Prosody Prediction

August 16, 2017

Hansjörg Mixdorff

Beuth University Berlin

mixdorff@beuth-hochschule.de

http://public.beuth-hochschule.de/~mixdorff/thesis/index.html

Overview

1) Introduction
2) Models in Prosody Research
3) Linguistic Background of Approach
4) Parameter Extraction
5) Statistical Modeling
6) Superpositional vs. Autosegmental Modeling of F0
7) Discussion

1) Introduction: Prosody

Working definition: *„The principles underlying the organisation of an utterance into a structure.“*

The most important prosodic features of speech and their measurable correlates:

•pitch -> fundamental frequency ***F0***

•quantity -> durations of phones, syllables words (...) ***D***

•loudness -> intensity ***I***

-> suprasegmental by nature

Measuring prosodic features

F0 contour ***f0(t)*** und intensity ***I(t)*** (‚envelope ‘) can be extracted directly from the speech signal, measuring durations presupposes the segmentation of the speech signal into *(phonetically or phonemically defined)* portions.

The estimation of a duration contour **D(t)** requires the establishment of the relationship between the duration of a particular segment and the durations of segments of the same type in the entire data base, the z-score, for instance.

$$z_i = (Dur_i - \mu_i) / \sigma_i$$

Information Coded in Prosodic Features

- Linguistic information
  - word accent / syllabic tone in tone languages
  - segmentation (*Phrasing, pausing*)
  - sentence mode (question vs. non-question)
  - focus (prominence), „accentuation“
- Para-linguistic information
  - attitude and intention of a speaker
  - sociolect, dialect

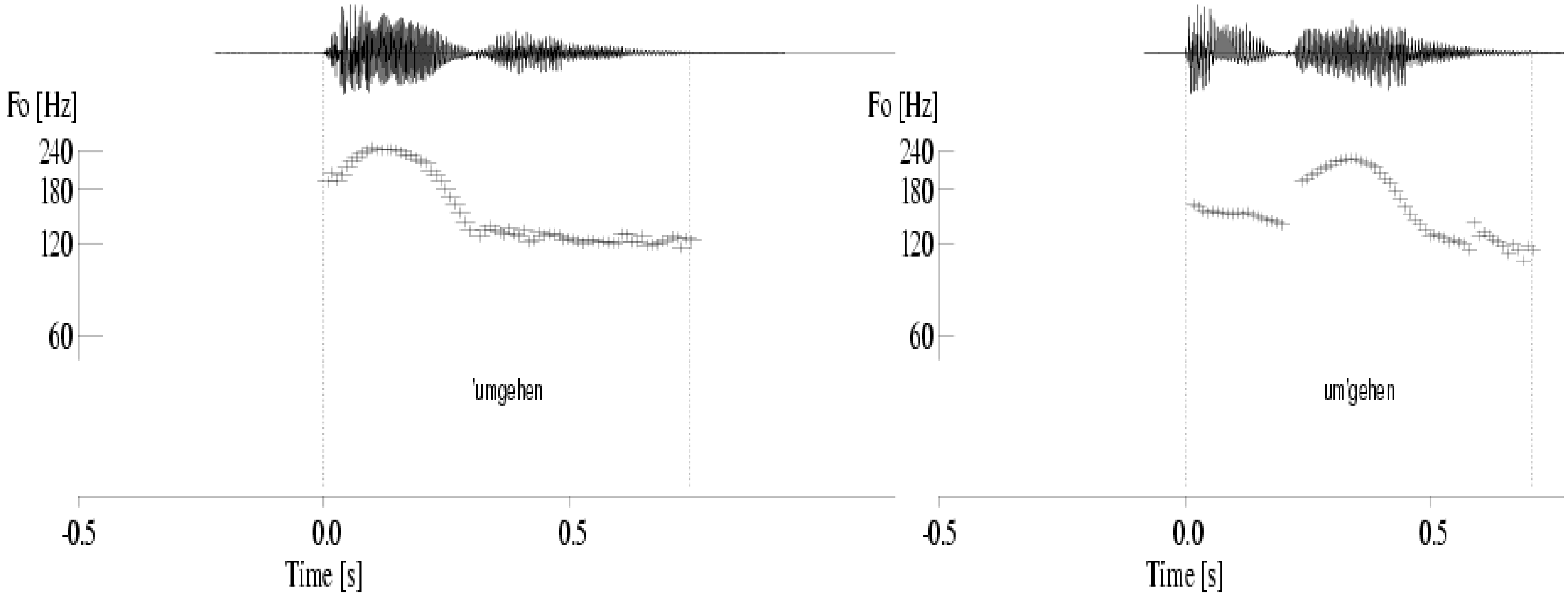
Consciously controlled

Information Coded in Prosodic Features (continued)

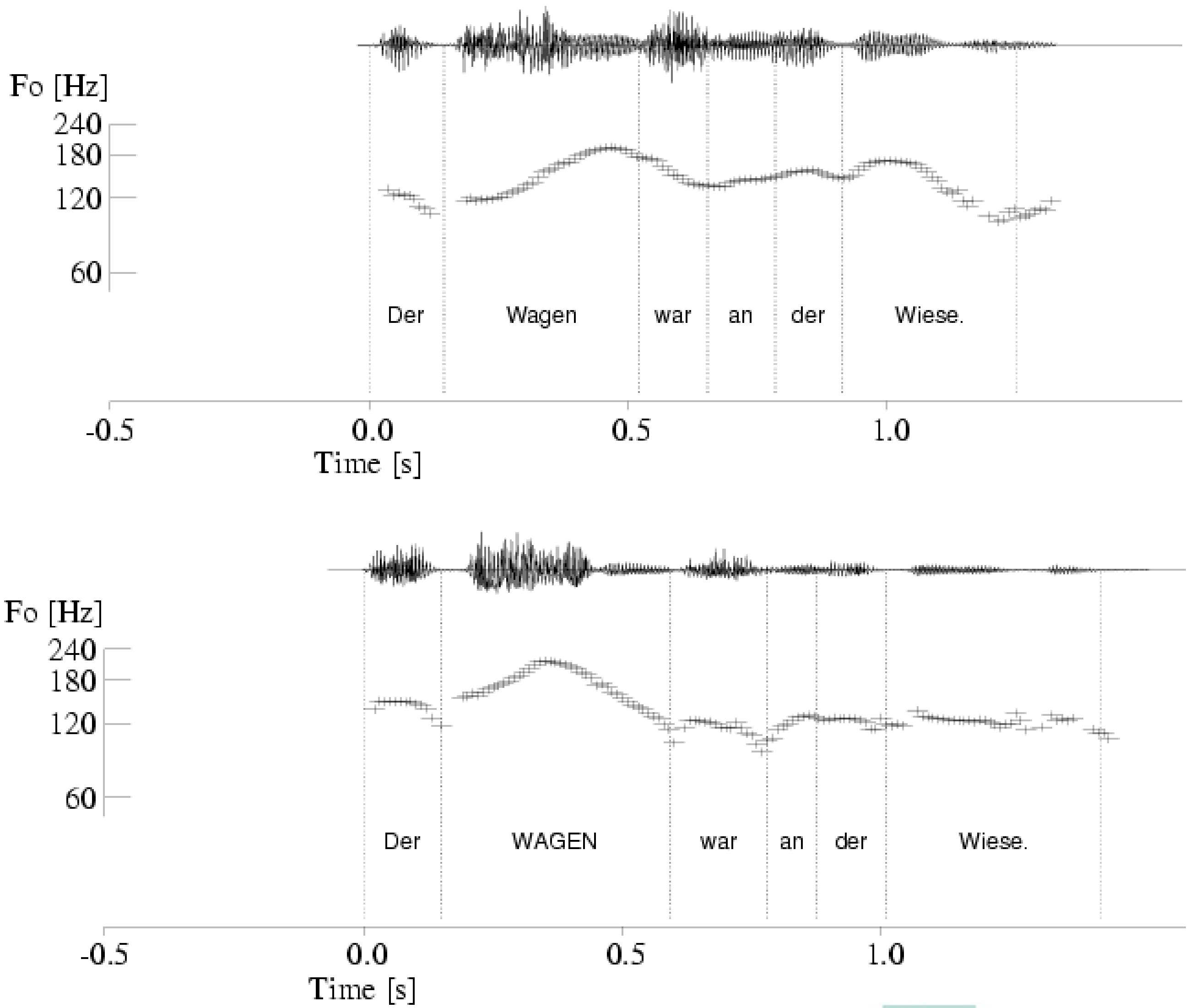
- Non-linguistic information
  - Age
  - Sex
  - health condition
  - emotional condition
  - *Speech of a single, human speaker*

Not consciously controlled

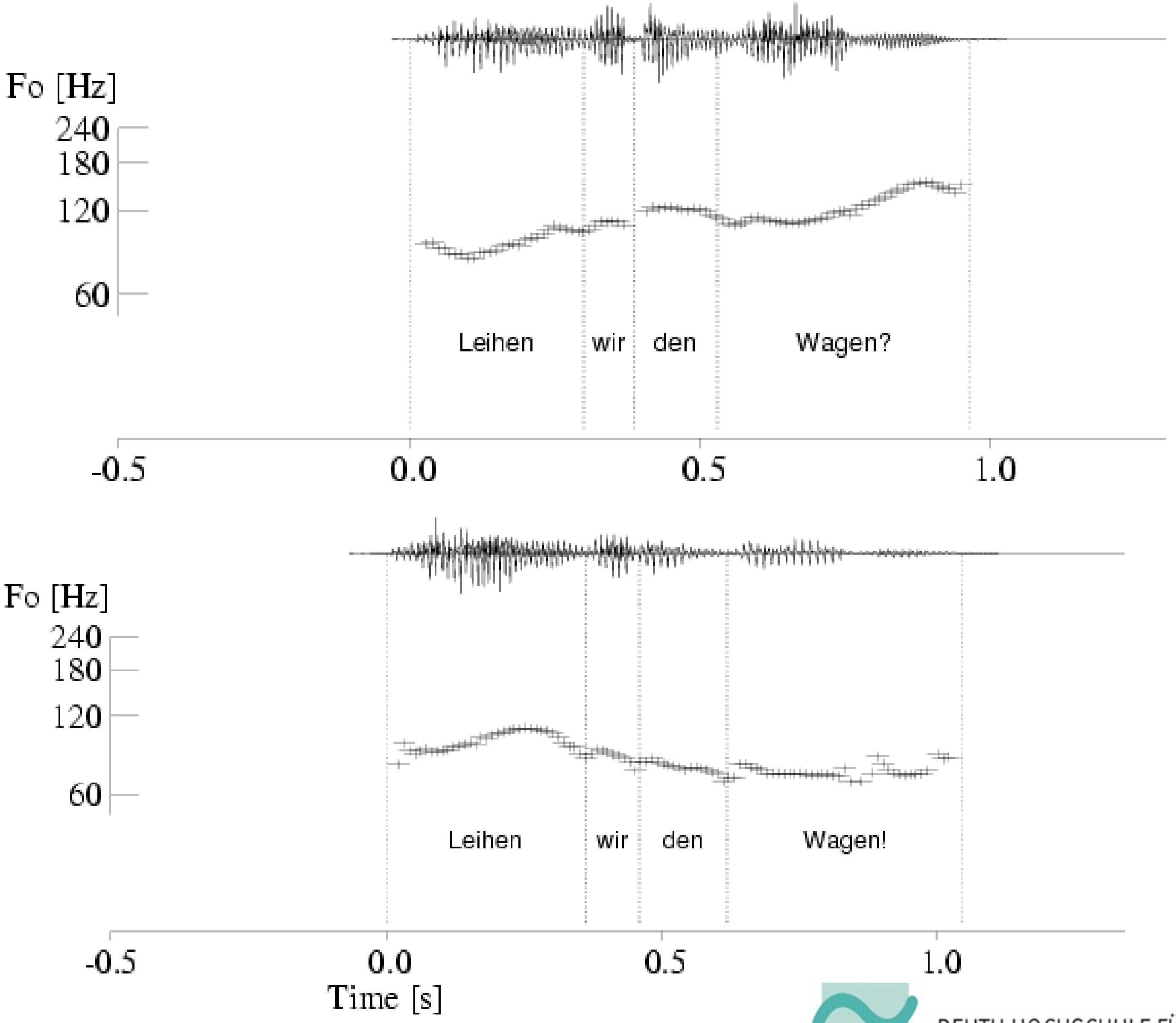
Word Accent Distinction



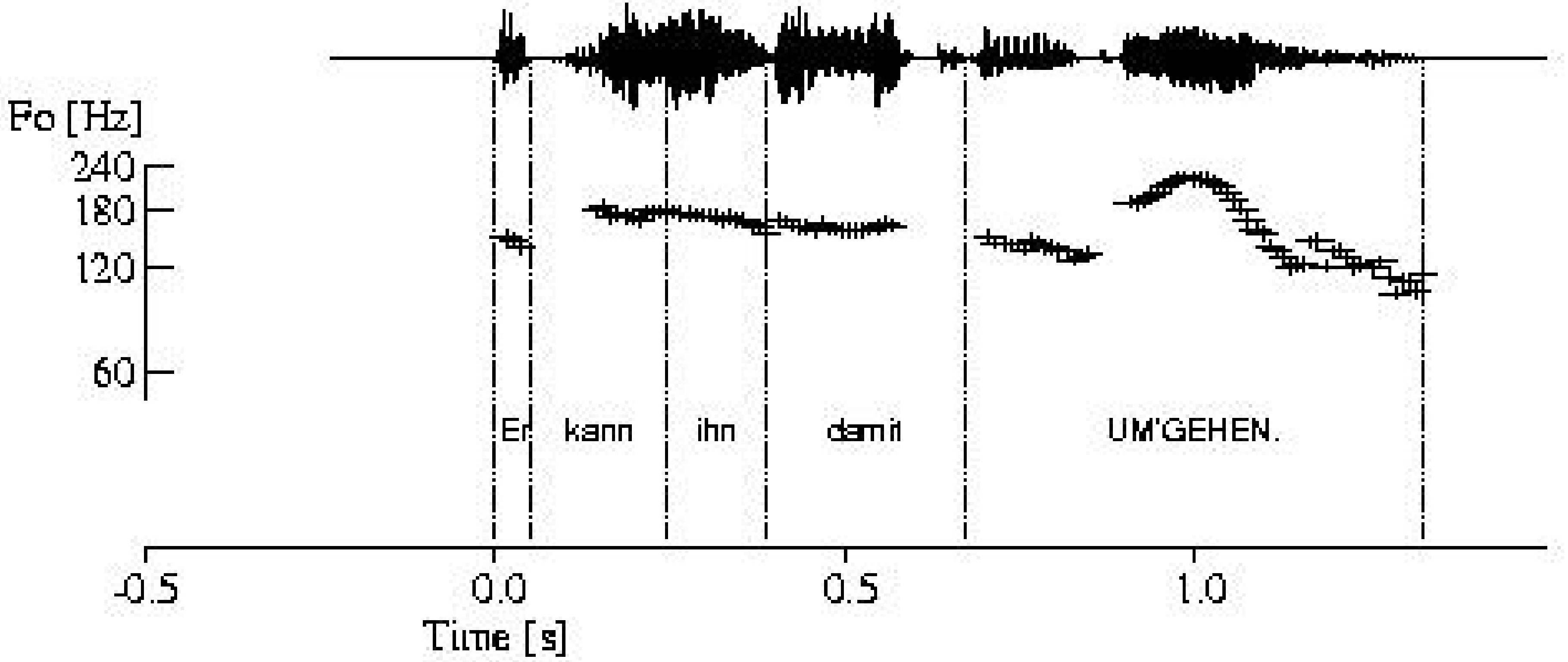
Broad vs. narrow focus



Sentence mode

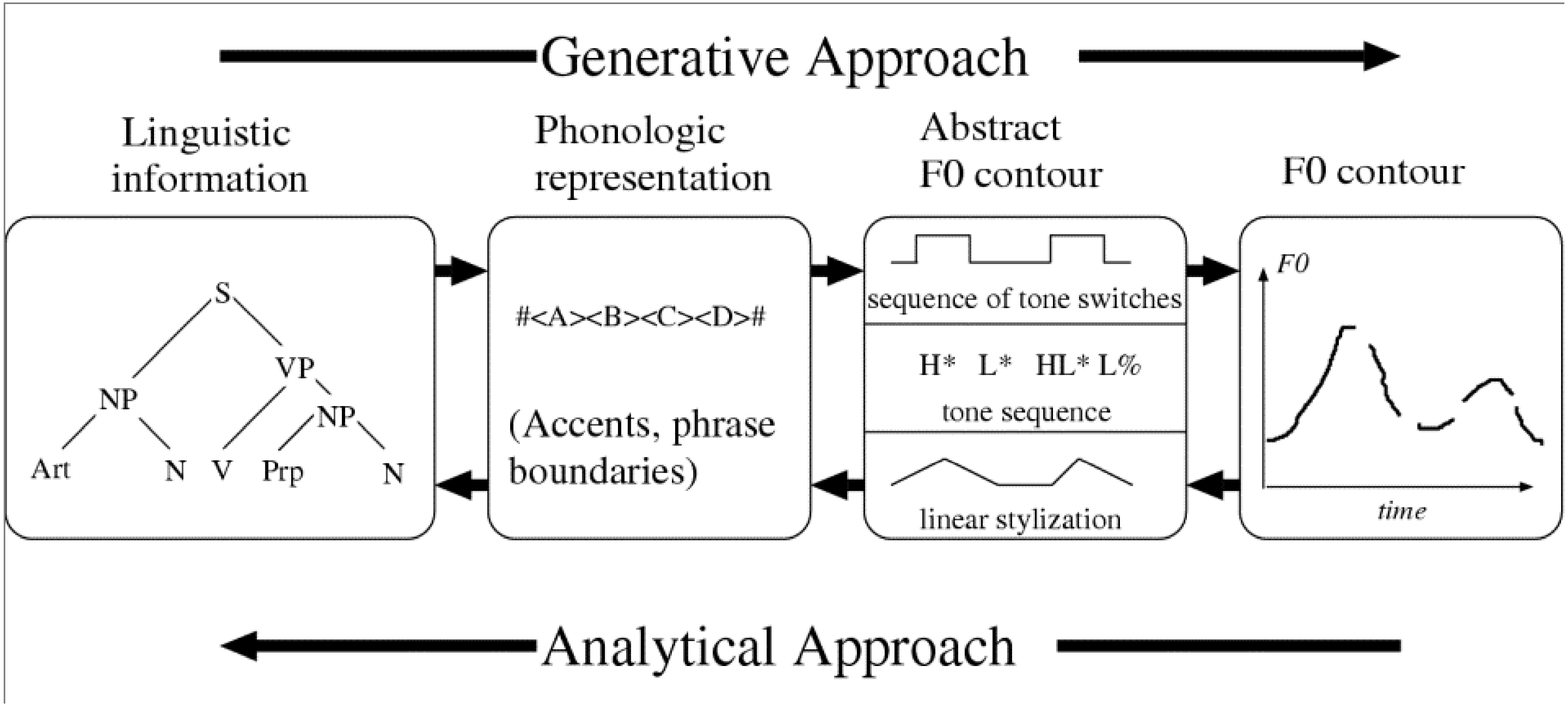


Manifestation of Multi-Dimensional Information in the Single-dimensional F0 Contour

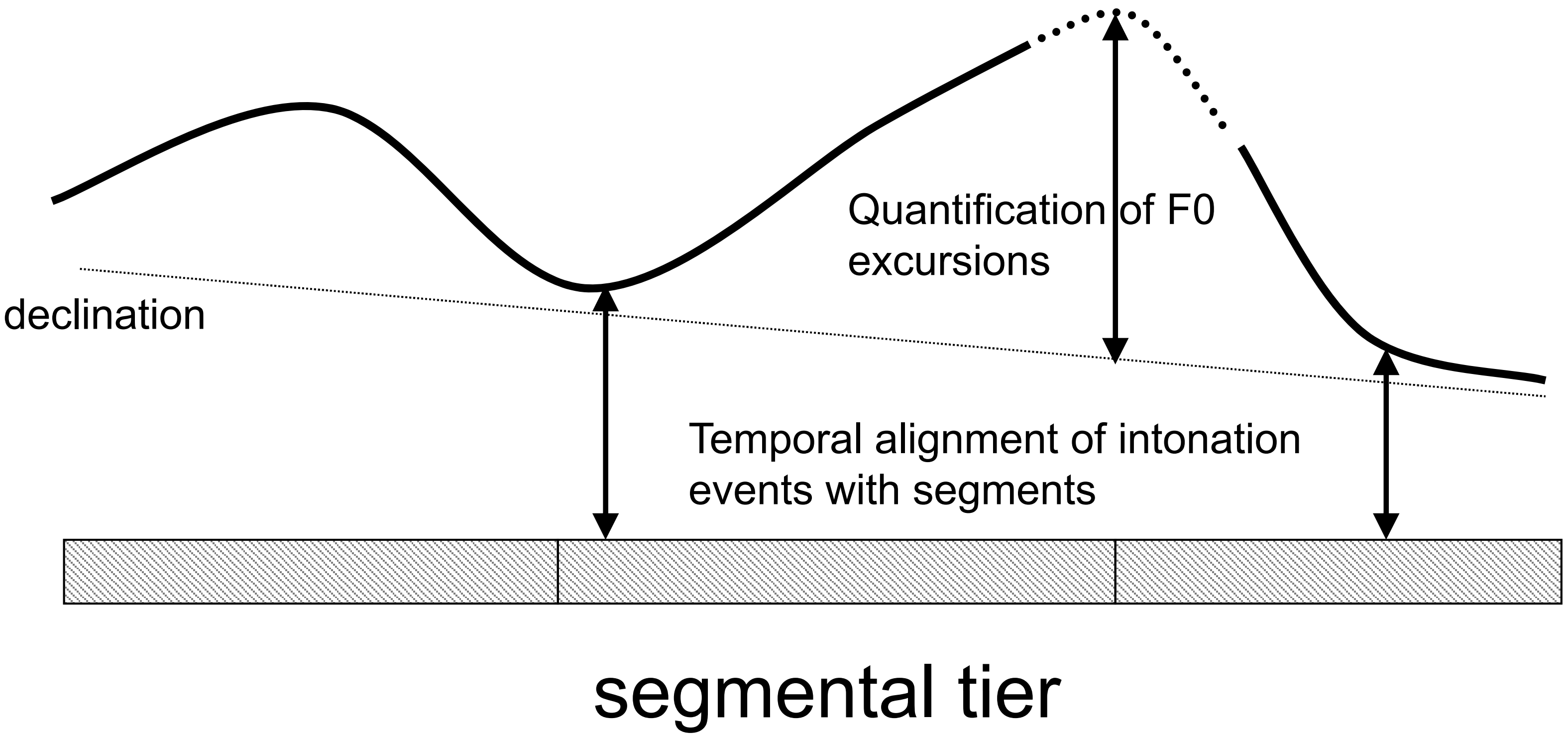


lexical accent, sentence mode, delimitation, focus...

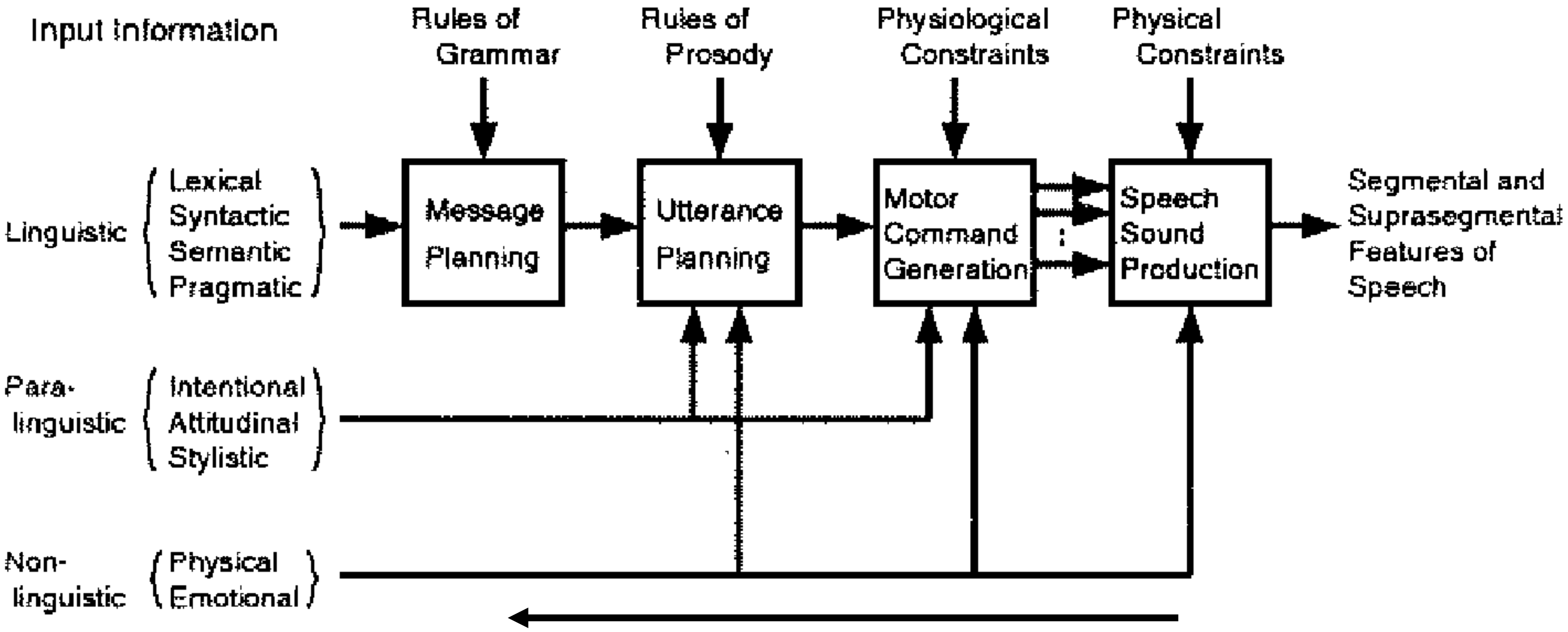
2) Models in Prosody Research



What we expect from a quantitative description of F0 contours

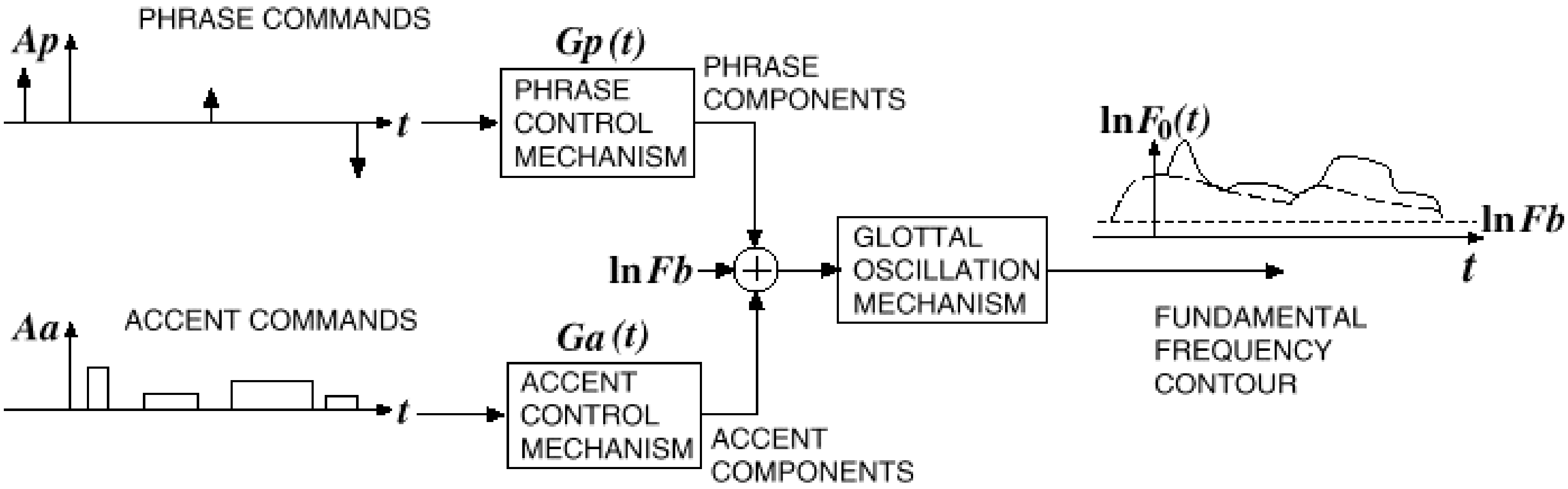

Quantification of F0 excursions
declination
Temporal alignment of intonation events with segments
segmental tier

Process of information coding (Fujisaki 1994)



We are also interested in the reversed process !

The Fujisaki Model

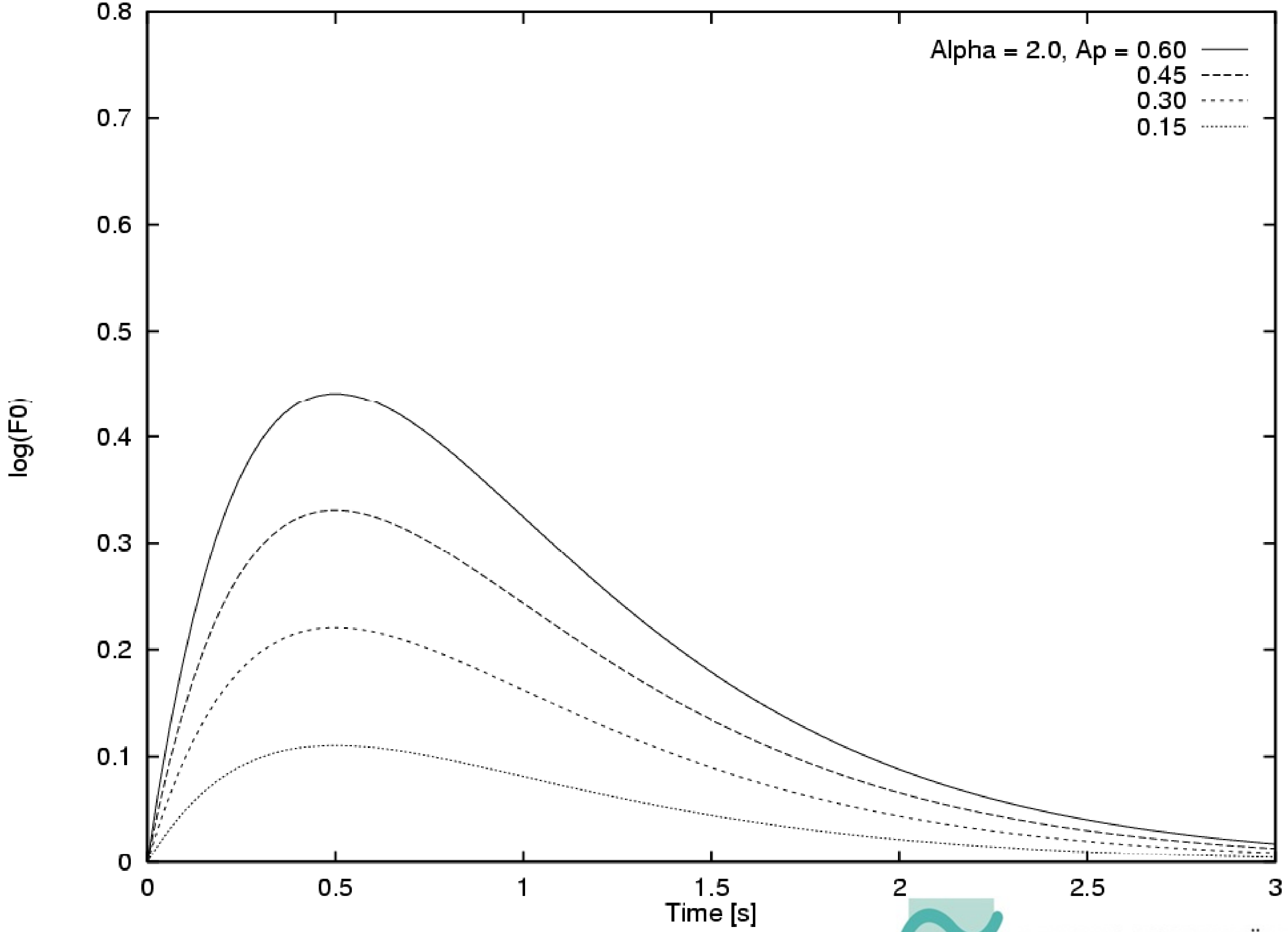


$$\ln F_0 = \ln Fb + \sum_{i=1}^{I} Ap_i Gp(t - T0_i) + \sum_{j=1}^{J} Aa_j [Ga(t - T1_j) - Ga(t - T2_j)]$$

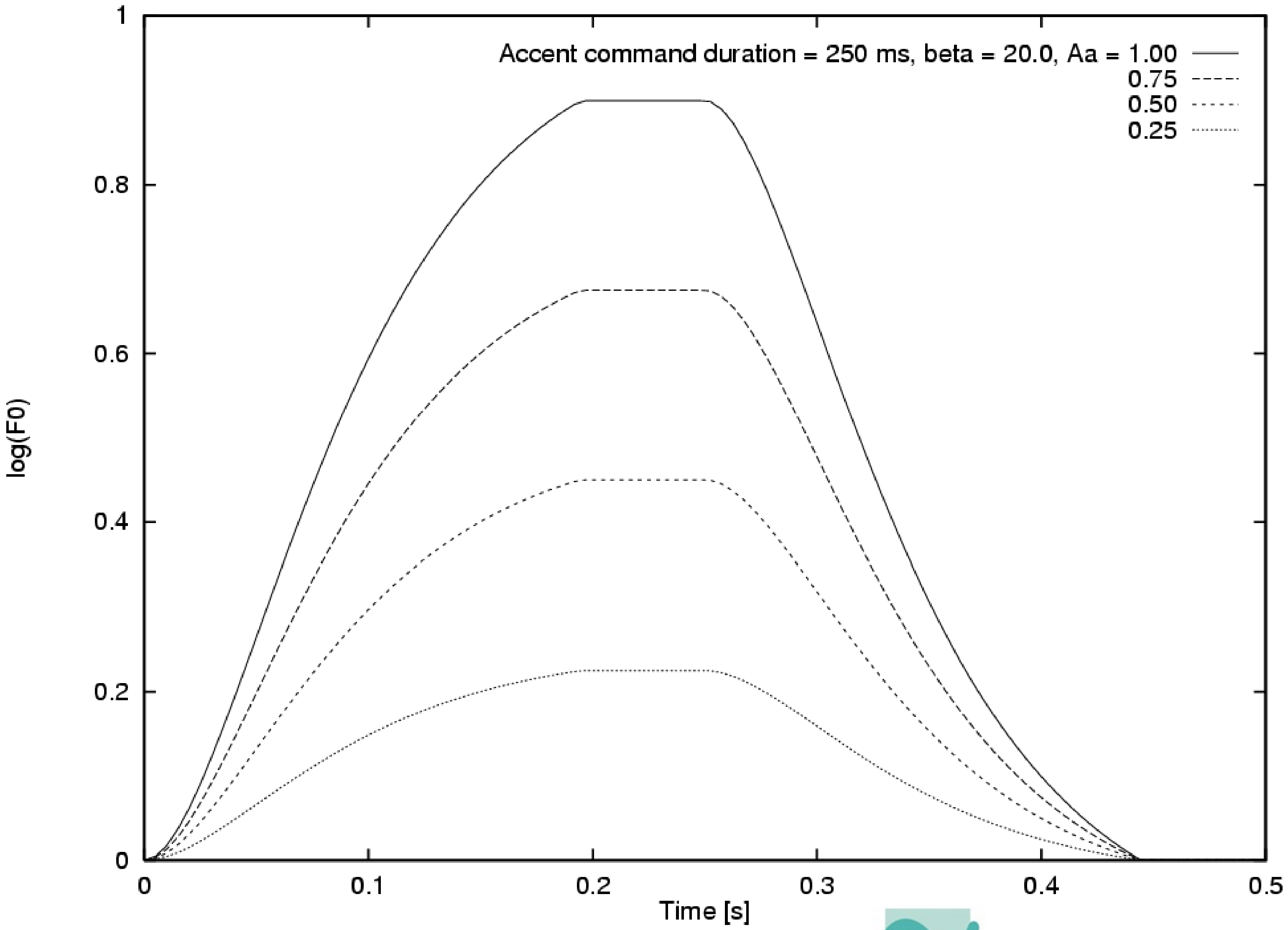
$$Gp(t) = \begin{cases} \alpha^2 \exp(-\alpha t), \text{for } t \geq 0 \\ 0, \text{for } t < 0 \end{cases}$$

$$Ga(t) = \begin{cases} \min[1 - (1 + \beta t)\exp(-\beta t), \gamma], \text{for } t \geq 0, \\ 0, \text{for t} < 0 \end{cases}$$

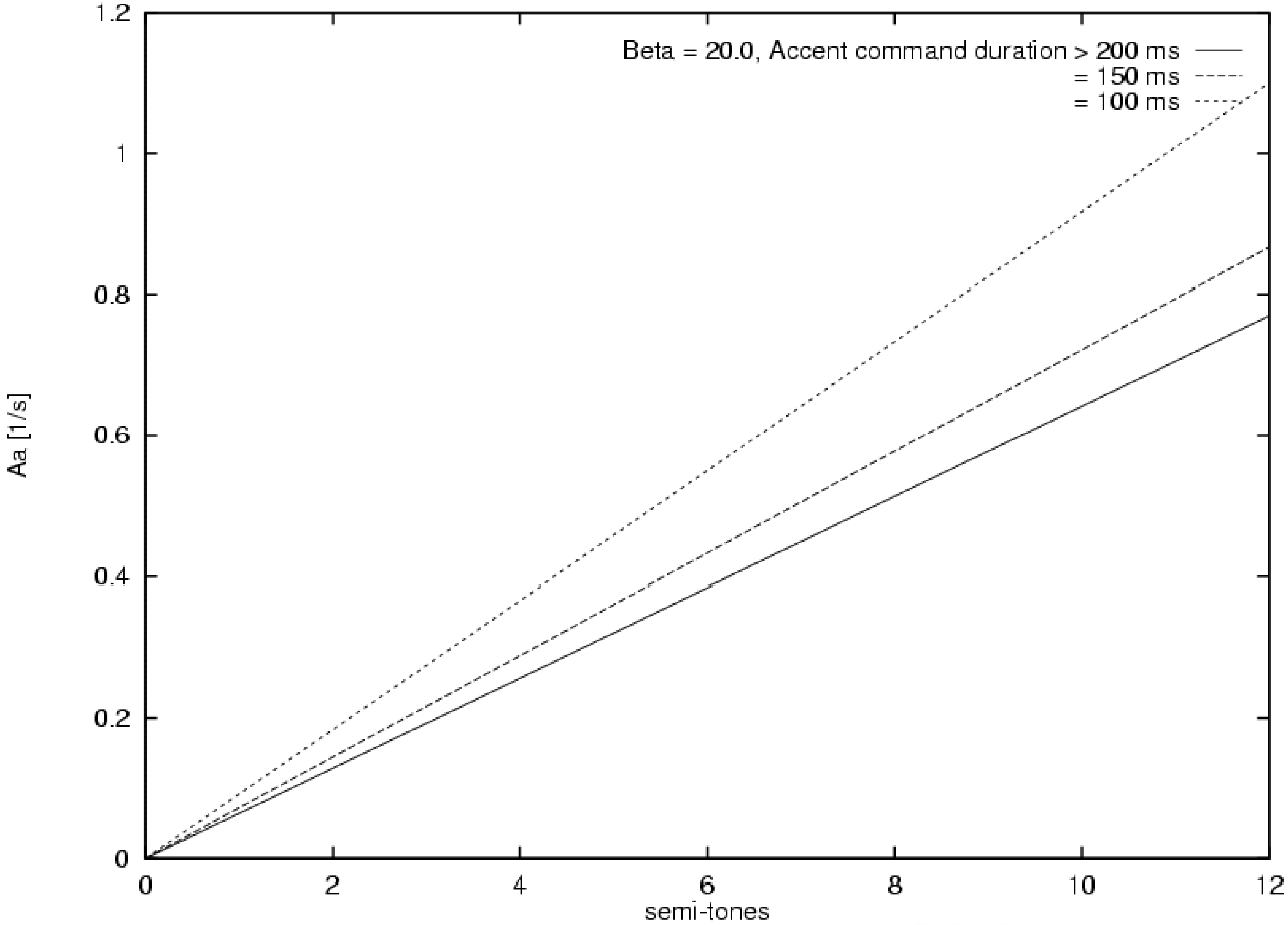
Characteristics of Phrase Component



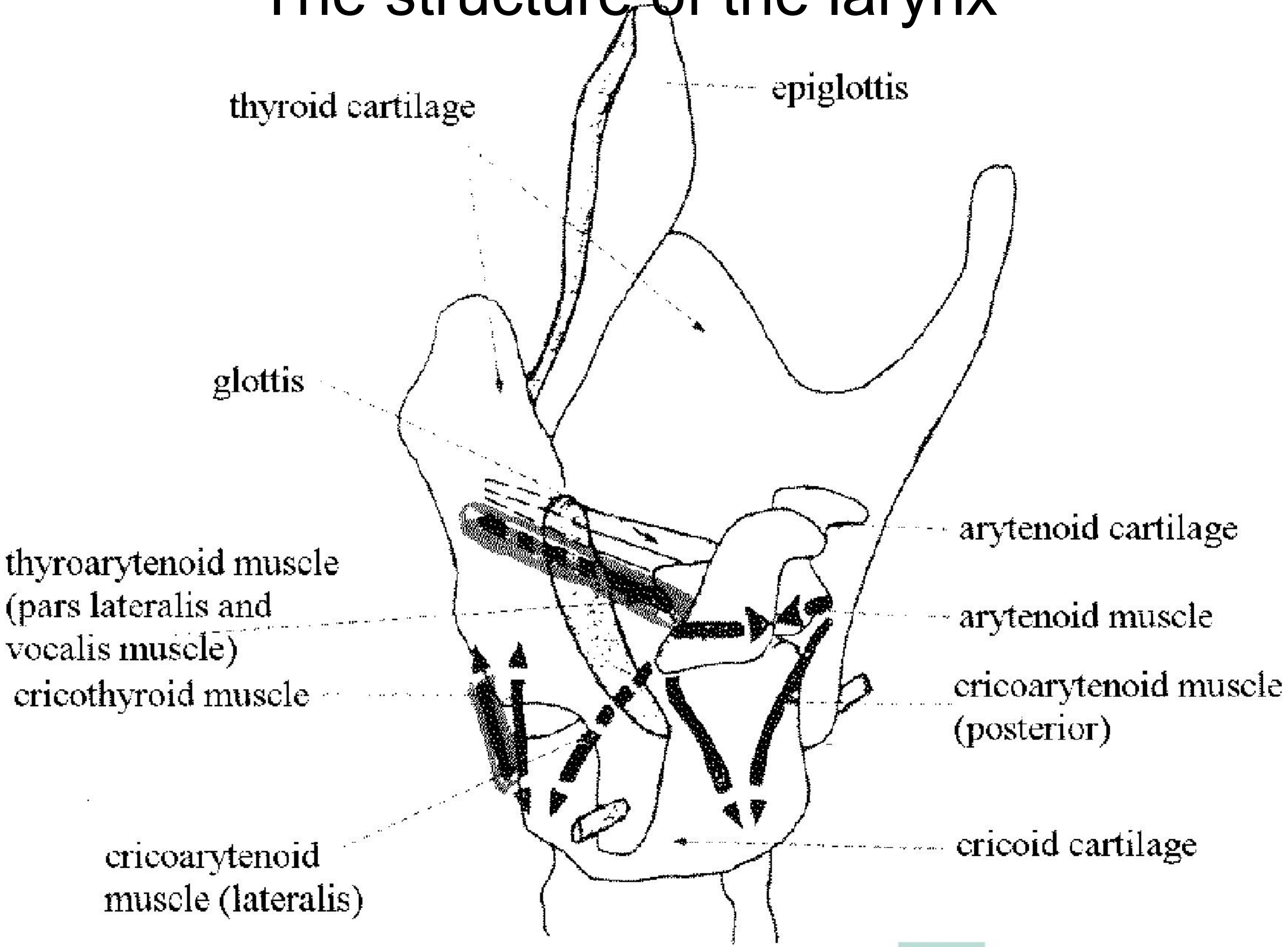
Characteristics of Accent Component



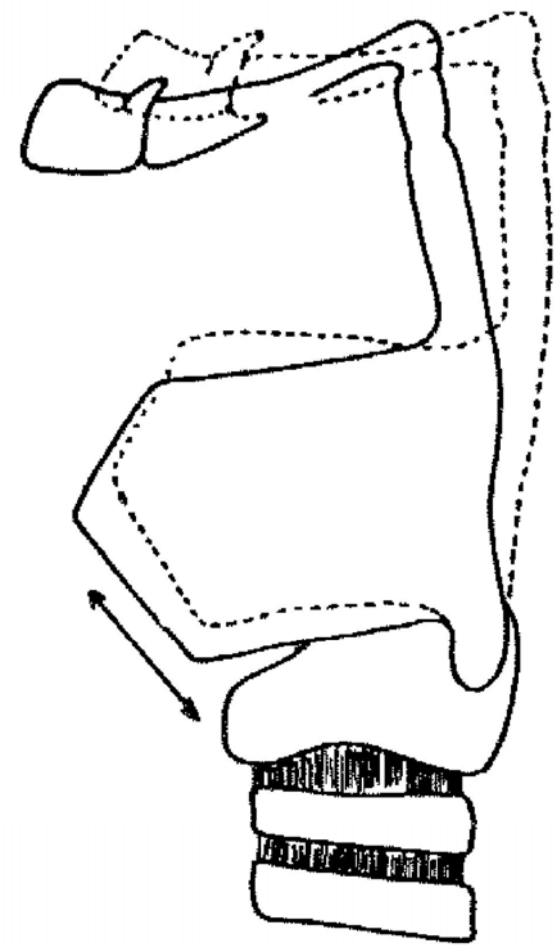
Accent Command Amplitude vs. Semi tone scale



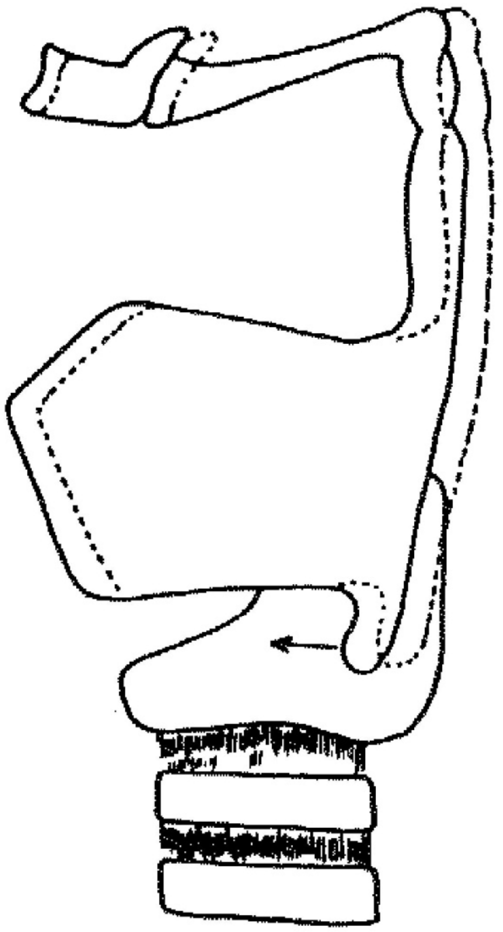
The Fujisaki model and its physiological interpretation: The structure of the larynx


thyroid cartilage
epiglottis
glottis
arytenoid cartilage
thyroarytenoid muscle (pars lateralis and vocalis muscle)
arytenoid muscle
cricothyroid muscle
cricoarytenoid muscle (posterior)
cricoid cartilage
cricoarytenoid muscle (lateralis)

Two degrees of freedom in the movement around the crico-thyroid joint



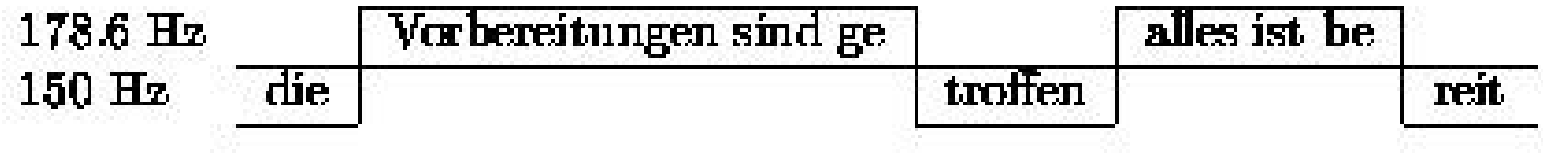
Rotation⇒accent component



Translation ⇒phrase component

3) Linguistic Background
(D.Eng. thesis, 1998)

- Phonologically relevant **tone switches** (Isačenko,1964)
- Perception experiments using simplified F0 contours

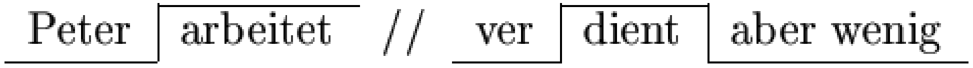


178.6 Hz

150 Hz

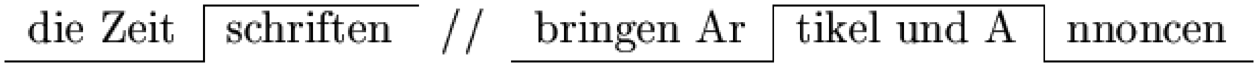
die Vorbereitungen sind ge troffen alles ist be reit

- ‘pitch interrupters‘ (//) at phrase boundaries



Peter arbeitet // ver dient aber wenig

“Peter works, but earns little.”



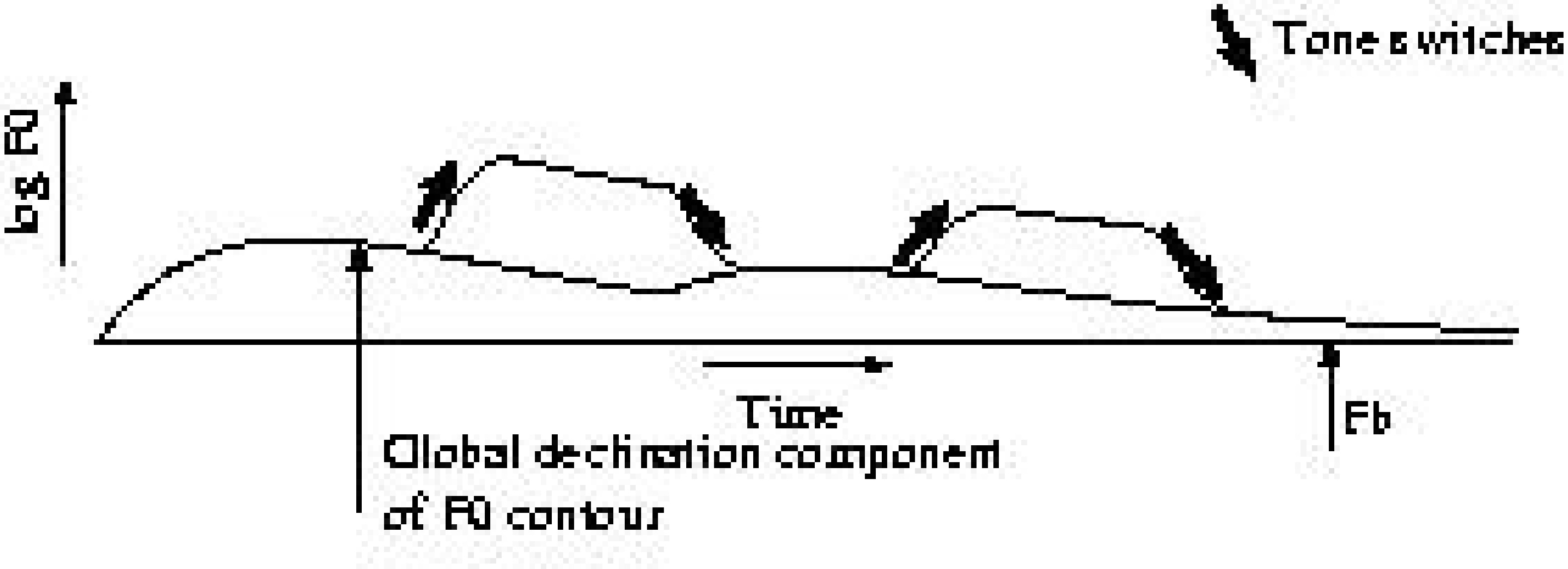
die Zeit schriften // bringen Ar tikel und A nnoncen

“The journals carry articles and advertisements.”

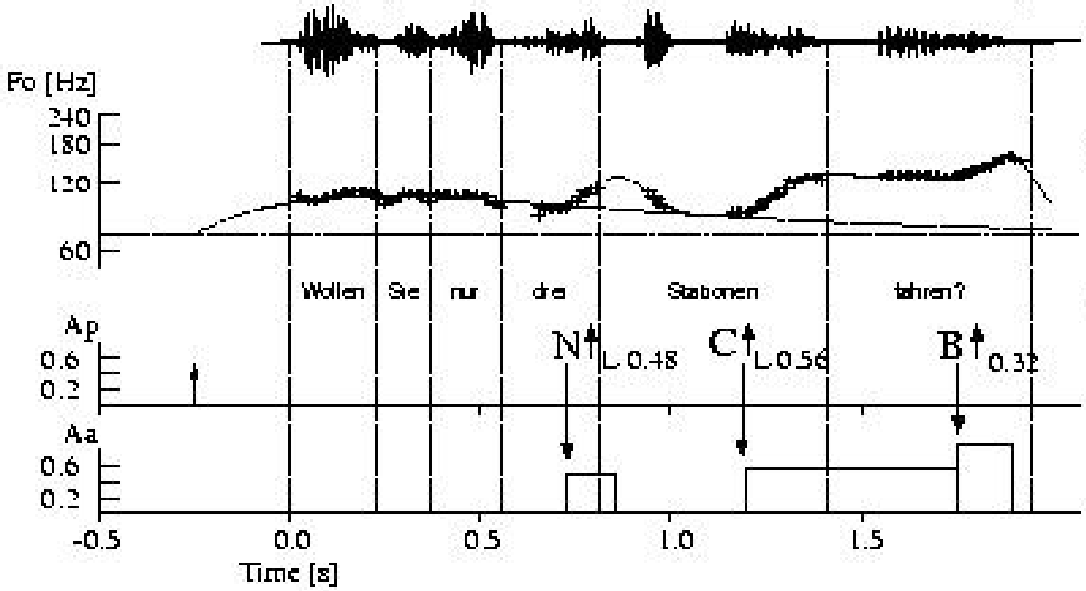
Linguistic Background (D.Eng. thesis, 1998)

Information intoneme I↓	Declarative-final accents, falling tone switch. Conveying a message.
Contact intoneme C↑	Question-final accents, rising tone switch. Establishing contact.
Non-terminal intoneme N↑	Non-final accents, rising tone switch. Signaling non-finality
Boundary tone B↑	Question-final boundary tone. Rising tone switch not necessarily connected to an accented syllable

Combining tone switches and Fujisaki model



Examples of Intoneme Assignment

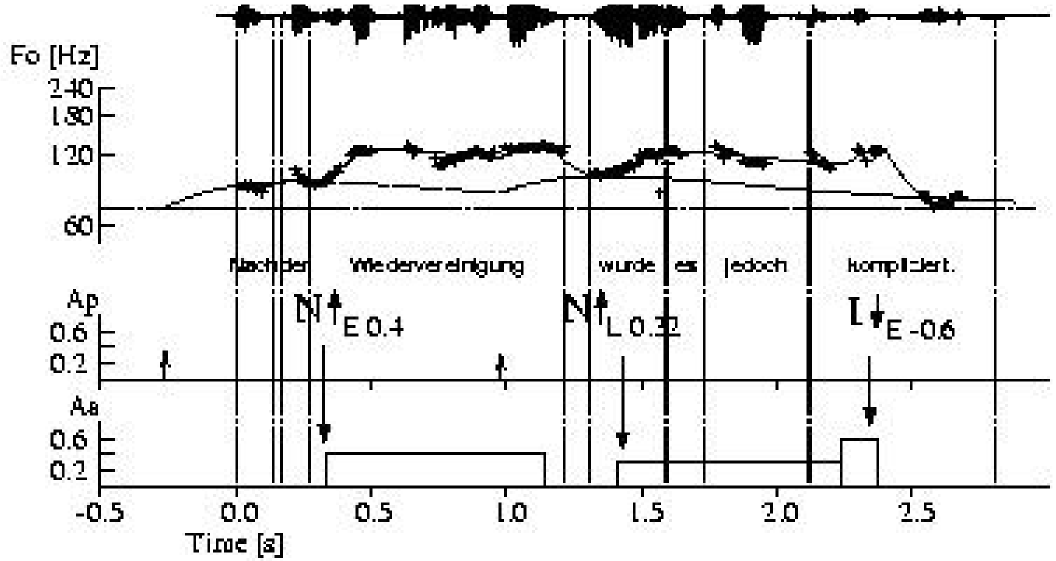


Key to figure:

$N\uparrow_{L\ 0.48}$ Non-terminal intoneme, rising tone switch late, dAa=0.48

$C\uparrow_{L\ 0.56}$ Contact intoneme, rising tone switch late, dAa=0.56

$B\uparrow_{0.32}$ Boundary tone, rising tone switch, dAa=0.32



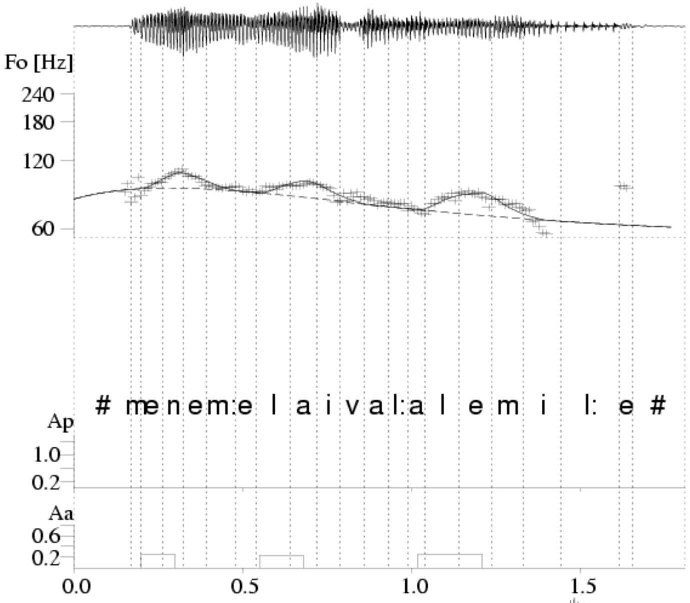
Key to figure:

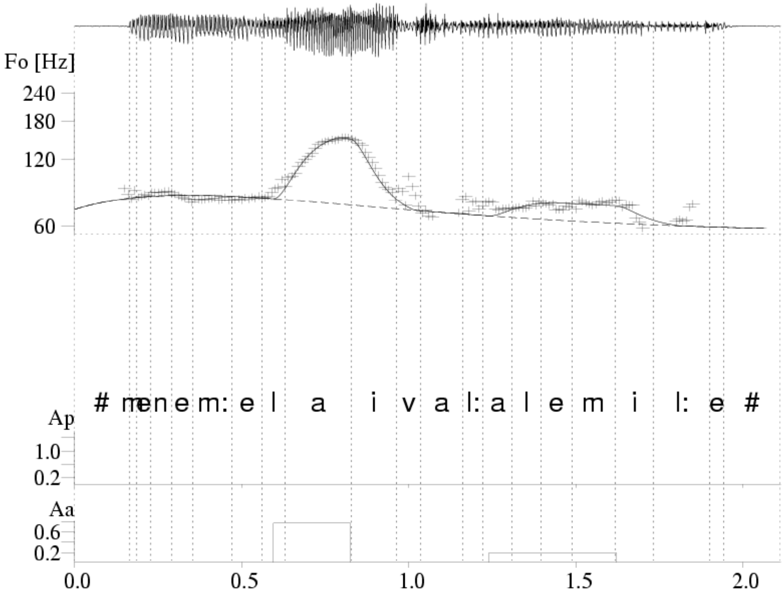
$N\uparrow_{E\ 0.4}$ Non-terminal intoneme, rising tone switch early, dAa=0.4

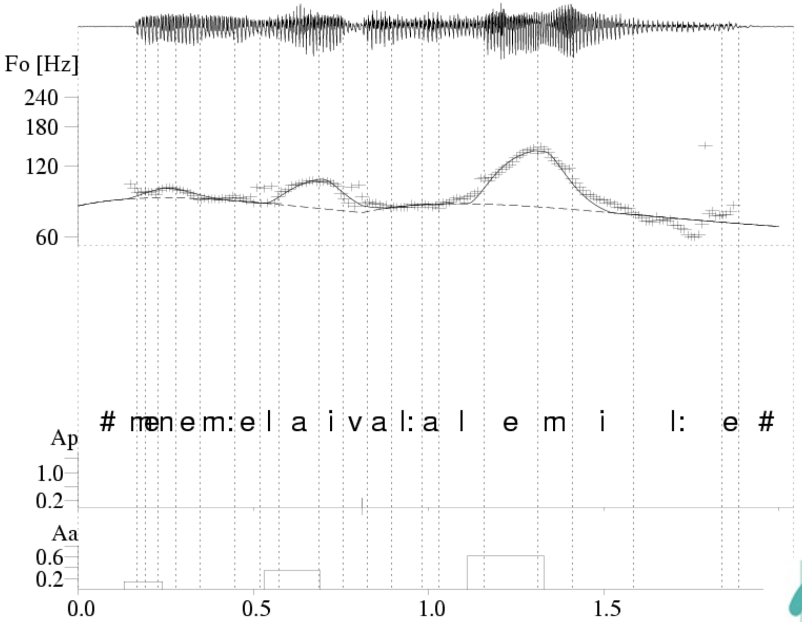
$N\uparrow_{L\ 0.32}$ Non-terminal intoneme, rising tone switch late, dAa=0.32

$I\downarrow_{E\ -0.6}$ Information intoneme, falling tone switch early, dAa=-0.6

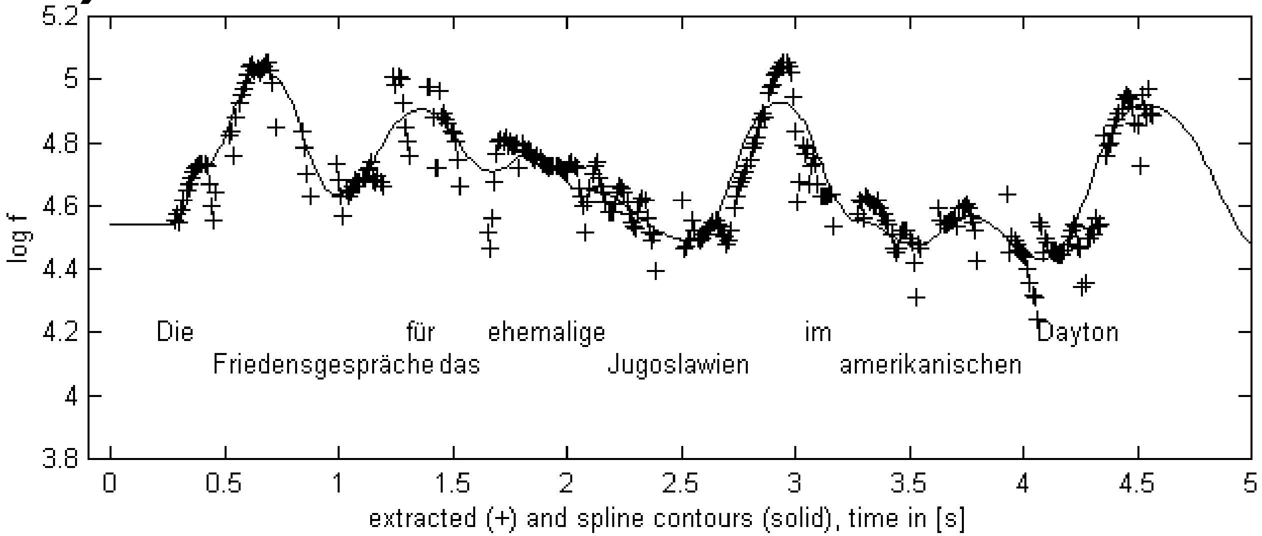
A Few Examples in Finnish



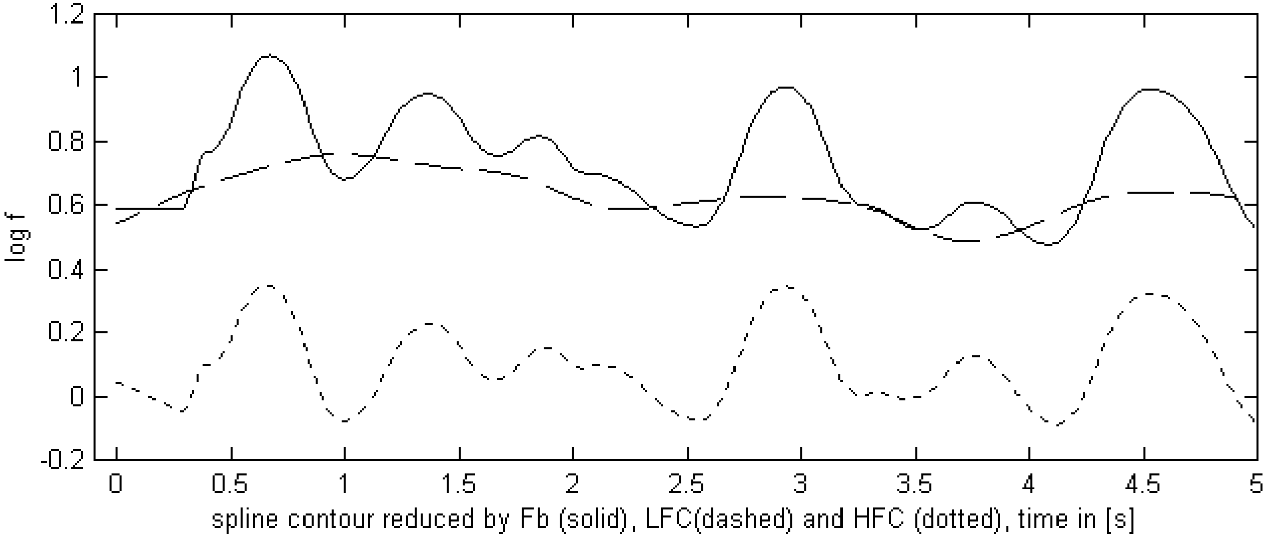




4) Parameter Extraction

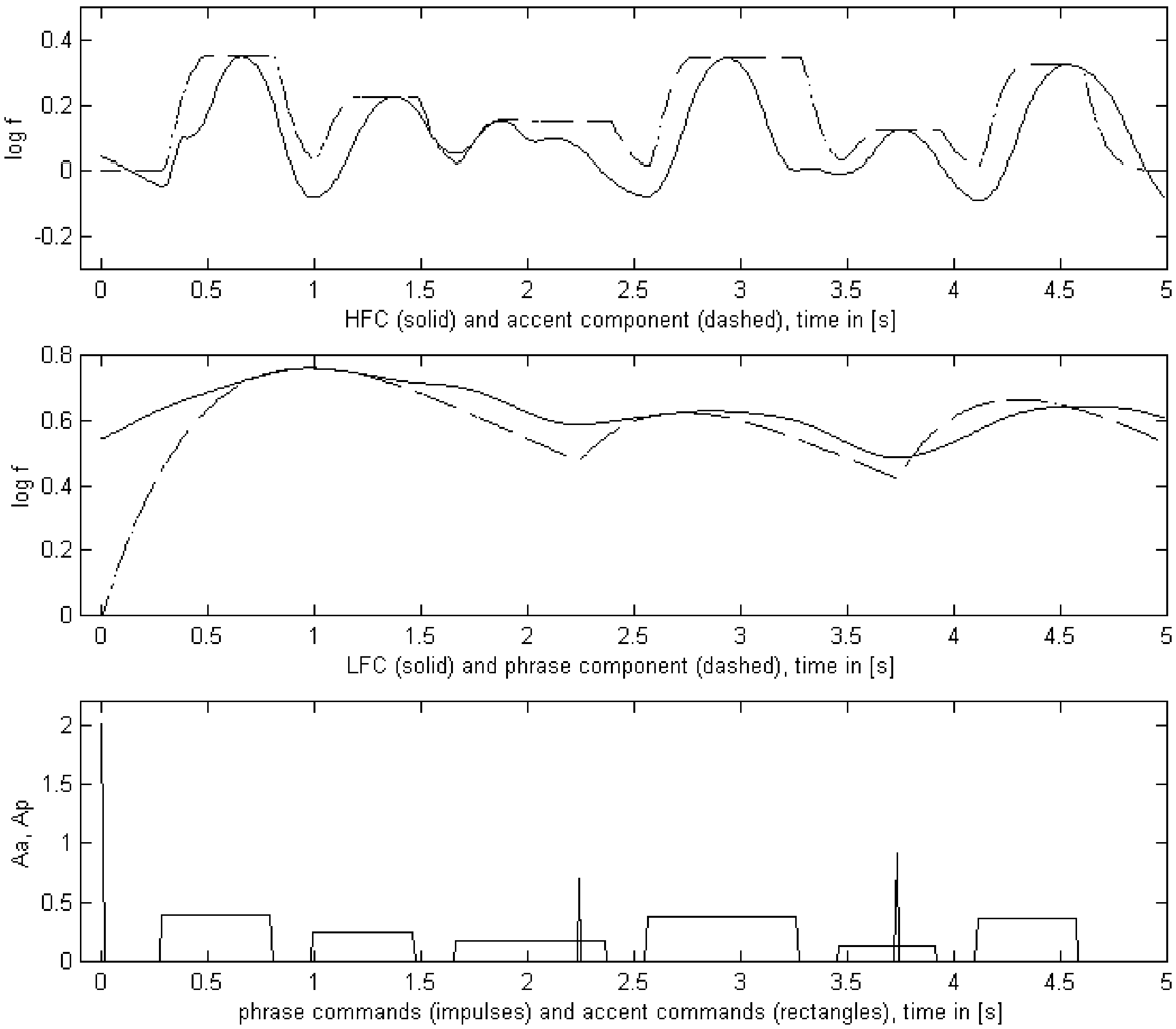


Stage 1



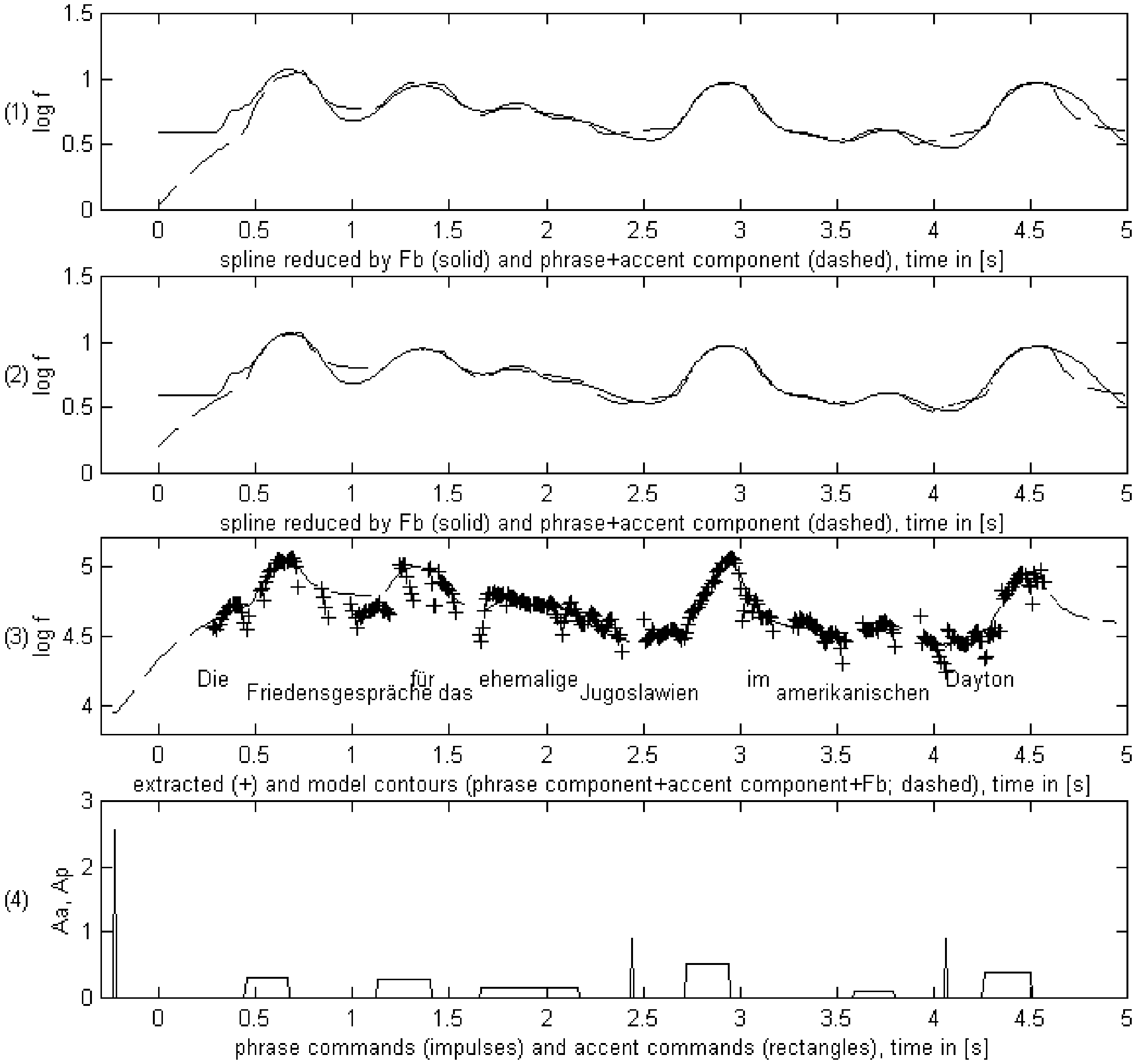
4) Parameter Extraction>

Stage 2

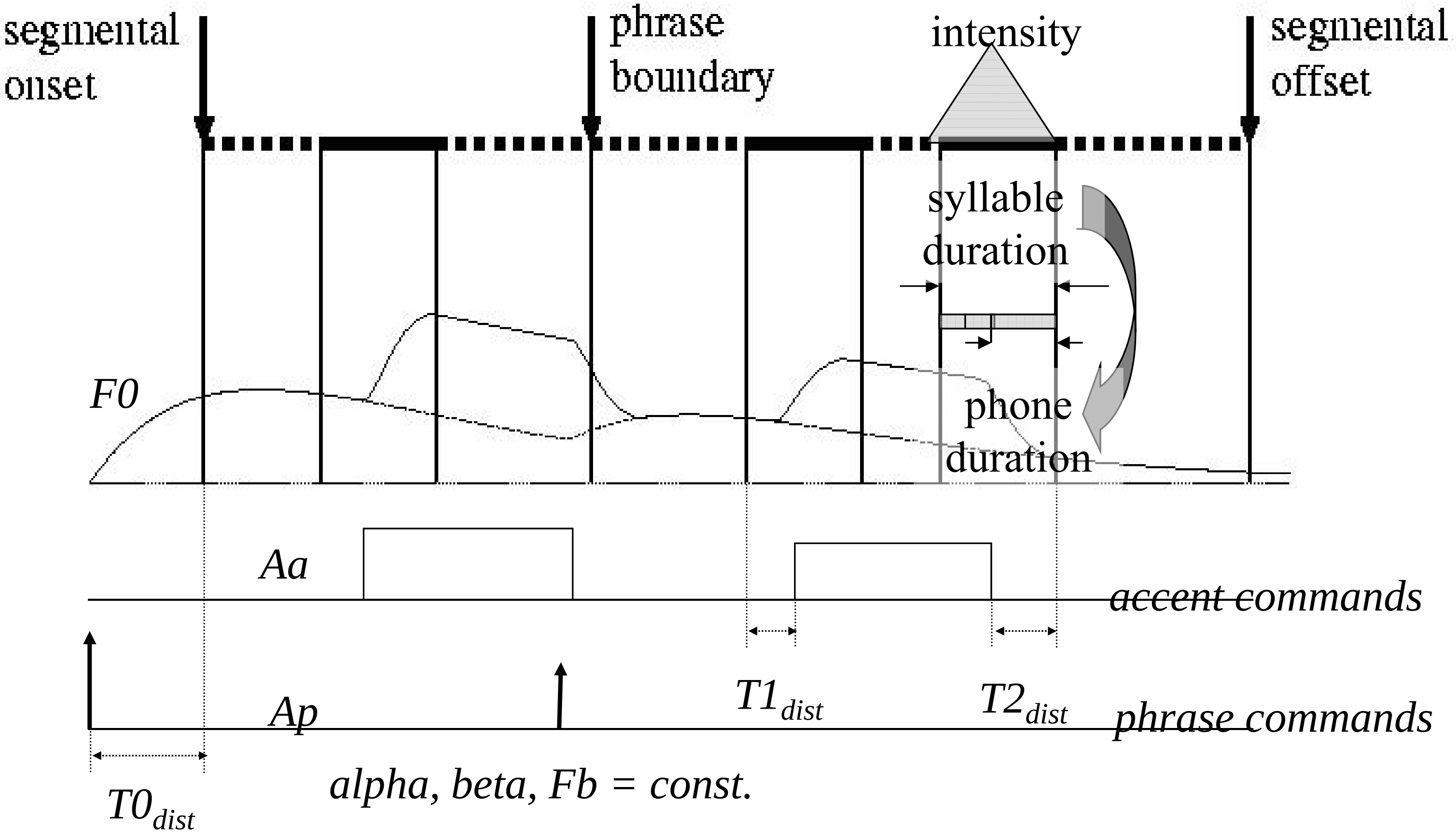


4) Parameter Extraction>

Stage 3



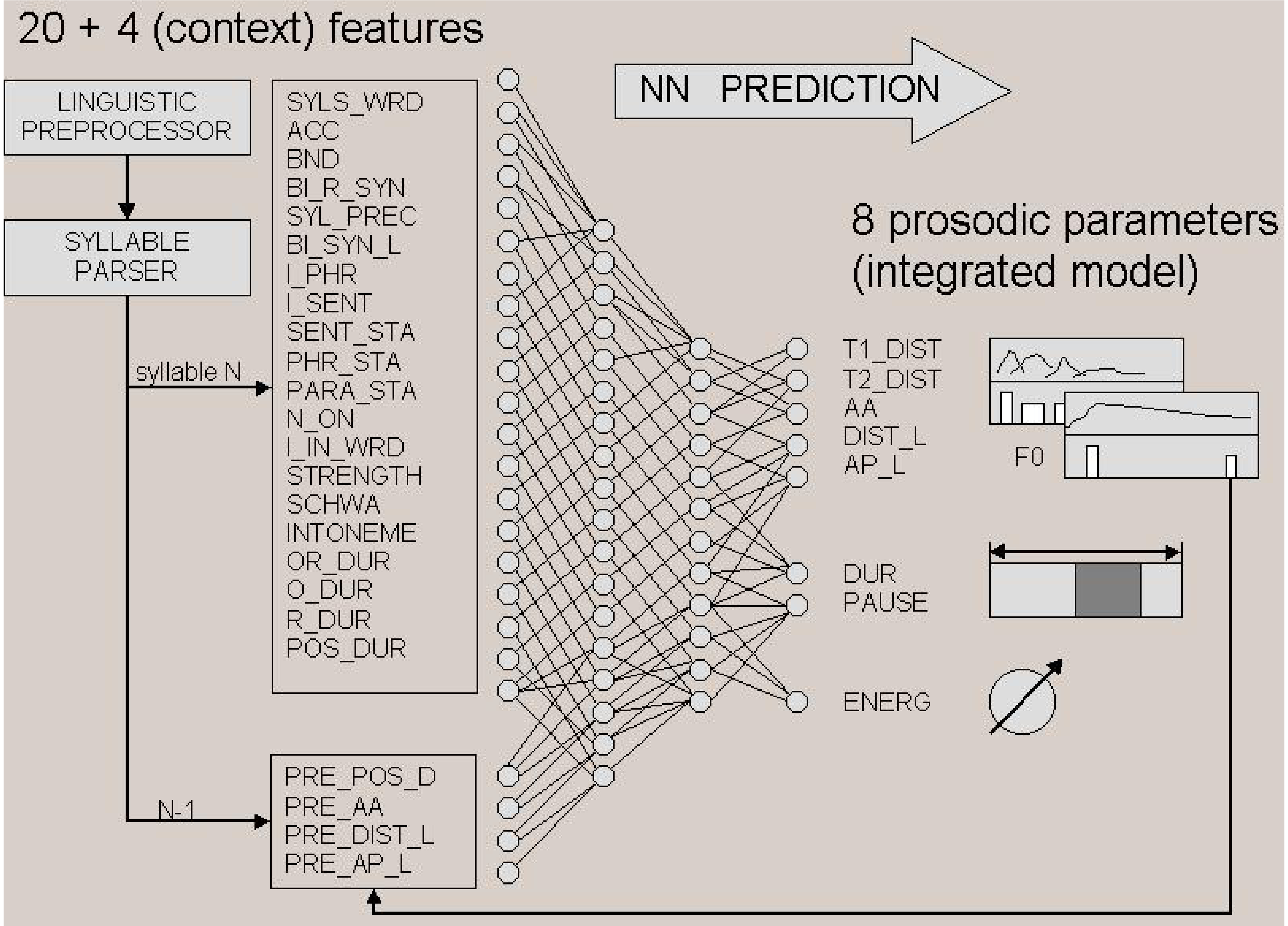
5) Statistical Modeling



alpha, beta, Fb = const.

5) Statistical Modeling>

FFNN structure
(Oliver Jokisch)

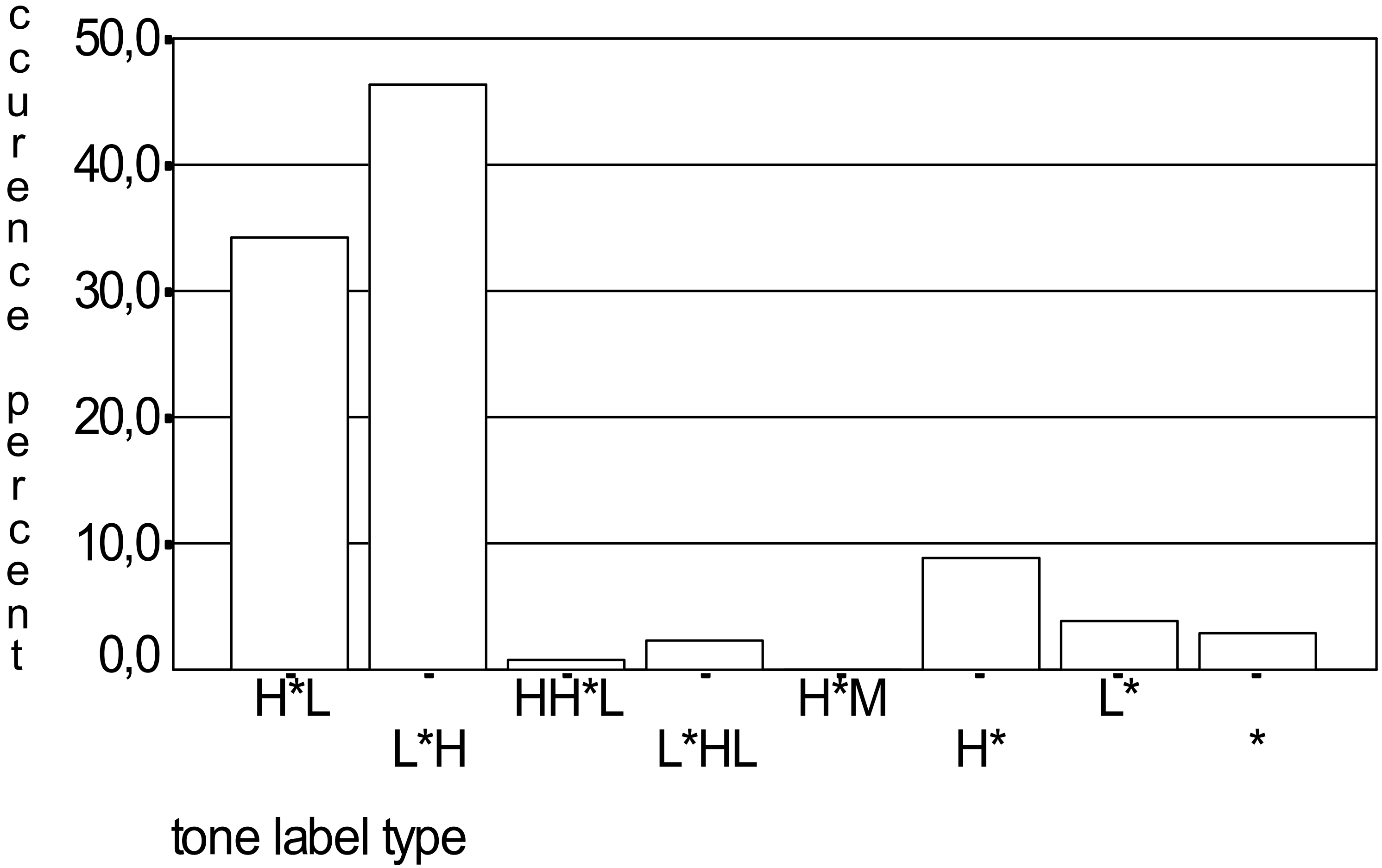


Input and Output Parameters of Prosodic Model

Output Parameter *out* of Model	Predictor Variable *in* of Model	*r (out,in)*	*N*
syllable duration	sum of duration means of phone classes in syllable	.640	13151
	boundary depth (right), 0=clitic, 1=word, 2=phrase, 3=sentence, 4=paragraph	.464	13151
	strength (0=unstressed, 1=stressed, 2=accented)	.349	13151
	nucleus schwa/non-schwa	-.191	13151
Aa	type of intoneme (tone switch class)	.257	3022
	part-of-speech	.128	3022
	phrase index in sentence	-.115	3022
$T1_{dist} = T1-t_{on}$	type of intoneme	.508	3022
	number of phones in syllable onset	.154	3022
$T2_{dist} = T2-t_{off}$	type of intoneme	.384	3022
	number of phones in syllable rhyme	-.198	3022
Ap	boundary depth (left)	.696	1047
	index of phrase in sentence	-.507	1047
	duration of preceding phrase	.320	1047
	Ap of preceding phrase command	-.184	1047
	duration of current phrase	.110	1047
$T0_{dist}= t_{on}-T0$	distance from preceding phrase command	.256	1047
intensity (mean frame power *rms* in syllable)	index of phrase in sentence	-.206	13151
	coda voiced	.141	13151
	index of syllable in phrase	-.124	13151
pause	boundary depth (left)	.622	1047
	index of phrase in syllable	-.376	1047

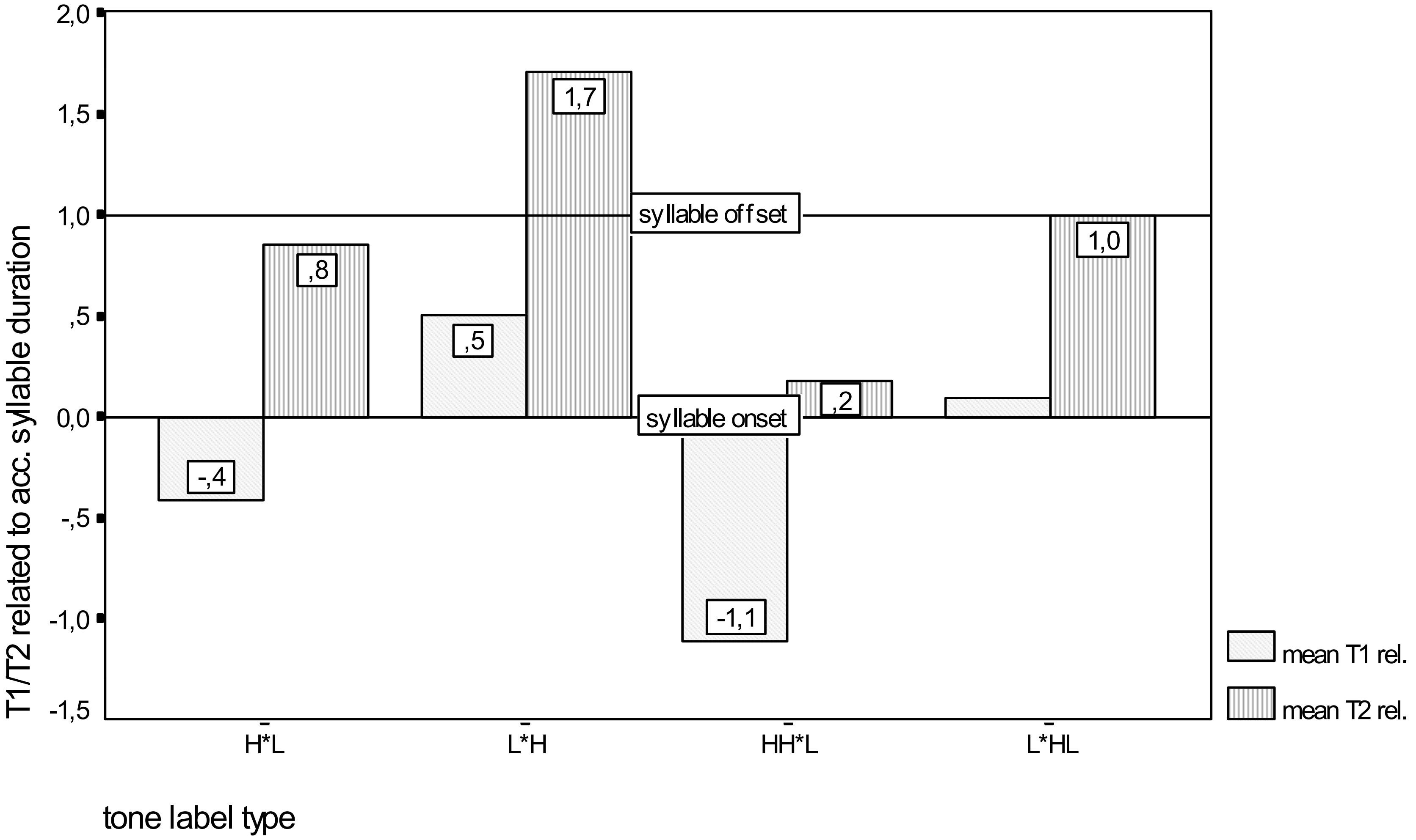
6) Superpositional vs. Autosegmental Modeling of F0>
Results of Comparision

Accentuation: Distribution of accent label types



H*L ⇔ I↓-intoneme L*H⇔ N↑-intoneme

Accentuation: Tone Labels and *T1/T2*



Results of Comparision

Phrasing: BIs vs. Phrase Commands

- BI4: 97 % aligned with phrase command, mean *Ap*: 1.32
- BI3: 57 % aligned with phrase command, mean *Ap:* 0.67
- only sentence boundaries: 100 %
- mean Ap for paragraph onset, sentence onset, intra-sentence boundaries: 2.28, 1.68, 0.8

7) Discussion and Conclusions

- The Fujisaki is applicable to any language as it is production-oriented and has a physiological interpretation
- The choice of parameters and threshholds needs to be guided by linguistic theory
- The quantitative model preserves the macro-intonation
- Can be used to perform unbiased first guess of ToBI accent labels or intoneme types
- Minor phrase boundaries require evaluation of additional cues, such as pre-boundary lengthening and short pauses