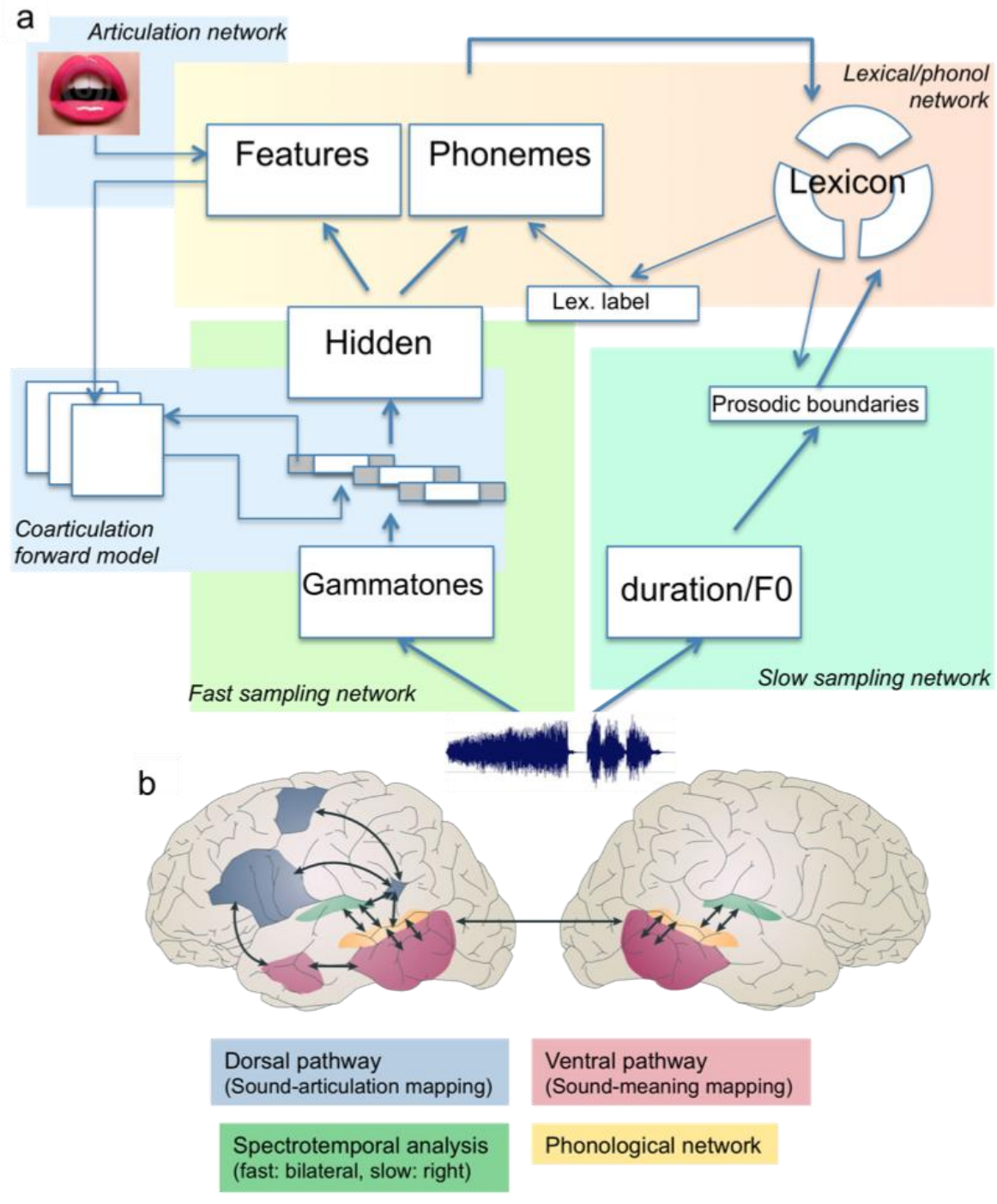


The parallel-integrative Bootstrapping scenario



Further questions

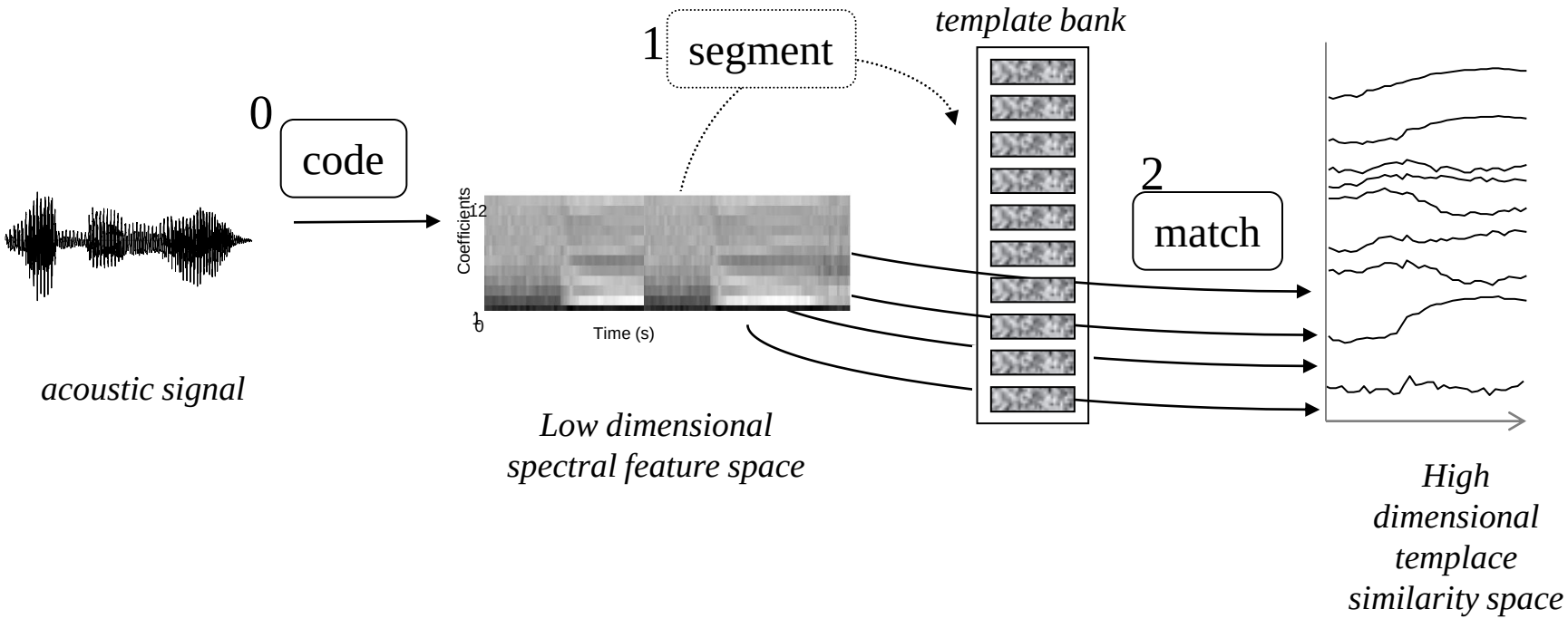
- What is the performance in a re size simulation?
- Is this linguistically plausible (what about other kinds of variations: free variations, insertions deletions, ..)
- Is this psychologically plausible? (do infants do it?)
- Is this neurally plausible?


a
Articulation network
Lexical/phonol network
Features
Phonemes
Lexicon
Lex. label
Hidden
Prosodic boundaries
Coarticulation forward model
Gammatones
duration/F0
Fast sampling network
Slow sampling network
b
Dorsal pathway (Sound-articulation mapping)
Ventral pathway (Sound-meaning mapping)
Spectrotemporal analysis (fast: bilateral, slow: right)
Phonological network

Sharon Peperkamp
Jean-Pierre Nadal
Luc Boruta
Sanjeev Kudanpur
Thomas Schatz
Andrew Martin
Isabelle Dautriche
Balakrishnan Varadarajan

Thank you

Coarse grained coding

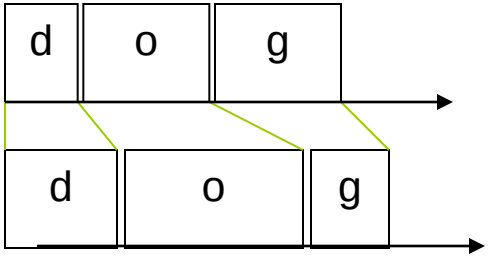


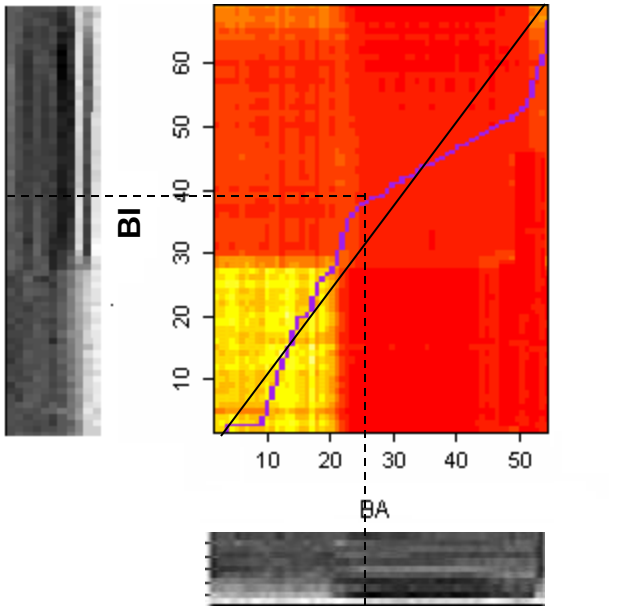
0 : continuous speech is *recoded* into spectral features

1 : incoming signal is *segmented* into syllabic-like templates (based on acoustic sonority). The templates are stored in an instance-based perceptual memory

2 : continuous speech is *matched* to all of the templates, using a running DTW algorithm. This yields simularity profile as a function of time for each stored template.

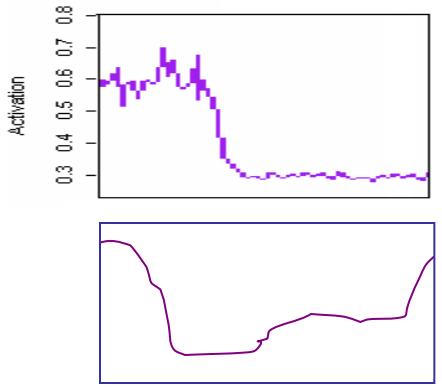
Step 2: Matching



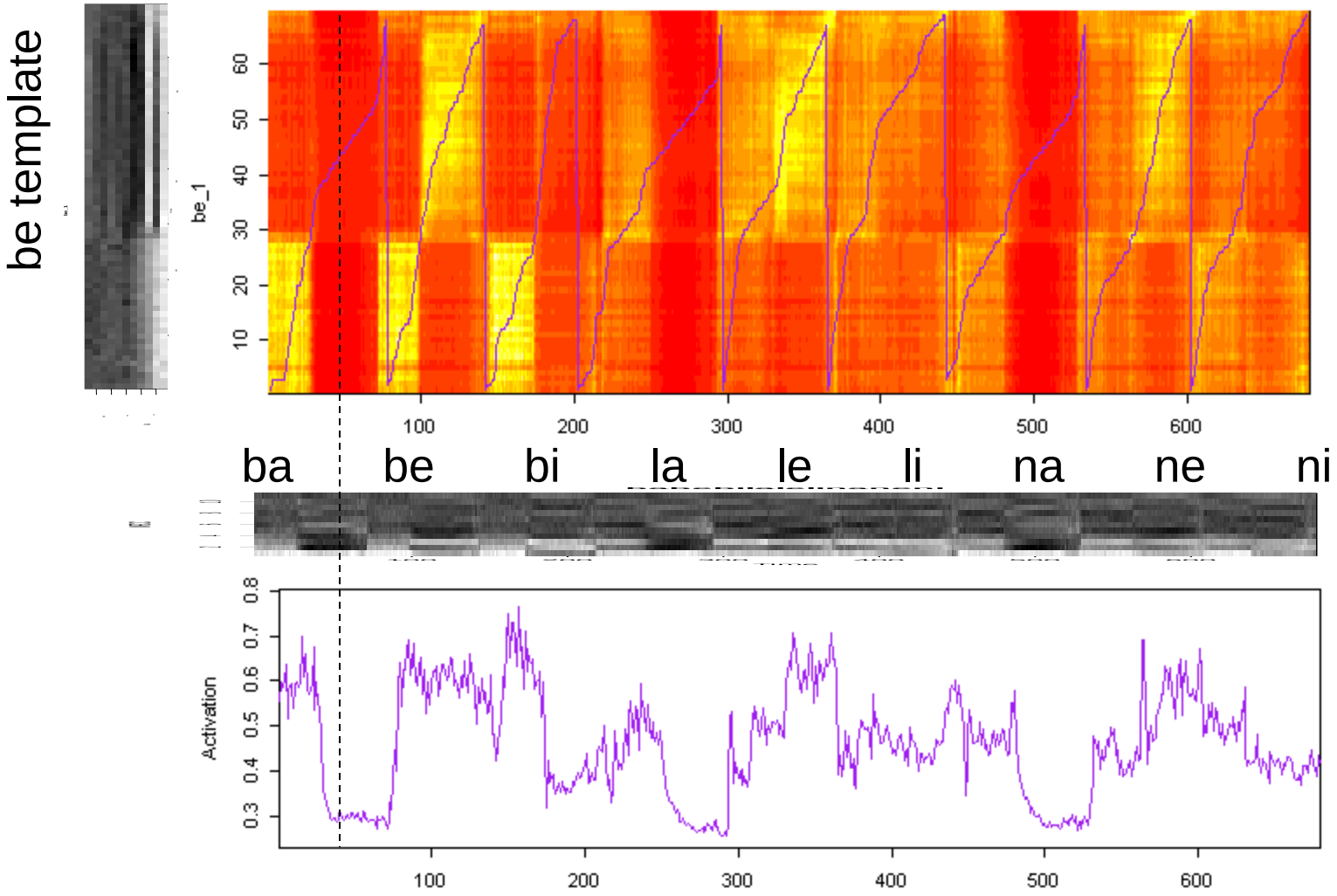


- align the input stimuli with each stored template using dynamic time warping

- generate a spectral match signal as a function of time *Spectral Similarity*=1/(1+d)



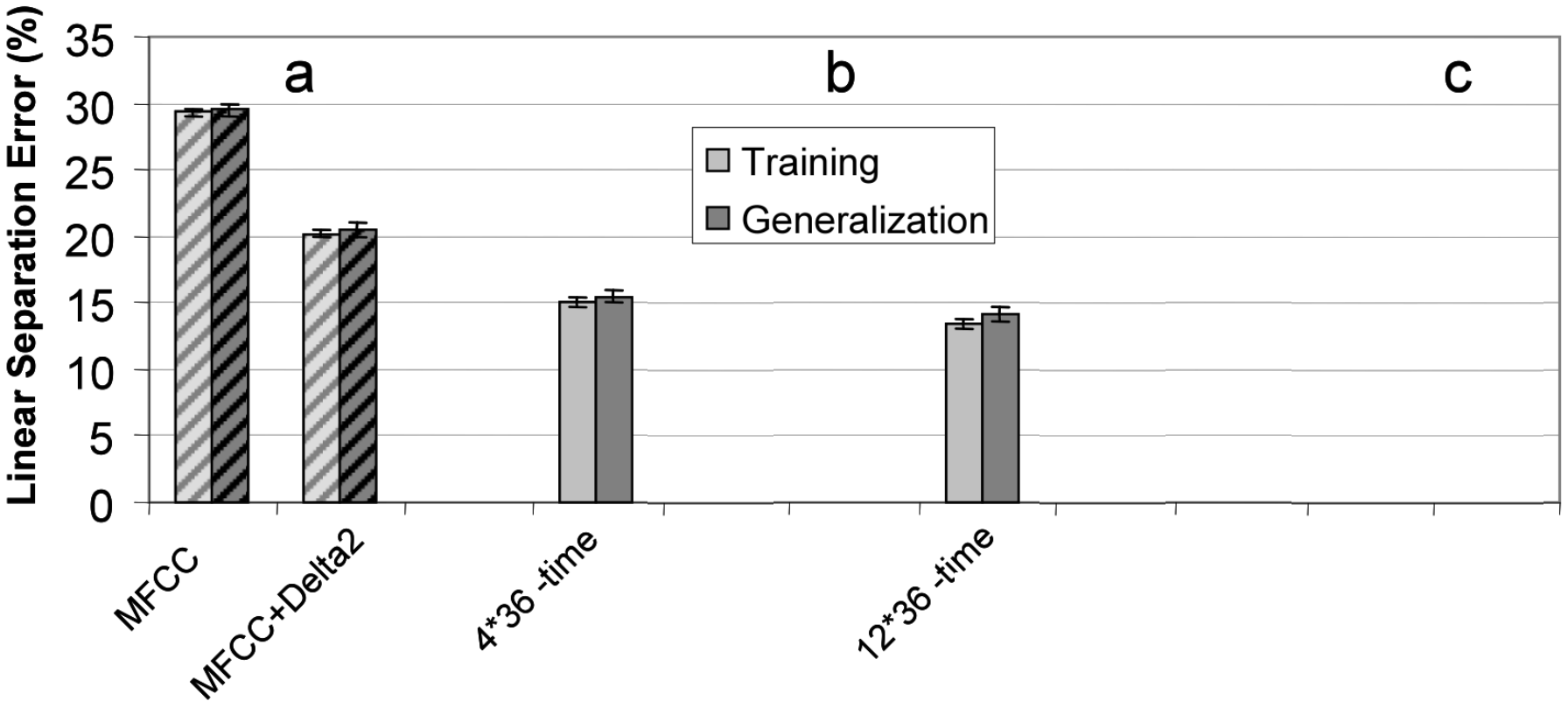
- generate a temporal match signal as a function of time (*Temporal Similarity* = |f ’-1|)

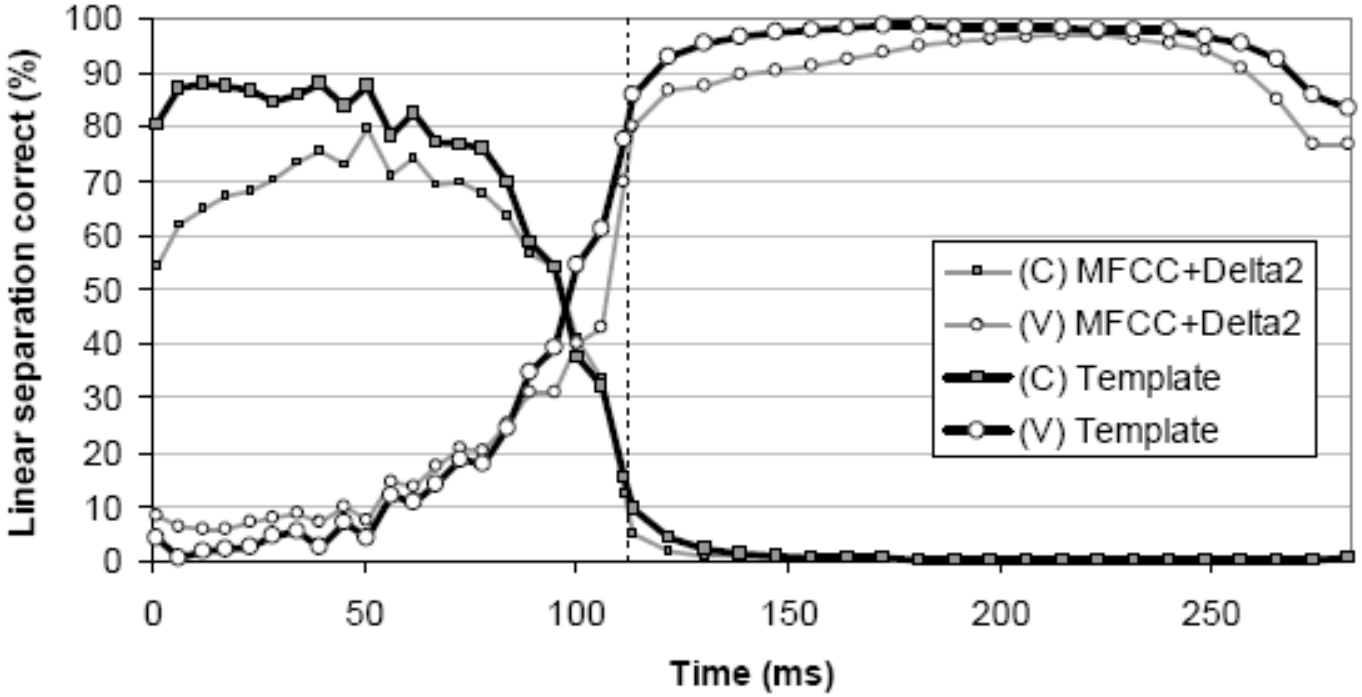

be template
be_1
10
20
30
40
50
60
100
200
300
400
500
600
ba
be
bi
la
le
li
na
ne
ni
Activation
0.3
0.4
0.5
0.6
0.7
0.8
100
200
300
400
500
600

Tests on pseudolanguages

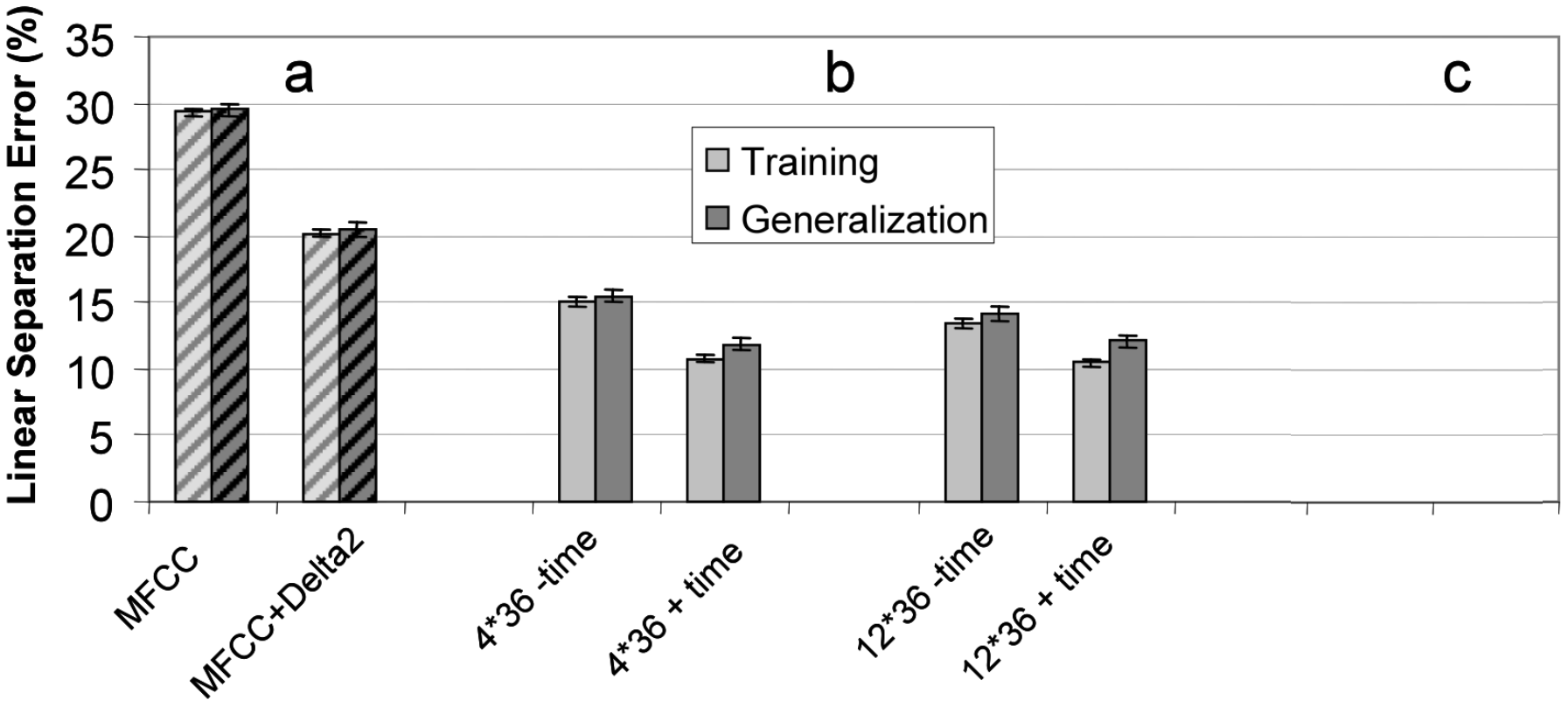
- *Monosyllabic language*
  - *36 syllables, 12 phonemes: /p t k s R m a e i o u y/*
  - *recorded 55 times by a single talker*
- *Polysyllabic language*
  - *16 syllables, 8 phonemes, /m d r / a u i e /*
  - *CV, CVCV, CVCVCV*
  - *recorded 40 times by a single talker*

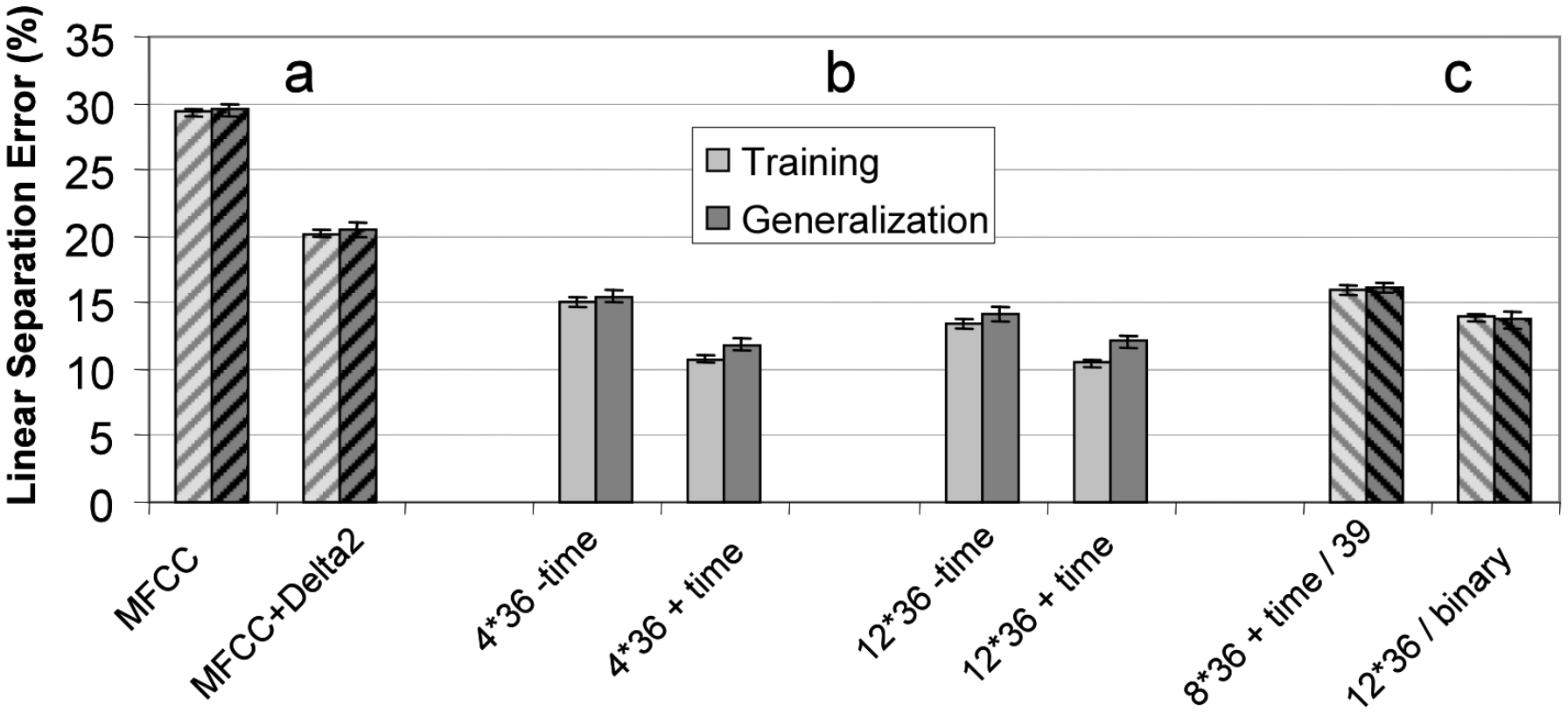
Q1: is the templatic code better than MFCC?




Linear separation correct (%)
100
90
80
70
60
50
40
30
20
10
0
0
50
100
150
200
250
Time (ms)
(C) MFCC+Delta2
(V) MFCC+Delta2
(C) Template
(V) Template

Q2: does the temporal information matter?



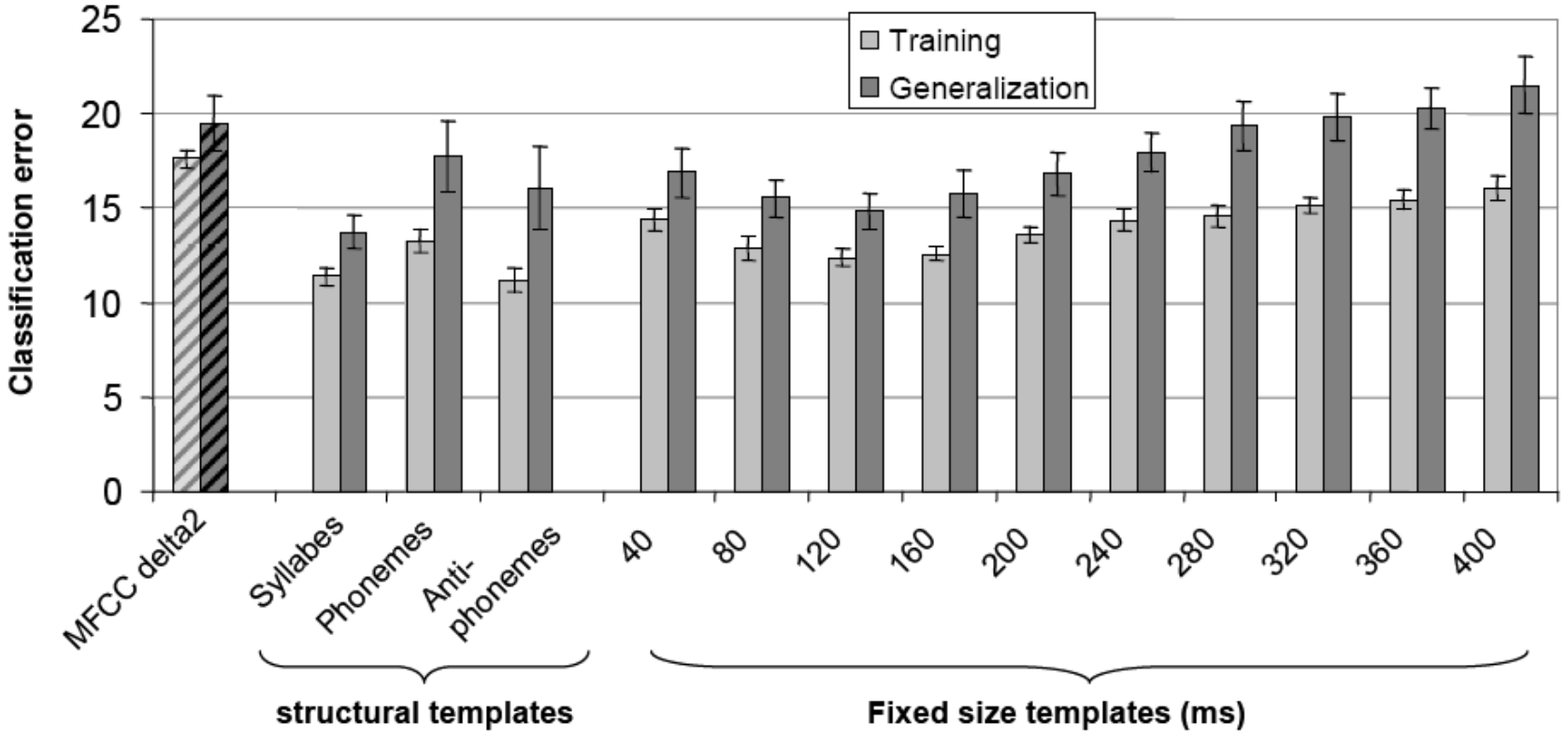

Linear Separation Error (%)
35
30
25
20
15
10
5
0
a
b
c
Training
Generalization
MFCC
MFCC+Delta2
4*36 -time
4*36 + time
12*36 -time
12*36 + time
8*36 + time / 39
12*36 / binary

Q3: is the templatic code language specific?

- « easy » monosyllabic language
  - 9 syllables, 6 phonemes: /s R m a e i/
- *« hard » monosyllabic language*
  - 9 syllables, 6 phonemes: /p t k u y o/

Code	Easy Language		Hard Language	
Baseline	Training	General.	Training	General.
MFCC + Delta 2	8.1% (0.3)	8.9% (0.7)	11.8% (0.3)	12.9% (0.7)
Appropriate Templates	4.8% (0.2)	5.2% (0.5)	8.2% (0.5)	10.2% (0.7)
Inappropriate Templates	8.0% (0.2)	9.4% (1.0)	14.0% (0.7)	16.6% (1.0)

Q4: what is the best code?



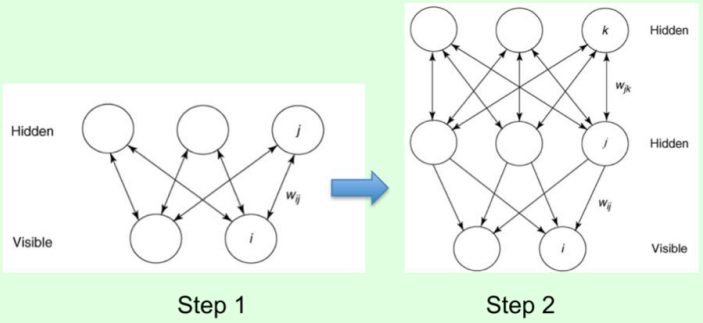
Conclusions

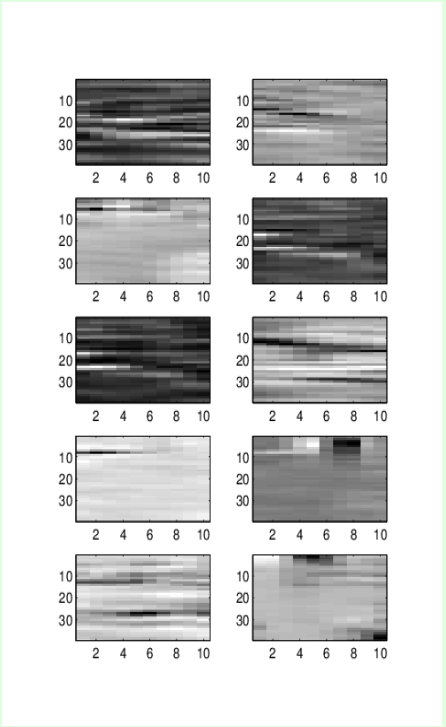
- Q1: Templatic representations are useful
  - Généralization improves (40-45% less errors than MFCC)
- Q2: Temporal similarity clearly a plus (even in a language with no length contrast)
- Q3: Templatic representations are attuned to the ‘native language’
- Q4: The best templates are not necessarily linguistically defined

Limits

- Proof of concept
  - Language
    - Read speech, mini language → large sample of conversational speech
  - Algorithm
    - Linear separation → Unsupervized clustering
    - Abrupt separation between template storing and exploitation→ incremental algorithms
  - Architecture
    - Deep Belief Networks, Sparse Dictionnaries

Deep Belief Network
Sparse Dictionnaries



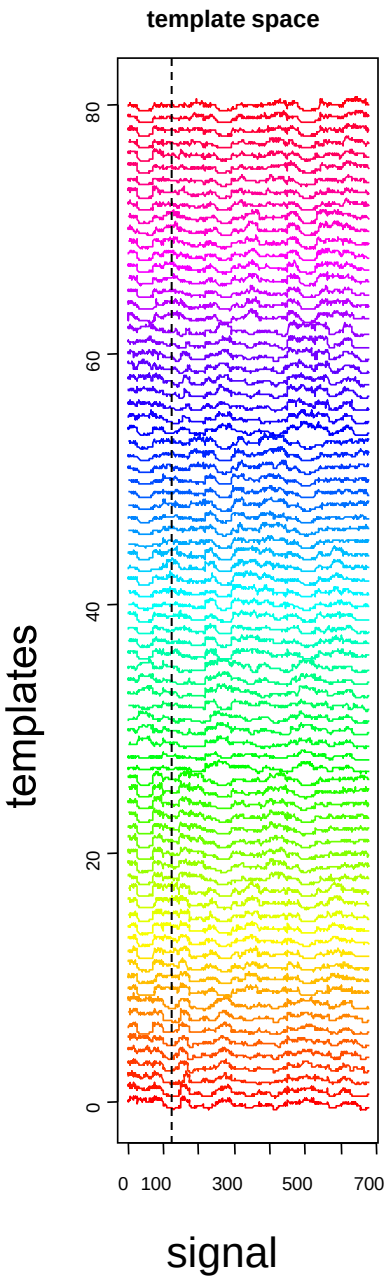


Hinton 2007; Schatz et al, in prep

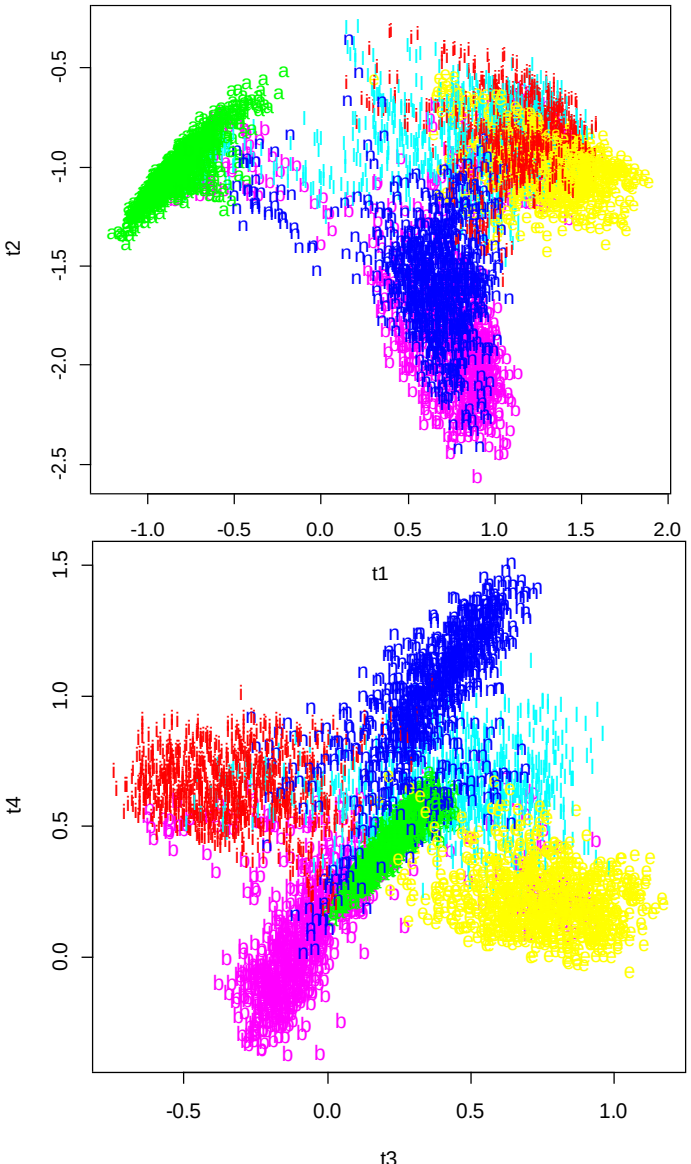
Sharon Peperkamp
Jean-Pierre Nadal
Luc Boruta
Sanjeev Kudanpur
Thomas Schatz
Andrew Martin
Isabelle Dautriche
Balakrishnan Varadarajan

Thank you

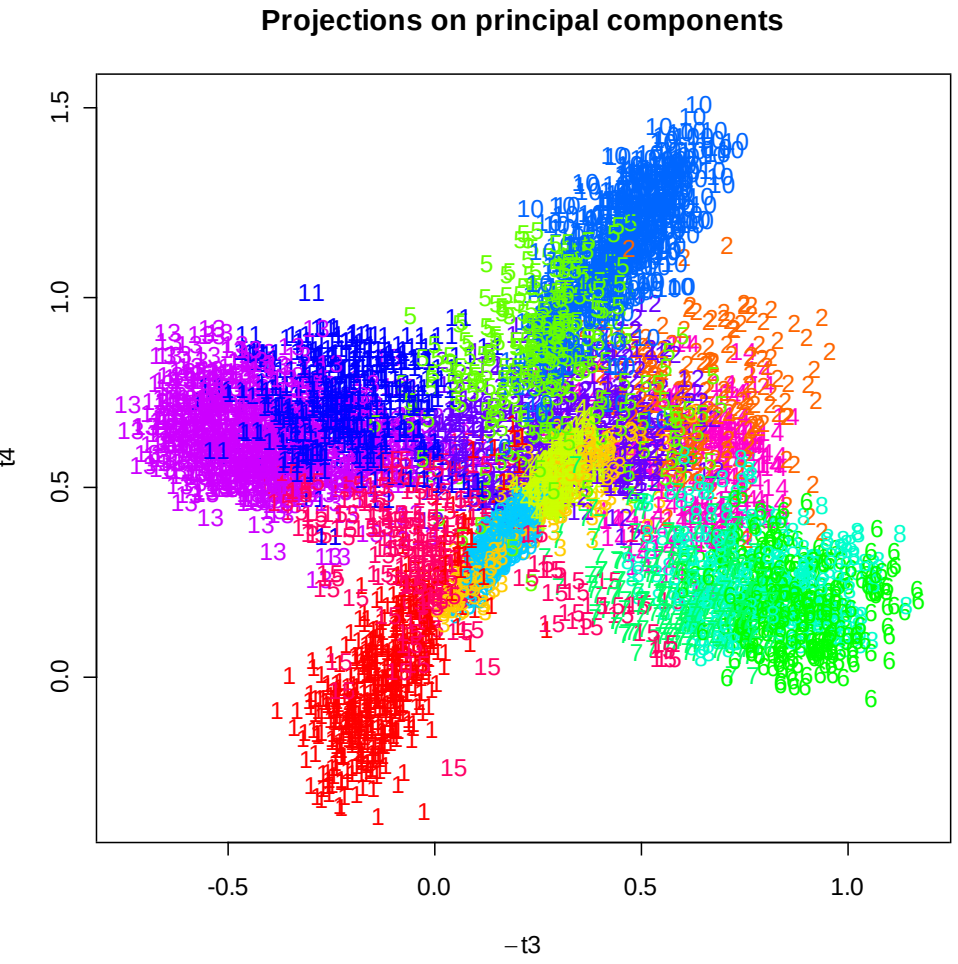
Patterns in template space

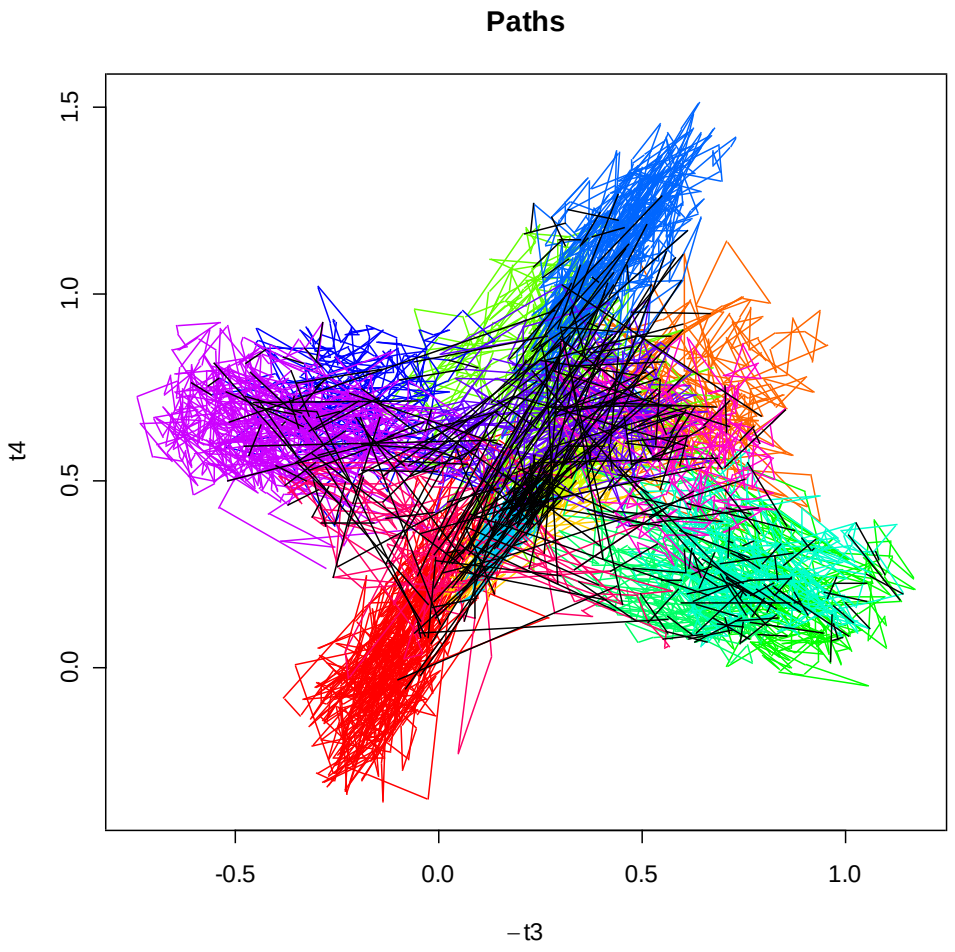


PCA (first 4 components)



Step 1a: initial overclustering

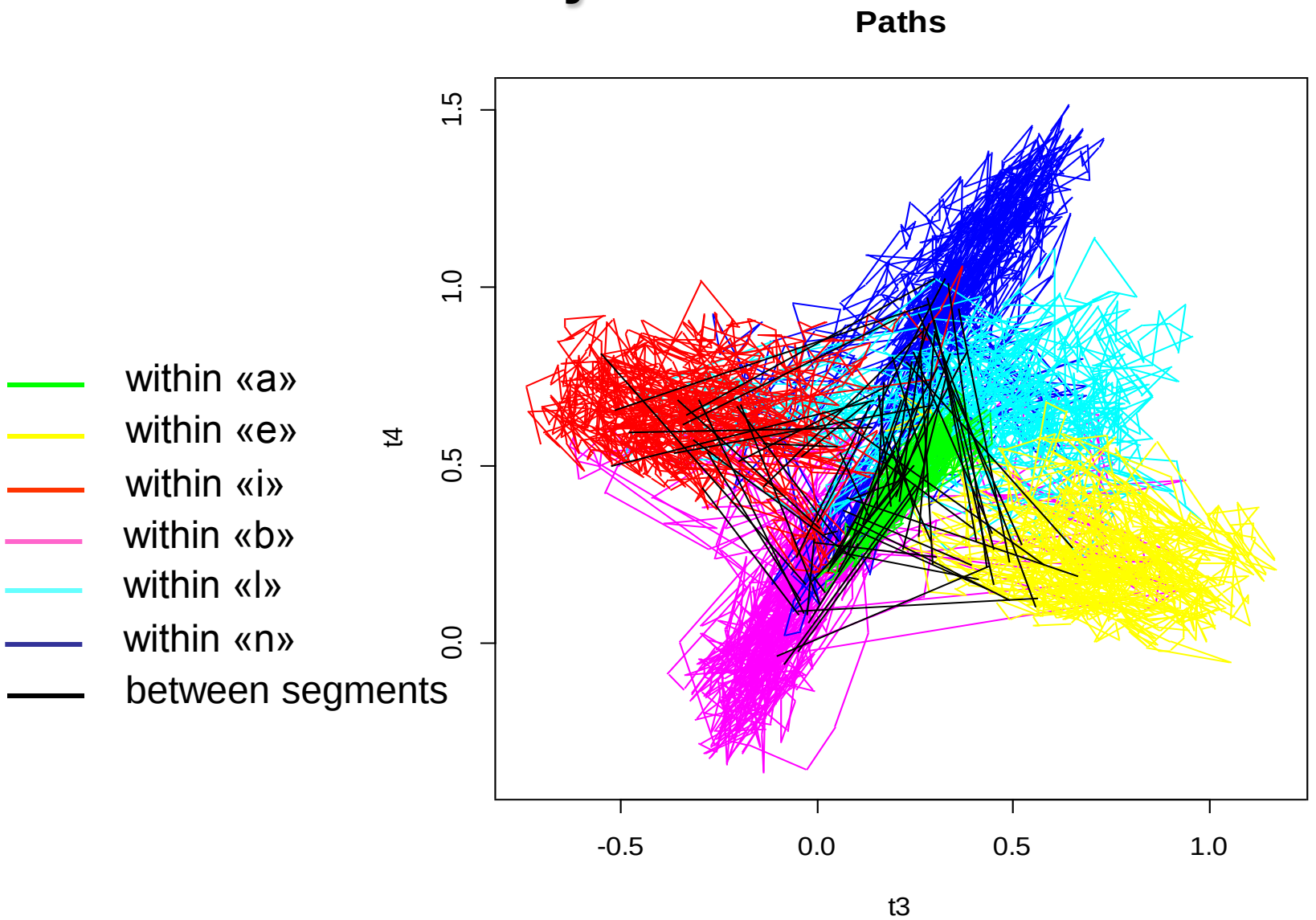




observations:

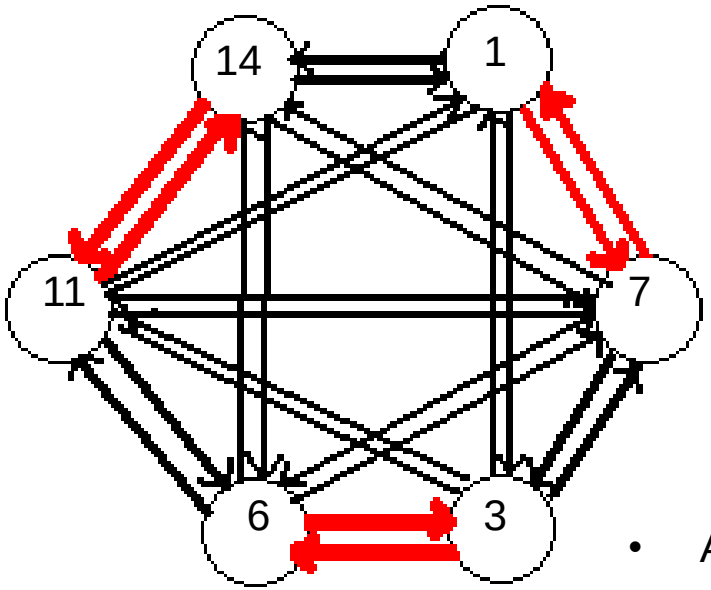
the clusters respect segments boundaries;
they create variants and context sensitive allophones
they trigger uneven between-cluster transitions

Trajectories:



Observation: *lots of local (within cluster) transitions, few long distance (between cluster) transitions*

Clusters as graphs

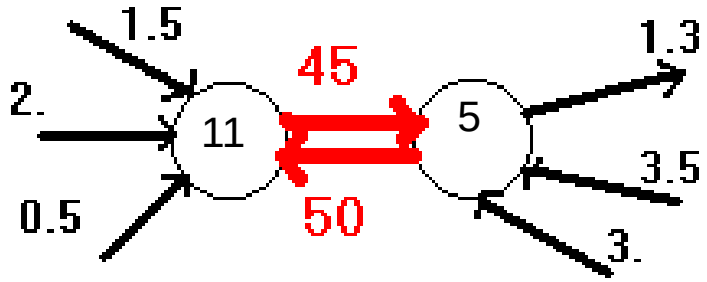


- nodes= clusters
- arrêtes=transition probabilities

- Aim of algorithm:
  - minimize sum of squares (between centers of clusters and examplars): *try to make lots of compact clusters*
  - minimize mutual information: *try to make an isotropic graph*

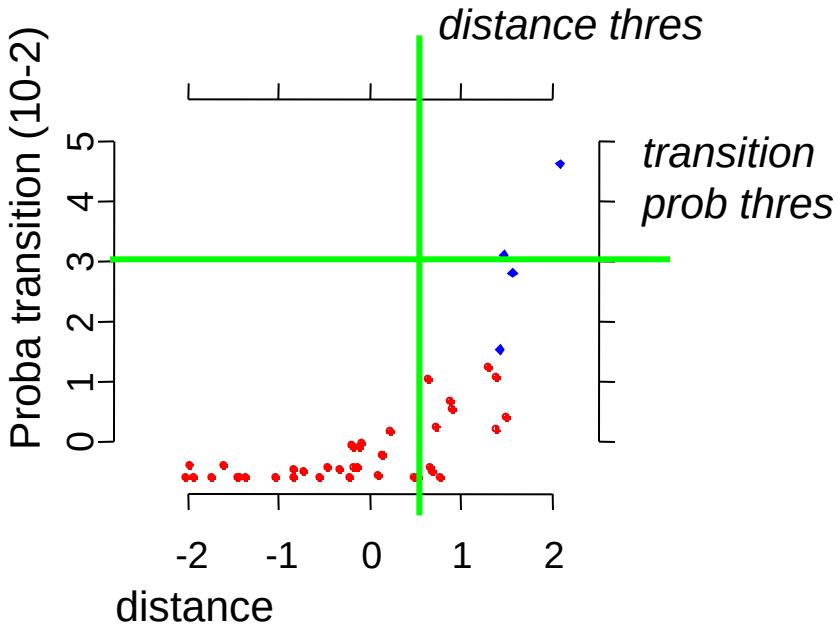
Step 1b: « Alternations » deletion

- Alternating sequences:
  11.5.11.5.11.5.11.5



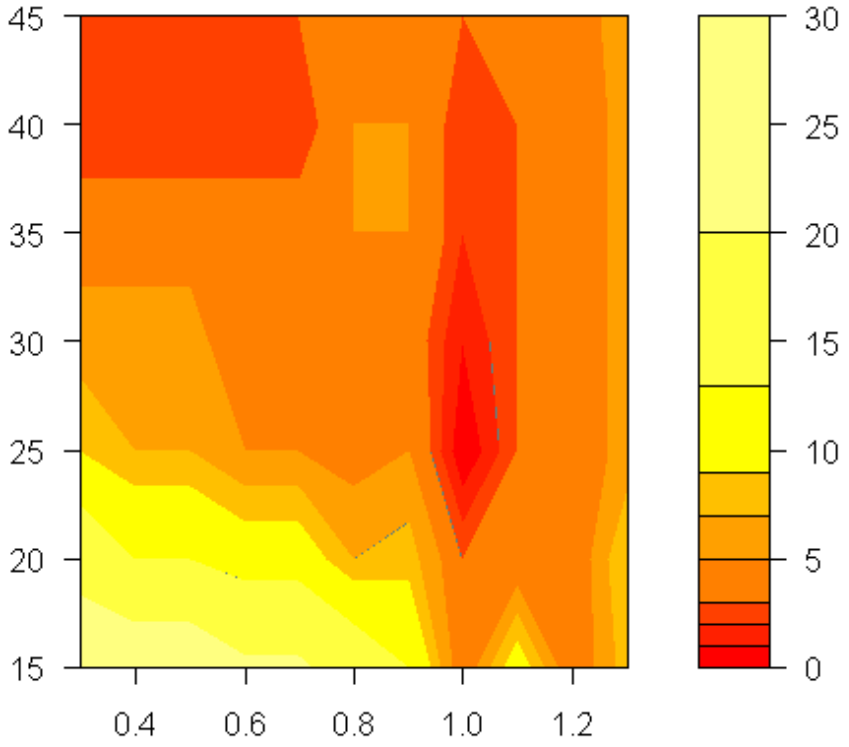
- *If the clusters are close enough, and the transition probability is large enough: merge the clusters*

Prelim tests on cluster reduction



Q4: how to adjust the clustering thresholds in a non-ad hoc way?

Nb of errors



cluster distance

General conclusion

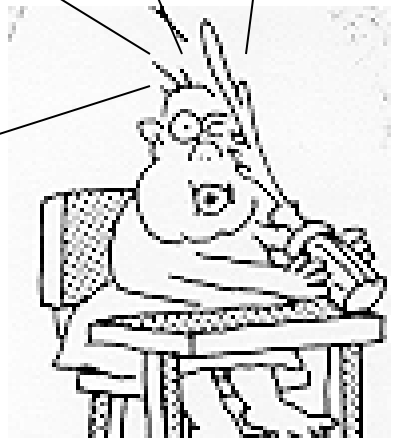
What is required in LAD?

0. overall architecture with acoustic, surface and underlying representations

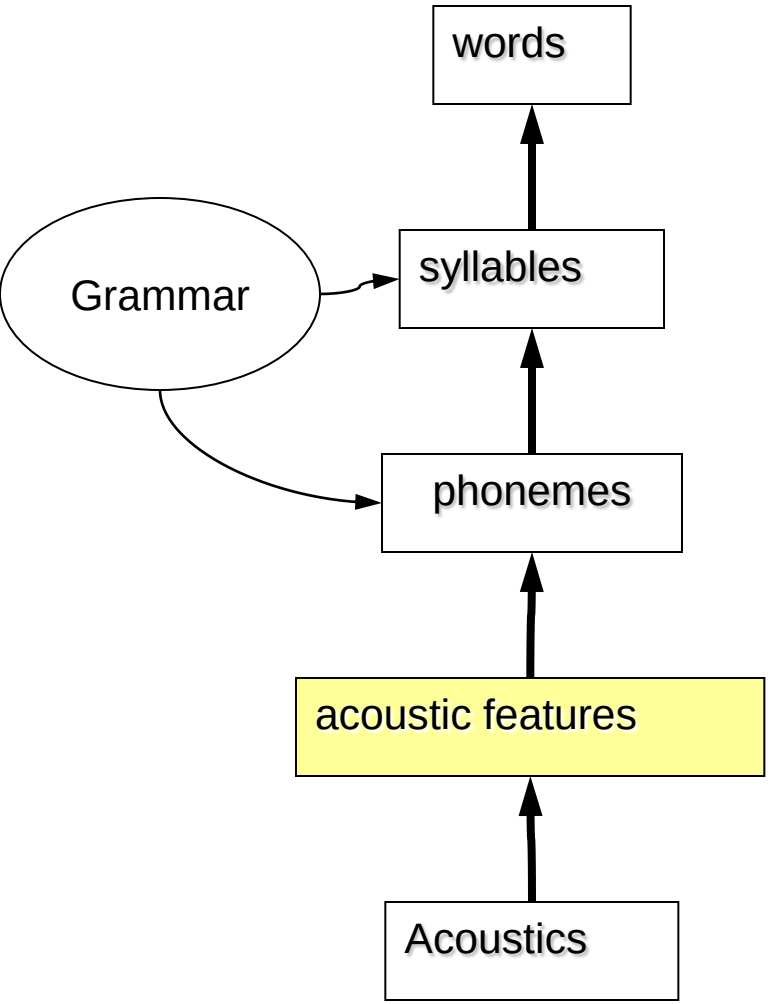
1. segmentation, storing and analysis of coarse grained acoustic templates

2. joint analysis of spectral and sequential information

3. information theoretic/distributional mechanism for the emergence of abstract phonemes

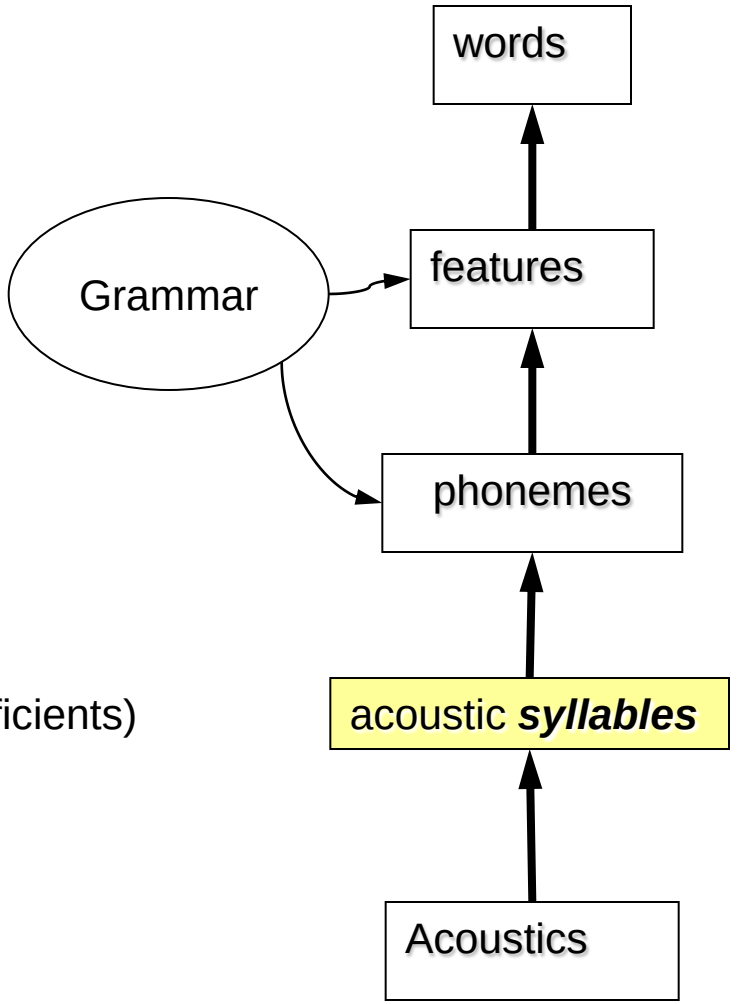


classical architecture



(MFCC coeficients)

proposed architecture



Early language acquisition timeline

	Suprasegments	Segments	Lexicon
Birth	language discrimination (rythmic cues) (Mehler *et al.* 1988)	fine grained universal phonetic discrimination	
1			
2			
3		language discrimination (phonetic cues) (Bosch, Sebastian-Galles)	
4	segmentation in intonational phrases (Hirsh-Pasek *et al*. 1987)		recognition of one's name
5		onset of vowel categories (Kuhl *et al*. 1992; Polka & Werker 1994)	
6	segmentation in phonological phrases (Gerken *et al*. 1994)		ability to segment frequent word forms (Jusczyk & Aslin 1995)
7			
8		phonotactic constraints (Friederici & Wessels 1993; Jusczyk *et al*. 1993, 1994)	
9	Frequent stress patterns		
10			function words (Shady 1996)
11		loss of nonnative consonantal contrasts (Werker & Tees 1984)	
12			recognition lexicon: ~20 words

(months)

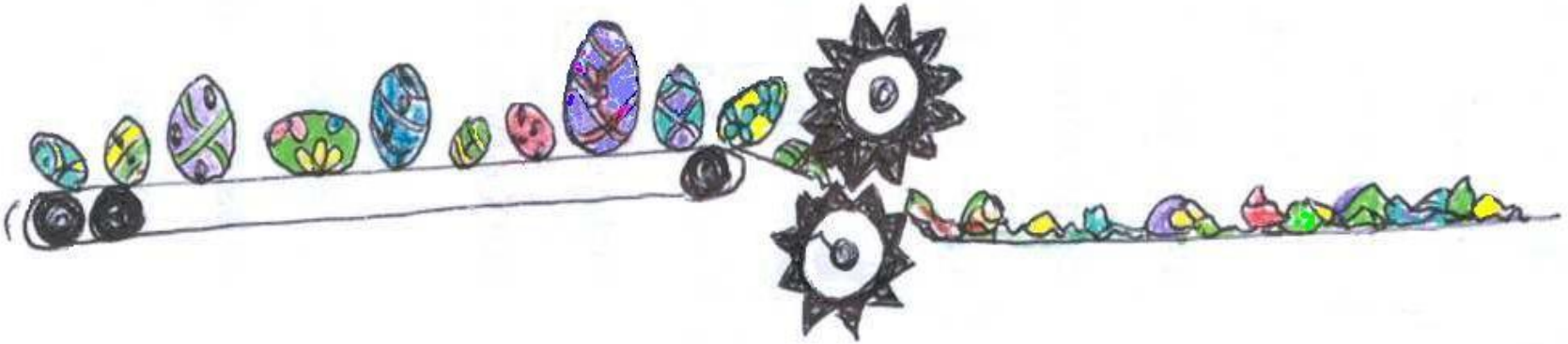
Early language acquisition timeline

	Suprasegments	Segments	Lexicon
Birth	language discrimination (rythmic cues) (Mehler *et al.* 1988)	fine grained universal phonetic discrimination	
1			
2			
3		language discrimination (phonetic cues) (Bosch, Sebastian-Galles)	
4	segmentation in intonational phrases (Hirsh-Pasek *et al*. 1987)		recognition of one's name
5		onset of vowel categories (Kuhl *et al*. 1992; Polka & Werker 1994)	
6	segmentation in phonological phrases (Gerken *et al*. 1994)		ability to segment frequent word forms (Jusczyk & Aslin 1995)
7			
8		phonotactic constraints (Friederici & Wessels 1993; Jusczyk *et al*. 1993, 1994)	
9	Frequent stress patterns		
10			function words (Shady 1996)
11		loss of nonnative consonantal contrasts (Werker & Tees 1984)	
12			recognition lexicon: ~20 words

(months)

Underlying forms

/uvspIlduRmIlk/

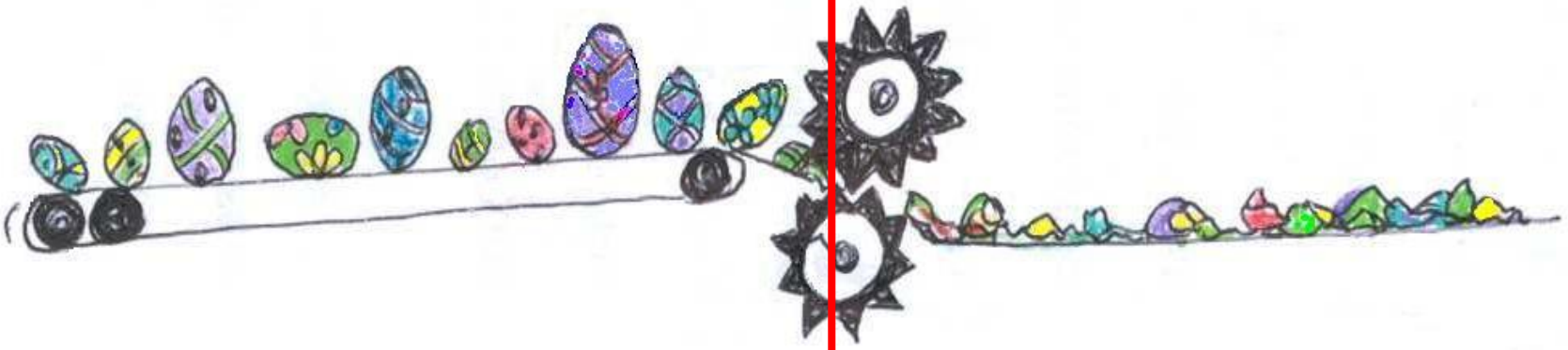


Acoustic signal

(C. Hockett 1955, p. 210)

Underlying forms

/uvspIlduRmIlk/



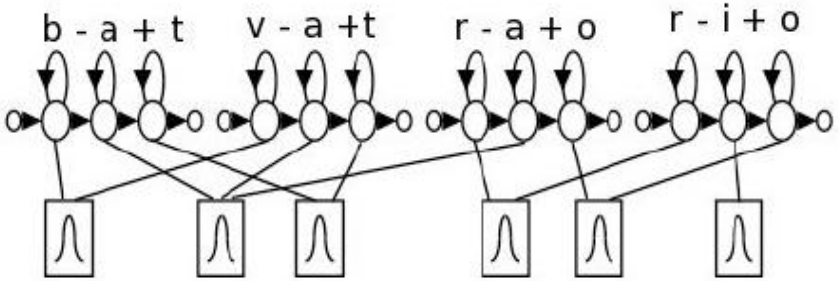
Acoustic signal

how?

when?

(C. Hockett 1955, p. 210)

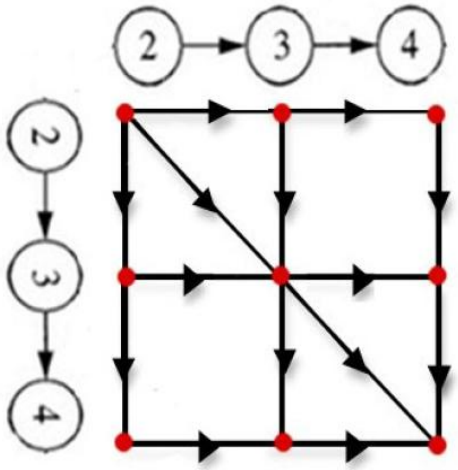
CSJ Corpus: 400 hours of annotated spontaneous speech



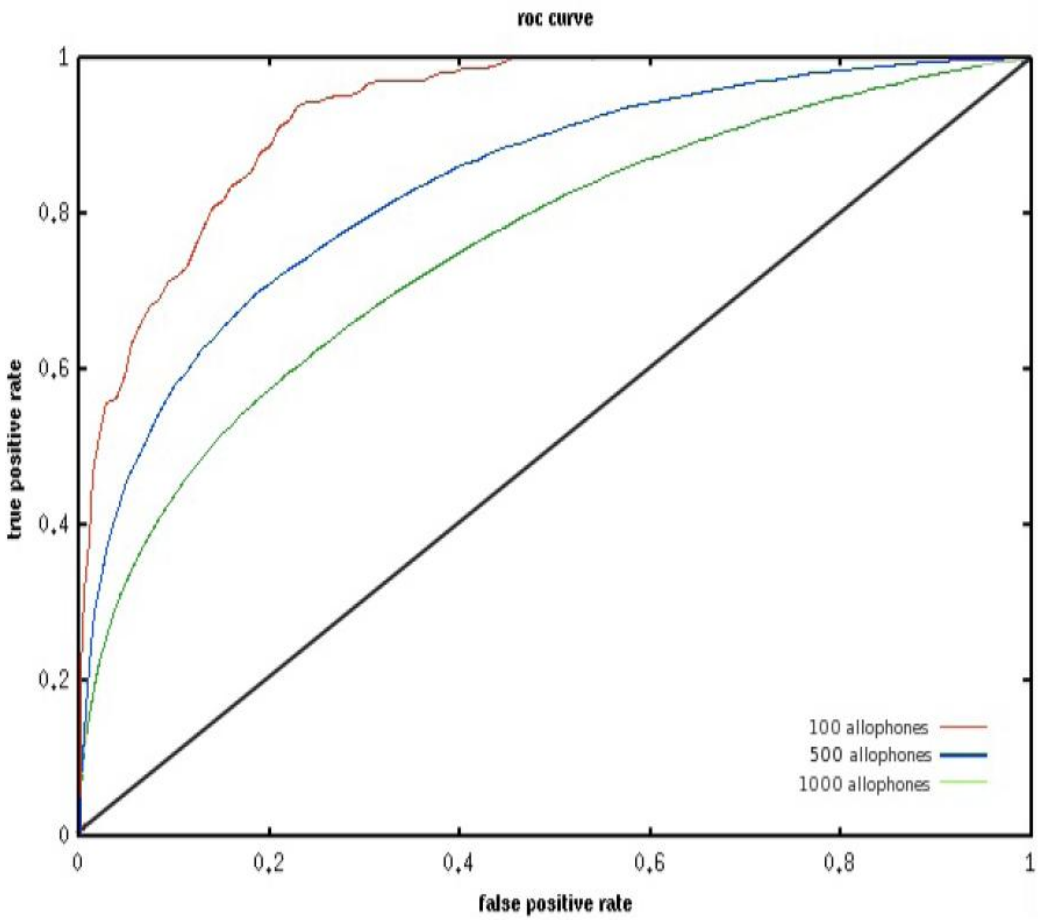
Nombre d'allophones	Taux de reconnaissance (%)
100	77.59
500	81.19
1000	82.68
1500	73.94

Ling. filter 1: acoustic distance

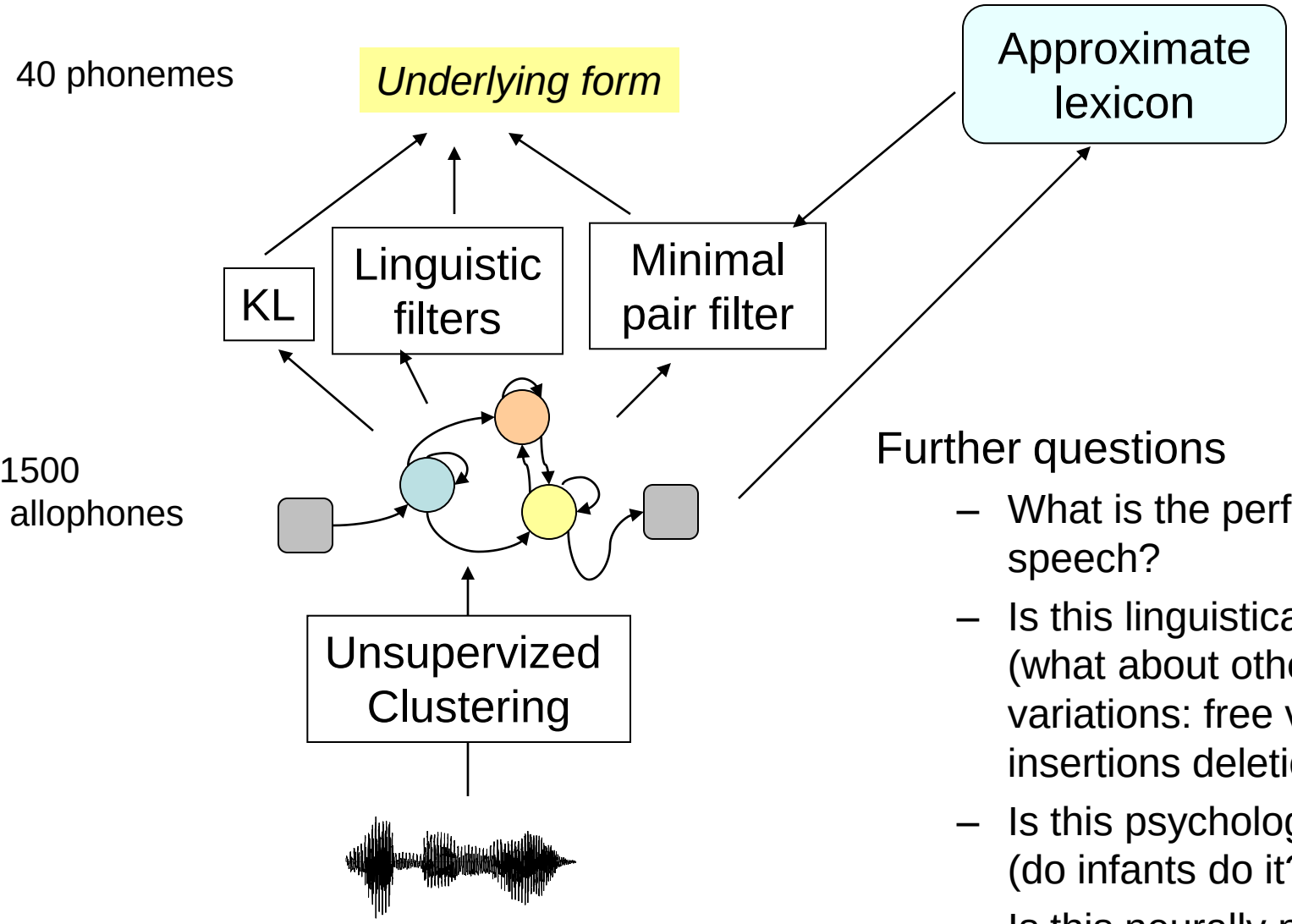
$$d_{KL}(f||g) = \int f(x)\frac{f(x)}{g(x)}dx$$



$$D_{acoustique} = \min_{c \in C} \sum_{I_c} \frac{D_{KL}(i,j)}{t_c + 1}$$



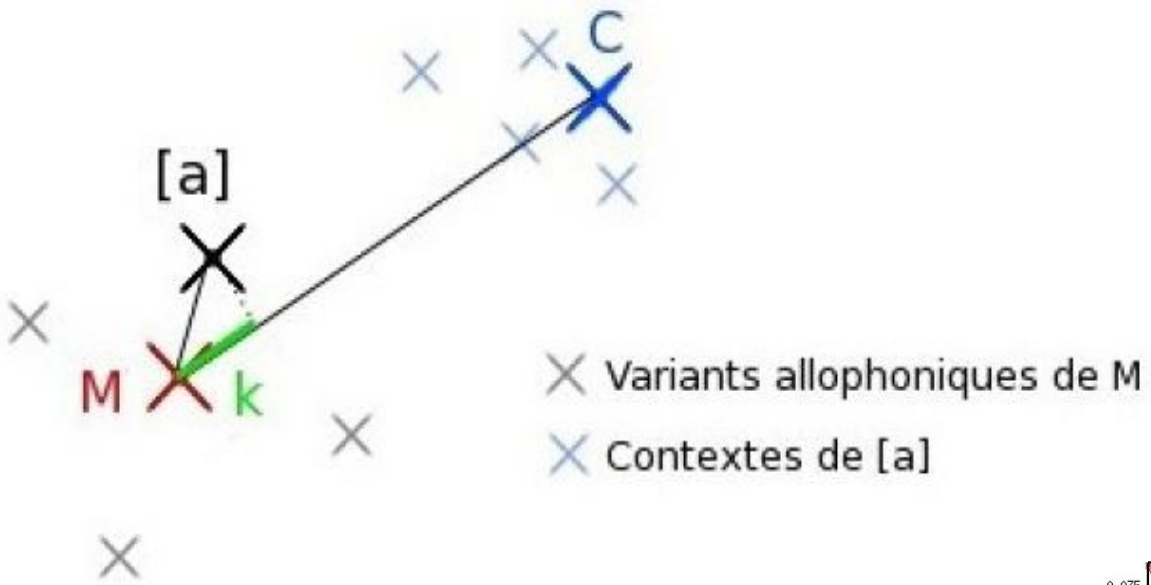
Summary

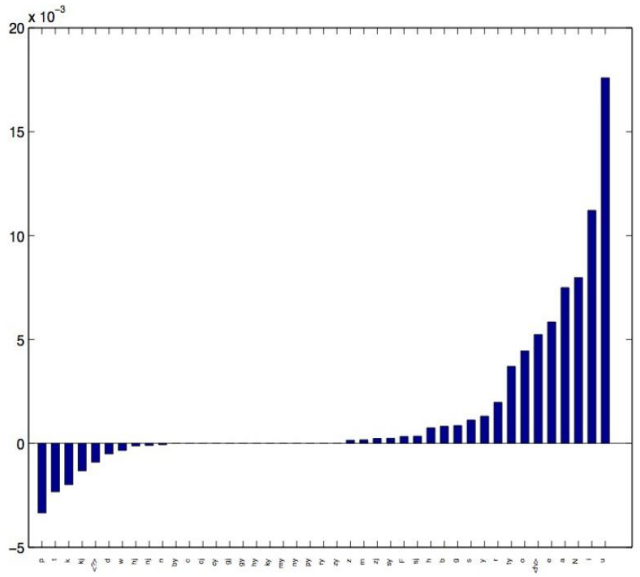


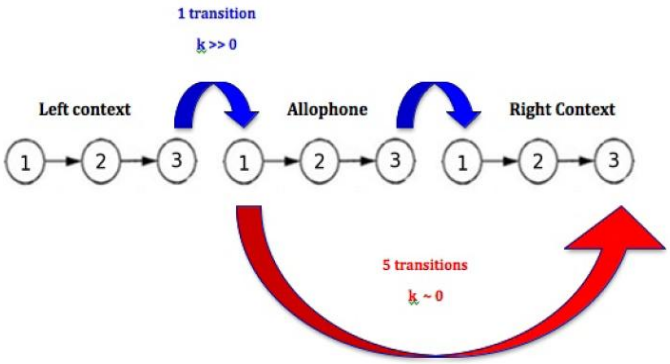
Further questions

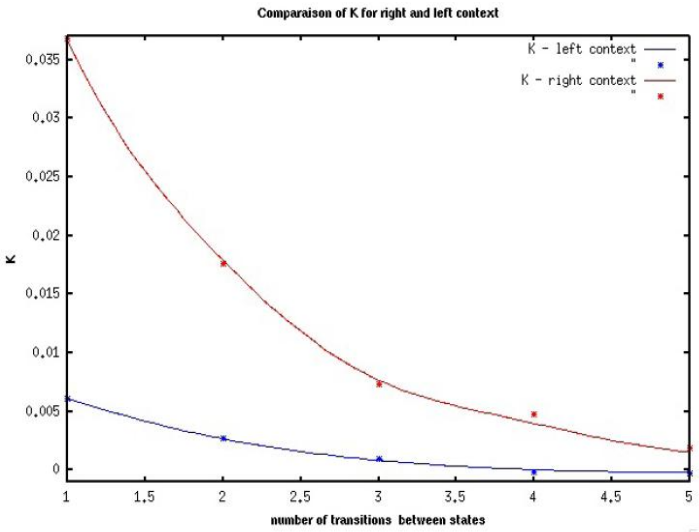
- What is the performance on real speech?
- Is this linguistically plausible (what about other kinds of variations: free variations, insertions deletions, ..)
- Is this psychologically plausible? (do infants do it?)
- Is this neurally plausible?

Ling filter 2: coarticulation model

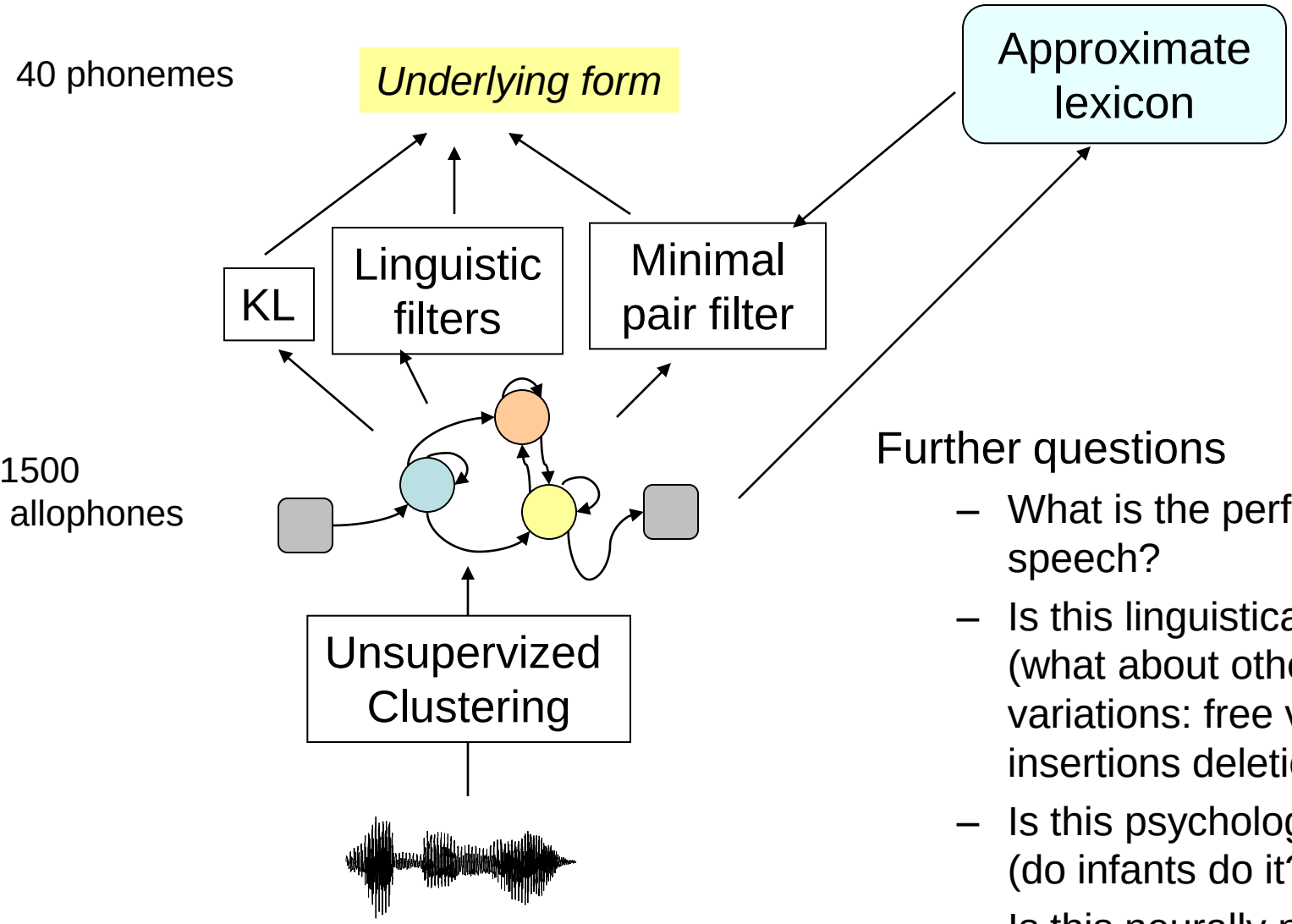








Summary



Further questions

- What is the performance on real speech?
- Is this linguistically plausible (what about other kinds of variations: free variations, insertions deletions, ..)
- Is this psychologically plausible? (do infants do it?)
- Is this neurally plausible?