FLST: Prosodic Models for Speech Technology

Bernd Möbius
moebius@coli.uni-saarland.de

http://www.coli.uni-saarland.de/courses/FLST/2014/

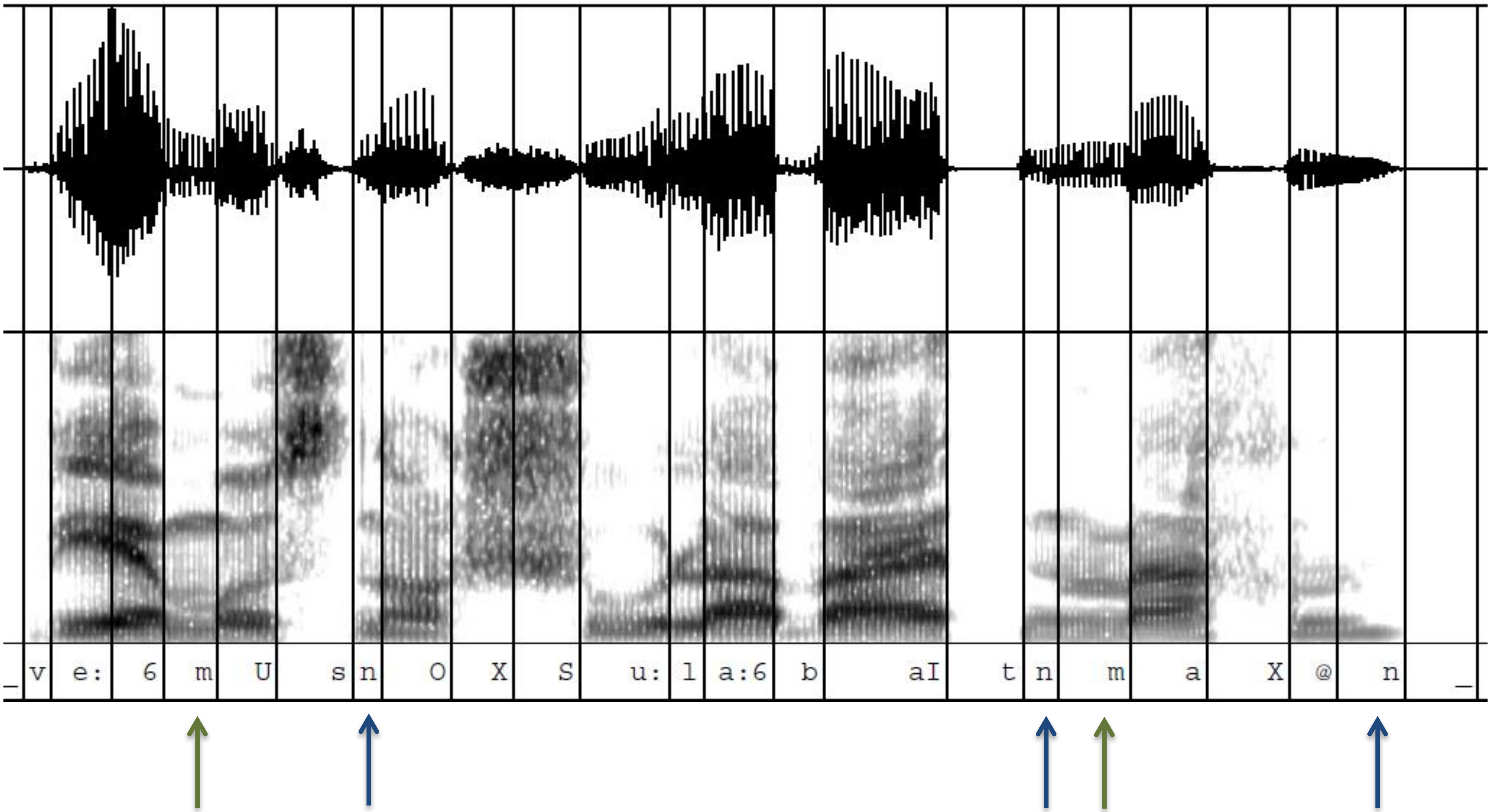
Prosody: Duration and intonation

- Temporal and tonal structure in speech synthesis
  - all synthesis methods
    - use models to predict duration and F0
    - models are trained on observed duration and F0 data
  - Unit Selection:
    - phone duration and phone-level F0 used in target specification
    - F0 smoothness considered
  - HMM synthesis: duration modeled by probability of remaining in the same state

Duration prediction

- Task of duration model in TTS:
  - predict duration of speech sound as precisely as possible, based on factors affecting duration
  - factors must be computable/inferrable from text

Duration prediction



Duration prediction

- Task of duration model in TTS:
  - predict duration of speech sound as precisely as possible, based on factors affecting duration
  - factors must be computable/inferrable from text
- Why is this task difficult?
  - extremely context-dependent durations, e.g. [ɛ] = 35 ms in *jetzt*, 252 ms in *Herren*
  - factors: accent status of word, syllabic stress, position in utterance, segmental context, …
  - factors define a huge feature space

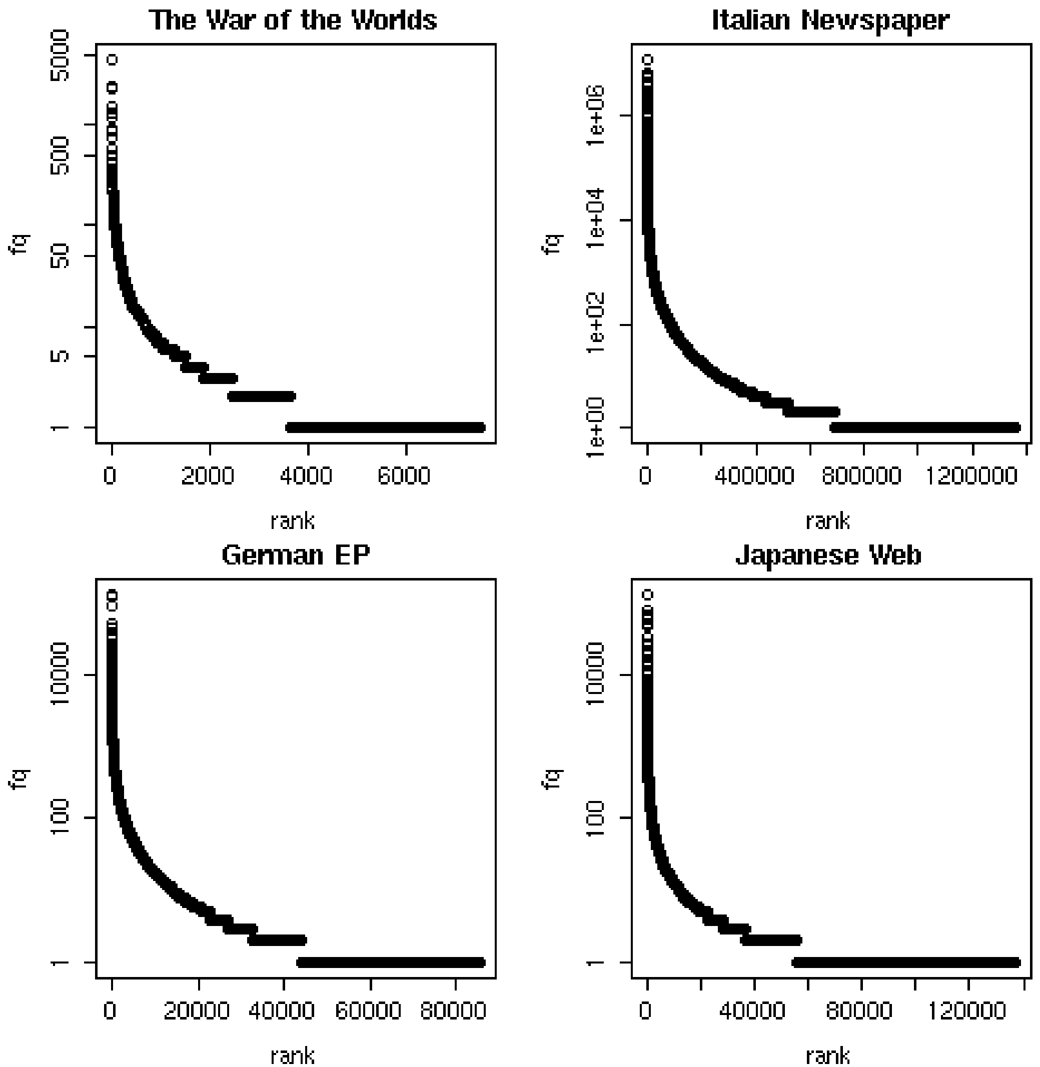
Duration models

- Automatic construction of duration models
  - general-purpose statistical prediction systems
    - Classification and Regression Trees [Breiman et al. 1984; e.g. Riley 1992]
    - Multiple regression [e.g. Iwahashi and Sagisaka 1993]
    - Neural Nets [e.g. Campbell 1992]
  - statistically accurate for training data
  - but often insufficient performance on new data

Data sparsity

- Why is this a problem?
  - **data sparsity**: feature space (>10k vectors) cannot be covered exhaustively by training data
  - **LNRE distribution**: large number of rare events - rare vectors must not be ignored, because there are so many rare vectors that the probability of encountering at least one of them in any sentence is very high

Data sparsity: word frequencies



Data sparsity

- Why is this a problem?
  - **data sparsity**: feature space (>10k vectors) cannot be covered exhaustively by training data
  - **LNRE distribution**: large number of rare events - rare vectors must not be ignored, because there are so many rare vectors that the probability of encountering at least one of them in any sentence is very high
  - vectors unseen in training data must be predicted by extrapolation and generalization
  - general-purpose prediction systems have poor extrapolation and are not robust w.r.t. missing data

Sum-of-products model

- Current best practice: Sum-of-products model

  [van Santen 1993, 1998; Möbius and van Santen 1996]

  - exploits expert knowledge and well-behaved properties of speech (e.g. directional invariance, monotonicity)
  - uses well-behaved mathematical operations (add./mult.)
  - estimates parameters even for unbalanced frequency distributions of features in training data

Sum-of-products model

- Sum-of-products model: general form
  [van Santen 1993, 1998]

$$\mathrm{DUR}(\vec{f}) = \sum_{i \in K} \prod_{j \in I_i} S_{i,j}(f_j).$$

K : set of indices of product terms

I_i : set of indices of factors occurring in i-th product term

$S_{i,j}$: set of parameters, each corresponding to a level on j-th factor

f_j : feature on j-th factor (e.g., f_1 = Vowel_ID, f_2 = ±stress, ...)

Sum-of-products model

- Sum-of-products model: specific form
  [van Santen 1993, 1998]

$$\mathrm{DUR}(V, C, P) \; = \; exp[S_{1,2}(C)S_{1,3}(P) + S_{2,3}(P) + S_{3,1}(V)]$$

V : vowel identity (15 levels)

C : consonant after V (2 levels: ±voiced)

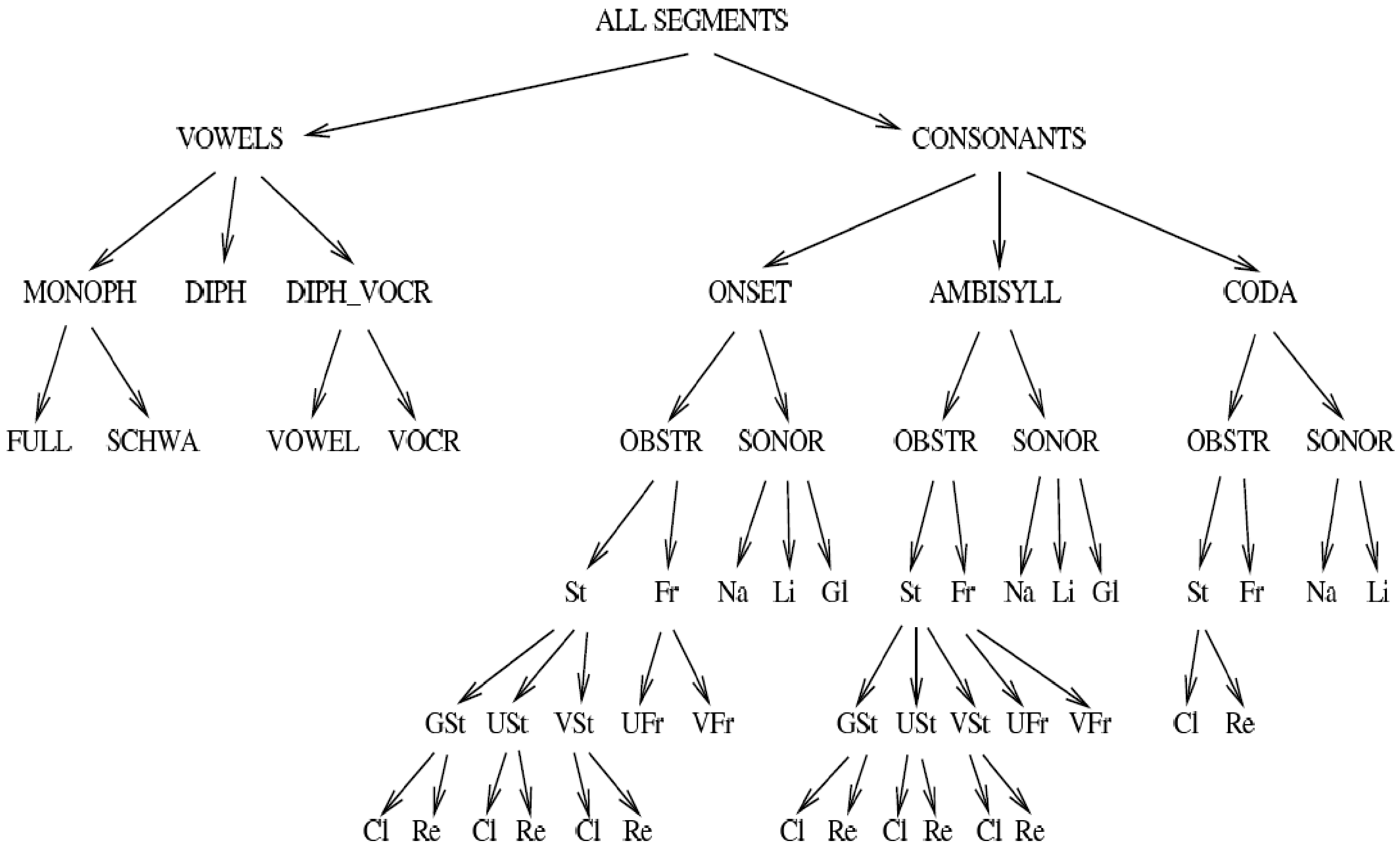
P : position in phrase (2 levels: medial/final)

here: 21 parameters to estimate (2+2 + 2 + 15)

Sum-of-products model

- SoP model requires:
  - definition of factors affecting duration (literature, pilot)
  - segmented and annotated speech corpus
  - greedy algorithm to optimize coverage: select from large text corpus a smallest subset with same coverage
- SoP model yields:
  - complete picture of temporal characteristics of speaker
  - homogeneous, consistent results for set of factors
  - best performance: $r = 0.9$ for observed vs. predicted phone durations (Engl., Ger., Fr., Dutch, Chin., Jap., …)

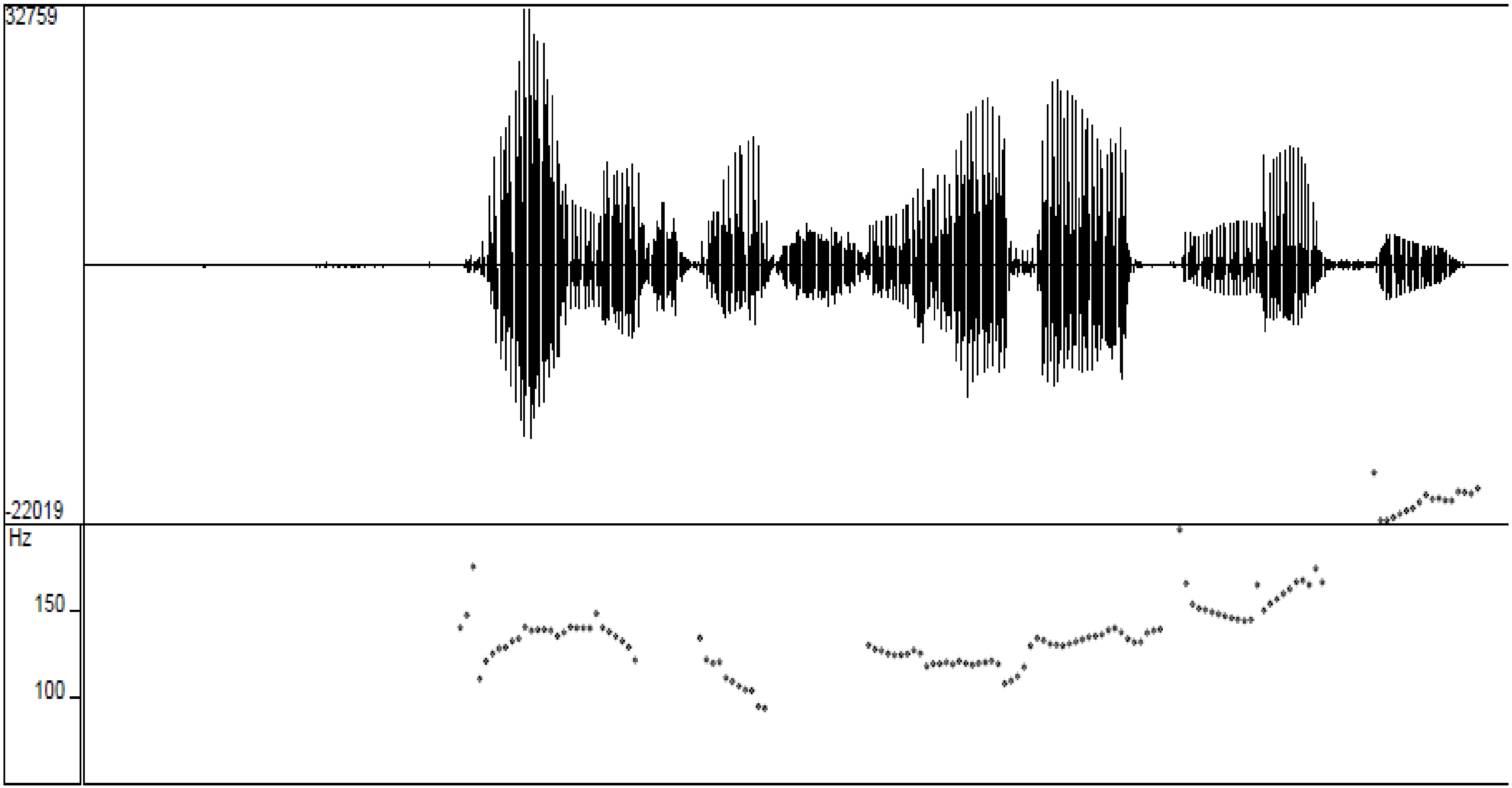
SoP model: phonetic tree



Intonation prediction

- Task of intonation model in TTS
  - compute a continuous acoustic parameter (F0) from a symbolic representation of intonation inferred from text

Intonation (F_0)



Intonation prediction

- Task of intonation model in TTS
  - compute a continuous acoustic parameter (F0) from a symbolic representation of intonation inferred from text
- Intonation models commonly applied in TTS systems:
  - ☑ phonological tone-sequence models (Pierrehumbert)
  - ☑ acoustic-phonetic superposition models (Fujisaki)
  - acoustic stylization models (Tilt, PaIntE, IntSint)
  - perception-based models (IPO)
  - function-oriented models (KIM)

Tone sequence model

- Autosegmental-metrical theory of intonation [Pierrehumbert 1980]
  - intonation is represented by sequence of high (H) and low (L) tones
  - H and L are members of a primary phonological contrast
  - hierarchy of intonational domains
    - **IP** – Intonation Phrase; boundary tones: H%, L%
    - **ip** – intermediary phrase; phrase tones: H-, L-
    - **pw** – prosodic word; pitch accents: H*, H*L, L*H, …

Pierrehumbert's model

- Finite-state grammar of well-formed tone sequences

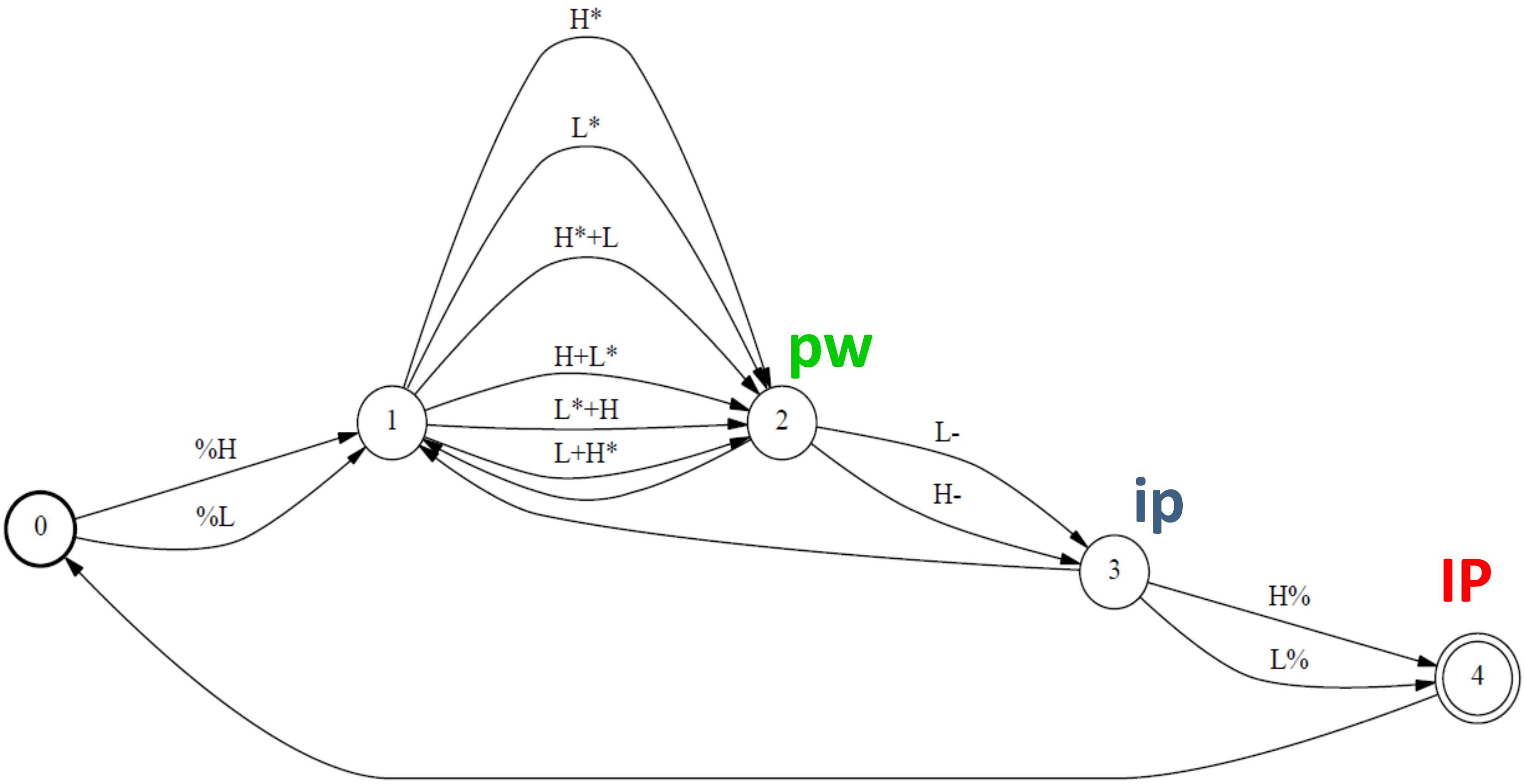
$$\left(\left\{ \begin{matrix} \%H \\ \%L \end{matrix} \right\} \left(\left\{ \begin{matrix} H* \\ L* \\ H*+L \\ H+L* \\ L*+H \\ L+H* \end{matrix} \right\}^{+}_{\textbf{pw}} \left\{ \begin{matrix} H- \\ L- \end{matrix} \right\} \right)^{+}_{\textbf{ip}} \left\{ \begin{matrix} H\% \\ L\% \end{matrix} \right\} \right)^{+}_{\textbf{IP}}$$

- Example [adapted from Pierrehumbert 1980, p. 276]

```
That's a remarkably clever suggestion.
             |                 |
%H           H*                H*L  L- L%
```

Pierrehumbert's model

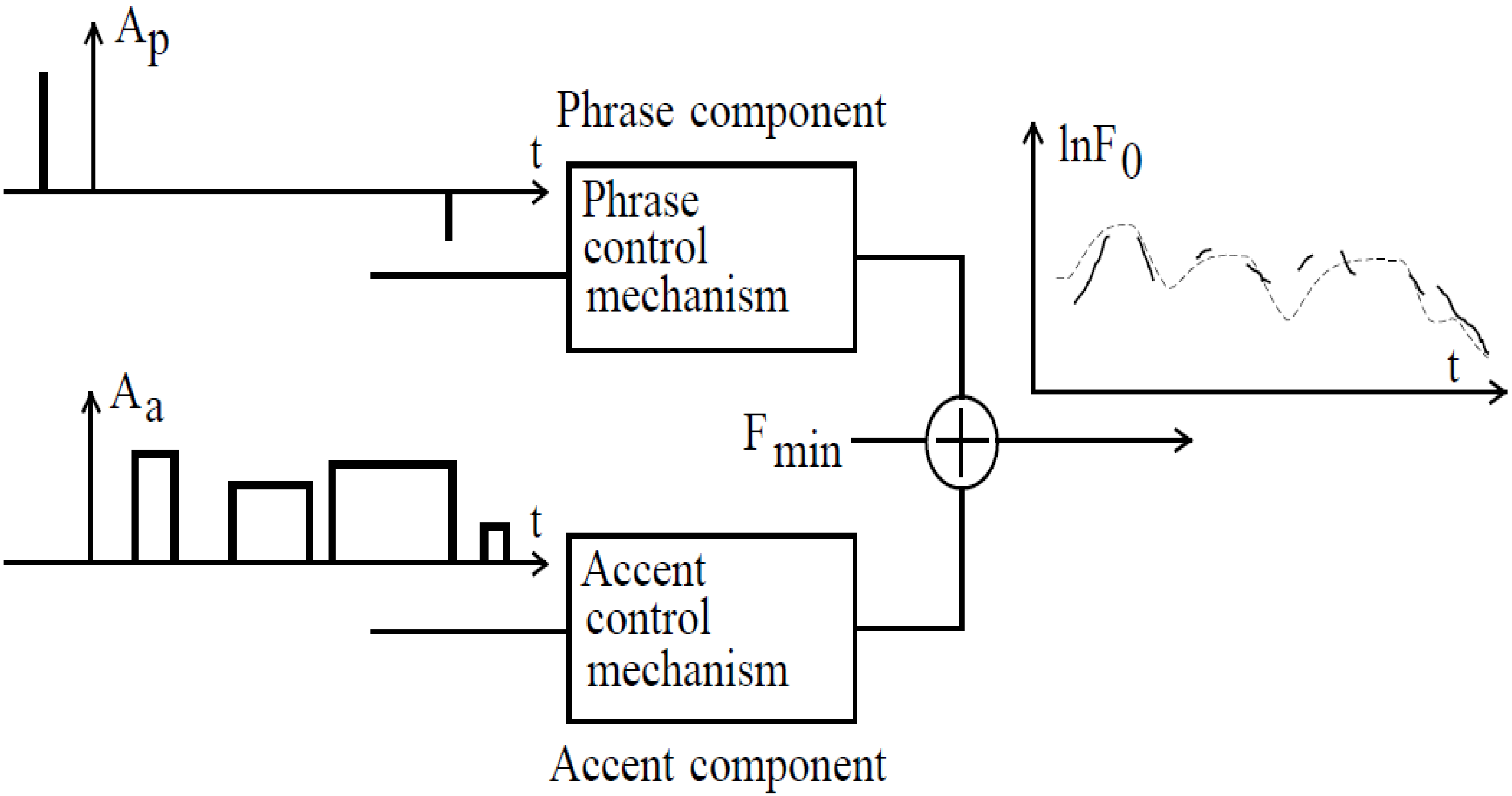
- Finite-state graph



ToBI: Tones and Break Indices

- Formalization of intonation model as transcription system [Pitrelli et al. 1992]
  - phonemic (=broad phonetic) transcription
  - originally designed for American English
  - limited applicability to other varieties/languages
    - language-specific inventory of phonological units
    - language-specific details of F0 contours
  - adapted to many languages (e.g. GToBI, JToBI, KToBI)
  - implemented in many TTS systems
    - abstract tonal representation converted to F0 contours by means of phonetic realization rules

Fujisaki's model

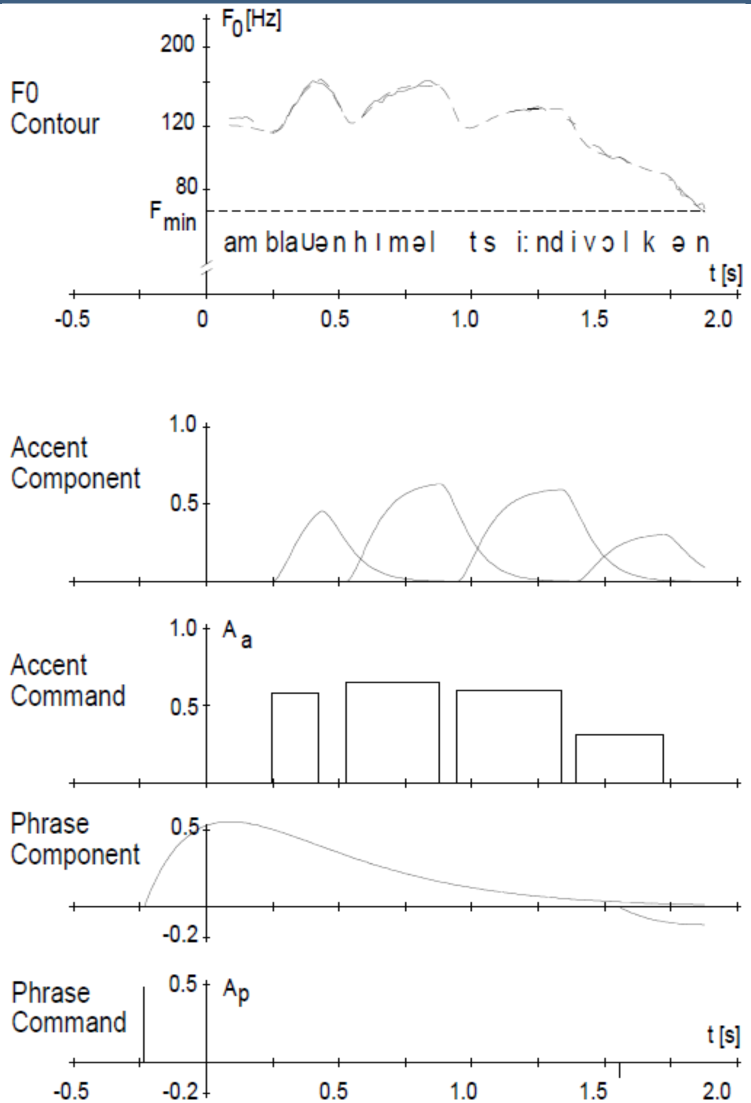


[Fujisaki 1983, 1988; Möbius 1993]

Fujisaki's model

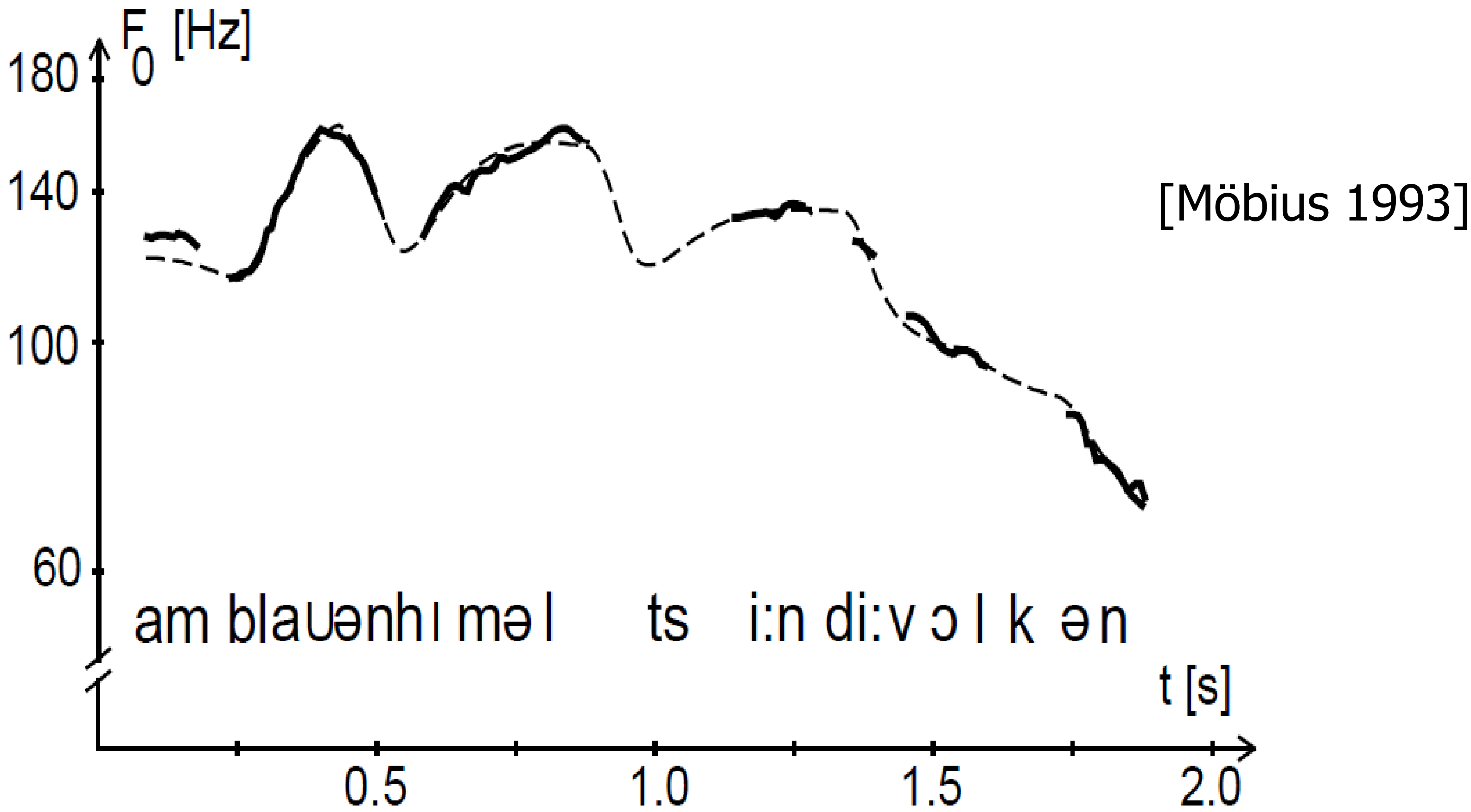
- Properties:
  - superpositional
  - physiological basis and interpretation of components and control parameters
  - linguistic interpretation of components
  - applied to many (typologically diverse) languages
- Origins:
  - Öhman and Lindqvist (1966), Öhman (1967)
  - Fujisaki et al. (1979), Fujisaki (1983, 1988), …

Fujisaki's model: Components



[Möbius 1993]

Fujisaki's model: Example



Approximation of natural F_0 by optimal parameter values within linguistic constraints (accents, phrase structure)

Comparison of models

- Tone sequence or superposition?
  - intonation
    - TS: consists of linear sequence of tonal elements
    - SP: overlay of components of longer/shorter domain
  - F0 contour
    - TS: generated from sequences of phonological tones
    - SP: complex patterns from superimposed components
  - interaction
    - TS: tones locally determined, non-interactive
    - SP: simultaneous, highly interactive components

F_0 as a complex phenomenon

- Main problem for intonation models: linguistic, paralinguistic, extralinguistic factors – all conveyed by F0
  - lexical tones
  - syllabic stress, word accent
  - stress groups, accent groups
  - prosodic phrasing
  - sentence mode
  - discourse intonation
  - pitch range, register
  - phonation type, voice quality
  - microprosody: intrinsic and coarticulatory F0

More on prosody in speech technology:
ASR (Wed Jan 28)

Thanks!