Analysis by Synthesis of Speech Prosody: from Data to Models.

Daniel Hirst

Laboratoire Parole et Langage,

CNRS & Université de Provence,

Aix en Provence, France

With the past, present and future collaboration of:

- *Caroline Bouzon*
- *Cyril Auran*
- *Saandia Ali*
- *Céline De Looze*
- *Anne Tortel*

Spoken vs. Written language

- Different backgrounds
- Different university departments
- Different conferences
- Different journals
- Engineers vs linguists

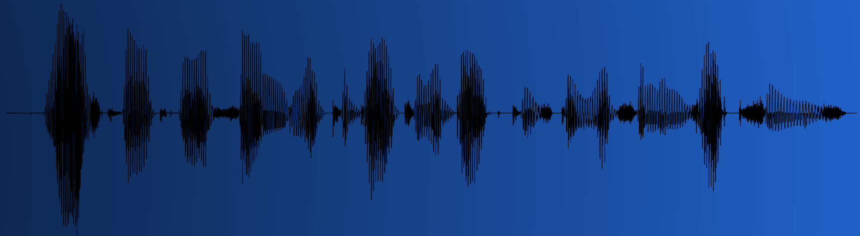
Automatic processing

- Yesterday

$$r_x(k)=\lim_{N\rightarrow\infty} E\left[\frac{1}{2\mathrm{N}+1}\sum_{n=-N}^{N} x(n+k)\,x^{\text{¿}}(n)\right]$$

- Today

Last week my friend had to go to the doctor's to have some injections. She is going to the far east for a holiday and needs to have an injection againnst cholera, tyhphoid fever, hepatitis A, polio and tetanus.



Text vs. Speech

- Processing by computers

	text	*speech*
– *Input*	keyboard/OCR	ASR
– *Storage*	1OO kB/h	100MB/h
– *Manipulation*	easy	hard
– *Output*	print	synthesis

Text vs. speech

- **Processing by humans**

	text	***speech***
– Input	eyes	ears
– Storage	???	???
– Manipulation	???	???
– Output	hands	mouth
	valuable resources	*preferred*

Text and speech… the missing link

- Speech carries extra information
  - Who is speaking
  - Prosody

- Speech = text + prosody

prosody and interpretation

verbal vs. non-verbal

what	*how*
intelligibility	naturalness

- OK. /əʊkeɪ/
- OK…
- OK?
- OK!
- ***OK*** OK!?
- OK :)

Smileys (emoticons)

:) :(;) :-/ :x :"> :p :-* :=((

affect and ambiguity

- He's very hard-working...
- Prosody sounds really interesting!
- She asked the man who lived there.
- Woman without her man is nothing.
- Sept cent vingt cinq mille six cent trente neuf

7 100 20 5 1000 6 100 30 9

720 5006 139

725639

ambiguity

- *Il semble que les policiers sont sur le point d'arrêter Spaggiari, mais il faudra qu'ils fassent vite pour trouver la cachette de l'ancien parachutiste.*

prosodic parameters (subjective)

- length
- pitch
- loudness
- quality

prosodic dimensions
(objective)

- time
- frequency
- intensity

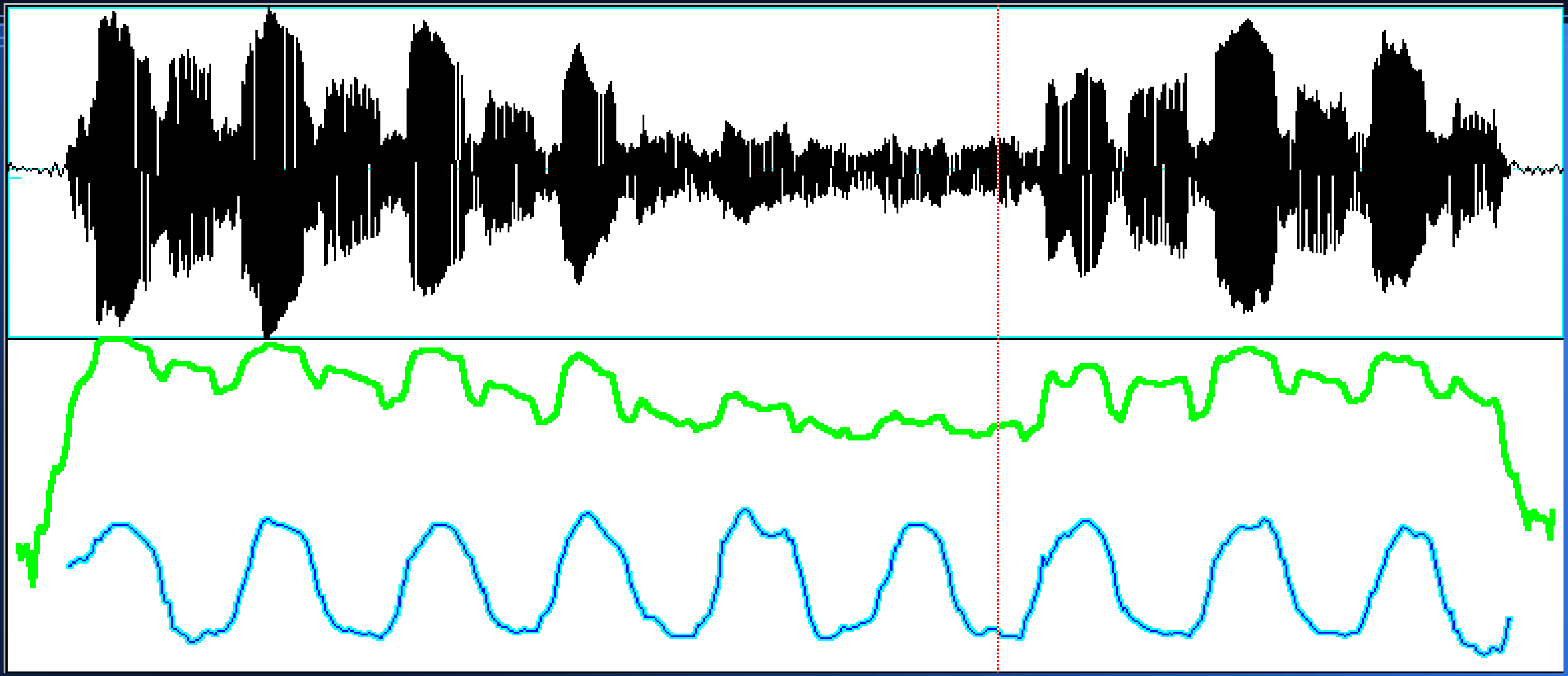
measuring length (duration)

- phonetic not acoustic parameter
  - timing of phonological unit (phoneme, syllable, word etc...)

measuring pitch

- pitch algorithms
  - autocorrelation (intonation research)
  - cross-correlation (voice research)
- octave errors (halving/doubling)
- two pass method (De Looze)

Measuring loudness



- 'ma ma 'ma ma 'ma ma 'ma ma …

Measuring loudness

- Intensity is not a robust indication of loudness in normal speaking conditions
- spectral tilt
  - more promising
  - no standard extraction algorithm

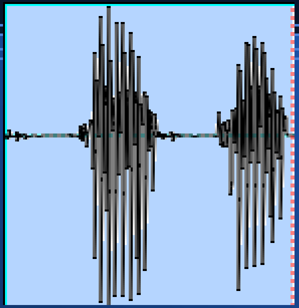
lexical prosody

- prosodic dimensions
  - time
  - frequency
  - intensity

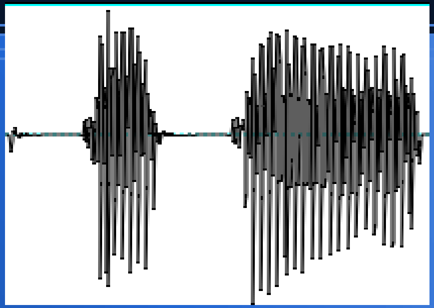
- lexical distinctions
  - quantity
  - tone
  - stress

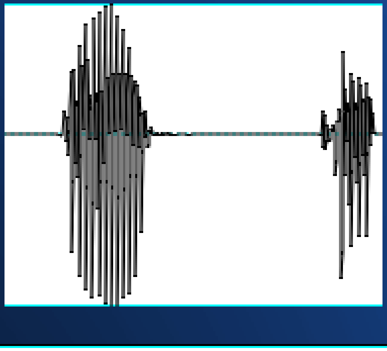
Quantity (Finnish)

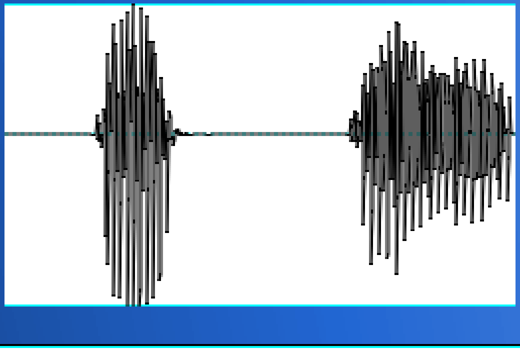
- taka



takaa

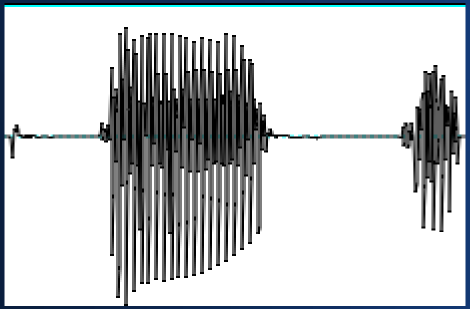


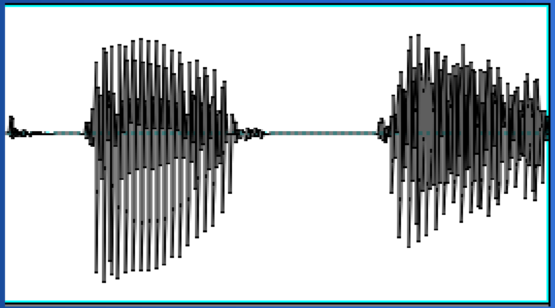




- takka

takkaa

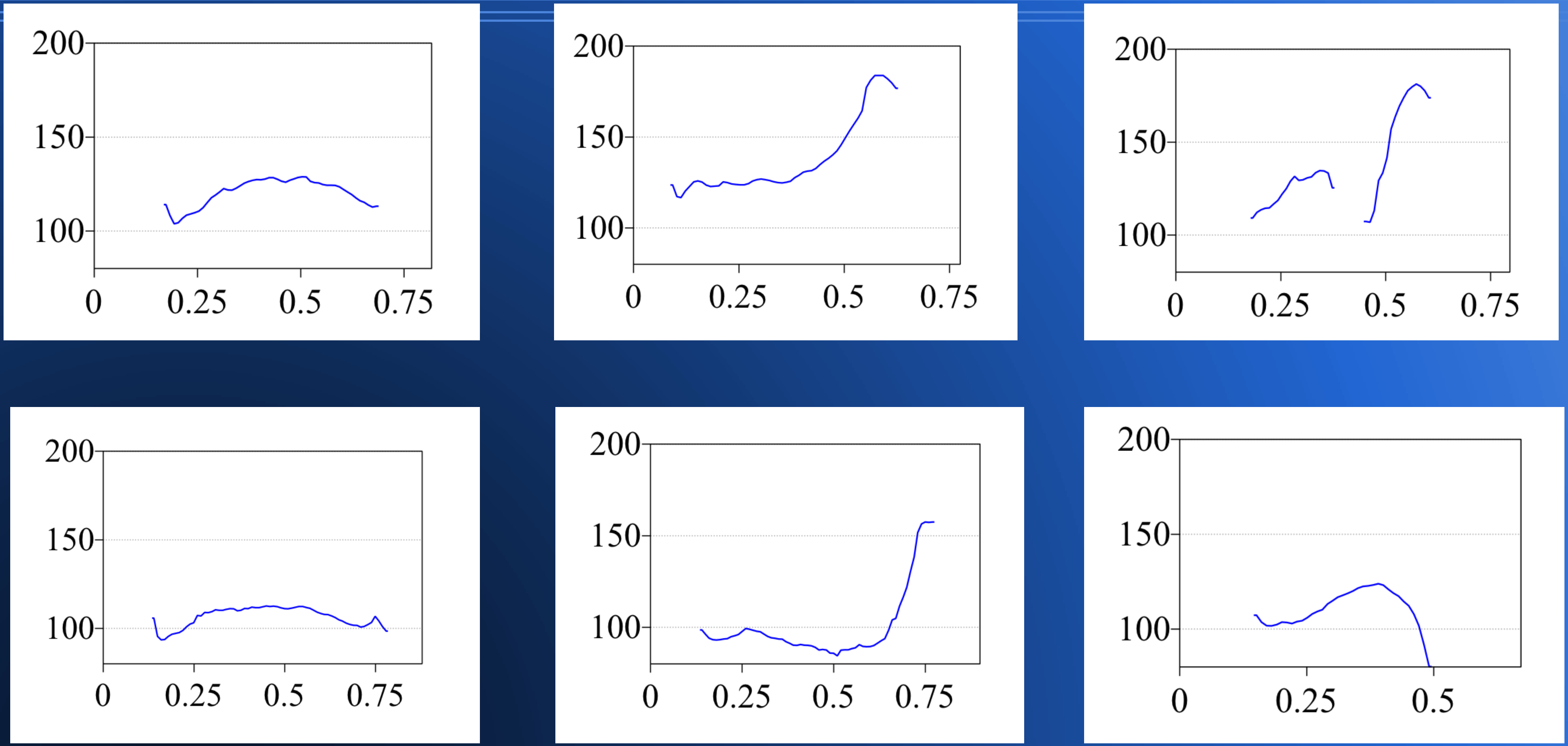




- taakka

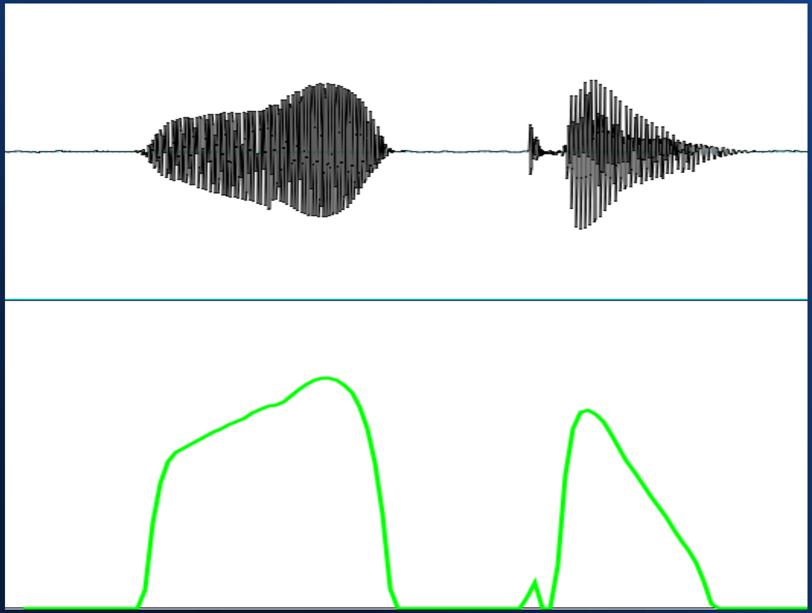
taakkaa

Tone (Vietnamese)

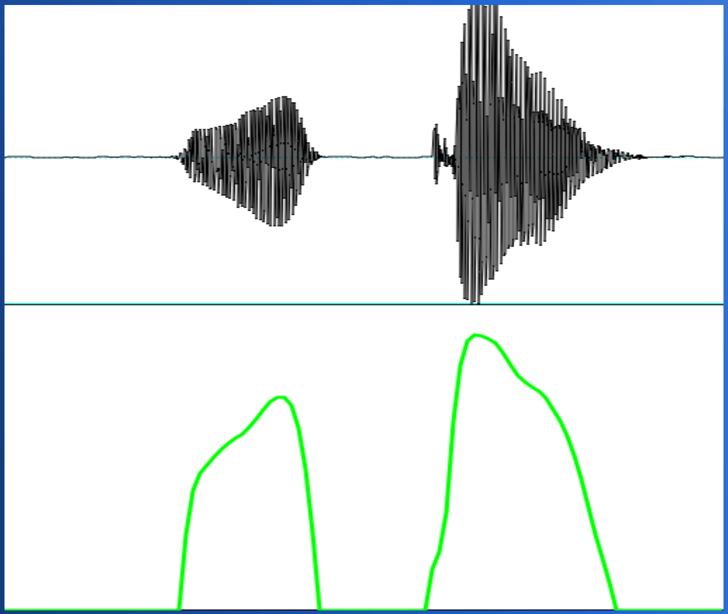


Stress (Russian)

мука /'muka/



мука /mu'ka/

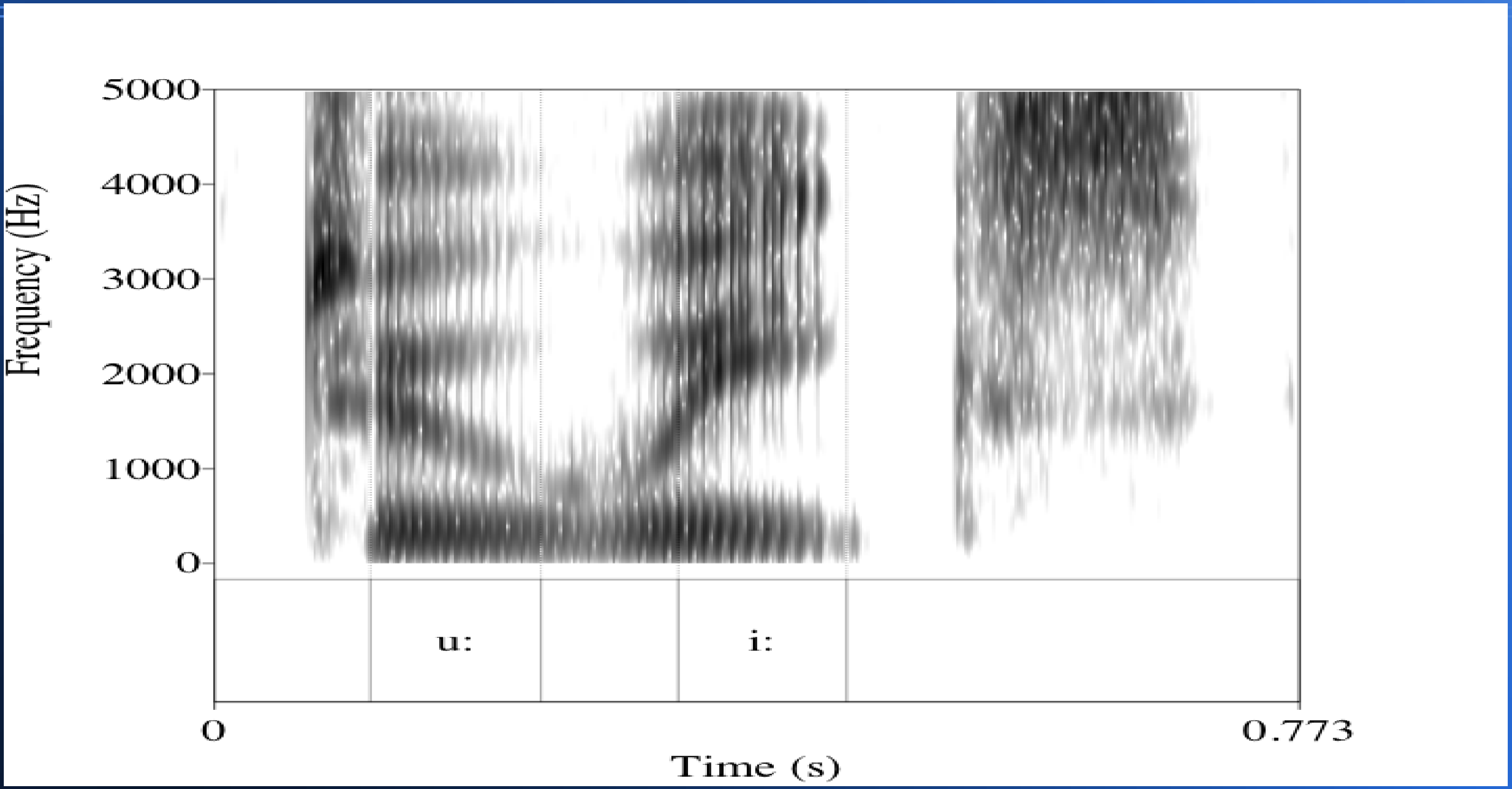


Lexical prosody and acoustics

lexical distinctions	prosodic dimensions
– quantity	duration
– tone	pitch
– accent	intensity

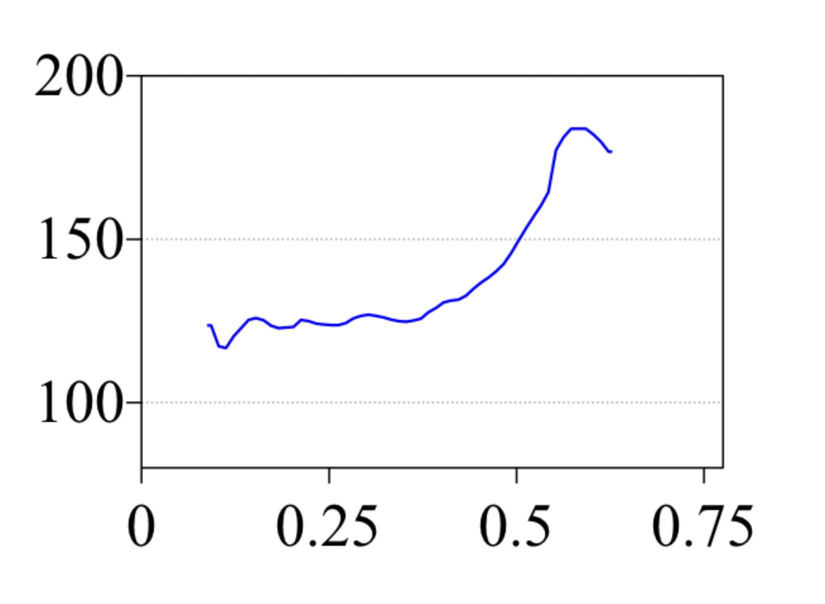
...not so simple!

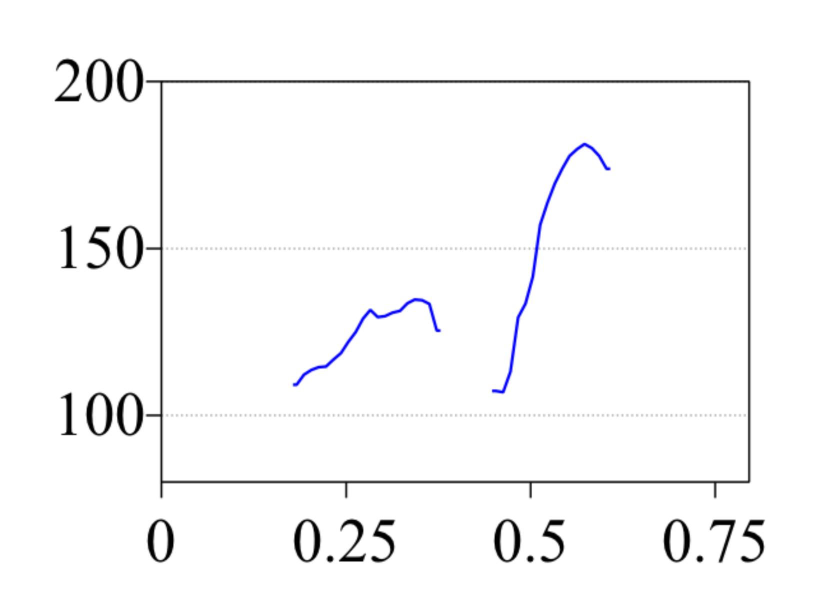
Quantity in English



Two weeks /tu: wi:ks/

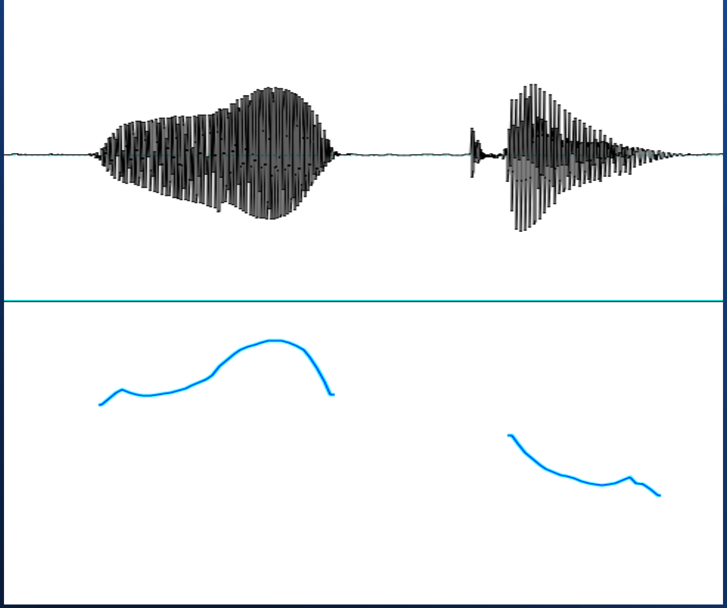
Tone (Vietnamese)

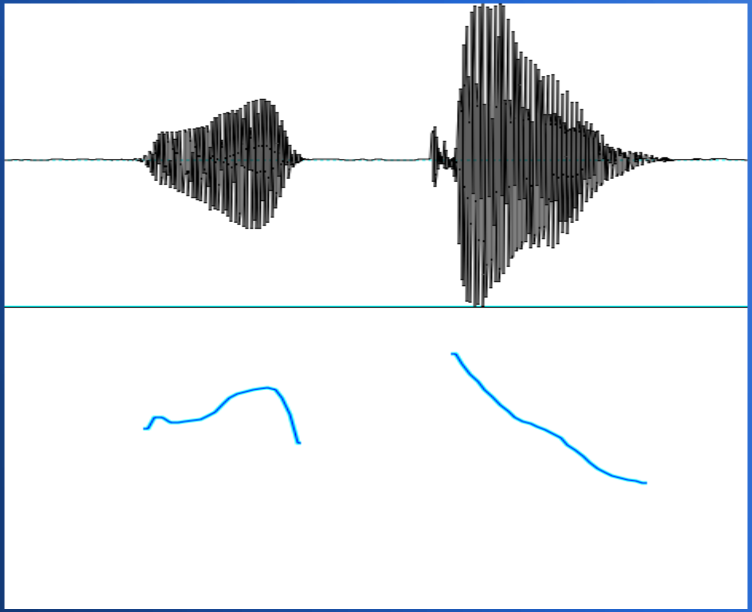




Stress (Russian)

мука /'muka/ мука /mu'ka/





дома /'doma/ дома /da'ma/

Pitch accent in Japanese – tone or stress?

/hasi desu/	*It's an edge*
/ha¬si desu/	*It's a bridge*
/hasi¬ desu/	*It's chopsticks*

phonemes and allophones

English:	port	sport
	/pɔ:t/	/spɔ:t/
	[p^{h}ɔ:t]	[spɔ:t]
French:	port	sport
	/pɔʁ/	/spɔʁ/
	[pɔʁ]	[spɔʁ]
Georgian	/p^{h}uri/ 'cow'	/puri/ 'bread'
	/k^{h}ari/ 'wind'	/kari/ 'door'
English	The Italian like sport.	The Italian likes Port.
	[ðiɪtæliənlaɪkspɔ:t]	[ðiɪtæliənlaɪksp^{h}ɔ:t]

Underlying and surface phonology

- "La science consiste à expliquer le visible compliqué par l'invisible simple."
- *Science consists in explaining the complicated visible by the simple invisible*.

Jean Perrin (1870-1942)

Lexical prosody in French

- No lexical quantity

  today but cf conservative French

  mettre /mɛtʁ/ ≠ maître /mɛ:tʁ/

  voler /vole/ ≠ collègue /kollɛg/

- No lexical tone

- No lexical stress

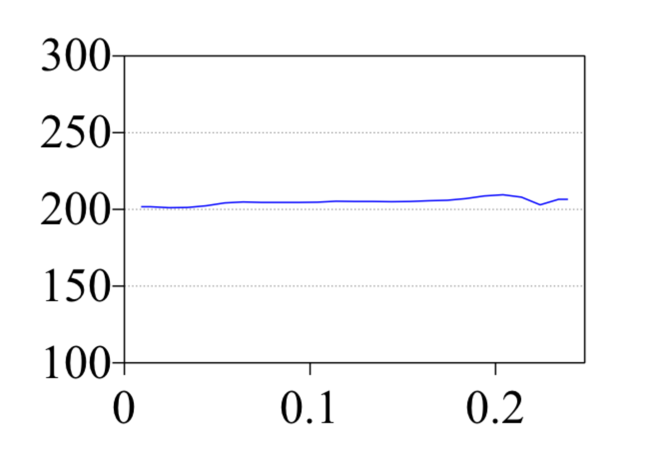
  in Standard French but cf Midi French:

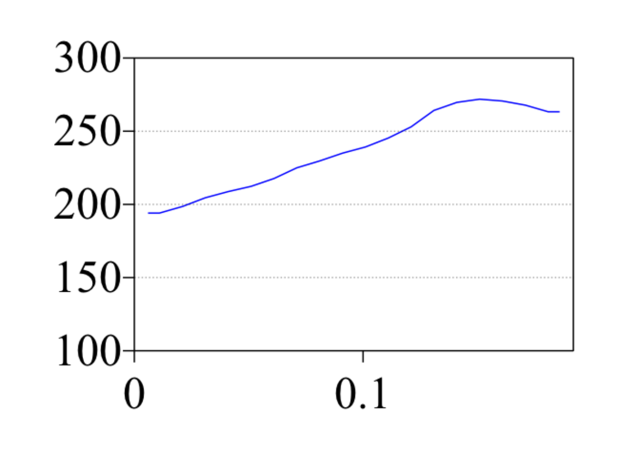
  boîte /'bwatø/ boîteux /bwa'tø/

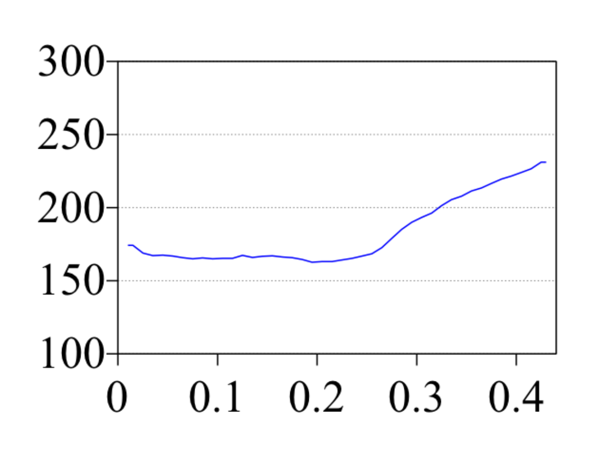
Non-lexical quantity in French:

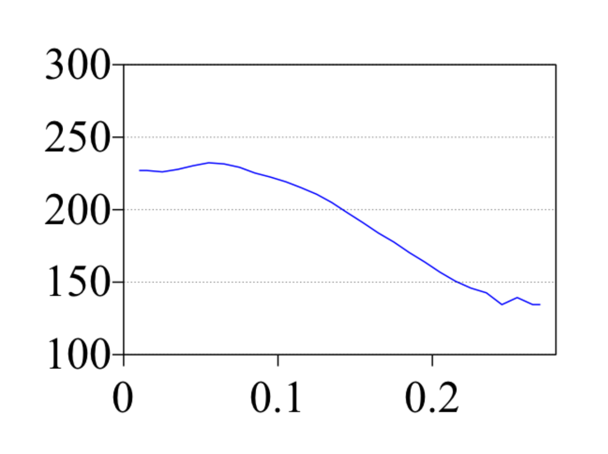
•	Il part tôt	[ilpaʁto]
	Ils partent tôt	[ilpaʁt:o]
•	Il a battu le chien	[ilabatyləʃjɛ̃]
	Il a abattu le chien	[ila:batyləʃjɛ̃]

Non-lexical tone in French









Non-lexical accent in French

- J’enlève son verre *(I take away his glass)*

  [ʒɑ̃'lɛvsɔ̃'vɛʁ]

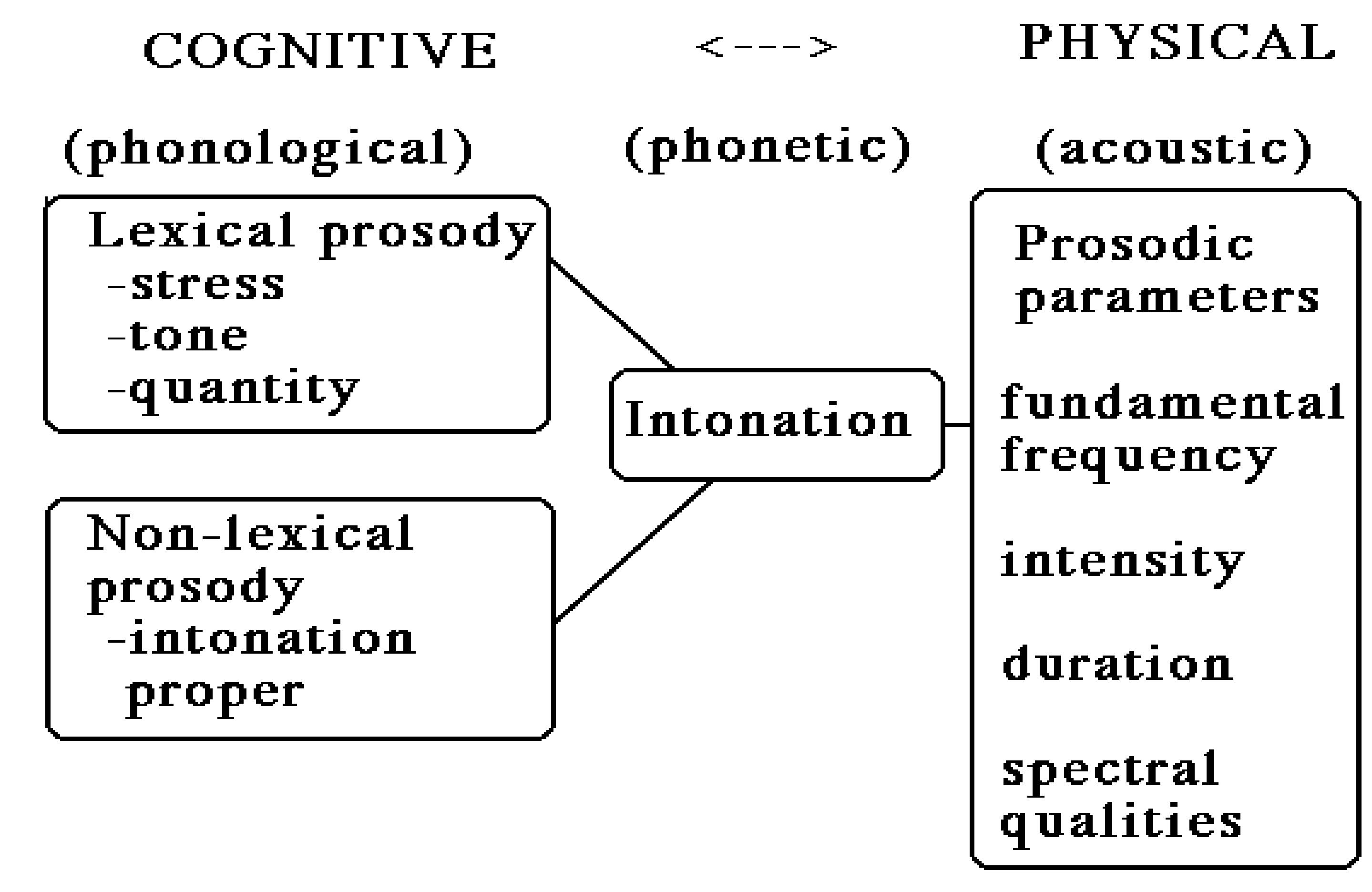
- Jean lève son verre *(Jean raises his glass)*

  ['ʒɑ̃'lɛvsɔ̃'vɛʁ]

Hypothesis

- All languages make distinctive use of quantity, tone and accent
- In some languages these are lexicalised

Prosody - abstract vs physical



COGNITIVE <---> PHYSICAL

(phonological) (phonetic) (acoustic)

Lexical prosody
- stress
- tone
- quantity

Non-lexical prosody
- intonation proper

Intonation

Prosodic parameters

fundamental frequency

intensity

duration

spectral qualities

Rhythmic typology

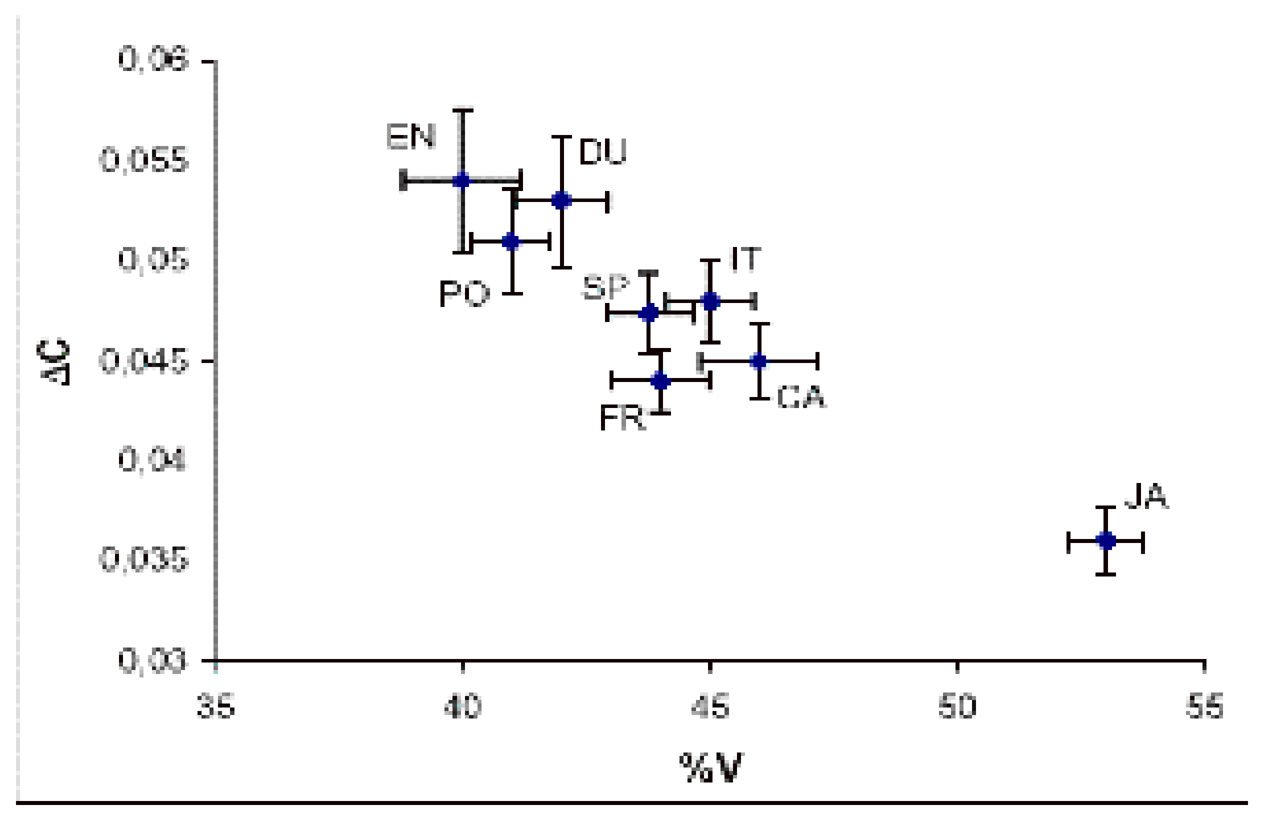
- Stress timing
  - *English, Russian, Arabic...*
- Syllable timing
  - *French, Telugu, Yoruba...*
- Mora timing
  - *Japanese, Tamil...*

experimental evidence

- Roach 1982
  - for (2 minutes each of)
    - *English, Arabic, Russian*
    - *French, Teluga, Yoruba*
  - no significant difference in variability of
    - interstress interval
    - syllable duration
- Dauer 1983, Bertinetto 1989

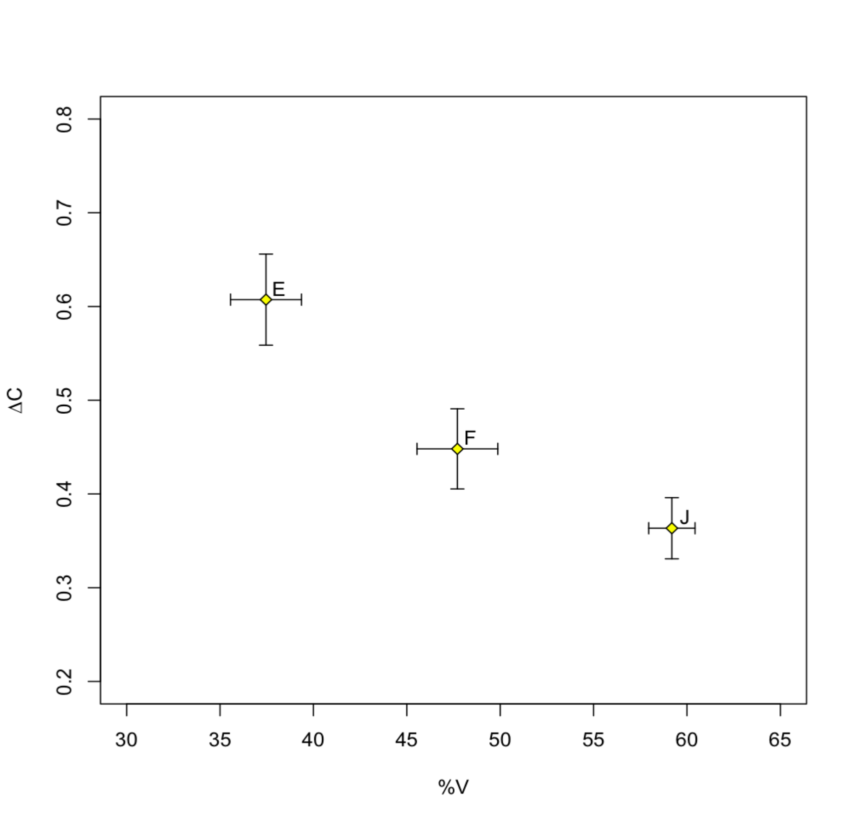
Vocalic and consonantal intervals

- A new metric - Ramus 1999

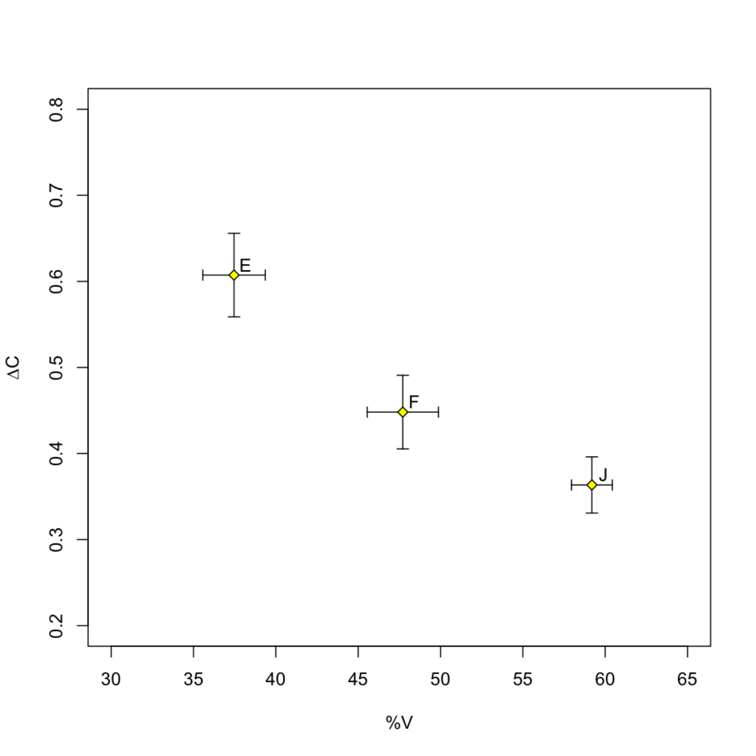


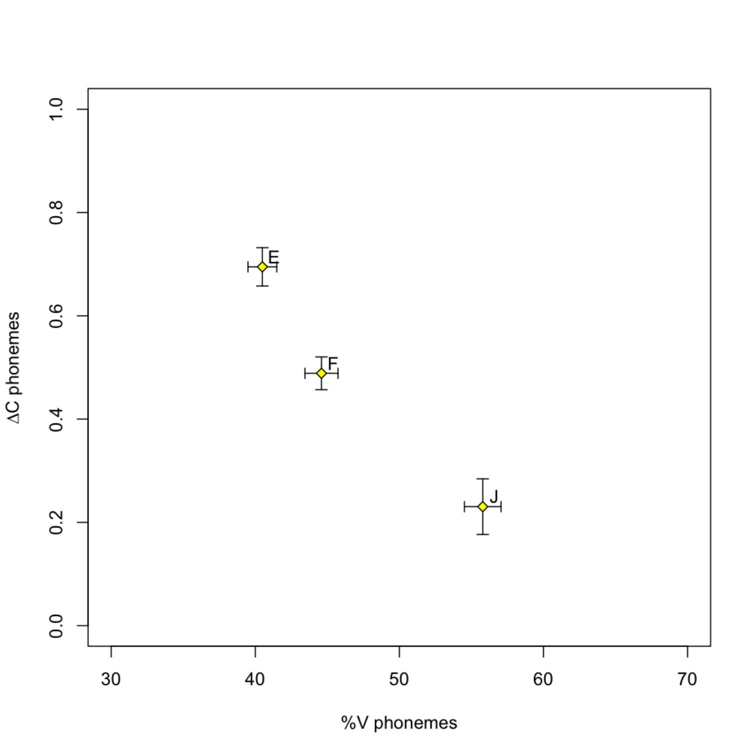
Replication on E, F and J

- 10 sentences each language (Eurom1 corpus)



Rhythm of speech or text?

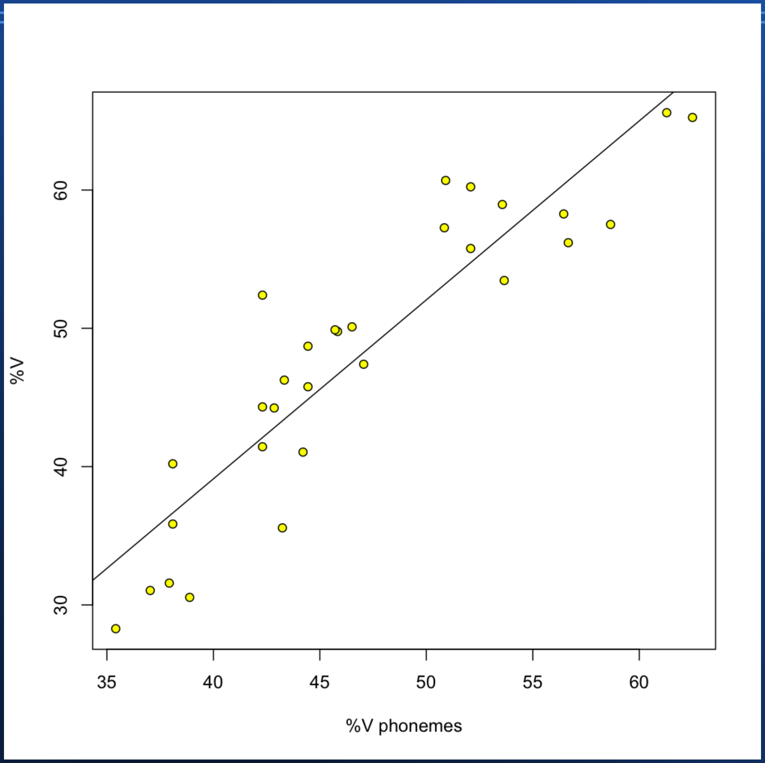


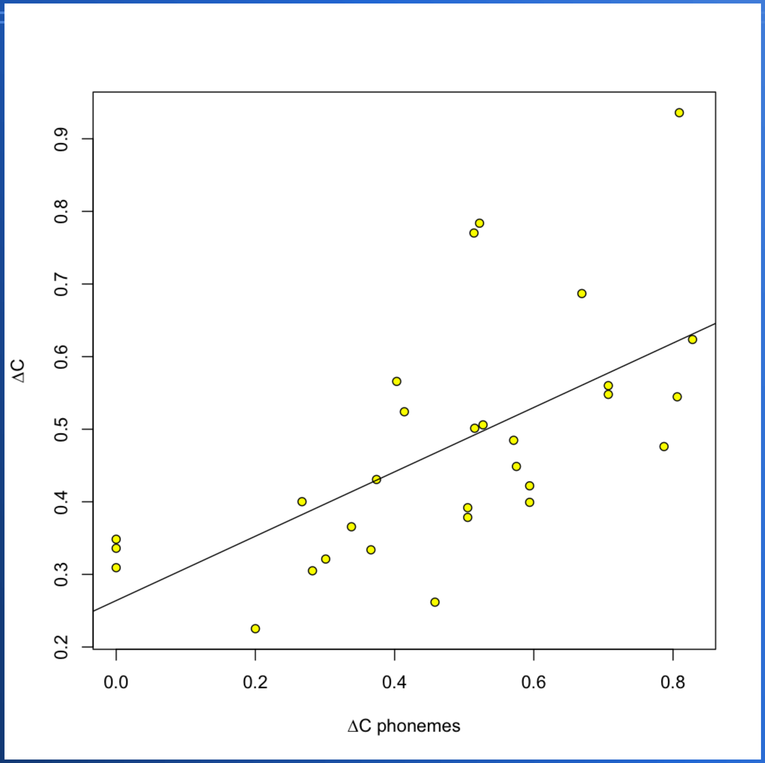


speech

text

%V ΔC for speech and text





speech, r = 0.911

text, r = 0.627

Rhythm types

- morse-code rhythm

  . – . . . – . . – . . .

- machine-gun rhythm

  – – – – – – – – – –

Linear model

- Faure, Hirst & Chafcouloff (1980)

  ISI = 220 + 140*nUS

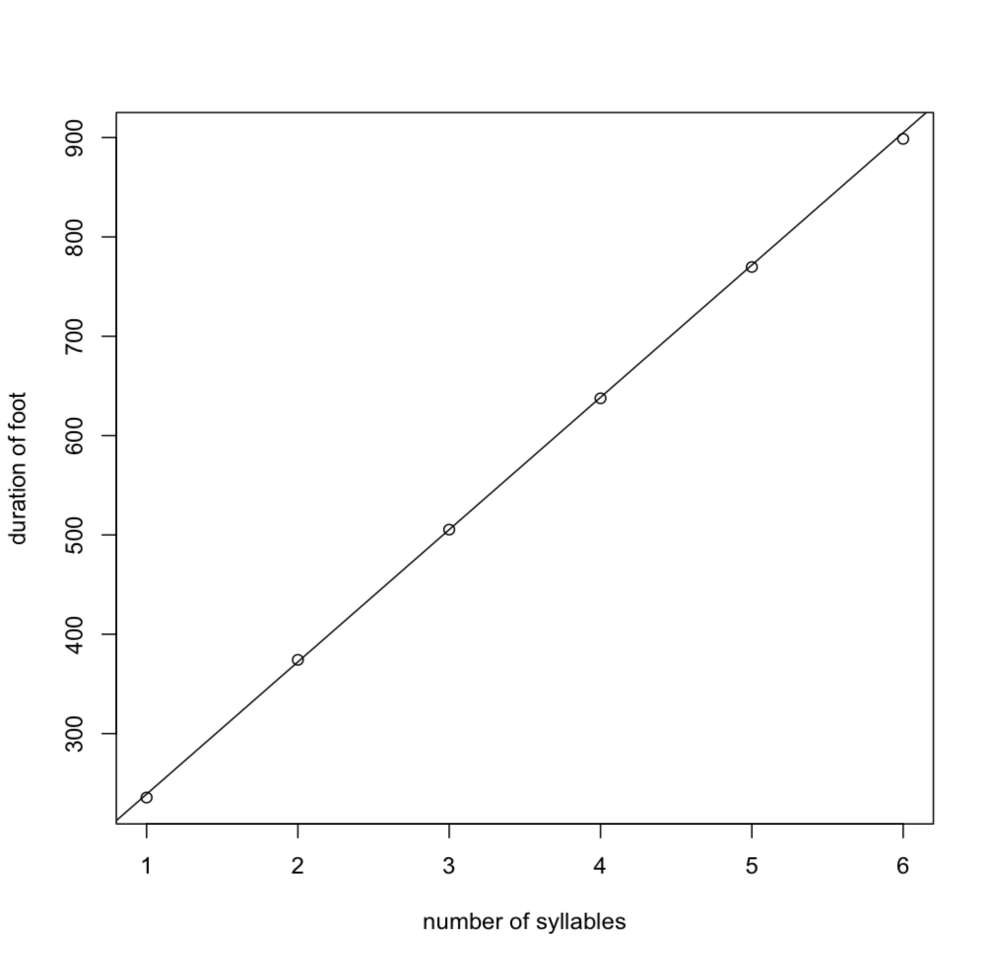
- Eriksson (1991)
  - *Spanish, Greek, Italian*

    ISI = 200 + 100*nUS

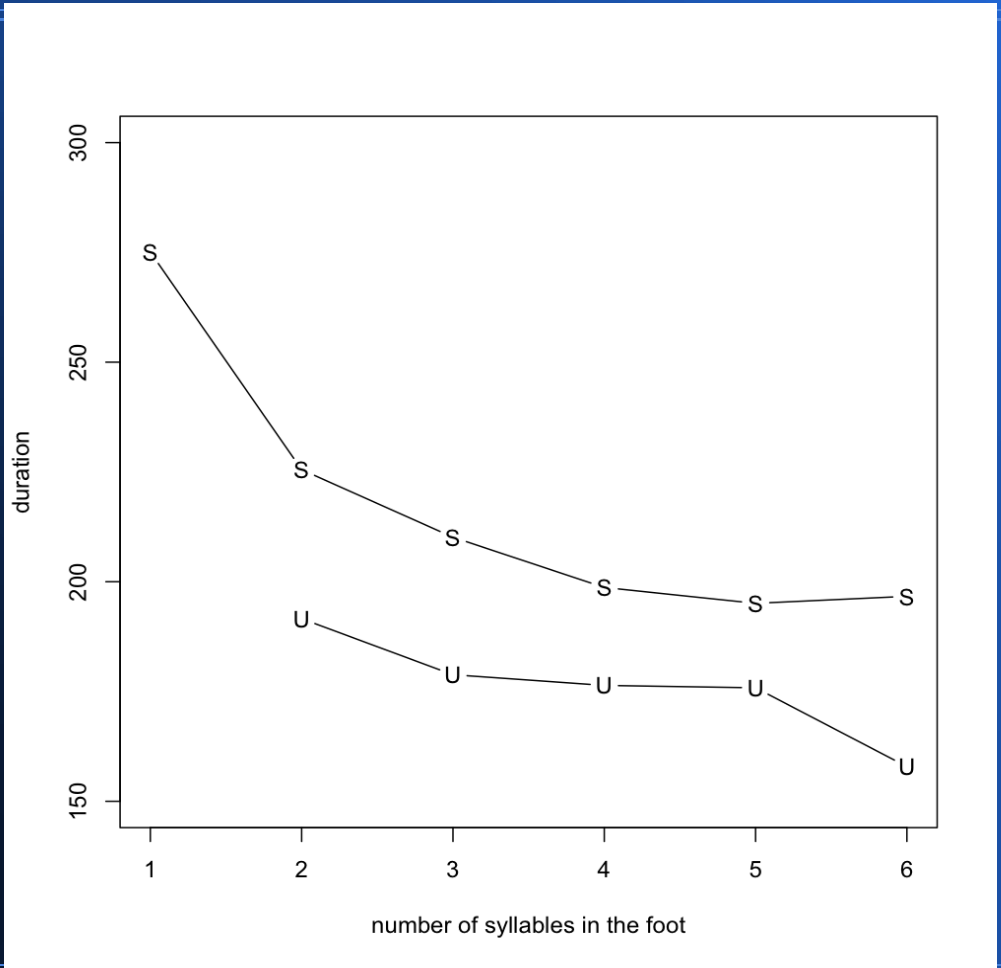
  - *English, Swedish, Icelandic*

    ISI = 300 + 100*nUS

duration of foot / number of syllables in foot



Mean duration of stressed, unstressed syllables / number of syllables in foot



Klatt’s “unsolved problem”

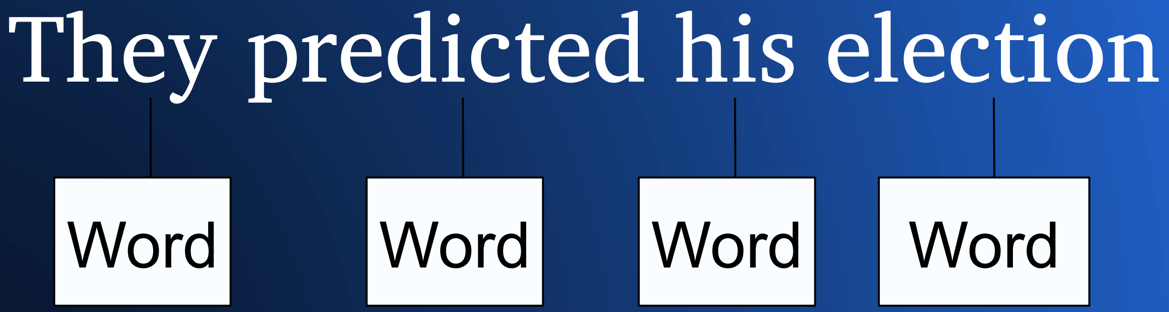
One of the unsolved problems in the development of rule systems for speech timing is the size of the unit (segment, onset/rhyme, syllable, word) best employed to capture various timing phenomena.

Klatt (1987) p.760

Prosodic structure of English

They predicted his election

Prosodic structure



They predicted his election

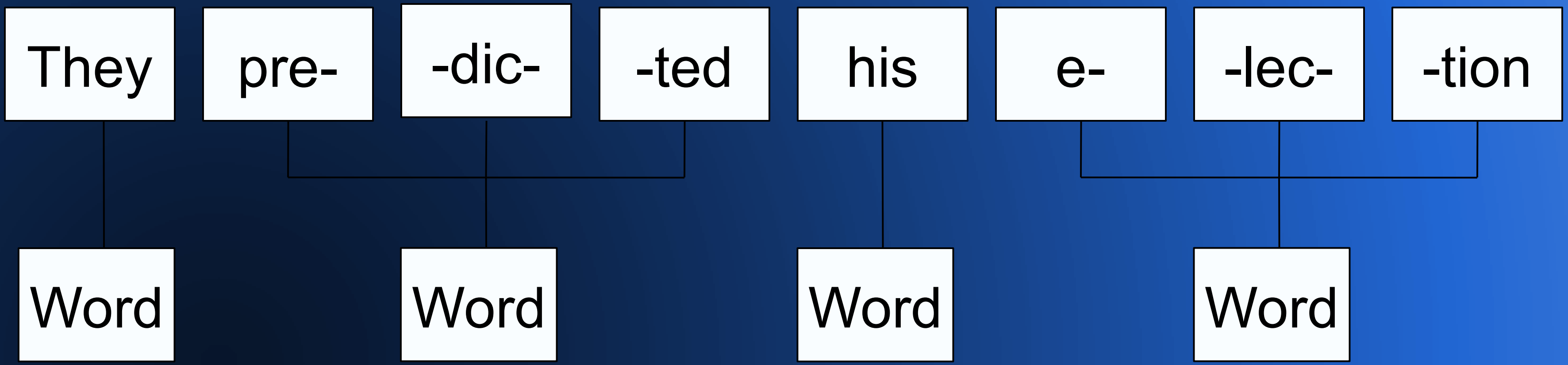
Word

Word

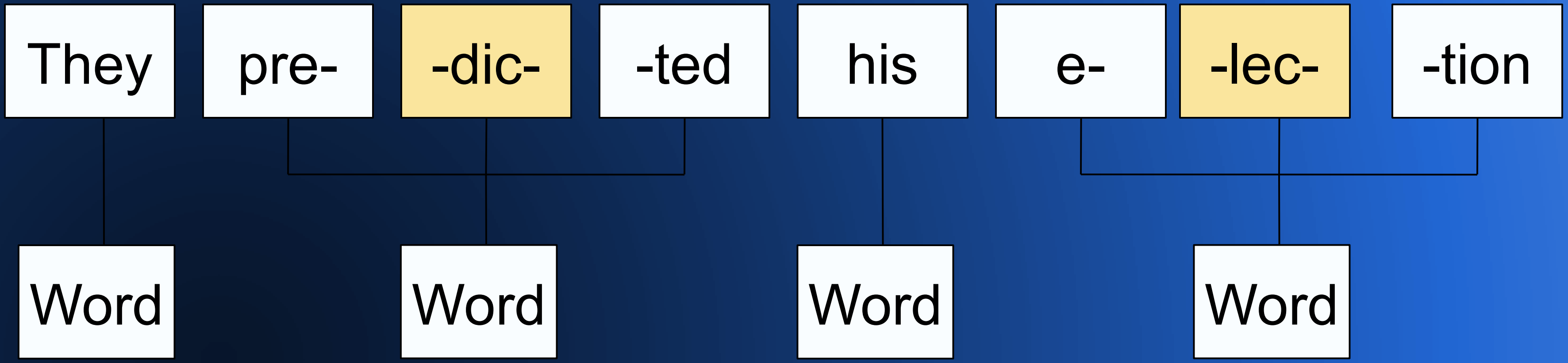
Word

Word

Prosodic structure



Prosodic structure



Prosodic structure

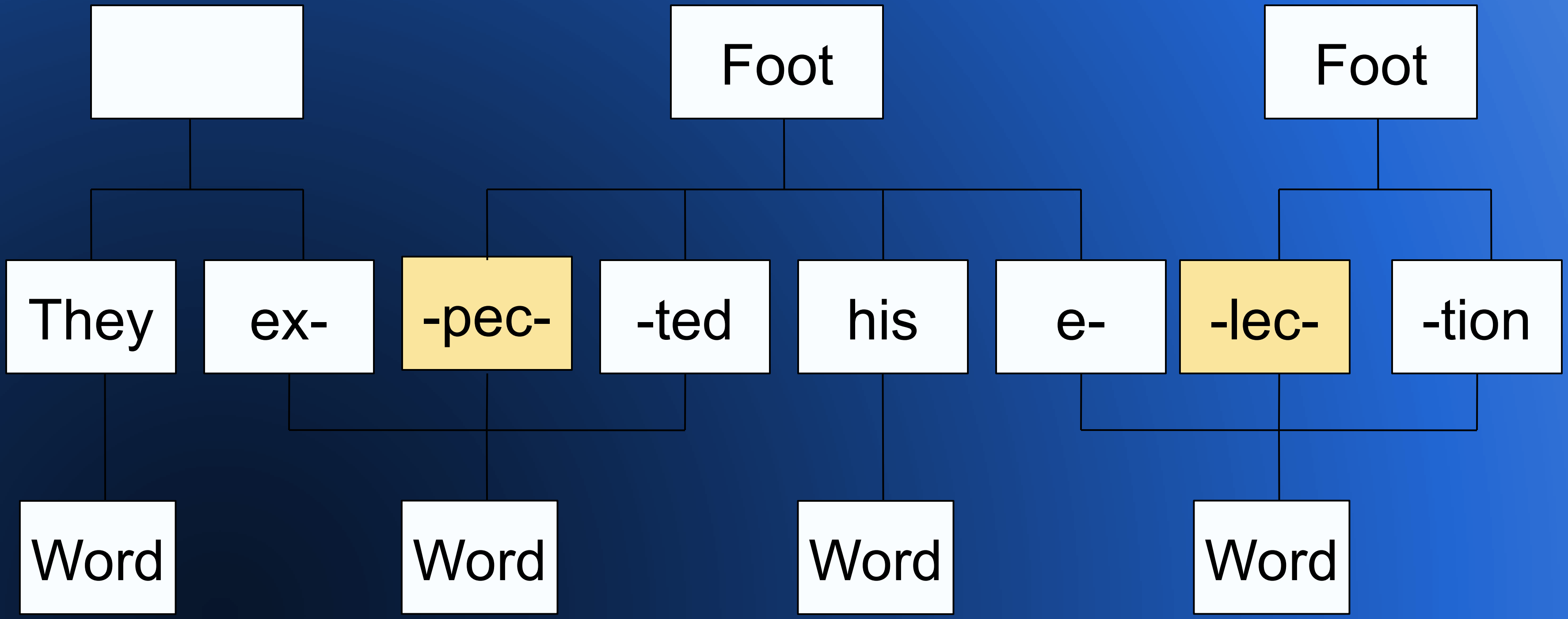
(stress-) foot (Abercrombie, Halliday):

= *sequence of syllables beginning with a stressed syllable and continuing up until the next stressed syllable*

s s S s S s s s S s s S s s S s s s

s s| S s| S s s s| S s s| S s s|S s s s

Prosodic structure



Foot

Foot

They ex- -pec- -ted his e- -lec- -tion

Word

Word

Word

Word

Prosodic structure

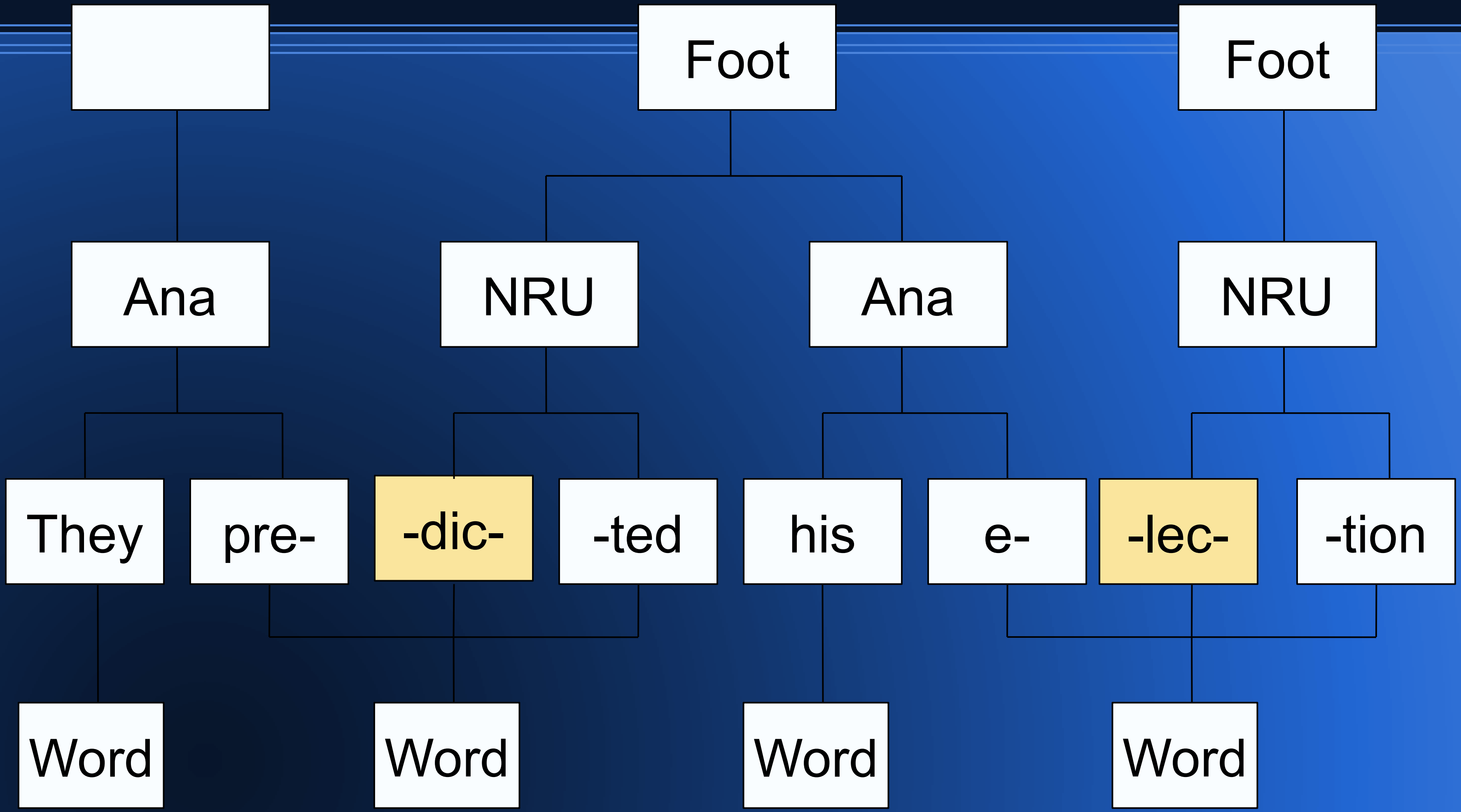
- Narrow rhythm unit (Jassem):

  sequence of syllables beginning with a stressed syllable and ending at the following word boundary

- Anacrusis (Jassem):

  sequence of unstressed syllables not included in a narrow rhythm unit.

Prosodic structure



Aix-Marsec database

- *SEC* (Spoken English Corpus)

  Knowles et al. 1996

- *Marsec* (Machine Readable SEC)

  Roach et al. 1993

- *Aix-Marsec*

  Auran, Bouzon & Hirst 2004

SEC

- 5.5 hours of “authentic” speech
- 53 speakers, c. 55000 words

SEC

- 5.5 hours of “authentic” speech
- c. 55000 words, 53 speakers
- Prosodic markup:tonetic stress marks

  (Knowles & Williams)

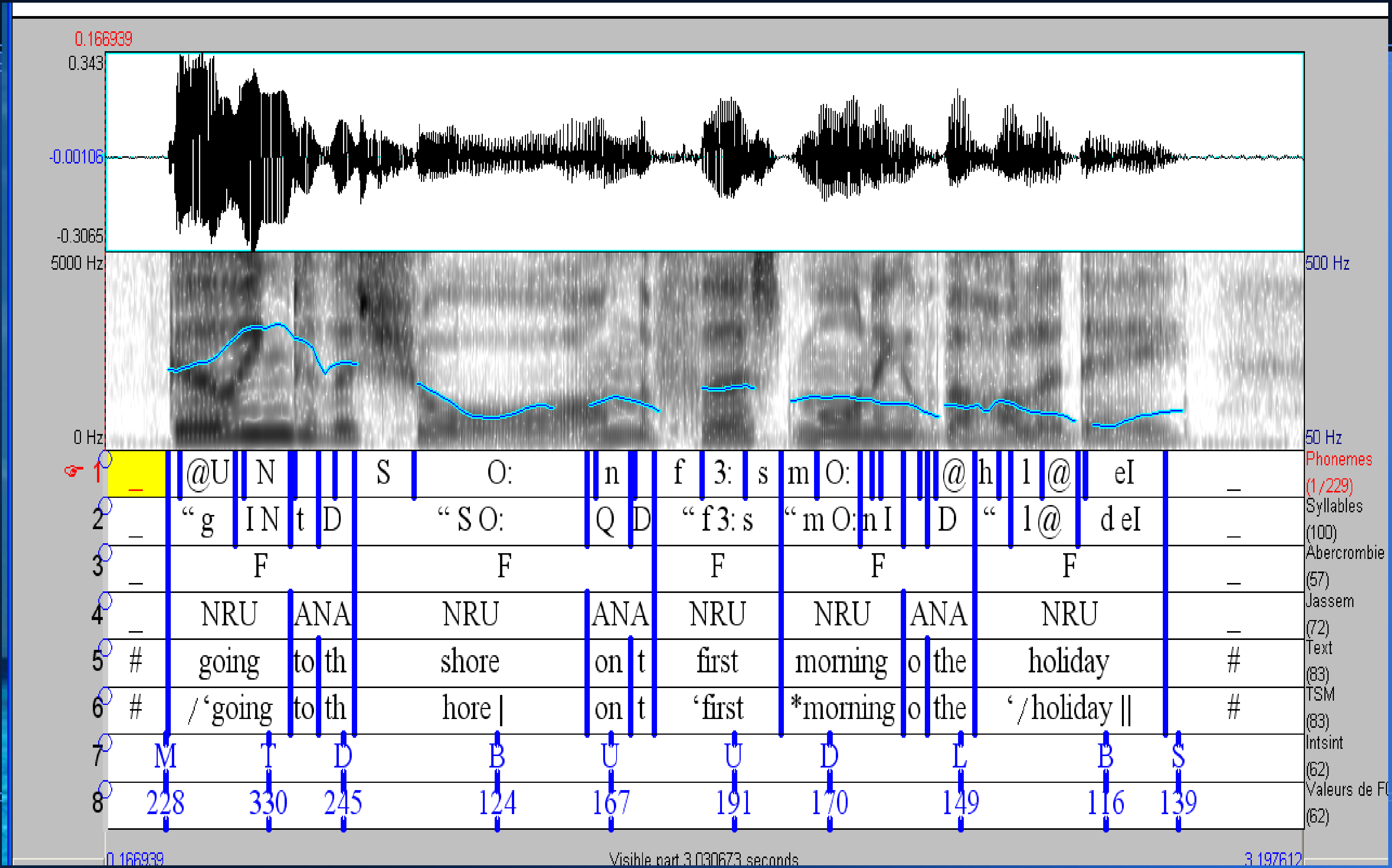
Marsec

- Tonetic stress markup > ASCII
  (Roach et al.)
- words aligned with signal

Aix-Marsec database

- Phonetic transcription
- Phonemes aligned with signal
- Prosodic structure (Praat TextGrids)
- Automatic analysis of intonation (Momel & INTSINT)
- Freely available from the authors

TextGrid from Aix-Marsec


0.166939
0.343
-0.00106
-0.3065
5000 Hz
0 Hz
500 Hz
50 Hz
1 _ @U N S O: n f 3: s m O: @ h l @ eI _
2 _ "g I N t D "S O: Q D "f 3: s "m O: n I D " l @ d eI _
3 _ F F F F F _
4 _ NRU ANA NRU ANA NRU NRU ANA NRU _
5 # going to th shore on t first morning o the holiday #
6 # /'going to th hore | on t 'first *morning o the '/holiday || #
7 M T D B U U D L B S
8 228 330 245 124 167 191 170 149 116 139
Phonemes
(1/229)
Syllables
(100)
Abercrombie
(57)
Jassem
(72)
Text
(83)
TSM
(83)
Intsint
(62)
Valeurs de F0
(62)
0.166939
Visible part 3.030673 seconds
3.197612

Hypothesis

- size of whole :: compression of parts

  If a prosodic constituent is involved in the planning of speech rhythm we should expect the size of the constituent to have a negative effect on the duration of the phonemes which make it up.

Method

- Linear correlation and regression
  - Independent variable:

    size of constituent (number of phonemes)
  - Dependent variable:

    mean lengthening/compression of phonemes

(Z score)

$$z_{i/p} = \frac{d_{i/p} - m_p}{s_p}$$

Results - 1

- Very significant negative correlation of lengthening of phonemes (Z-score) with number of phonemes in
  - Word
  - Foot
  - Narrow Rhythm Unit

↓

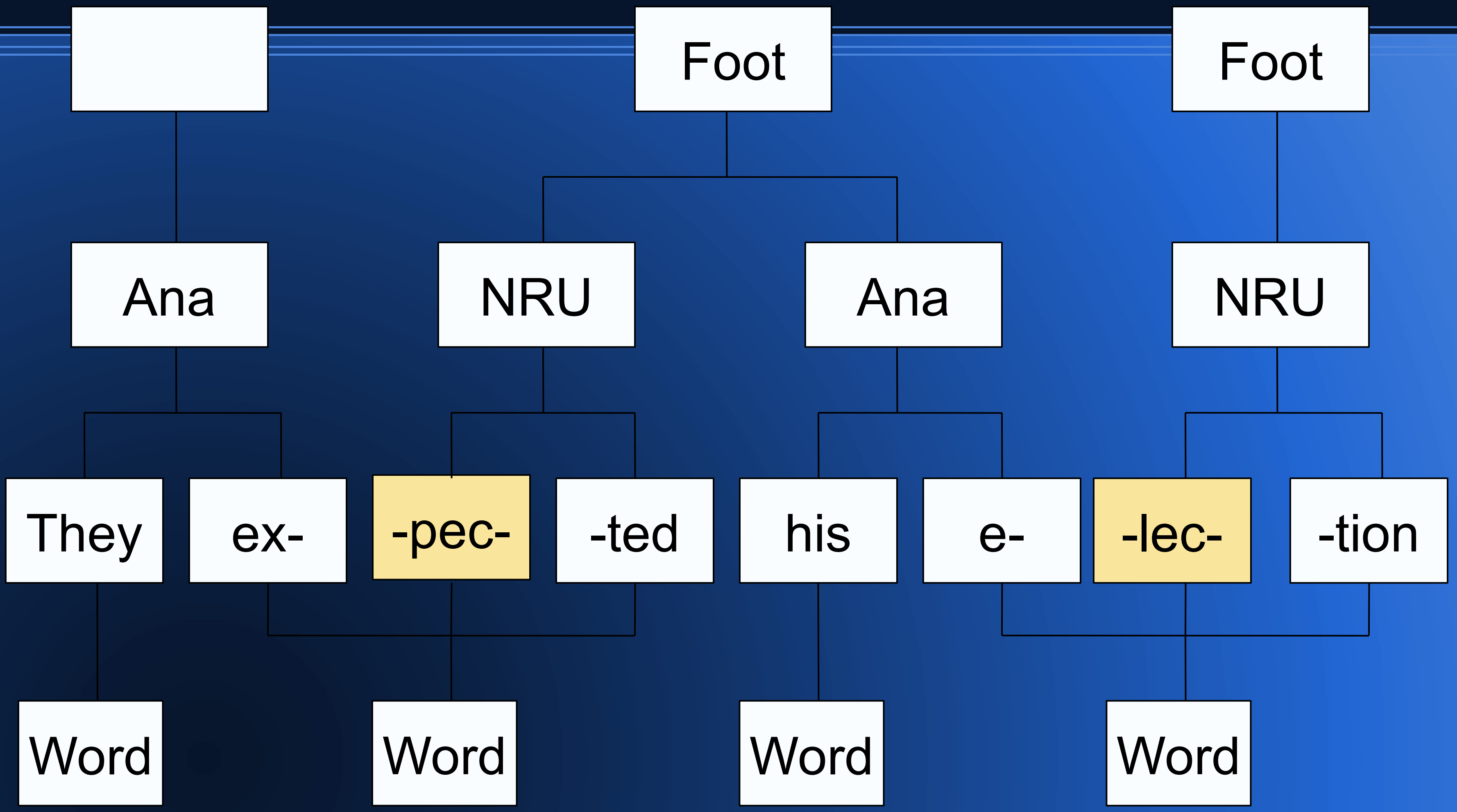
Results - 2

- Little or no correlation of lengthening/compression of phonemes (Z-score) with number of phonemes in:
  - Syllable
  - Anacrusis

Interpretation

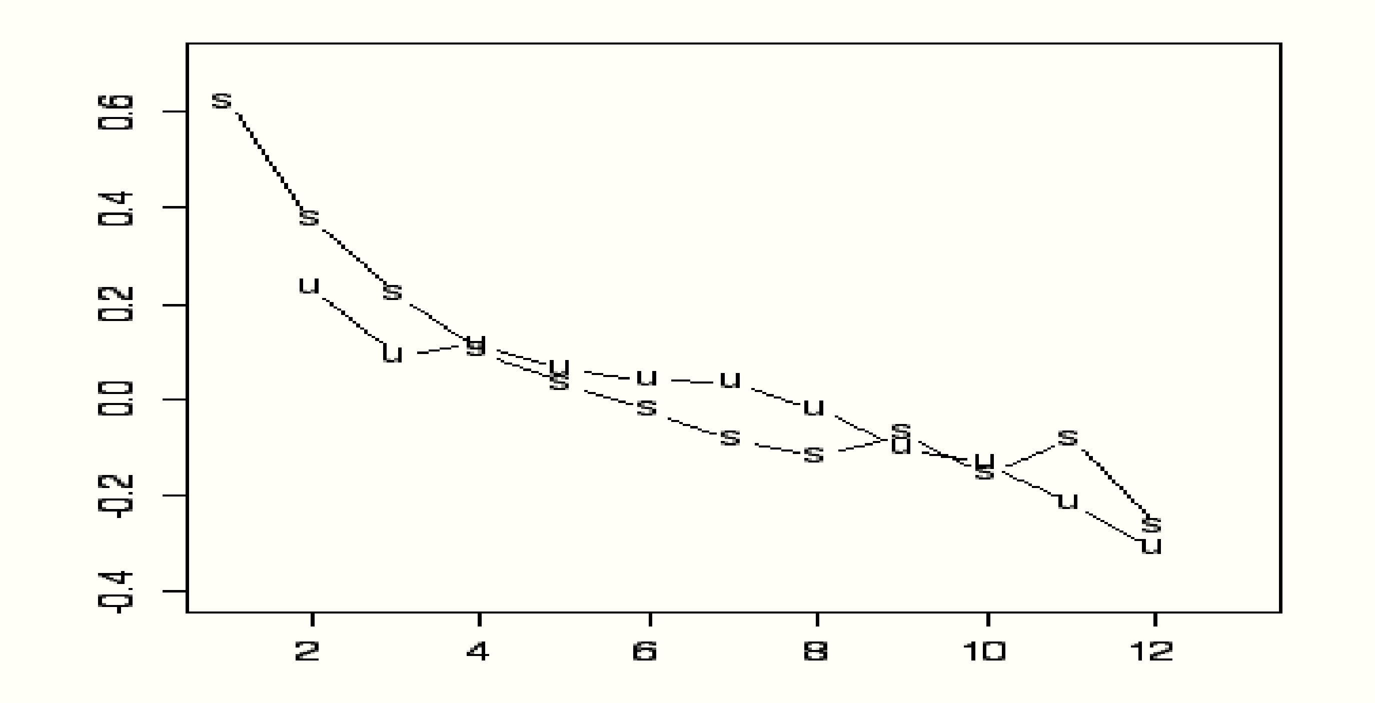
- Syllable and anacrusis have little effect on the lengthening of English phonemes
- Word, foot and narrow rhythm unit play significant role (in that order)

Prosodic structure

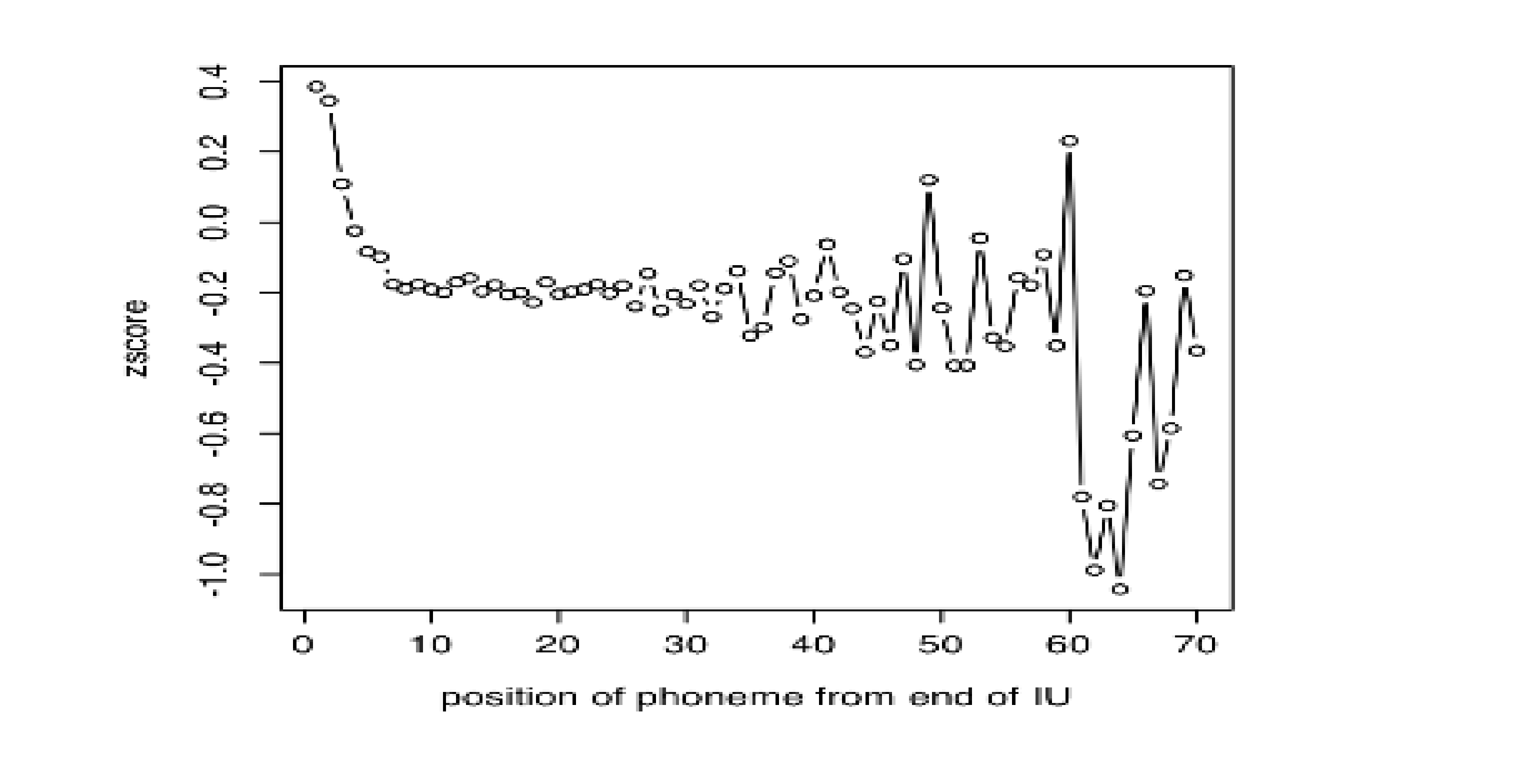


Results - 3

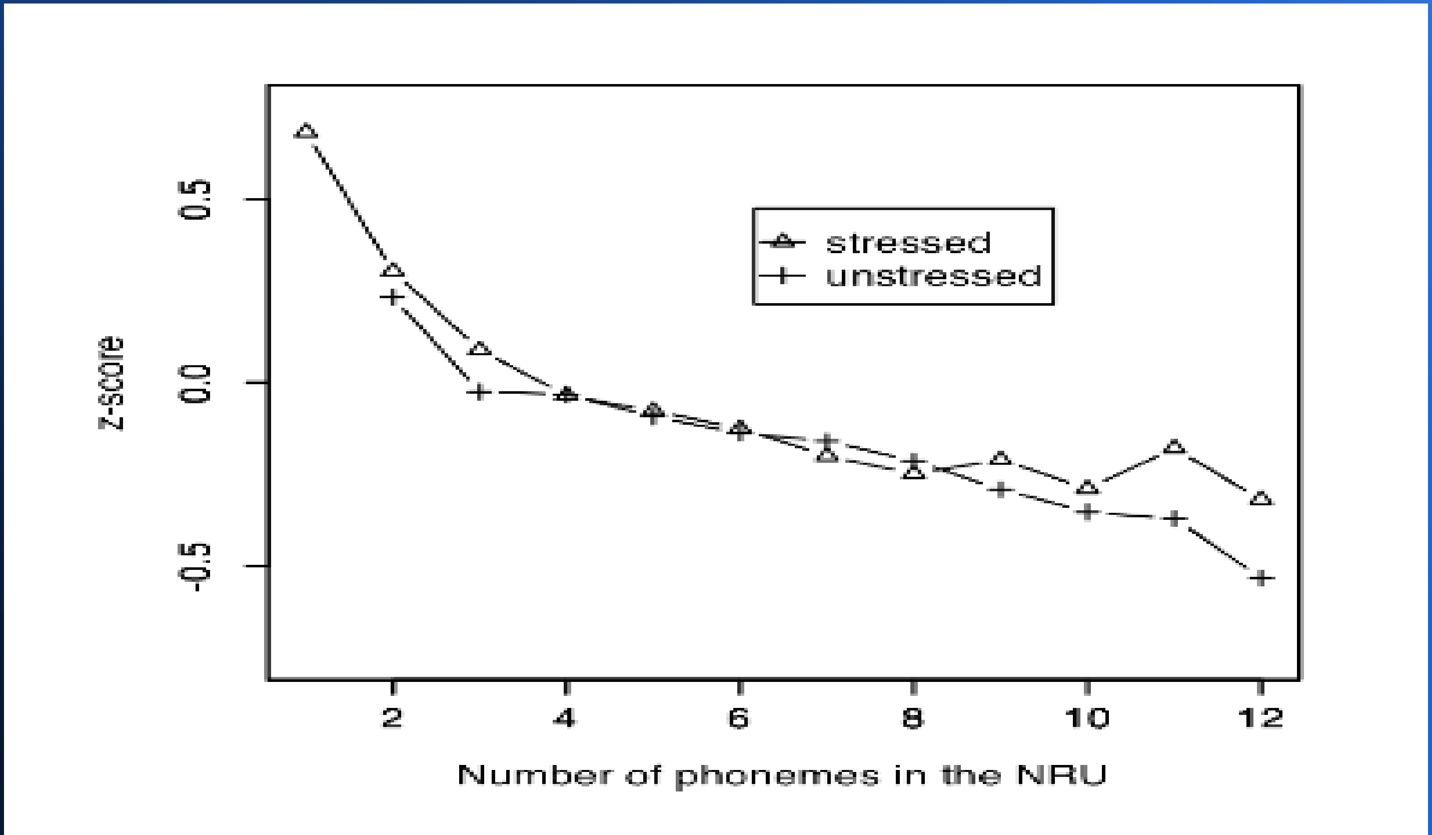
- No simple effect of stress !!!



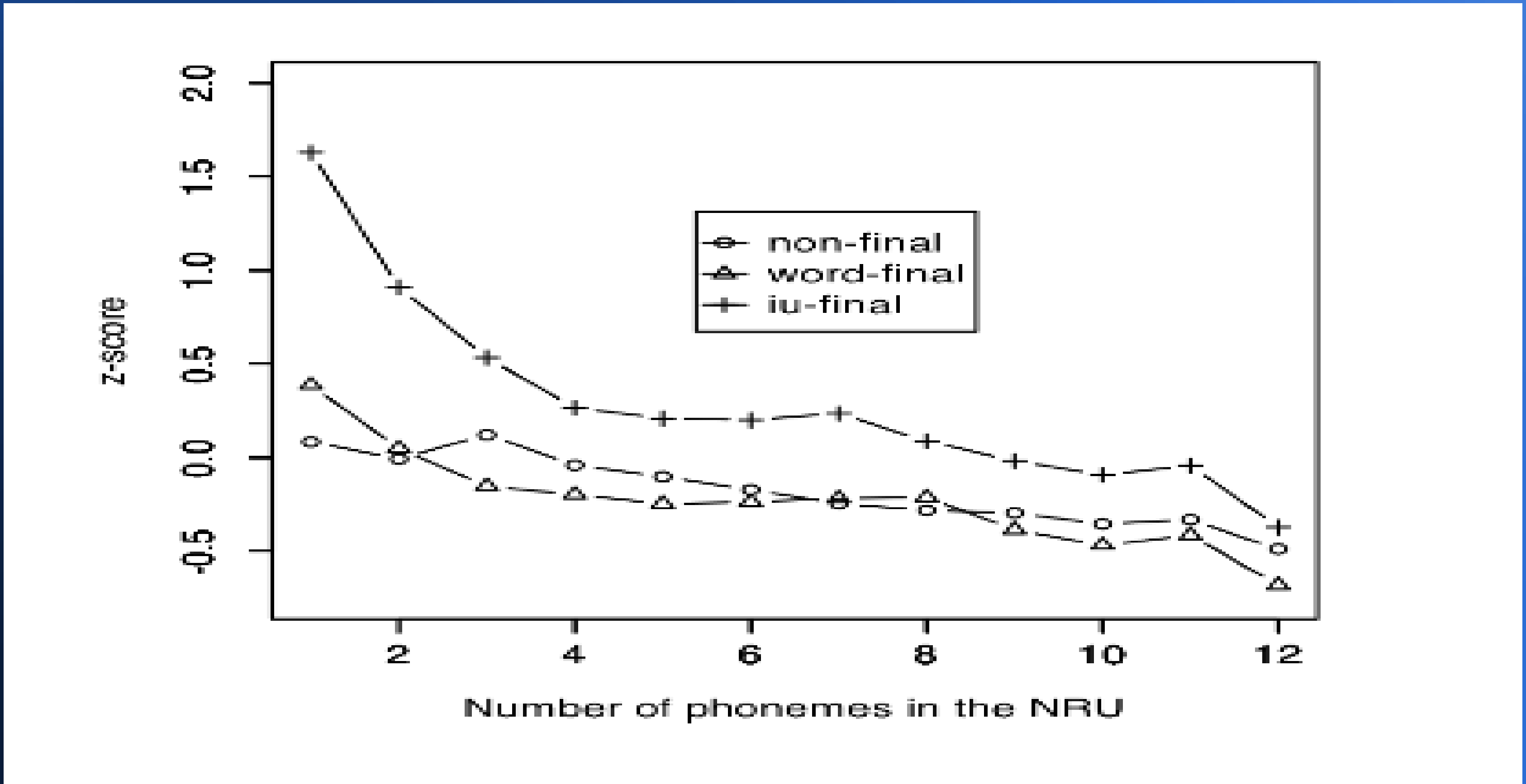
Final lengthening



Excluding last two phonemes of intonation unit



Word-final lengthening?



Conclusions

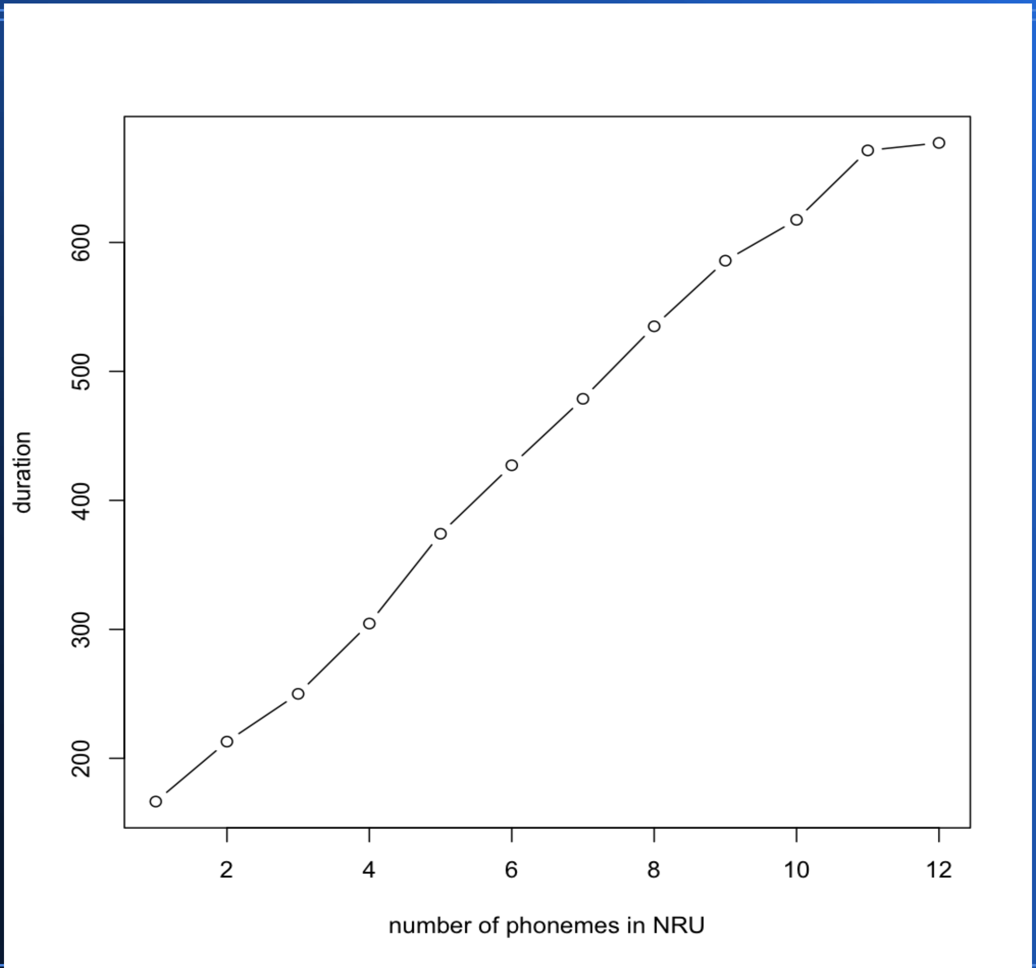
- No compression at level of syllable

  (cf Jassem et al. 1978)
- Phonemes in stressed syllable have NO specific lengthening

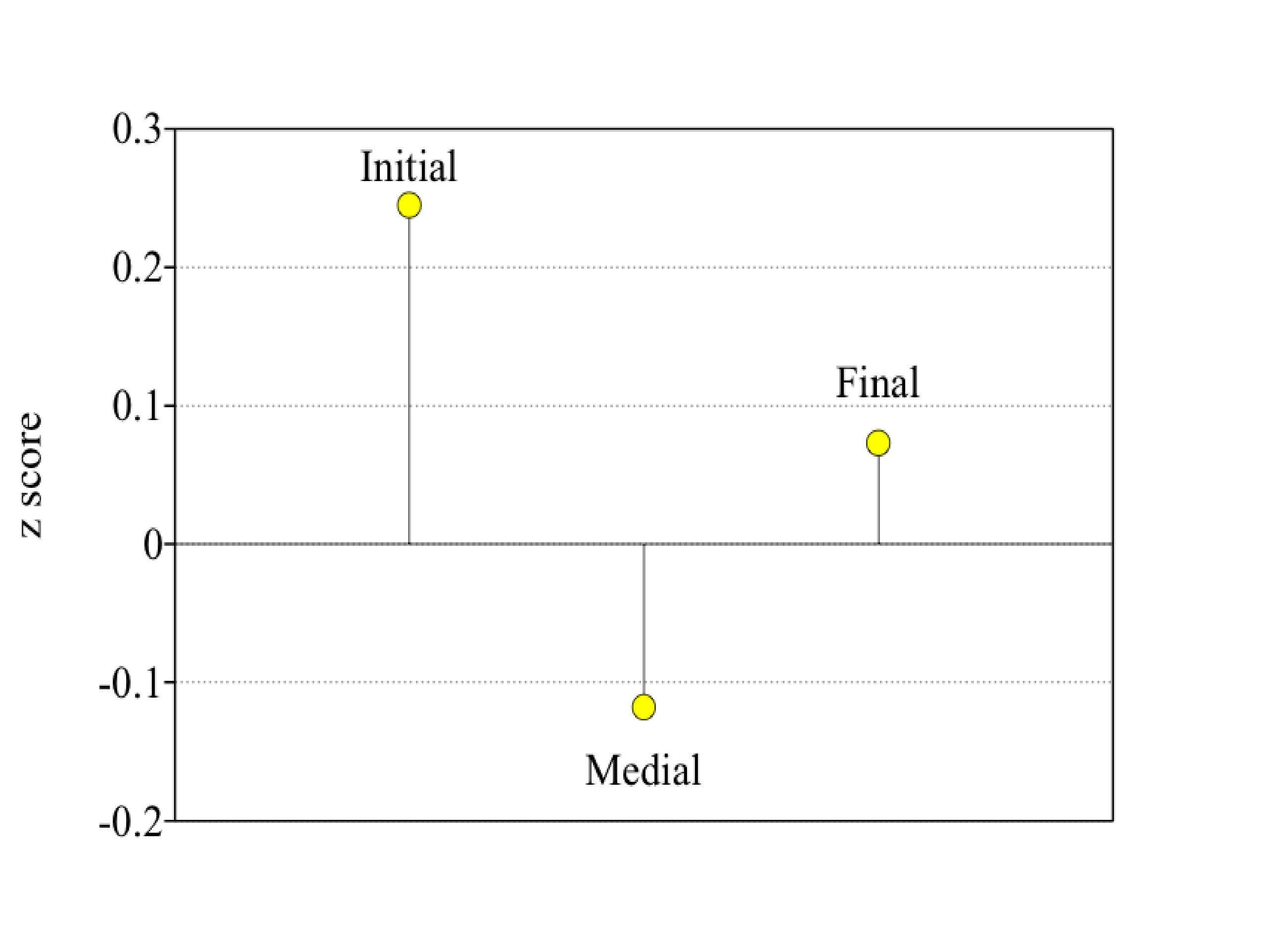
  (cf Jassem 1952!)
- The solution to Klatt's unsolved problem is the Narrow Rhythm Unit (for English)

  (cf Jassem 1952!!!)
- No evidence for specific word-final lengthening

Duration of NRU / number of phonemes in NRU



mean z-score of phoneme / position in NRU



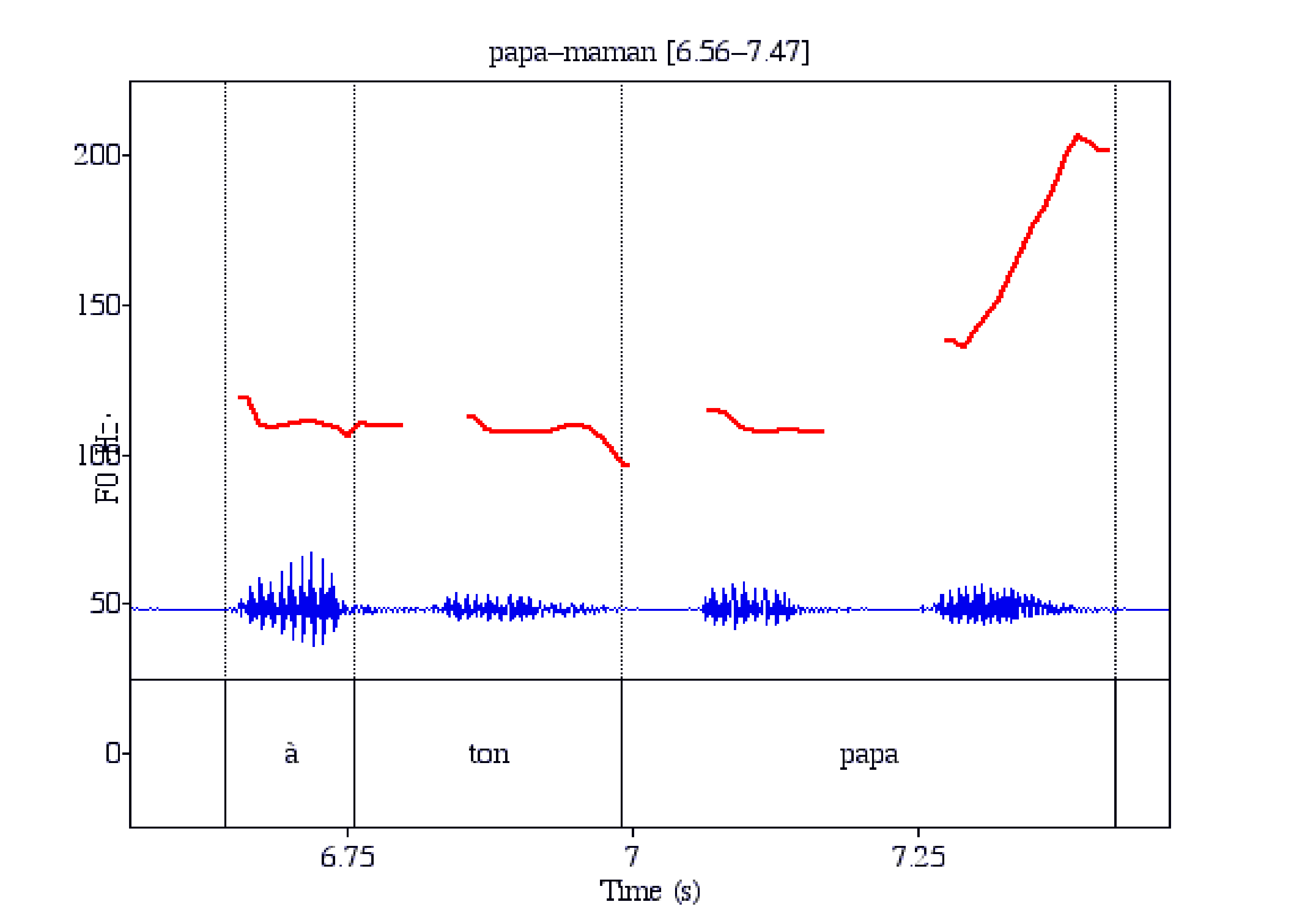
modelling speech melody

- Perception models

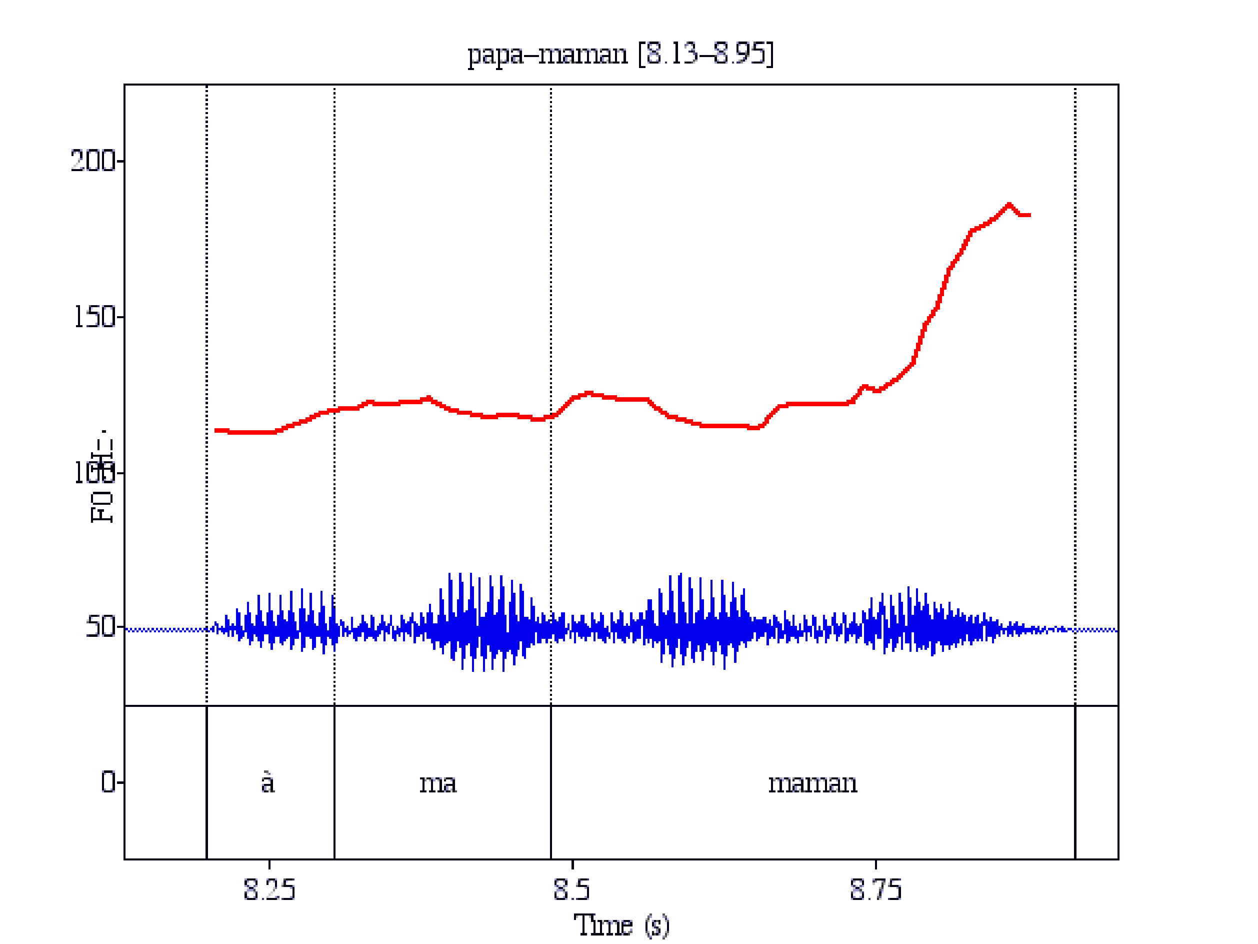
- Production models

- Acoustic models

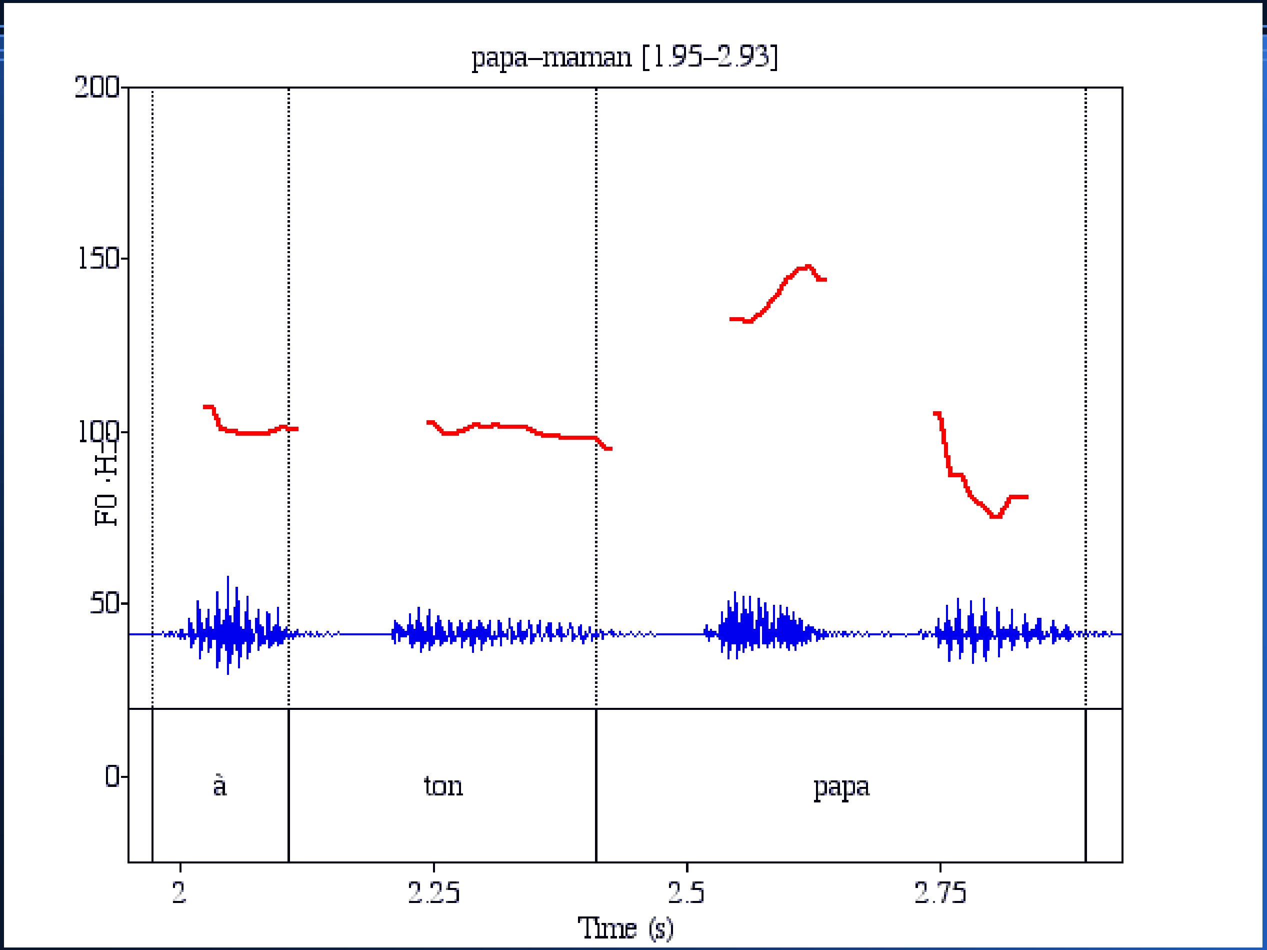
Raw f0


papa–maman [6.56–7.47]
200
150
100
50
0
F0 (Hz)
à
ton
papa
6.75
7
7.25
Time (s)

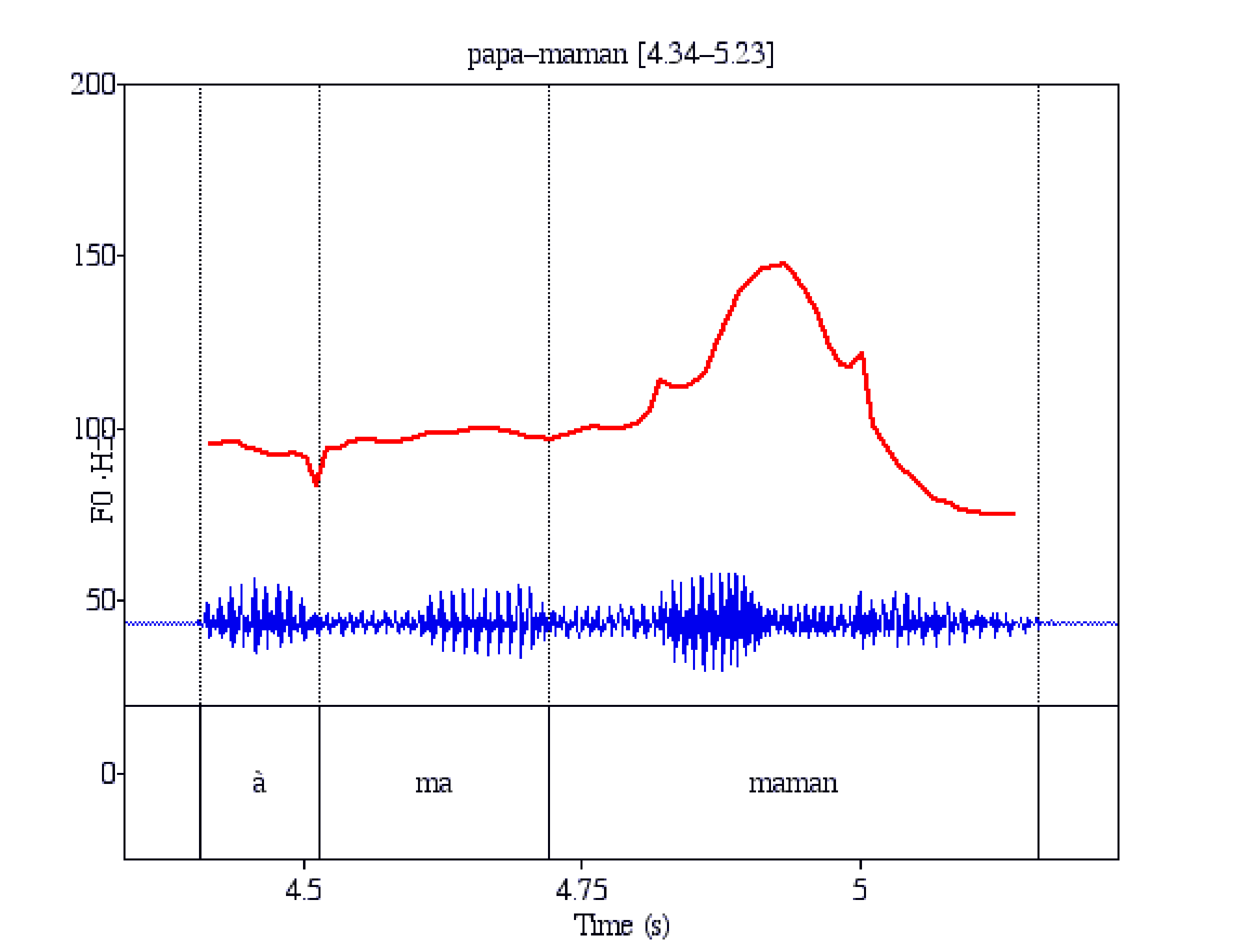
Raw f0


papa–maman [8.13–8.95]
200
150
100
50
0
F0 (Hz)
à
ma
maman
8.25
8.5
8.75
Time (s)

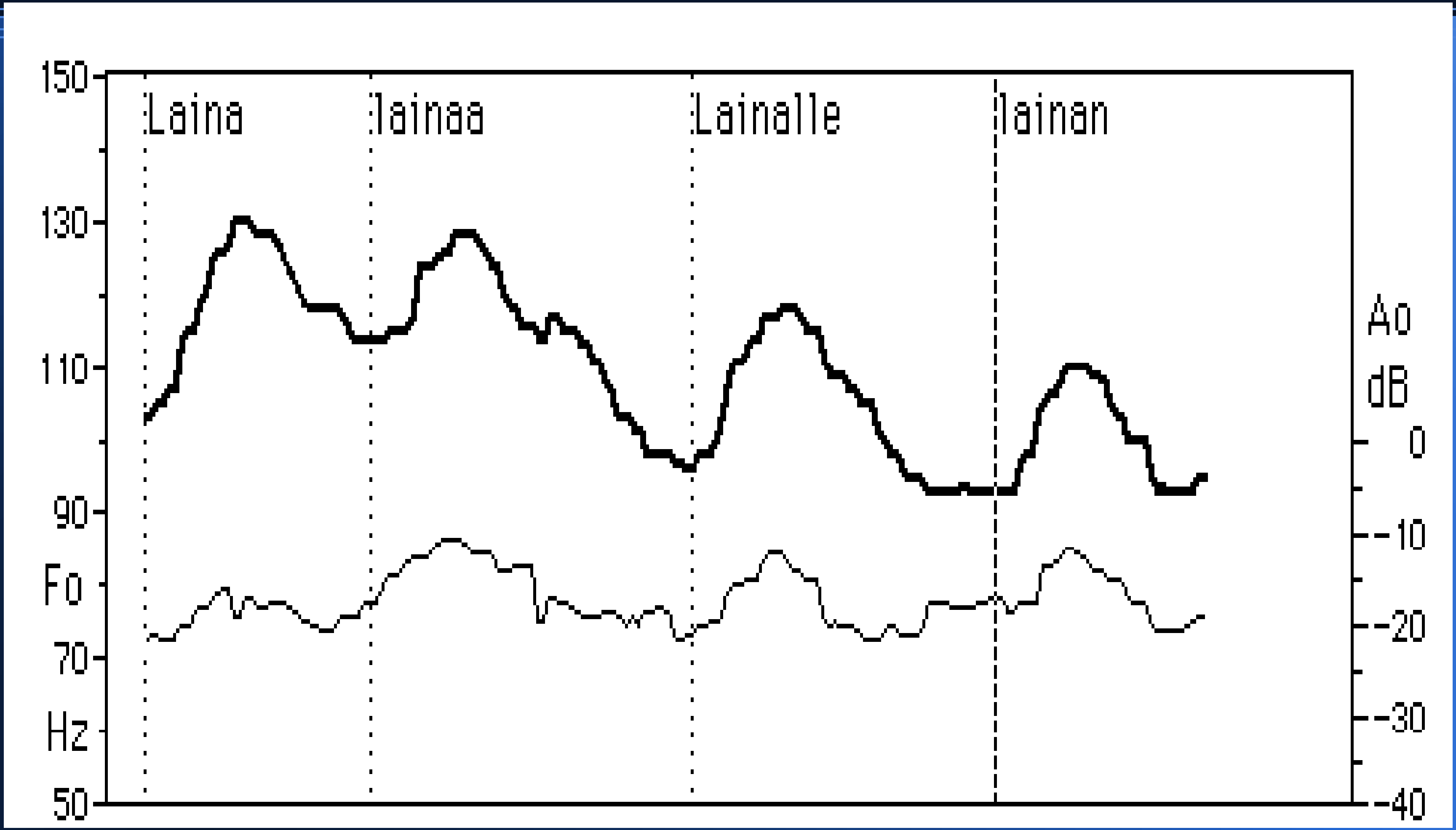
raw f0


papa–maman [1.95–2.93]
200
150
100
50
0
F0 (Hz)
à
ton
papa
2
2.25
2.5
2.75
Time (s)

Raw f0

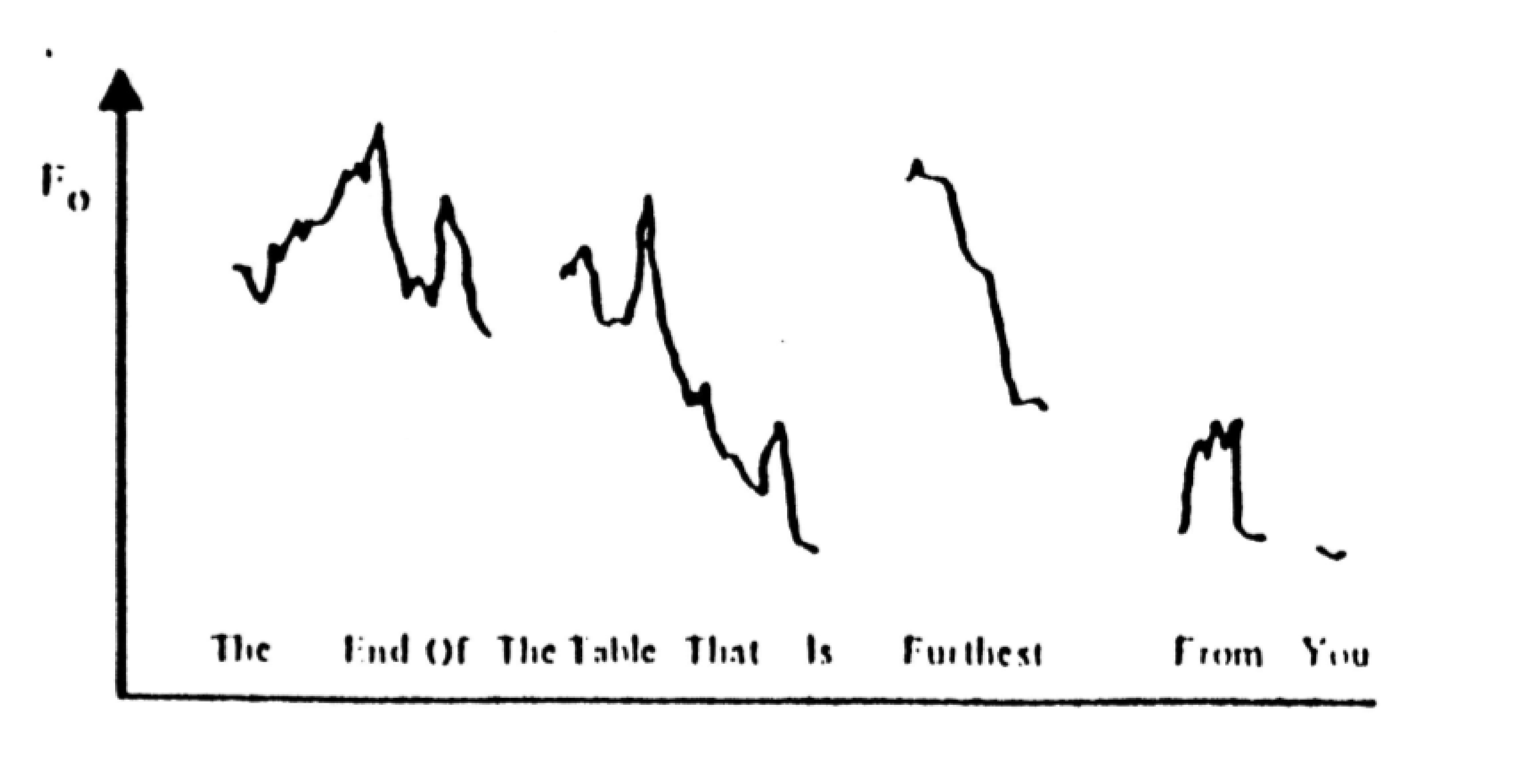

papa–maman [4.34–5.23]
200
150
100
50
0
F0 ·Hz·
à
ma
maman
4.5
4.75
5
Time (s)

Finnish

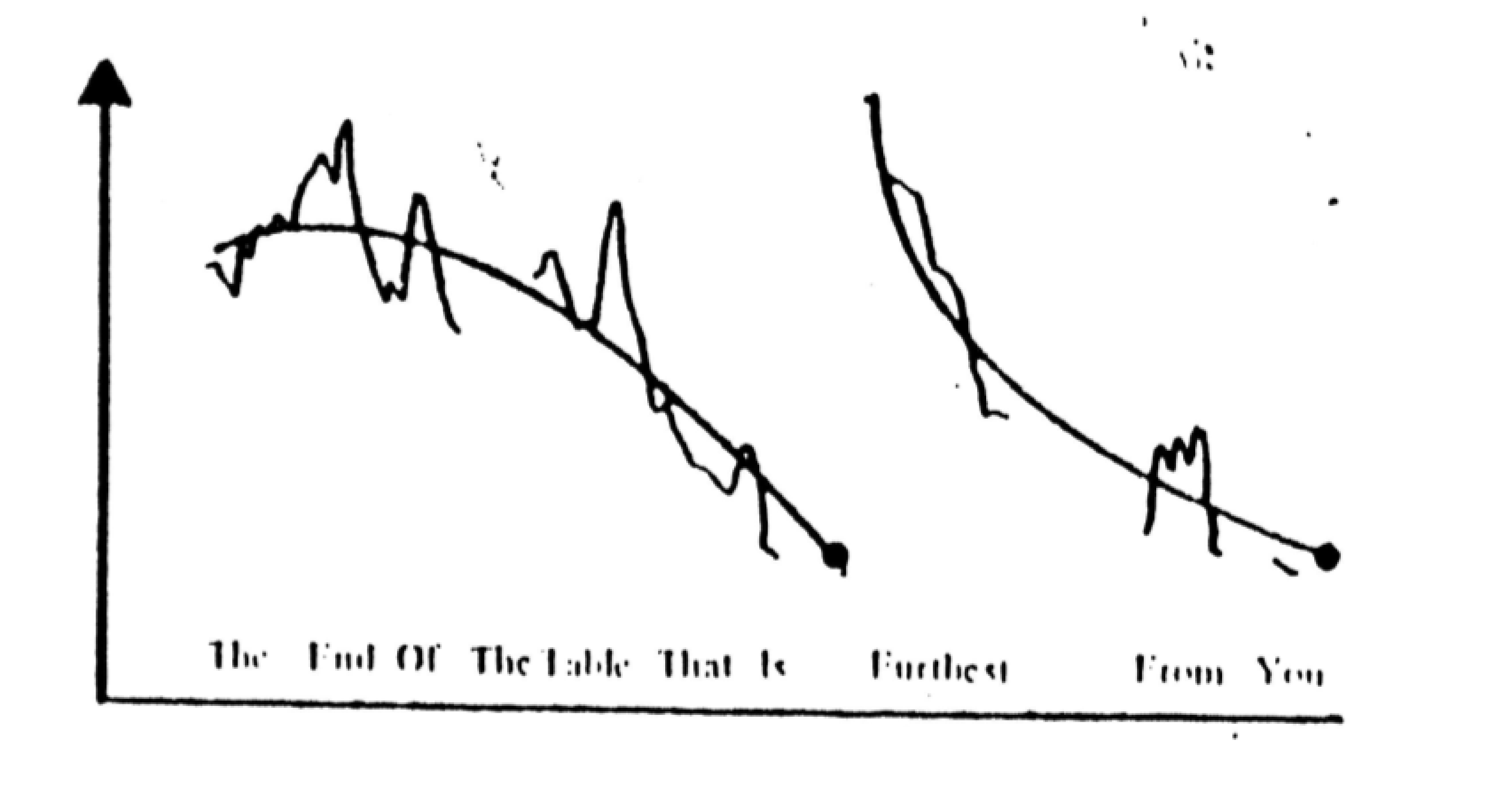


150
130
110
90
Fo
70
Hz
50
Laina
lainaa
Lainalle
lainan
Ao
dB
0
-10
-20
-30
-40

Kloker 1975



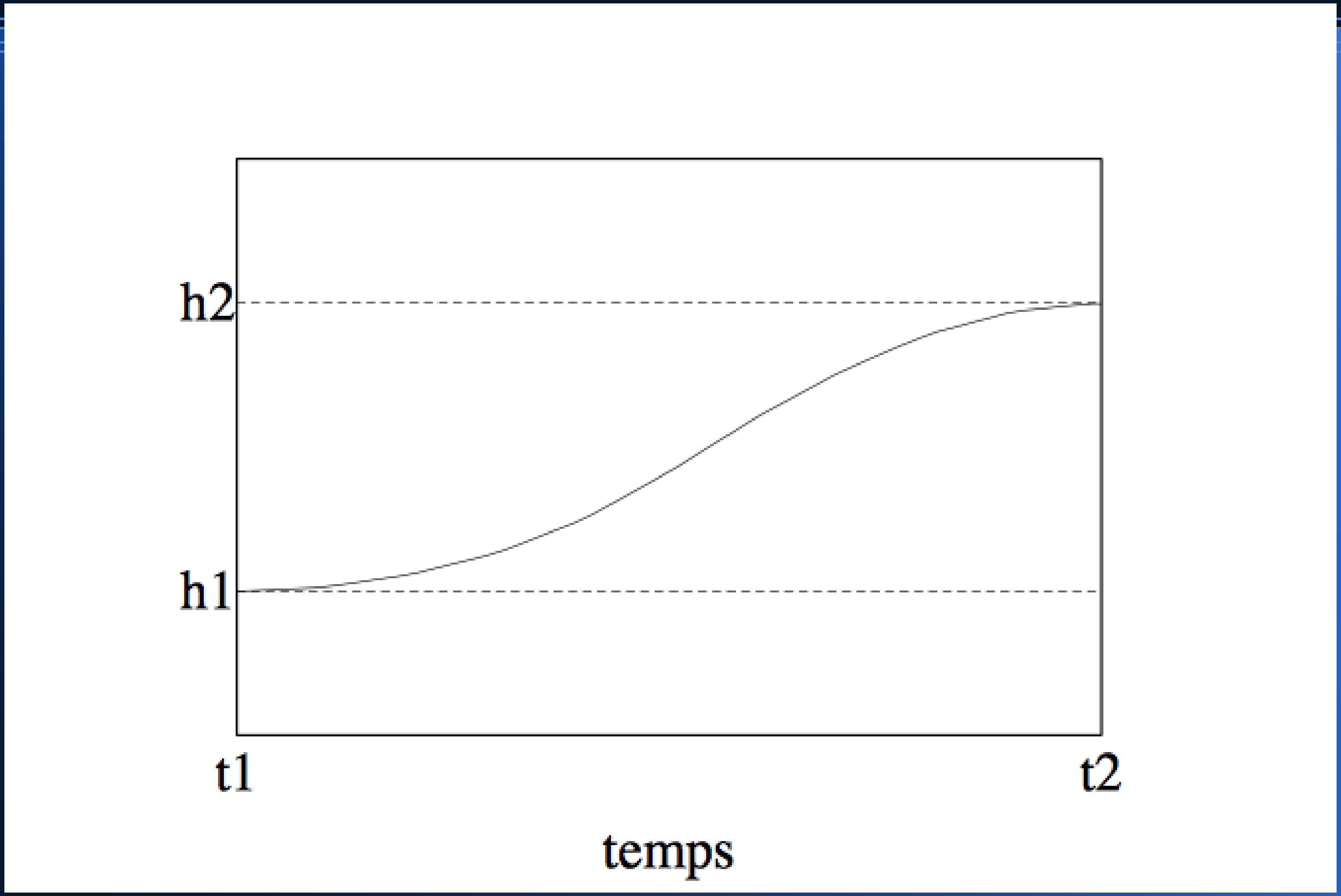
Gamma function: $y = at^{b}e^{ct}$



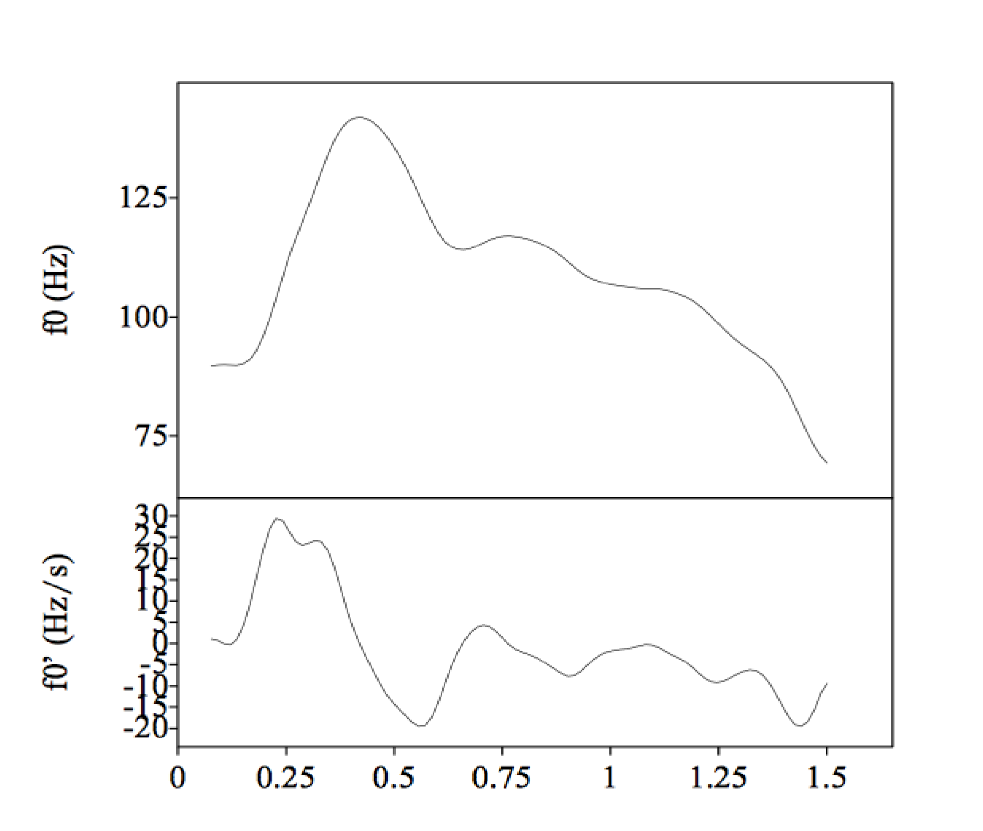
Hirst's law

An acoustic model should not depend on which end of the table you are talking about.

f0 transition



First derivative of raw f0



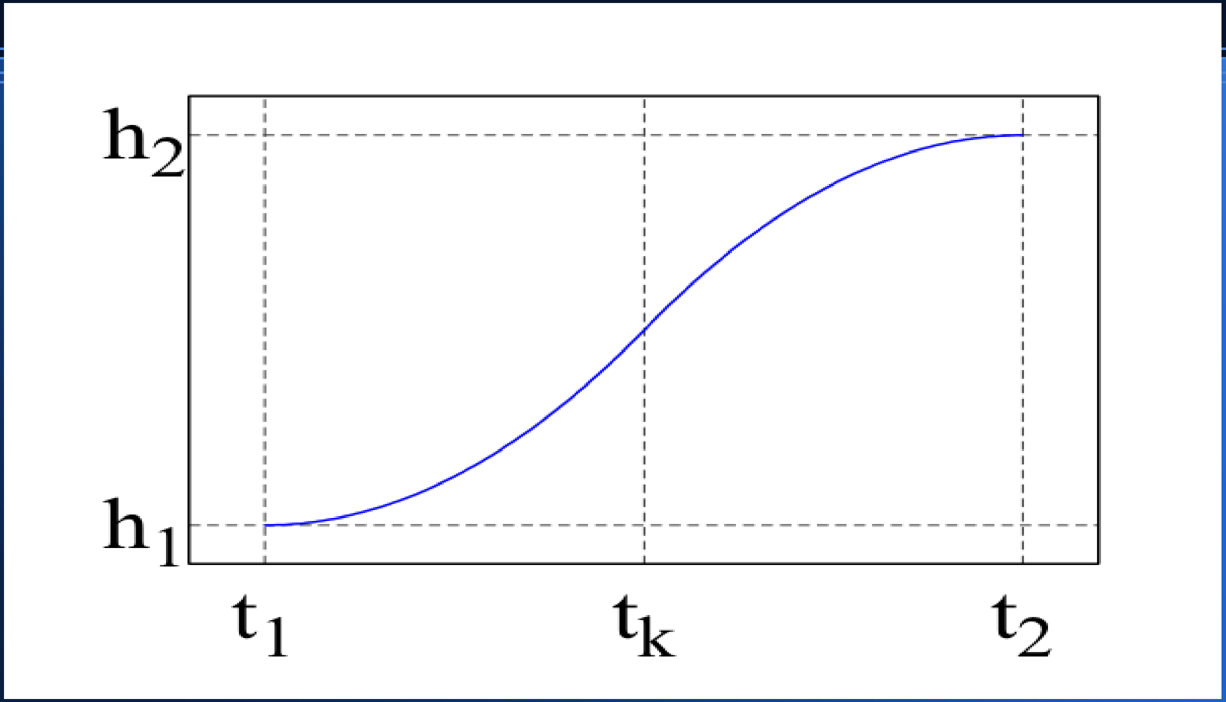
But who stole Jane's bicycle? (ma'ma'ma...)

Quadratic spline function

- Spline function
  - Sequence of functions of degree n, derivatives of which up to n-1 are everywhere continuous

- Quadratic spline
  - Sequence of targets linked by two quadratic functions ($y = ax^2 + bx + c$)

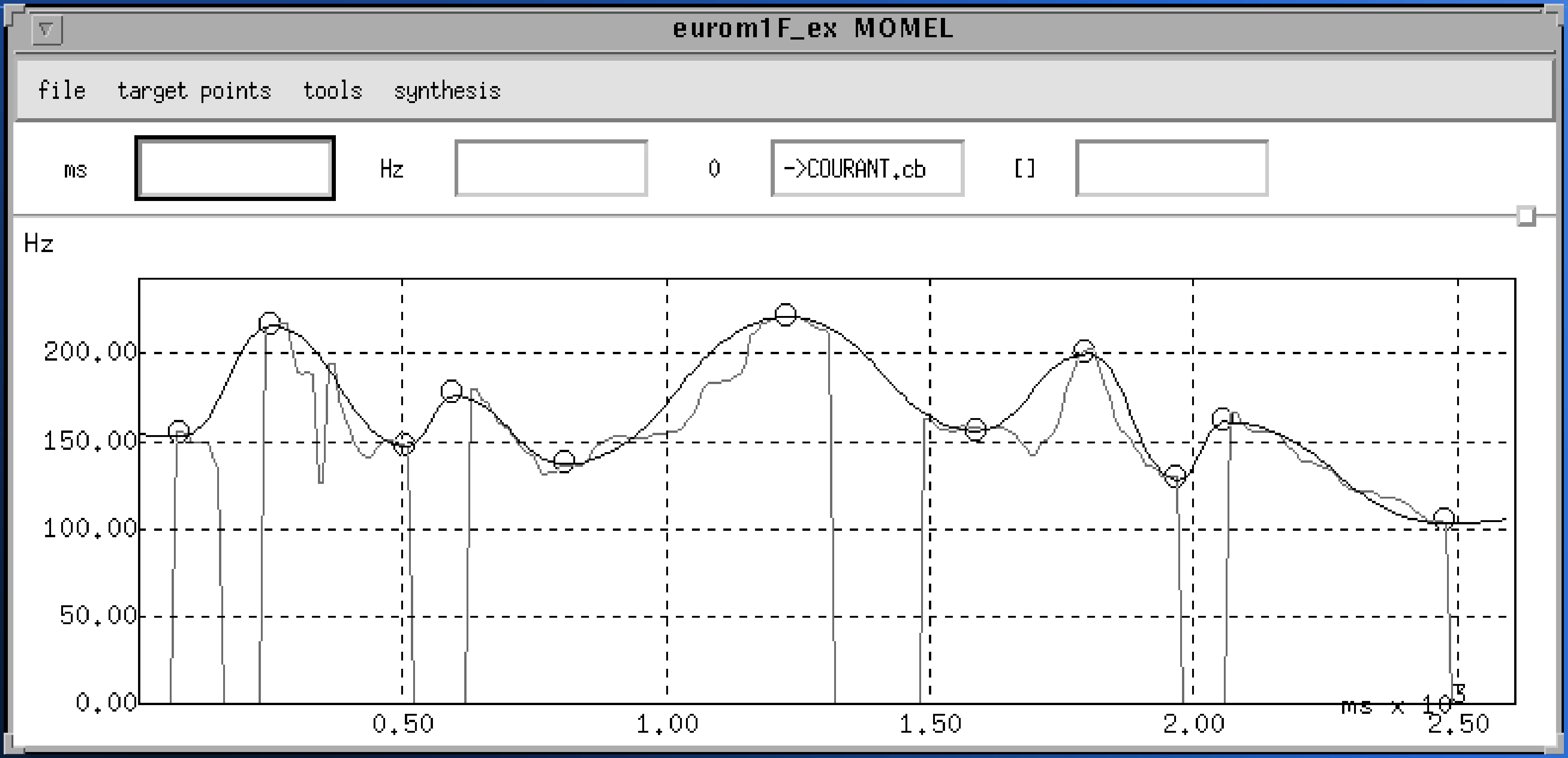
Quadratic spline function



$$y = h_1 + \frac{(h_2-h_1)(x-t_1)^2}{(t_k-t_1)(t_2-t_1)}$$

$$y = h_2 + \frac{(h_1-h_2)(x-t_2)^2}{(t_k-t_2)(t_1-t_2)}$$

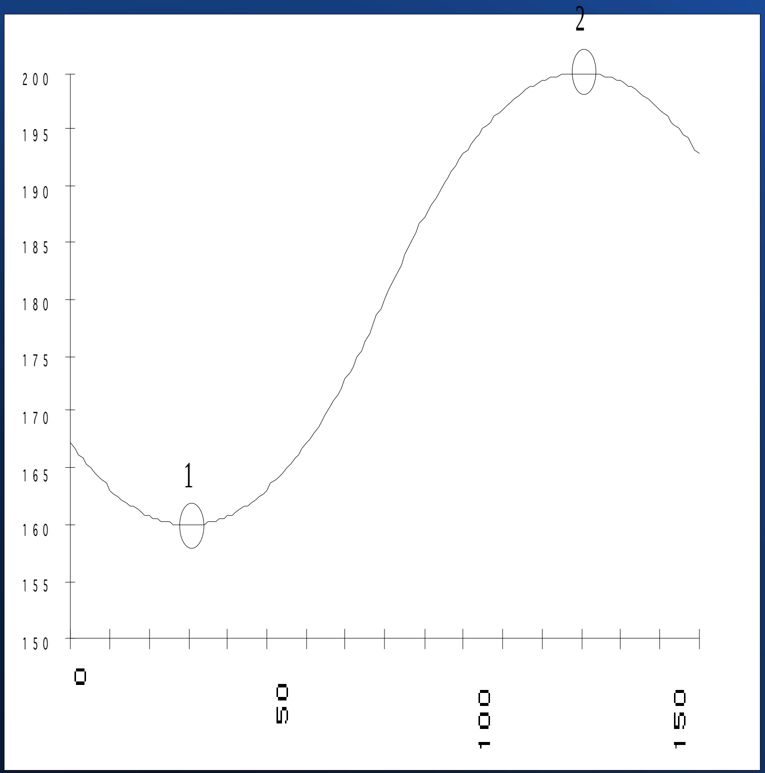
Quadratic spline function

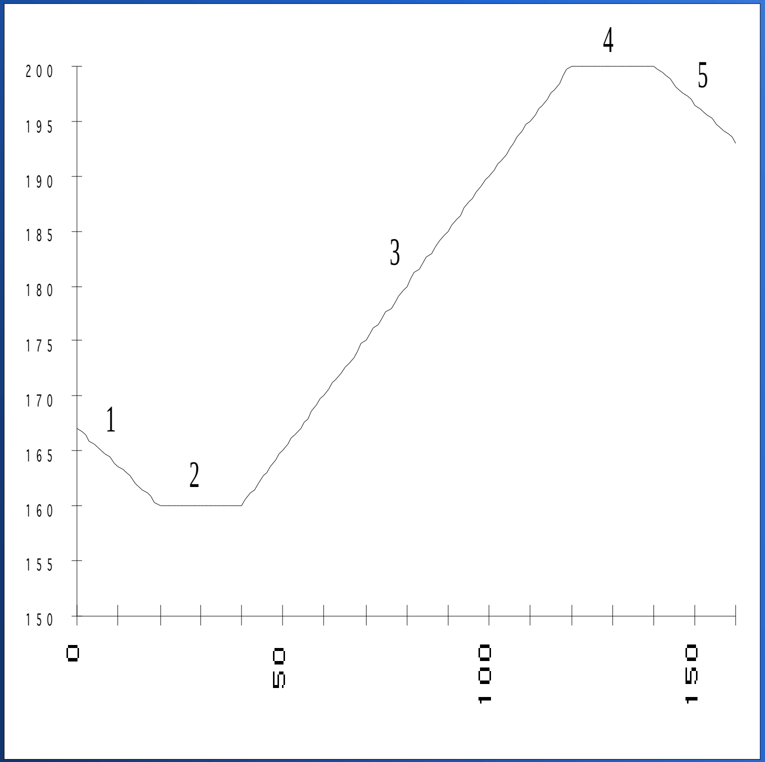


Il faut que je sois à Grenoble, Samedi vers quinze heures

Curves vs. straight lines

- 't Hart 1991





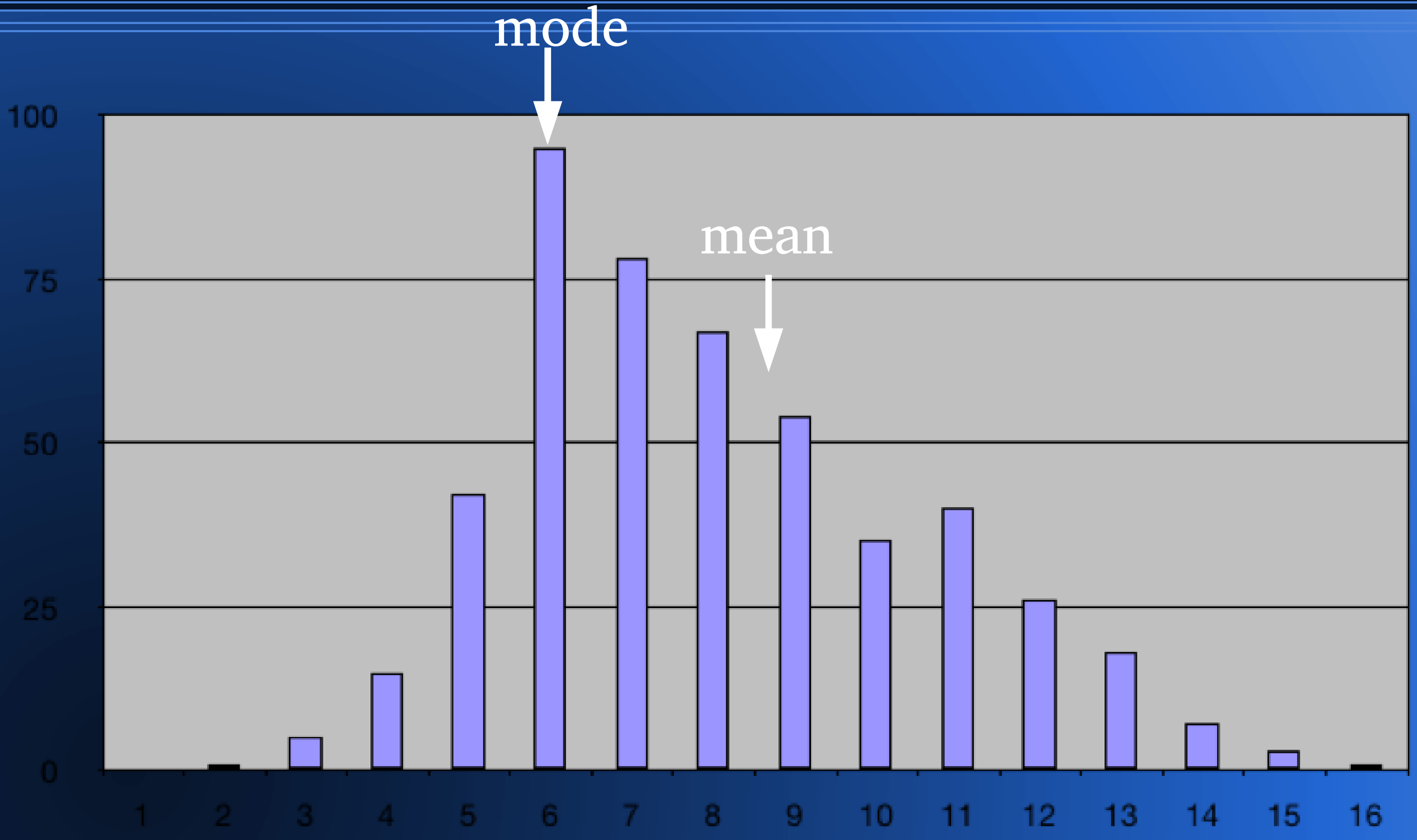
Automatic Momel

- Hirst & Espesser 1993

  Asymmetric quadratic modal regression

  - Modal
  - Quadratic
  - Asymmetric

Mean and Mode



Mean and Mode

- *Mean*

  value minimising sum of squares of differences from data

- *Mode*

  value minimising number of cases more than Δ from data

Generalise to function

- *Linear regression*

  function minimising sum of squares of differences from data

- *Modal regression*

  function minimising number of cases more than Δ from data

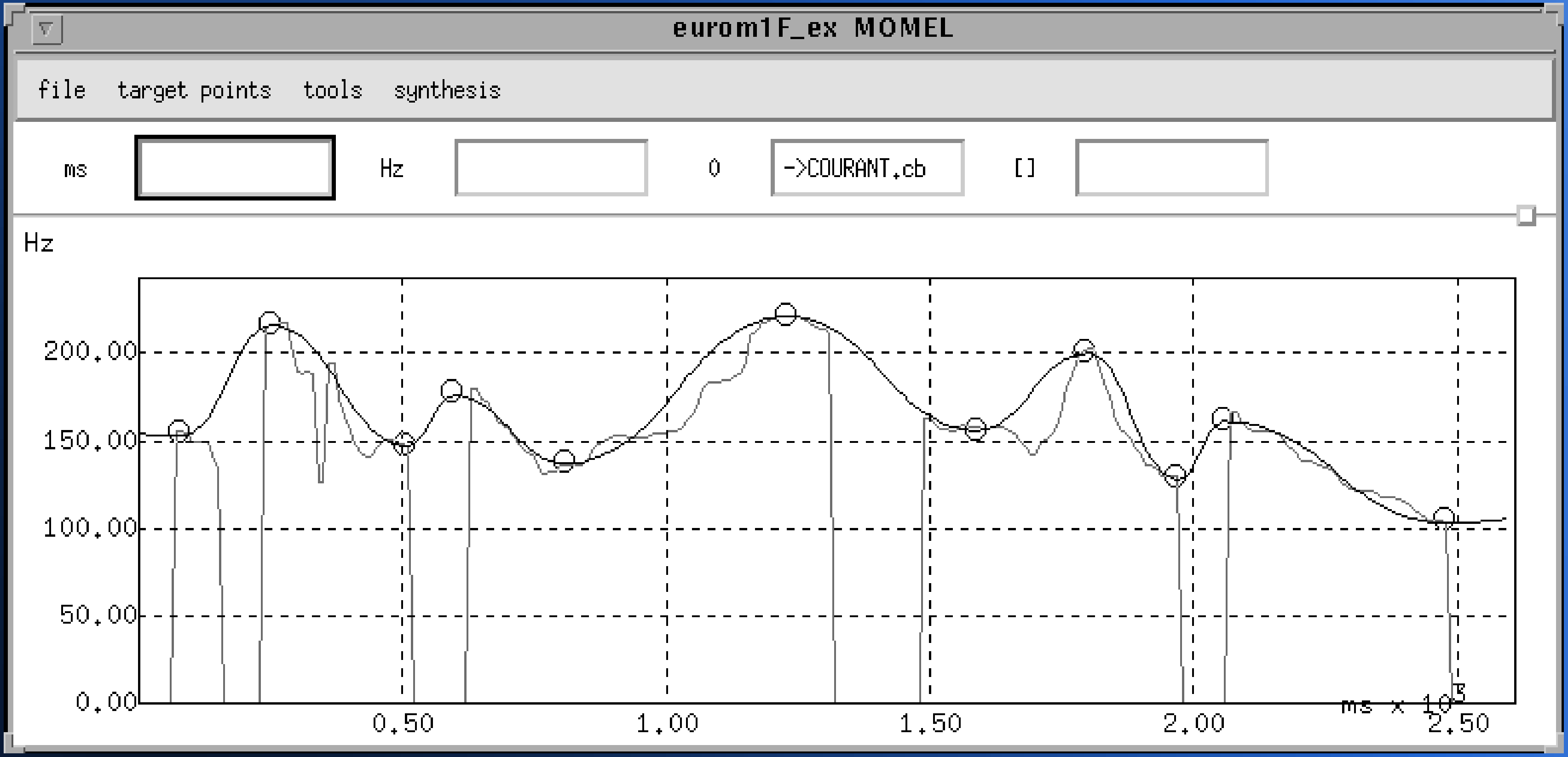
Asymmetric regression

- *no* values more than Δ above the function
- Minimise number of values more than Δ below it
- Here, function is

$$f = at^2 + bt + c$$

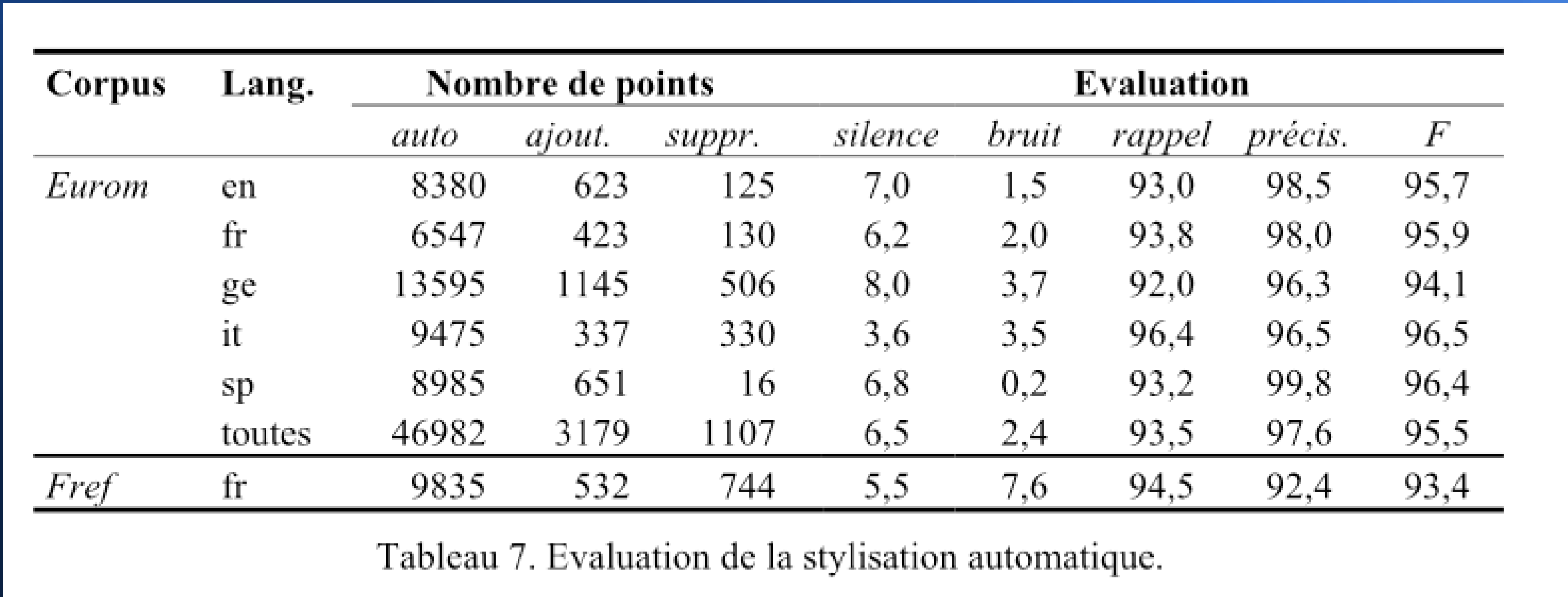
Momel

- Hirst & Espesser 1993



Evaluation of Momel

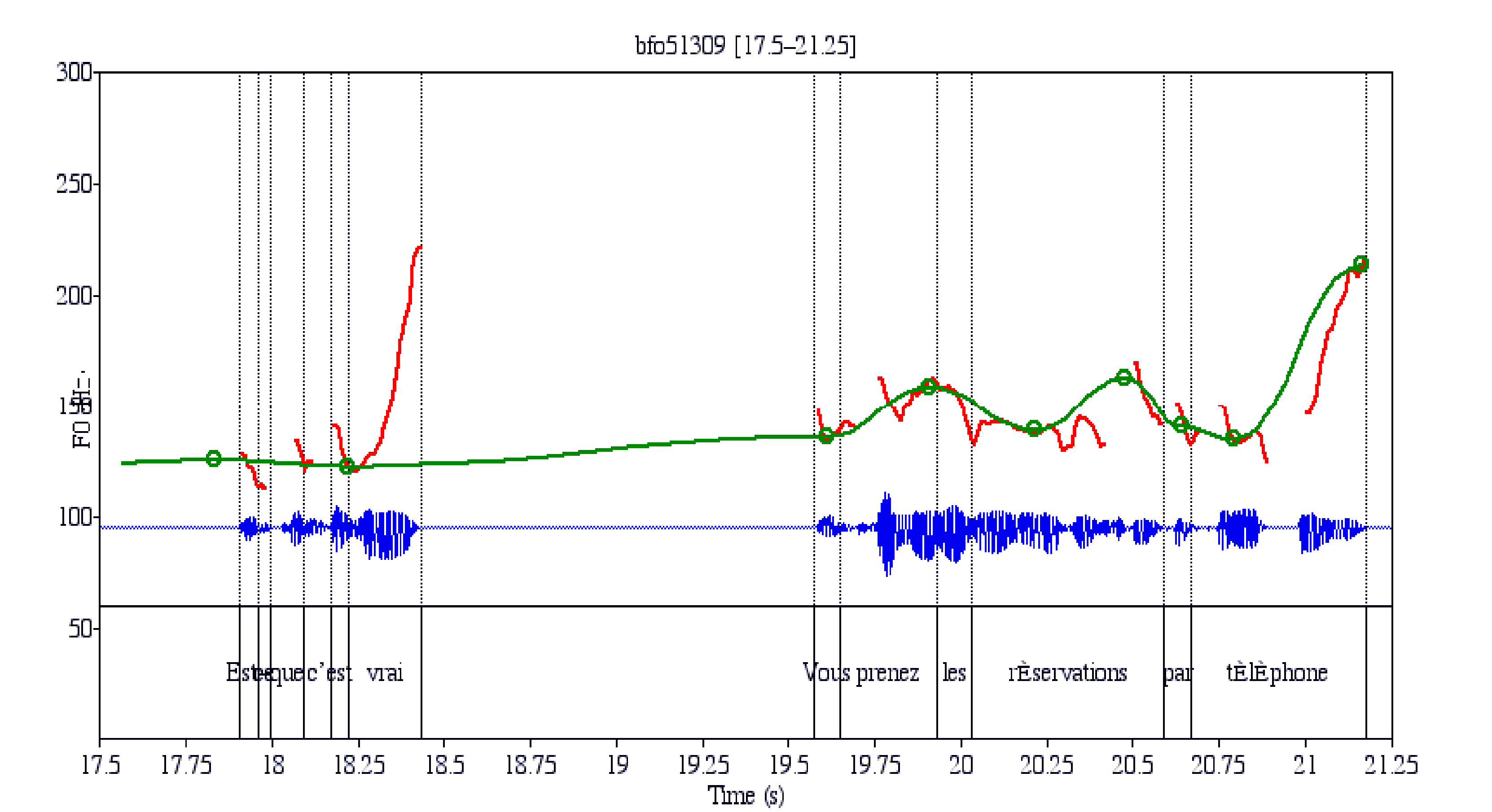
- Estelle Campione, 2001



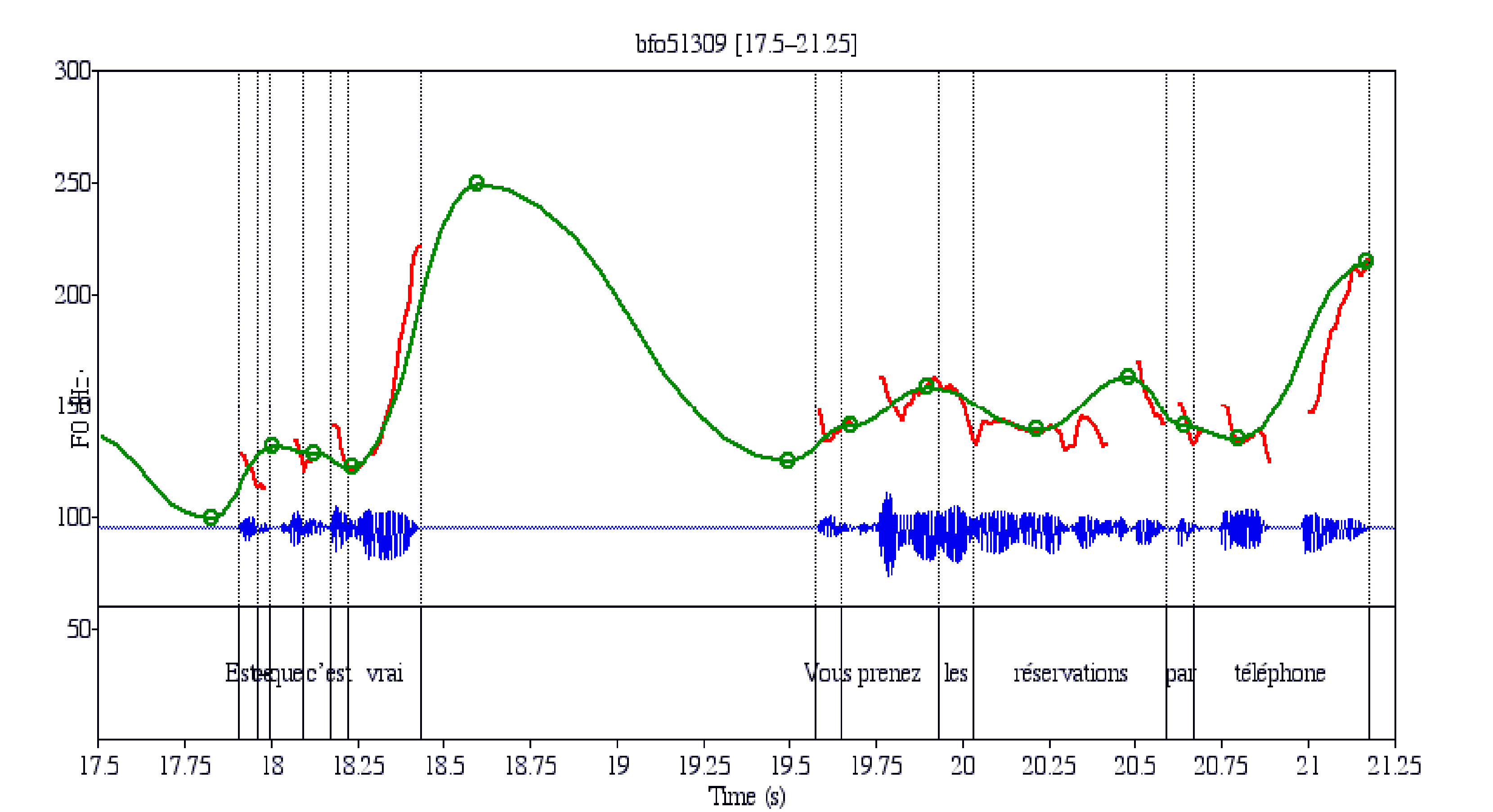
Corpus	Lang.	Nombre de points			Evaluation				
		auto	*ajout.*	*suppr.*	*silence*	*bruit*	*rappel*	*précis.*	*F*
Eurom	en	8380	623	125	7,0	1,5	93,0	98,5	95,7
	fr	6547	423	130	6,2	2,0	93,8	98,0	95,9
	ge	13595	1145	506	8,0	3,7	92,0	96,3	94,1
	it	9475	337	330	3,6	3,5	96,4	96,5	96,5
	sp	8985	651	16	6,8	0,2	93,2	99,8	96,4
	toutes	46982	3179	1107	6,5	2,4	93,5	97,6	95,5
Fref	fr	9835	532	744	5,5	7,6	94,5	92,4	93,4

Tableau 7. Evaluation de la stylisation automatique.

Improved algorithm



Improved algorithm



Momel – theory neutral?

- Theory friendly
- used for
  - Fujisaki model (Mixdorff)
  - ToBI (Maghbouleh, Wightman & Cambell, Cho (K-ToBI)
  - INTSINT

INTSINT

- An *INternational Transcription System for INTonation*
- Based on minimal pitch contrasts in descriptions of intonation patterns
- Used in *Hirst & Di Cristo 1998* for 9 different languages
  - *British English, Spanish, European Portuguese, Brazilian Portuguese, French, Romanian, Russian, Moroccan Arabic and Japanese*
- Extension for duration and rhythm

Basic INTSINT

- ***Absolute tones***

  T(op) M(id) B(ottom)

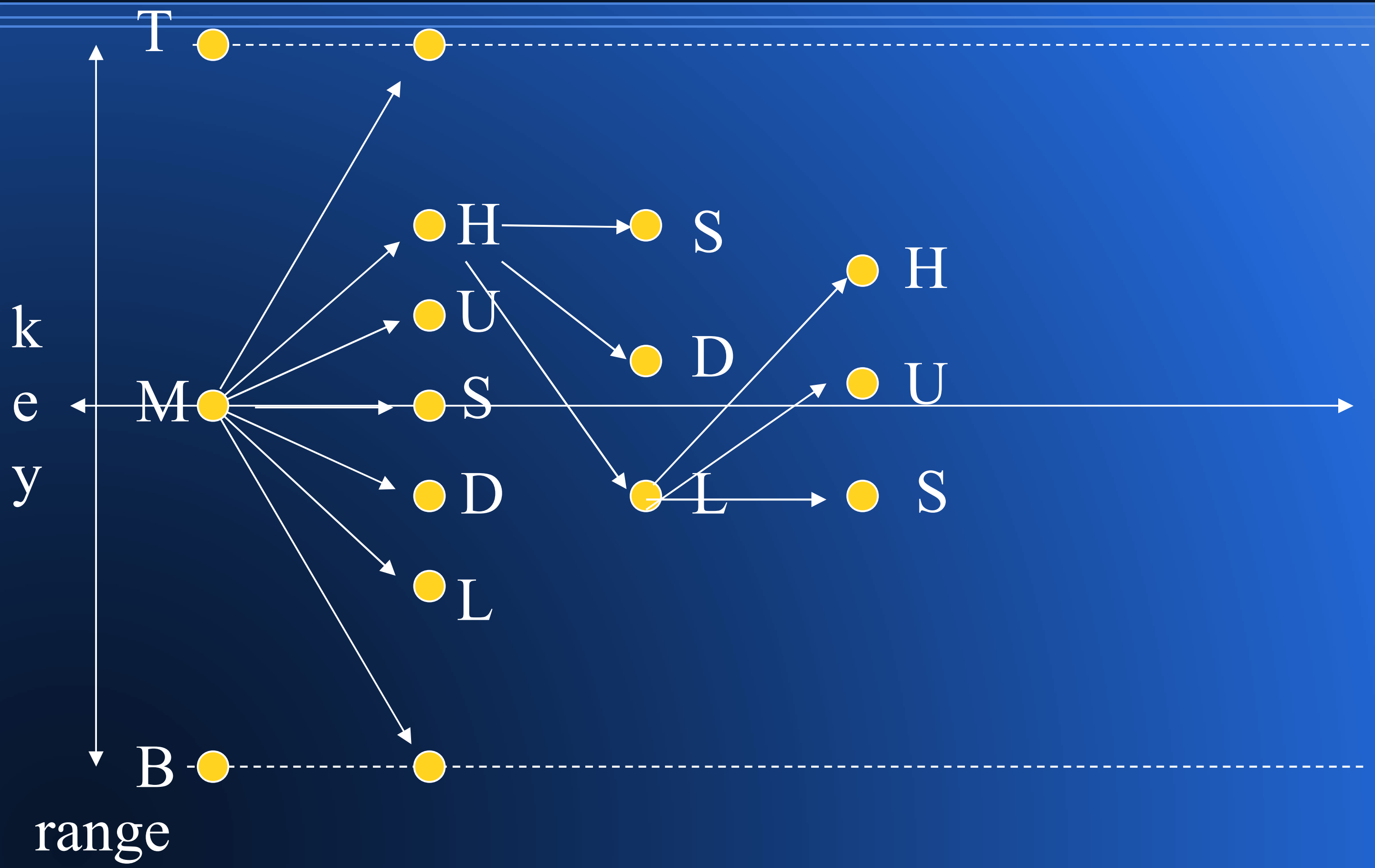
- ***Relative tones***

  H(igher) S(ame) L(ower)

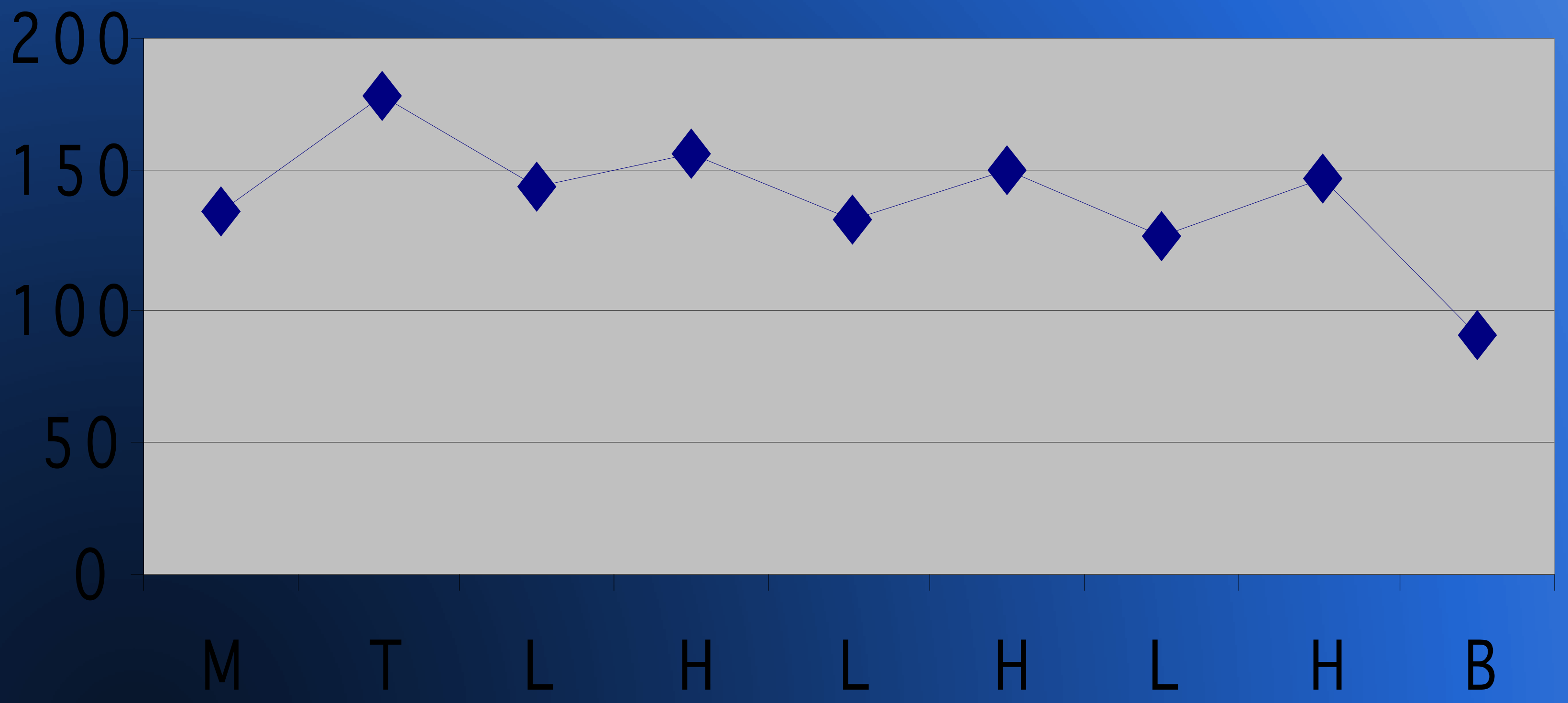
- ***Iterative relative tones***

  U(pstepped) D(ownstepped)

2 speaker parameters: *Hirst 2005*



downdrift



Intsint to Momel

key : k (Hertz), range: r (octaves)

- T = $k * \sqrt{2}^{r}$
- M = k
- B = $k/\sqrt{2}^{r}$
- H = $\sqrt{}(P * T)$
- S = P
- L = $\sqrt{}(P * B)$
- U = $\sqrt{}(P * \sqrt{}(P * T))$
- D = $\sqrt{}(P * \sqrt{}(P * T))$

Momel to Intsint

Perl script

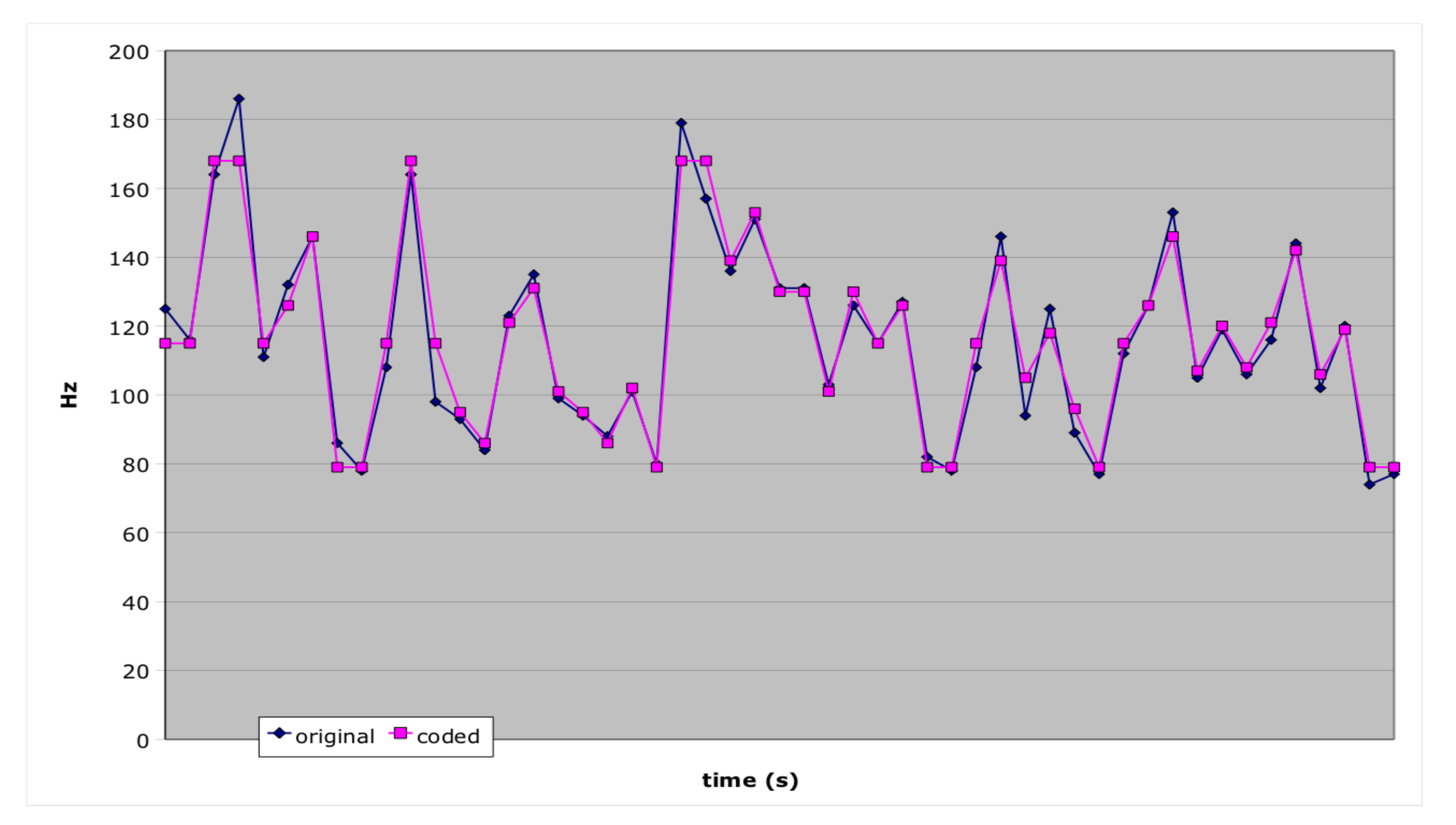
Optimal coding of target points within parameter space:

- range = 0.5…2.5 octaves (step: 0.1)
- key = mean ±50 Hz (step: 1)

output

```
; French.intsint created on Mon Aug 24 10:25:05 2009 by intsint.pl 2.11
; from French.momel
;    27 values   mean = 297
<parameter range=1.4>
<parameter key=309>
0.469 B 190 190
0.989 M 354 309
1.081 H 429 394
1.464 L 252 274
2.014 T 500 502
2.353 L 275 309
```

original vs coded targets



variety of intonation systems

- prosodic forms are universal
- prosodic functions are quasi-universal
- variety of intonation systems is from the mapping between function and form

analysis by synthesis

- Prosodic functions -->
- Underlying (abstract) phonological representation -->
- Surface phonological representation (discrete phonetic) (INTSINT) -->
- Phonetic (continuous) representation (Momel) -->
- Acoustic output

Non-emphatic intonation

	Pre-head	Head + Body		Nucleus + Tail
English US	[M	[H L] [H L]	…	[H B]]
			…	[H B]]
English UK	[M	[H] [D]	…	[D B] H]
			…	[D B] H]
French	[M	[S H] [L H]	…	[D B]]
			…	[D H]]

Parametric model

	TU	*IU(+term)*	*IU (-term)*
English	[Ss$_0$]	TU$_1$	TU$_1$
	[HL]	[L L]	[LH]
French	[s$_0$S]	TU$_1$	TU$_1$
	[LH]	[LL]	[LH]

Sample derivation

- *Functional representation*

  |But she 'didn't 'say she was 'coming 'home on °Saturday +

- *Underlying phonological representation*

```
[But she [didn't] [say she was] [coming] [home on] [Saturday]]

[L       [H    L ] [H         L][H    L ] [H     L ] [H  L]  H]
```

- *Surface phonological representation*

```
[But she [didn't] [say she was] [coming] [home on] [Saturday]]

[M       [  H   ] [    D      ] [  D    ] [  D     ] [ D  B]  T]
```

- *Phonetic representation*

```
[But she [didn't] [say she was] [coming] [home on] [Saturday]]

[ 127      151      133          120      112       106 90 180 ]
```

- *Acoustic representation…*

Thank you for listening

If you have any questions we don't have time for now

daniel.hirst@lpl-aix.fr