Measuring and synthesising expressivity: Some tools to analyse and simulate phonostyle

J.-Ph. Goldman - University of Geneva

EMUS Workshop
05.05.2008

Outline

1. Expressivity
   - What is, how to characterize expressivity ?
   - How do we produce and perceive expressivity ?
2. Analysing phonostyle
   - Application: radiophonic vs. read-aloud styles
3. Simulating phonostyle

Goals

- Answer these questions:
  - What is expressivity / what is phonostyle ?
  - Is it a set of features along speech
    or sporadic, specific, temporally-targeted signs ?
  - Which perception of expressivity do we have?

- Figure out attested prosodic objects like:
  - Prominence
  - Major Intonation Units
- And make a form and function categorization but with
  - no theoretical model (data injection)
  - no (or few) language dependency

1.Expressivity: functions of prosody

phonosyntax

phonostyle

- **Linguistics**
  - Lexical, syntactic, semantic and discourse organisation
- **Para-linguistics**
  - Speech style
    - spontaneous, semi-prepared, repeated, read
    - familiar, didactic, formal…
  - Behaviour: irony, incredulity, excitation, seriousness,…
  - Languages habits
    - idiolect, dialect, sociolect
- **Extra-linguistics**
  - Emotions : joy, anger, sadness, surprise,…
  - Identity : physiology and control of larynx, articulators…

Inherent

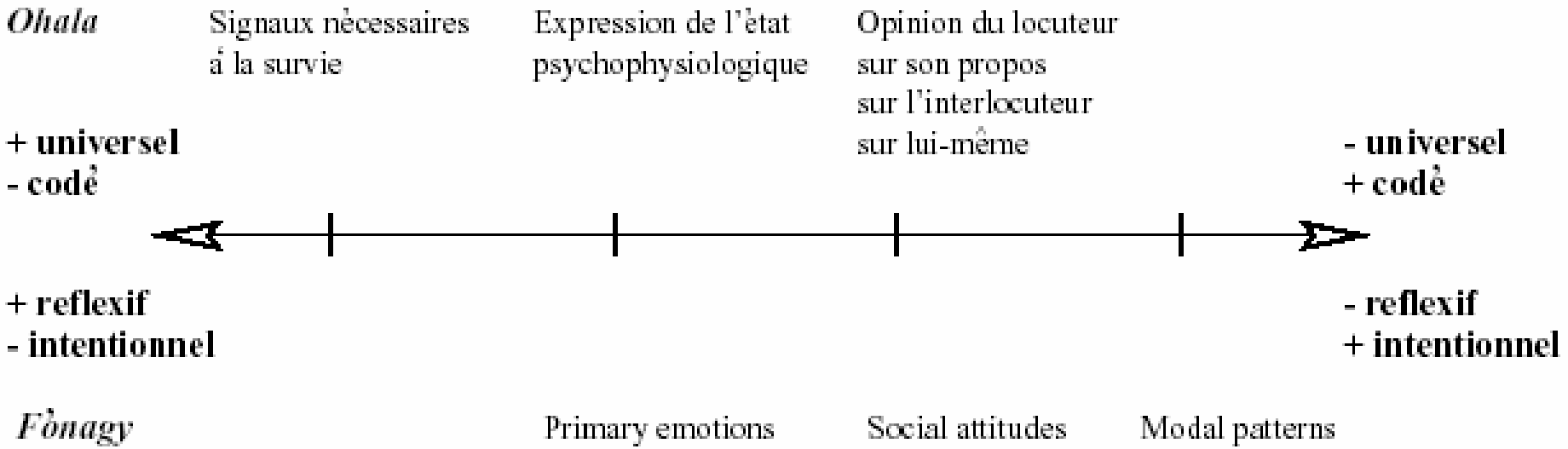
Conscious

Unconscious

Anchored

\+ ..... Variable ..... –

1. E. is not in or out or para-linguistics : it's all over, it's gradual



- Dimensions/parameters of expressivity:
  - Control (reflexive ←→ intentional)
    - Consciousness
    - Actable
  - Conventional (universal ⇔ coded)
  - Variable over various periods of time

1. Which parameters for expressivity and which period of time ?

Parameters

- Language
- Segmental
- Prosodic (acoustic)
  - F0
  - Time
  - Intensity
  - VQ
  - Articulation
  - …

- Which prosodic objects to consider ?
  - Syllable
  - Intonation group
  - Periods
  - Larger units?...
  - Extra speech objects: pause, breath

→ Which span of study?

→ Which minimum duration?

→ Which scattering ?

→ Which agreement across speakers ?

2.Analysing

- Framework: *atheoretical semi-automatic approach*
  - Manual annotation for bootstrapped training
  - Automatic annotation tools
    → time-saving, reproducible, coherent
  - Bidirectional validation by error diagnosis
    → corpus development , theory tackling

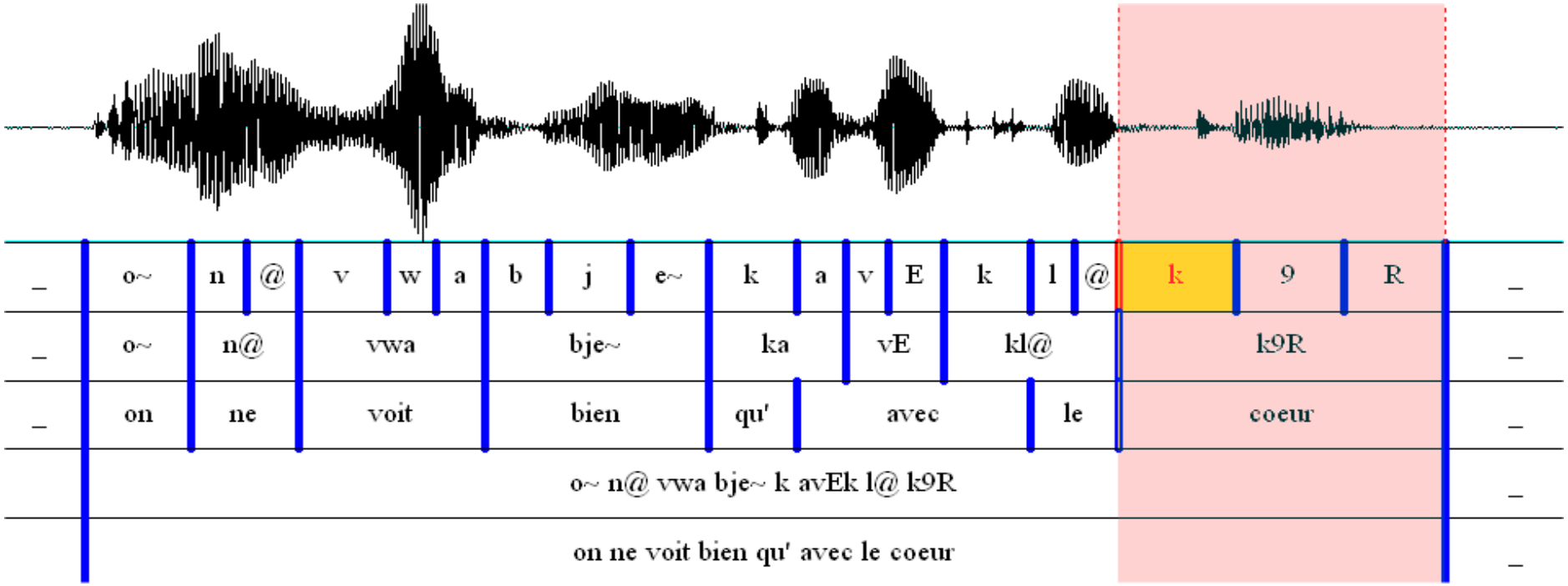
- Data annotation tools

1. Segmentation	EasyAlign	(1)
2. F0 stylisation	ProsoGram	(2)
3. Prominence detection	ProsoProm	(3)
4. Lexical annotation	Lex tool	(4)

- Global and local prosodic measurements
  5. ProsoReport

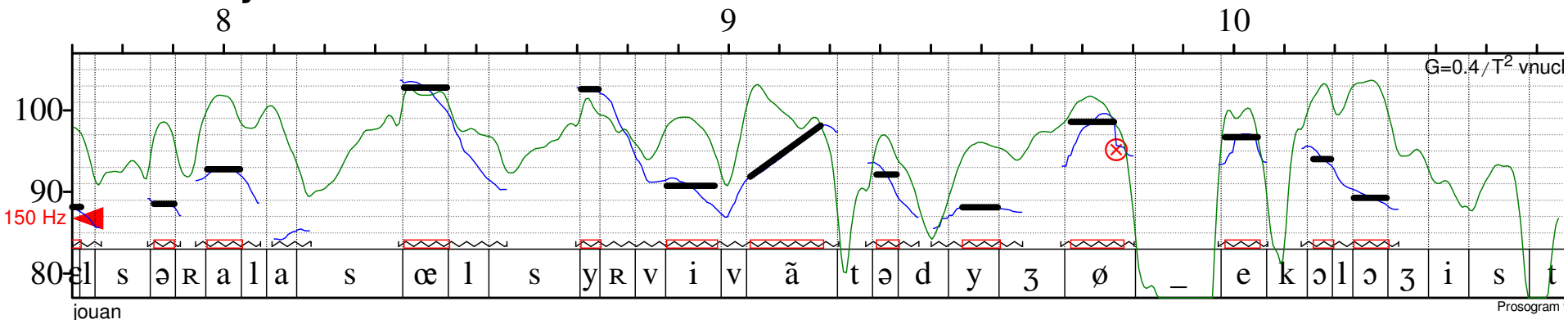
2.Analysing : Segmentation

- *EasyAlign* : praat plugin (Goldman 2008)
  - Multi-tier annotation : phones, syllables, words
  - Ergonomic for the non-specialist



2. Analysing : F0 stylisation

- *ProsoGram* (Mertens - 2004)
- Stylization of F0 to obtain a transcription of intonation based on
  - the simulation of *tonal perception*
  - a syllable-size segmentation motivated by phonetic, acoustic or perceptual properties.
- How
  1. **Select/segment** the voiced portion of the segm. unit, that has sufficient intensity/loudness (using difference thresholds relative to the local peak).
  2. **Stylize** the F0 of the selected time intervals.

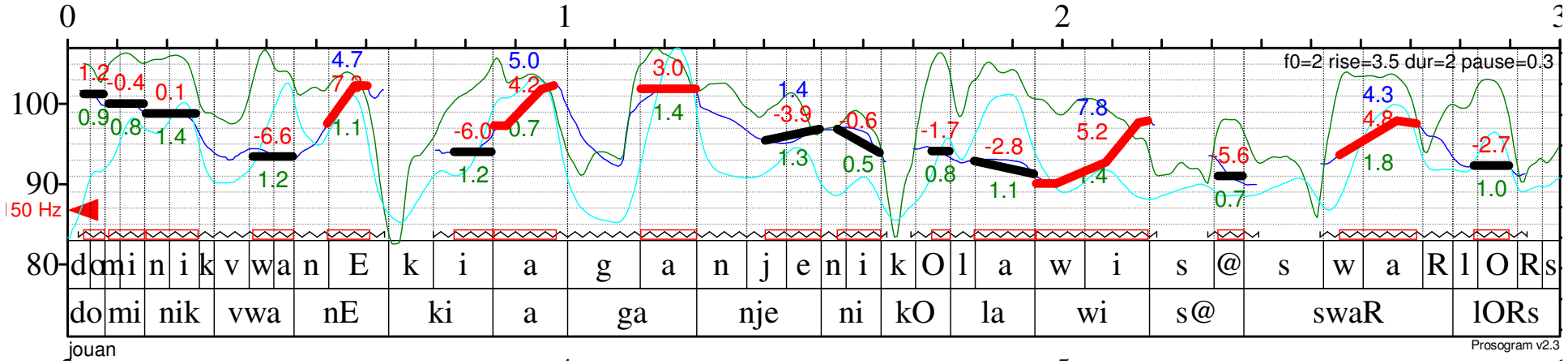


Elle sera la seule survivante du jeu écologiste

- Do we need these *just noticeable differences* for naturalness ? for expressivity ? for communication ?

2. Analysing: Prominence detection

- *ProsoProm* (Goldman - 2008)
  - Prominence detection based on
    - segmentation + perceptive stylisation
    - acoustic parameters (F0 variations, Relative F0, Relative duration)
    - threshold decision



  - Training and validation:
    - with a 60 minutes manually annotated corpus

- Which relativatisation span ? Which decision strategy ?

2. Analysing : Lexical annotation

- Which syntax-discourse⇔prosody interface ? How to map them?
- Any linguistics within prosody? Lexical item intrinsic properties
  - Lexical accent
    - not in French but….
      - lexical/clitic distinction
      - post-tonic schwa
      - Enclitic words (*prenez-en, vas-y, mange-le, avez-vous*)
  - Lexicalized emphasis (i.e. degree adverb)
  - Polysyllabicity
  - Syntactic structure

➔ goal: mark up lexical-word initial and final syllable

Which linguistic, prosodic distinction between lexical and functionnal words (+/- clitic) are we looking for ?

2. Analysing : ProsoReport

- Goal: to combine segmentation and f0 stylisation to produce a detailed prosodic report on *phonostylistic variables*
- Compare ProsoReport of various recordings with additive optional components
  - Basic ProsoReport (+align, +styl)
    - Global and local pitch and duration statistics
  - Full ProsoReport (+prom +lex)
    - Same stats
      - for prom/non-prom syllables
      - for initial, penultimate, final syllables
      - For prom/non-prom x initial, penultimate, final syllables

2. Analysing : ProsoReport

Alignement syllabique et stylisation mélodique

syllabes	○ nombre de syllabes	
durée	○ durée de parole	s. (avec pauses)
	○ durée d'articulation	s. (sans pauses)
	○ taux d'articulation	% (articul./parole)
	○ débit de parole	syll./s. (pour parole)
	○ débit d'articulation	syll./s. (pour articul.)
	○ durée moyenne syllabes + déviation std	s.
	○ durée moyenne nuclei[4] syl. + dév. std	s.
	○ proportion des nuclei pdt l'articulation	%
f0 globale	○ moyenne + déviation standard	ST (relatif à 1 Hz)
	○ étendue (inter quantile range)	ST
f0 tons	○ proportion des tons statiques,	%
	○ prop. des tons dynamiques montants,	%
	○ prop. des tons dynamiques descendants	%
	○ prop. des tons dynamiques complexes	%
f0 mvt	mouvement intra-nuclei des tons	
	○ dynamiques (excluant les statiques)	ST/syll. et ST/sec.
	○ mouvement absolu intra-nuclei (pour tous les tons)	ST/syll. et ST/sec.
	○ mouvement absolu inter-nuclei (i.e mvt moyen entre les tons)	ST/syll. et ST/sec.
	○ agitation mélodique	ST/syll. et ST/sec.
intensité	○ intensité moyenne dans les nuclei	dB
	○ intensité moyenne hors des nuclei	dB
	○ différence d'intensité entre nuclei et non-nuclei	dB

2. Analysing : ProsoReport

alignement + stylisation + proéminences

Détection de proéminences

syllabes	proportion de syllabes proéminentes	%
	proportion de syllabes non proémi-nentes	%

durée, f0, intensité,…des syllabes proéminentes vs non proéminentes

alignement + stylisation + lex

Annotation morphosyntaxique

syllabes	proportion de syllabes en position ini-tiale de mot lexical	%
	proportion de syllabes en position fi-nale de mot lexical	%

durée, f0, intensité,…des syllabes initiales vs finales

alignement + stylisation + lex + proéminences

Annotation morphosyntaxique et détection de proéminences

syllabes	proportion de syllabes proéminentes en position initiale de mot lexical	%
	proportion de syllabes proéminentes en position finale de mot lexical	%

2. Analysing : radio vs. read style study

- Main hypotheses
  - Show specific properties of radio (feature) style (Callamand 1973; Léon; Burger & Auchlin 2007)
  - In this study, compared to "neutral" reading (Goldman & Auchlin 2006)
    H1: hyper articulation
    H2: greater melodic register and fidgetiness (not jitter!)
    H3: more initial accents

- Corpus
  - 3 radio features (~2mn each, 1♂ 2♀, from Radio France) R
  - same texts read-aloud by one female reader L

	L-j	L-a	L-g	L-total	R-total	FR-j	FI-a	FI-g
Nombre de syllabes	772	644	525	**1941**	**1901**	755	638	508
Durée de parole	161,295	138,925	122,19	**422,41**	**384,40**	149,675	123,21	111,51

2. Radio study:

H1: articulation ratio

	L-j	L-a	L-g	Lmoy	R-moy	FR-j	FI-a	FI-g
Durée de parole	161,2	138,9	122,2	**140,8**	**128,1**	149,6	123,2	111,5
Durée d'art.	127,8	110,9	99,9	**113**	**112,4**	132,8	110,6	93,7
Taux d'art.	79,2	79,8	81,8	**80,2**	**87,5**	88,8	89,8	84

➔ radio style uses less silence in proportion

H1: nuclei ratio

	L-j	L-a	L-g	L-moy	R-moy	FR-j	FI-a	FI-g
Durée noy.	60,09	51,767	41,747	**153,6**	**138,5**	56,588	47,342	34,574
% noy/ artic.	47	46,7	41,8	**45,17**	**40,77**	42,6	42,8	36,9

➔ radio nuclei are smaller relatively to syllabe duration

2. Radio study:
H2: register

	L-j	L-a	L-g	L-moy	R-moy	FR-j	FI-a	FI-g
Étendue f0 max-min	15,8	10,9	23	**16,57**	**18,87**	23	15,4	18,2
95%-5%	8,6	8,4	8,7	**8,57**	**12,90**	13,8	12,2	12,7

H2: intra- et inter-syllabic melodic movements

	L-j	L-a	L-g	L-moy	R-moy	FR-j	FI-a	FI-g
Mouvements intra-nuclei des tons dyn.	18,2	20	20	**19,40**	**25,67**	26,3	26,1	24,6
Intra-mvt	2	3	2,9	**2,63**	**4,40**	4,5	5,3	3,4
Inter-mvt	13,3	14,5	14,9	**14,23**	**17,23**	18,4	15	18,3
Agitation	15,3	17,5	17,8	**16,87**	**21,63**	22,9	20,3	21,7

➔ cumulated tonal path (in ST/sec.)

2. Radio/read study:

H3 : more initial accents

	L-j	L-a	L-g	L-moy	R-moy	FR-j	FI-a	FI-g
Prom.	29,5 (228/772)	33,2 (214/644)	34,5 (181/525)	**32,40**	**37,03**	35,6 (269/755)	35,9 (229/638)	39,6 (201/508)
Prom/i	19,7	18,1	26,3	**21,37**	**31.43**	30,1	30.3	33,9
Prom/f	58,6	64,5	58,8	**60,63**	**59.6**	59,3	59.4	60,1

3. Simulating

- Scheme

  A→A+ : manipulate A's voice

  ..→ A+ : synthesize A's voice

  B+→A+ : map B onto A

1. Which units, span to consider?
2. Which parameters to transfer?
3. Which precision for stylisation ?

- Analyse
  1. Segmentation and annotation
  2. Acoustic observations
     - Pitch variations
     - Durations
     - …
  3. Make a prosodic report

- Simulate / synthesize
  1. Segment and annotate A (and B)
  2. (observe B)
  3. Inject rules or B's params

3. Simulating

- How far can we go with ProsoCopy ?
  - Take an expressive voice (A)
    Hélène Jouan - Edito France Inter – 03/2007

    Dominique Voynet qui a gagné, Nicolas, oui, ce soir, lorsque le président du conseil constitutionnel énumèrera la liste des candidats officiels, et bien, elle sera la seule survivante du jeu écologiste.
  - Take a "neutral" voice (B)
    Acapela Synthesis – Claire's voice
  - Manipulate B to inject A's pitch and phone durations

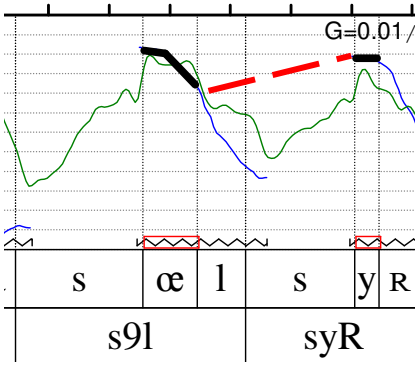
Phoneme sequence identity required.
…how to handle pause (insertion and deletion) ?

3. Simulating

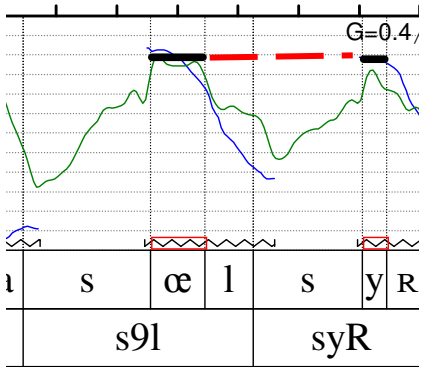
- Limits of segmentation and stylisation

"elle sera la seule survivante"

glissando=0.01/T²

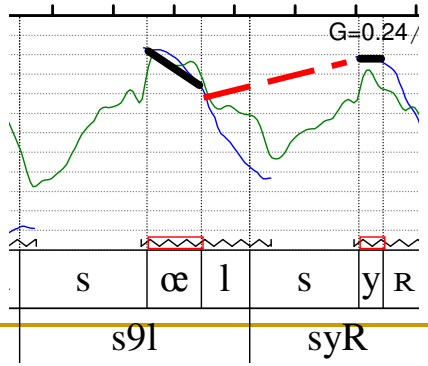


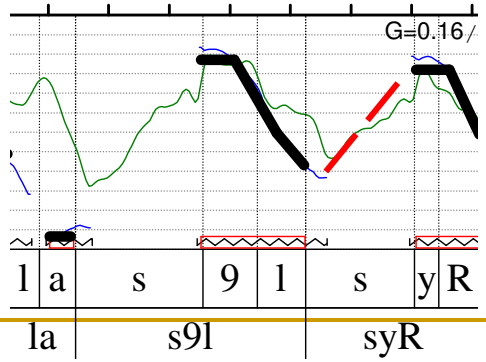
full



glissando=0.40/T²

glissando=0.24/T²





glissando=0.16/T²
allow nuclei in onset and coda

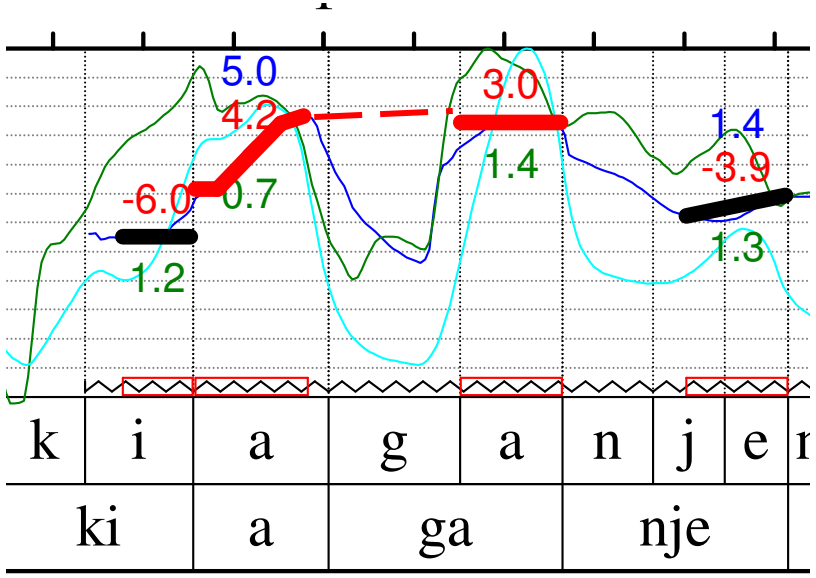
full

3. Simulating

- Limits of segmentation and stylisation

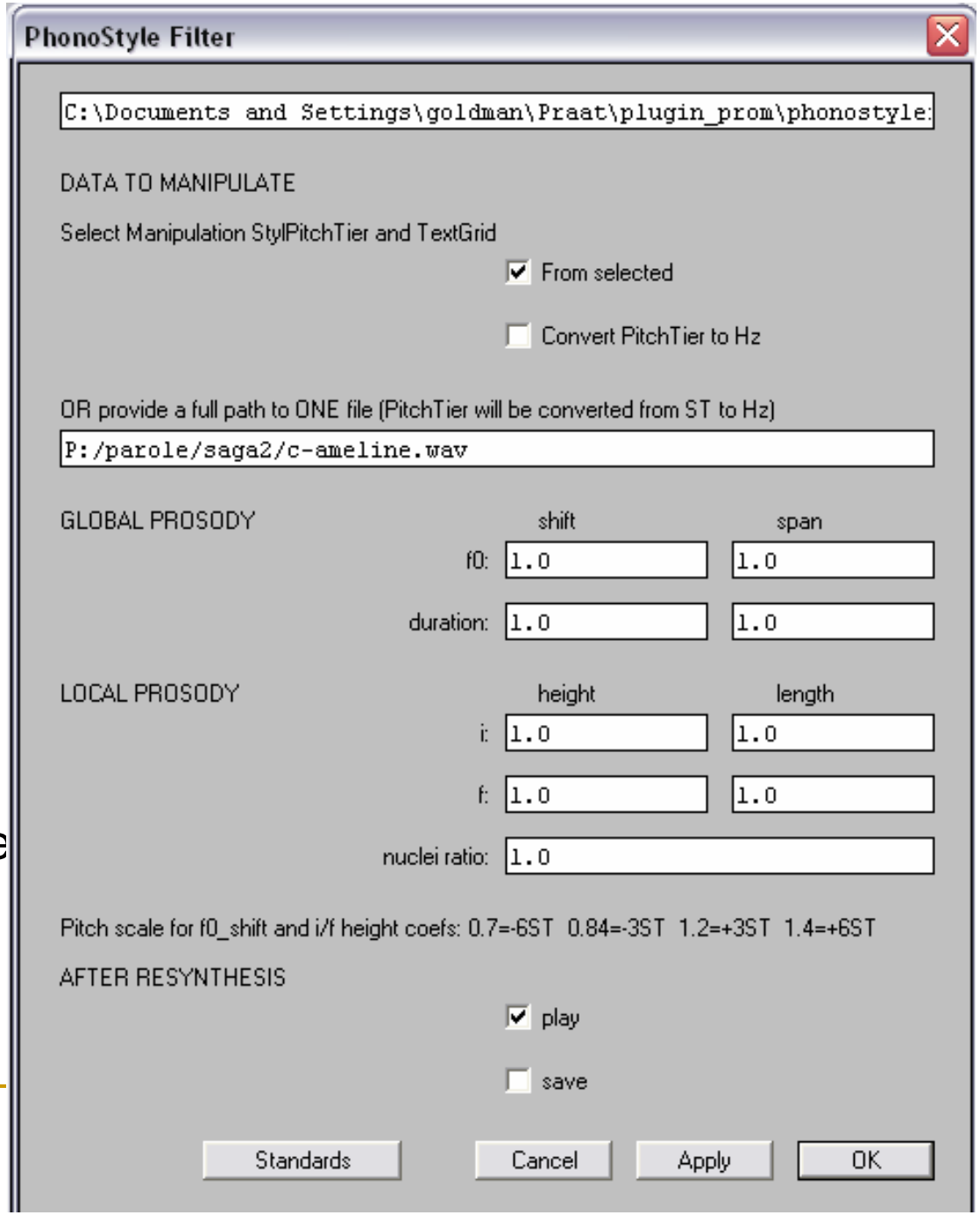
“qui a gagné”

Prosocopy
Original



3.Simulating

- PhonoStyleFilter : prosodic manipulation
  - Globally
    - Shift F0 or modify F0 span
    - Shift duration (mean syllable duration)
  - Locally
    - +3ST x1.1 on “i” syllable
    - +0ST x1.3 on “f” syllable
    - nuclei proportion
      - more nuclei…
      - less nuclei



Conclusion

- Analysing phonostylistic properties in speech
  - Tools to extract targeted acoustic parameters
    - To model on speaker
    - To compare speakers
  - But… is acoustic pertinence equivalent to perception?

- Simulating
  - Tool to apply phonostylitic cues globally and locally with linguistically motivated selection of units
  - Limits of stylisation