Poster presented at 14th ICPhS, San Francisco, 1999

INTONATION MODELLING IN PROSYNTH

JILL HOUSE, JANA DANKOVIČOVÁ, MARK HUCKVALE

UNIVERSITY COLLEGE LONDON, UK

http://www.phon.ucl.ac.uk/project/prosynth.htm

EPSRC grant no. GR/L52109

ABSTRACT

ProSynth uses a hierarchical prosodic structure (implemented in XML) as its core linguistic representation. To model intonation we map template representations of F0 contours onto this structure. The template for a particular pitch pattern is derived from analysis of a labelled speech database. For a falling nuclear pitch accent this template has three turning points: two which define the F0 peak and one marking the end of the F0 fall. Statistical analysis confirmed that the alignment and shape of the template are sensitive to the properties of the structure and provided quantitative values for F0 synthesis. Our results suggest that phonetic interpretation of the nuclear pitch accent is best related to the accented Foot rather than to the accented syllable. The F0 information is integrated with temporal and segmental information to determine parameter values for synthesis.

INTRODUCTIONs

Hypothesis

Use of a hierarchical prosodic structure to model and integrate timing, intonation and fine acoustic detail will make synthesis more natural and robust.

Aim for modelling F0

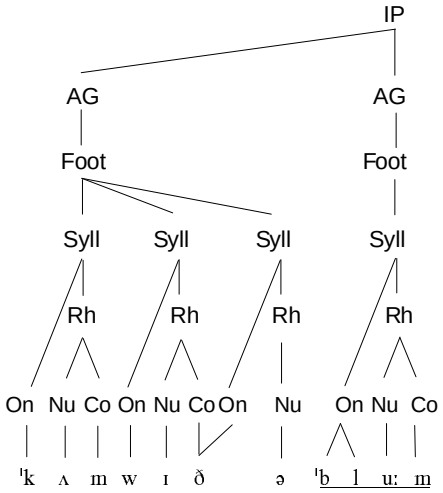
To identify and model the systematic variation that is related to aspects of the structure.

ProSynth principles

- non-linear linguistic representation (hierarchical prosodic structure)
- declarative principles for one-step phonetic interpretation
- phonological and phonetic information is distributed across nodes in structure as attributes and parameter values
- phonetic interpretation may be sensitive to information at any level
- system-independent description of the linguistic structures
- open computational architecture for synthesis (using XML)

Prosodic hierarchy

- IP (intonation phrase) consists of one or more AGs (accent groups: domain of pitch accent configuration)
- AGs consist of one or more Feet (rhythmical units)
- each Foot contains one or more syllables
- accented syllable = leftmost syllable in leftmost Foot of an AG
- last accented syllable in IP = IP nucleus
- relationships between units at the same level are determined by headedness



PROCEDURE

Material

- male speaker, Southern British English
- medium size database (458 utterances) exemplifying a subset of possible structures
- selected structures:
  - up to two AGs
  - AG with up to two Feet
  - Feet up to two syllables
  - controlled for Onset and Rhyme type in the IP nuclear syllable
- falling IP nuclear contour (declarative) H* L- L%
- automatic segmentation, hand-corrected
- F0 calculated from simultaneously recorded laryngograph signal

Example utterances (IP nucleus underlined)

- 1 AG

  do you 'mind
  get a 'pint
  in a 'line
  with a 'rope
  be'low

  to re'mind us
  with a 'needle
  they were 'hopeful

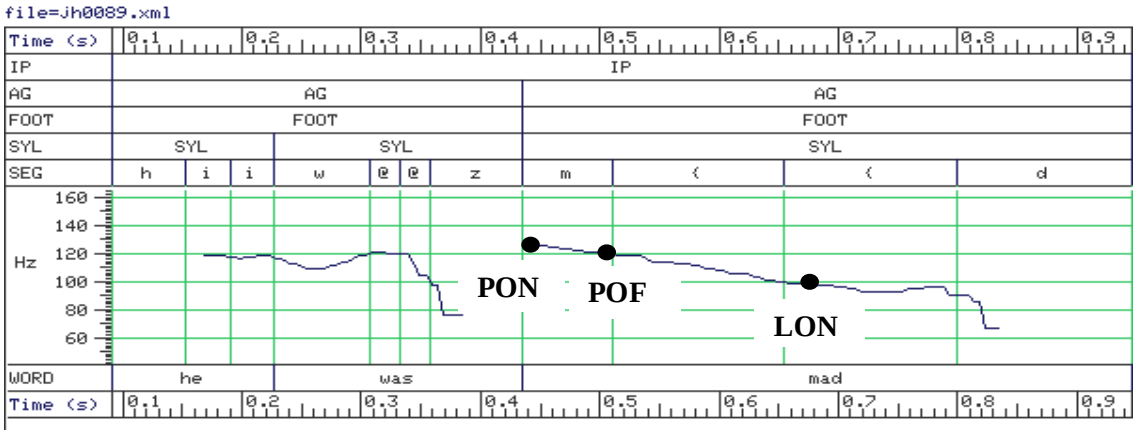
- 2 AGs

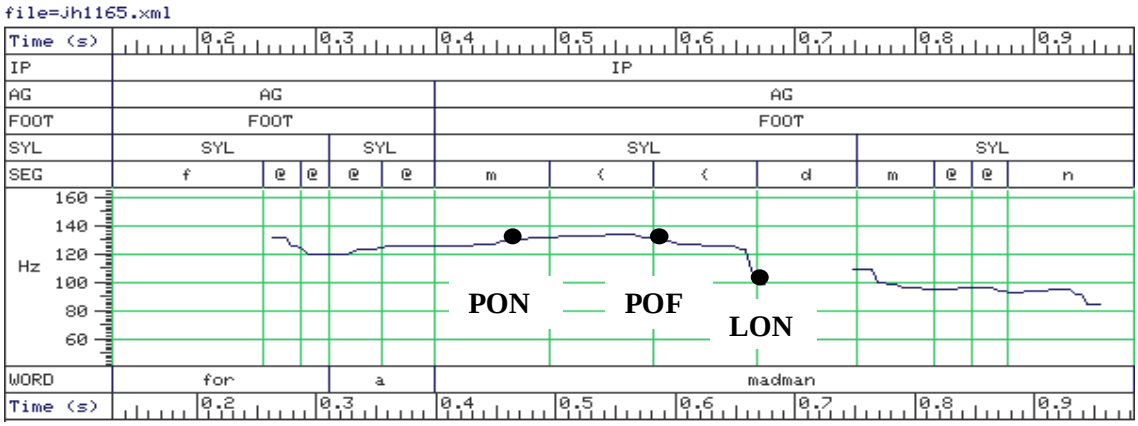
  'come with a 'bloom
  a 'man in a 'room
  a 'face in a 'crowd

Stages in the analysis

1. visual analysis

- identifying the minimum number of turning points (defining the template) within IP nucleus
- observation of regularities in alignment of template to structure





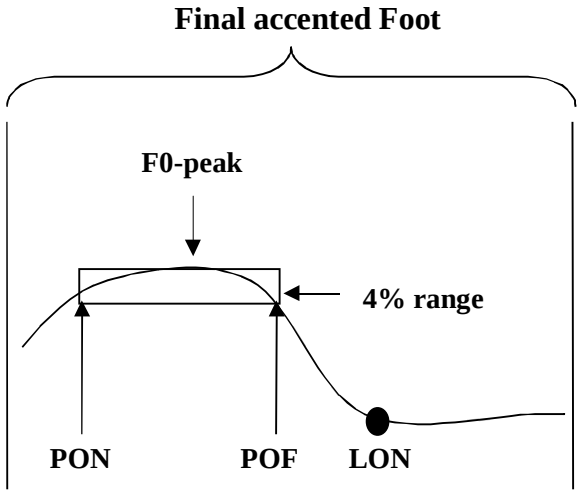
- turning points
  - two points for the peak (many peaks were really plateaux)
    - PON – peak onset
    - POF – peak offset
  - LON – level onset
    (the point from which the low tone spreads till the end of voicing)

2. informal auditory verification (MBROLA)

3. automatic identification of PON, POF and LON and temporal alignment with respect to the beginning of accented syllable

Method

- absolute F0 peak located
- PON and POF located by finding the range of times around the peak where F0 value was within 4% range
- LON - earliest point at which the F0 contour dipped 75% down from the peak and the mean value of final 50 ms



4. statistical analysis

- analysis of variance (General Linear Model) on the temporal alignment of PON, POF and LON:
  - alignment of PON and POF expressed in terms of:

    (i) distance from the beginning of Foot in proportion to accented syllable duration
    (ii) distance from the beginning of Foot in proportion to Foot duration
    (beginning of accented syllable = beginning of Foot)

  - peak duration
  - alignment of LON expressed as a distance from the beginning of the Foot in proportion to Foot duration

- using factors:

Onset type
- approximant
- nasal
- devoiced sonorant in cluster ('clnovoi')
- voiced sonorant in cluster ('clvoi')
- voiced obstruent
- voiceless obstruent
- empty Onset

Coda type
- sonorant
- voiced obstruent
- voiceless obstruent
- empty Onset

Foot type
- NOTAIL (monosyllabic)
- TAIL (polysyllabic)

RESULTS OF THE STATISTICAL ANALYSIS

1. PON and POF alignment

(i) distance from the beginning of syllable (Foot) in proportion to syllable duration

PON

Overall model (75% variance explained)
Significant factors ($p < 0.001$)

- Onset type
- Foot type
- Onset type*Foot type

NOTAIL (67% variance explained)

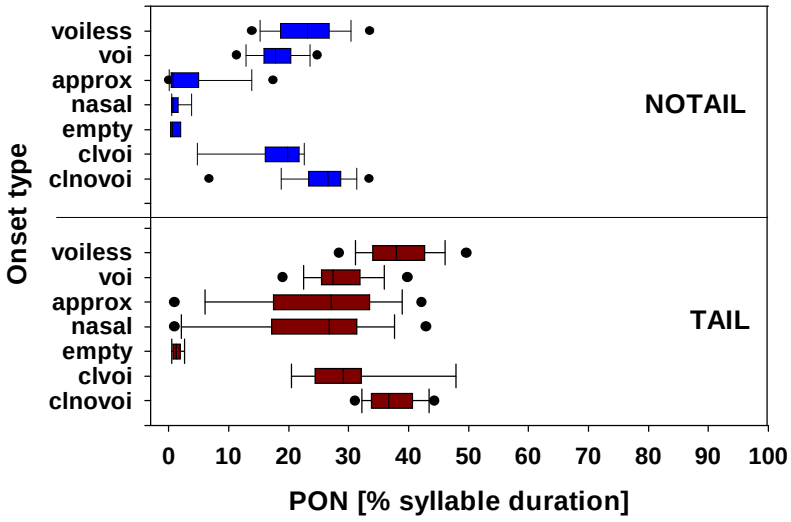
Significant factor ($p < 0.001$)

- Onset type
  (empty, nasal and approximants
  vs. all other Onset types)

TAIL (45% variance explained)

Significant factor ($p < 0.001$)

- Onset type
  (empty vs. nasal and approximants
  vs. others)



(dots = 5 and 95 percentiles
whiskers = 10 and 90 percentiles
boxes = 25 and 75 percentiles
line in the box = median)

Fig. 1 PON as a function of Onset type

POF

Overall model (74% variance explained)
Significant factors ($p < 0.001$)

- Onset type
- Coda type
- Foot type
- Onset type*Foot type
- Coda type*Foot type

NOTAIL (29% variance explained)
Significant factors ($p < 0.001$)

- Onset type (empty vs. others)
- Coda type (empty vs. others)

TAIL (38% variance explained)
Significant factors ($p < 0.001$)

- Onset type (empty vs. others)
- Coda type (voiceless vs. others)
- Onset type*Coda type

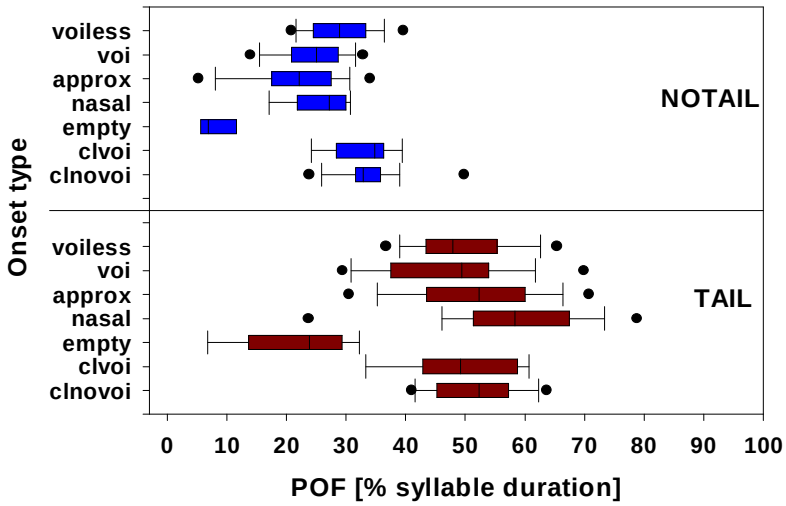


Fig. 2 POF as a function of Onset type

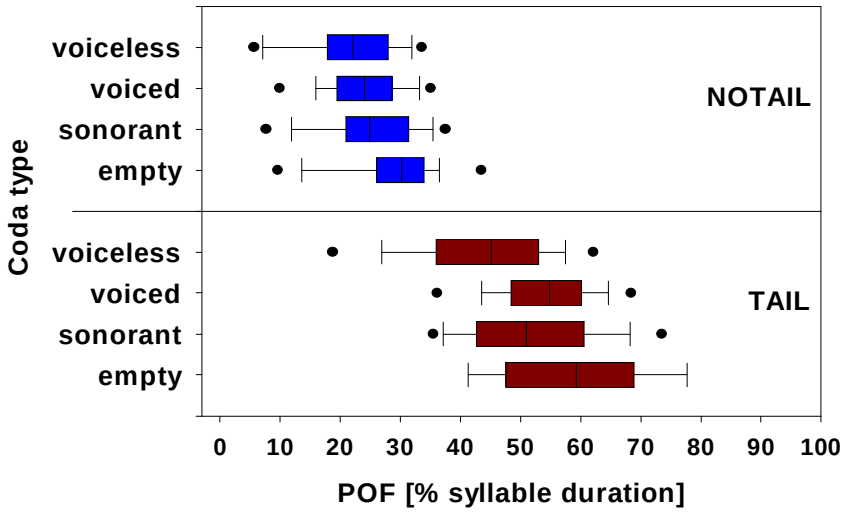


Fig. 3 POF as a function of Coda type

(ii) distance from the beginning of syllable (Foot) in proportion to Foot duration

- identical statistical analysis was carried out for PON and POF in relation to Foot duration
- results for NOTAIL Feet are the same as in (i) since Foot = syllable

PON

TAIL (50% variance explained)

Significant factor ($p < 0.001$)
- Onset type (empty vs. nasal, approximants and voiced vs. others)

POF

TAIL (30% variance explained)

Significant factors ($p < 0.001$)
- Onset type (empty vs. others)
- Coda type (voiceless vs. others)

2. Peak duration (PON–POF distance)

(i) related to syllable duration (Fig. 4)

- consistent rightward shift in alignment of both PON and POF in TAIL Feet
- proportional peak duration longest in syllables with sonorant Onsets (nasals and approximants)
- peak duration across all Onset types in TAIL feet takes a larger proportion of the syllable

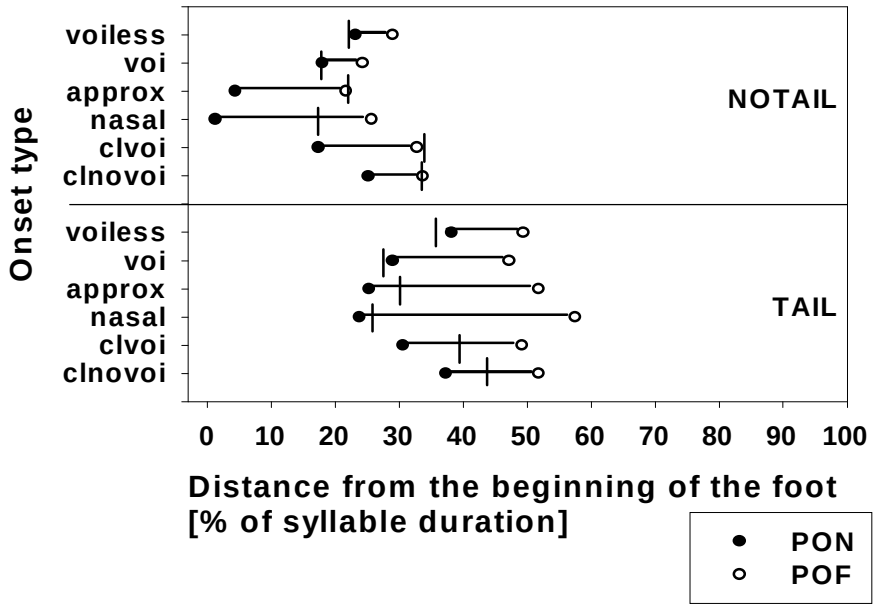


Fig. 4 Mean peak duration as a function of Onset type (related to syllable)
(vertical lines = mean values for the beginning of Rhyme)

(ii) related to Foot duration (Fig. 5)

- no consistent rightward shift in alignment of PON and POF in TAIL Feet
- peak durations in TAIL and NOTAIL Feet occupy comparable proportions of Foot
- longer peaks still observed in syllables with sonorant Onsets

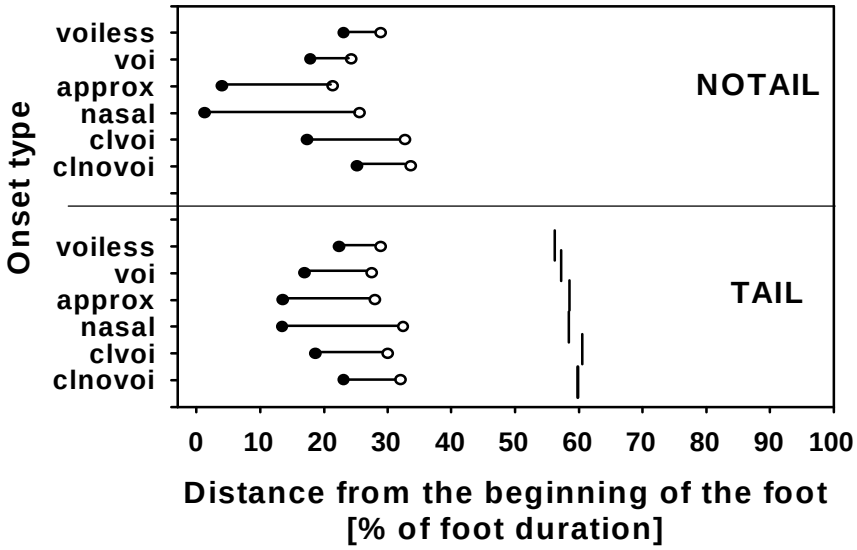


Fig. 5 Mean peak duration as a function of Onset type (related to Foot)
(vertical lines = mean values for syllable boundary)

3. LON alignment

- related to Foot duration

NOTAIL
Significant factor ($p < 0.001$)
- Coda type (voiceless vs. others)

TAIL
No significant factors – LON across all Feet was about 50% of Foot duration

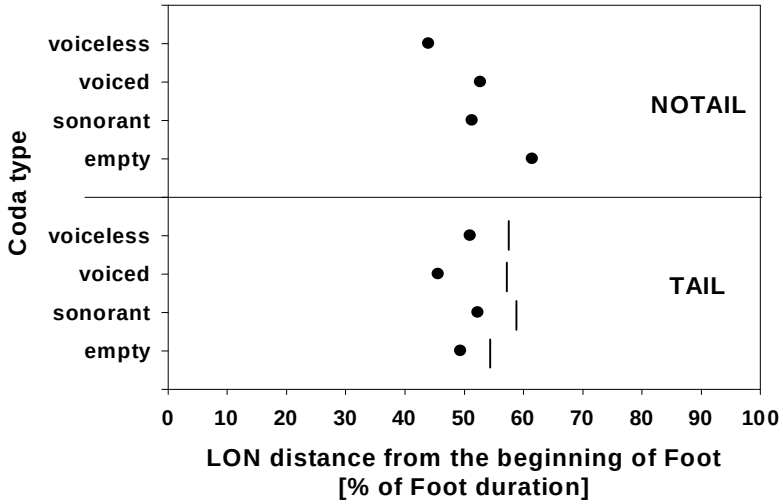


Fig. 6 LON as a function of Coda type

MODELLING F0 TURNING POINTS FOR SYNTHESIS

- temporal alignment for PON, POF and LON, based on the statistical analysis, is now specified at Foot level on the prosodic hierarchy
- phonetic interpretation is sensitive to the identified structural constraints
- F0 values for PON, POF and LON are (for now) based on the visual analysis and auditory evaluation using MBROLA

SUMMARY AND DISCUSSION

- It is important to model both **Peak Onset** and **Peak Offset** (thus recognizing peak duration) to achieve natural sounding synthesis
- Findings about F0 peak alignment reported in the literature sometimes relate to our findings for Peak Onset and sometimes for Peak Offset
- Relating Peak Onset and Peak Offset to ***Foot*** duration (rather than syllable duration) reduces variability in their alignment and peak duration
- **Level Onset** (end of F0 fall) seems to have a consistent anchor point (around the mid-point of the Foot)

Preliminary results from **perceptual testing** (in progress) indicate that correct modelling F0 turning points leads to faster comprehension in a task involving true/false judgements.

Future work

- Extending analysis to IP nuclear Accent Groups (AGs) consisting of (i) single tri-syllabic Foot and (ii) two Feet
- Analysis and modelling of pre-nuclear AGs
- Analysis and modelling of other nuclear pitch accents (e.g. rising tones)
- Perceptual testing on (i) the minimum number of F0 turning points for pre-nuclear and nuclear AGs templates and (ii) alignment of these templates within the prosodic structure