Speech, Hearing and Language: work in progress

Volume 11

Intonation modelling in ProSynth

Jill HOUSE, Jana DANKOVICOVA, and Mark HUCKVALE

Department of Phonetics and Linguistics
UNIVERSITY COLLEGE LONDON

Intonation modelling in ProSynth*

Jill HOUSE, Jana DANKOVICOVA, and Mark HUCKVALE

Abstract
ProSynth uses a hierarchical prosodic structure (implemented in XML) as its core linguistic representation. To model intonation we map template representations of F_0 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 define the F_0 peak and one marks the end of the F_0 fall. Statistical analysis confirmed that the alignment and shape of the template are sensitive to the properties of the structure and also provided quantitative values for F_0 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. In determining parameter values for synthesis, we conclude that F_0 information should be integrated with temporal and segmental information.

1. Introduction

1.1 Hypothesis

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

1.1.1 Aim for modelling F0

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

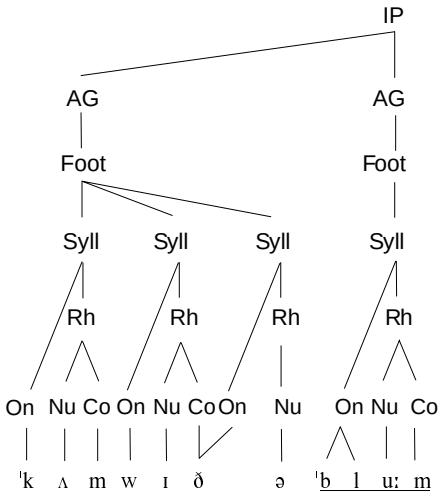
1.1.2 ProSynth principles

- use of a 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)

1.1.3 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

* This paper is adapted from a poster presented at the 14th International Congress of Phonetic Sciences, 1999, San Francisco. It extends and updates the paper published in the ICPhS Proceedings online at: http://synth.phon.ucl.ac.uk/prosynth/ucl_icphs99.pdf.



- relationships between units at the same level are determined by headedness

2. Procedure

2.1 Material

- male speaker, Southern British English
- medium size database (458 utterances) exemplifying a subset of possible structures
- selected structures:
  - ⇒ up to two AGs
  - ⇒ AGs 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
- F_0 calculated from simultaneously recorded laryngograph signal

Example utterances (IP nucleus underlined)

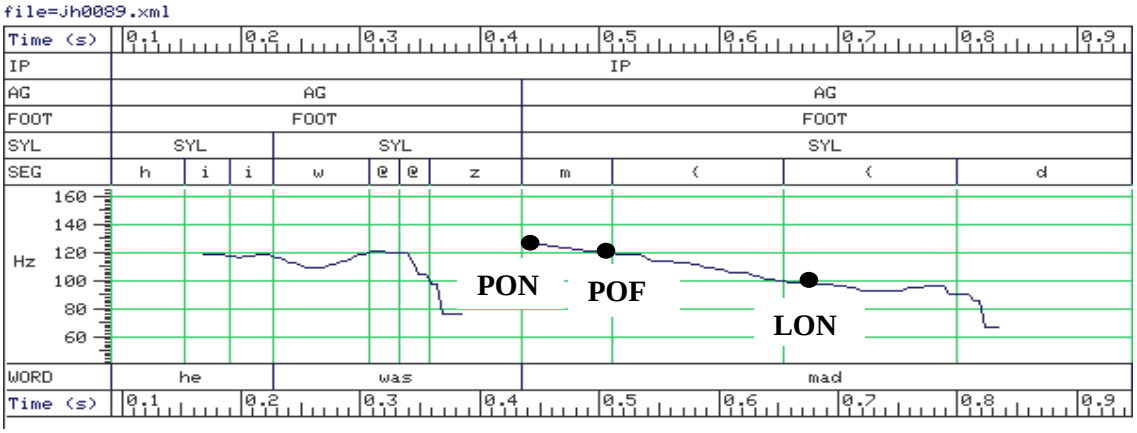
1 Accent Group		**2 Accent Groups**
do you 'mind	to re'mind us	'come with a 'bloom
get a 'pint	with a 'needle	a 'man in a 'room
in a 'line	they were 'hopeful	a 'face in a 'crowd
with a 'rope		
be'low		

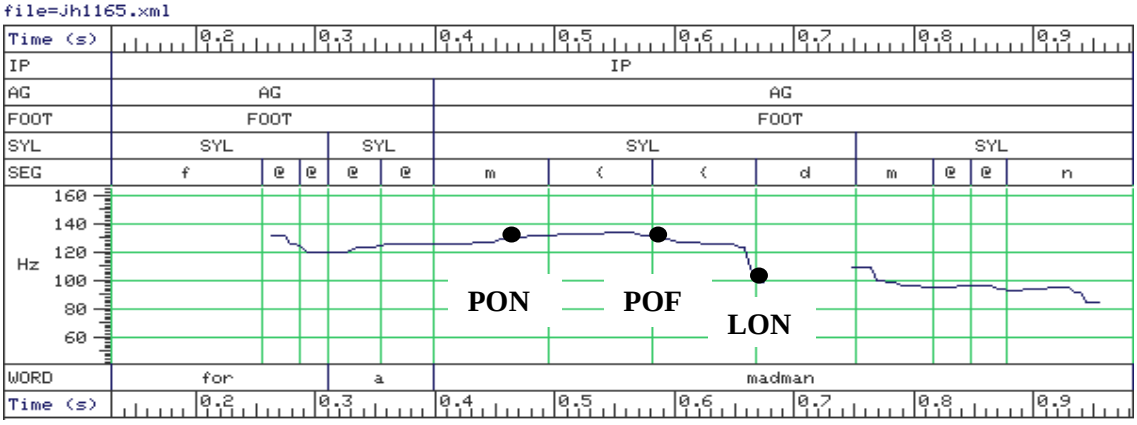
2.2 Stages in the analysis

2.2.1 Visual analysis

- identify the minimum number of turning points (defining the template) within IP nucleus

- observation of regularities in alignment of template to structure





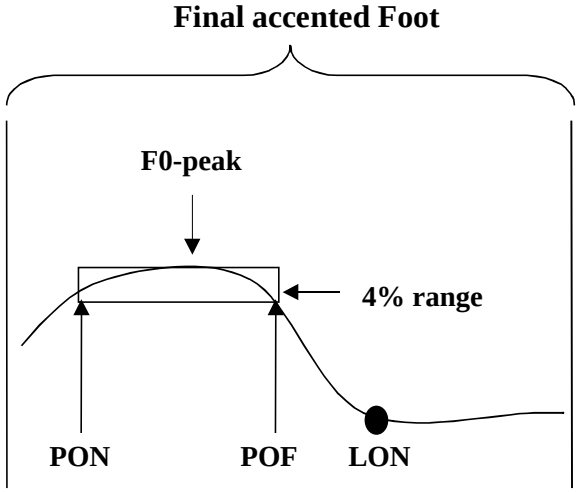
Turning points

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

2.2.2 Informal auditory verification (MBROLA)

2.2.3 Automatic identification of peak onset, peak offset and level onset, and temporal alignment with respect to the beginning of accented syllable - procedure

- Absolute F_0 peak located
- Peak onset and peak offset located by finding the range of times around the peak where F_0 value was within 4% range
- Level onset identified as earliest point at which the F_0 contour dipped 75% down from the peak and the mean value of final 50 ms



2.2.4 Statistical analysis

- analysis of variance (General Linear Model) on the temporal alignment of peak onset and offset and level onset
- alignment of peak onset and offset 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)
- alignment of level onset expressed as a distance from the beginning of the Foot in proportion to Foot duration
- peak duration

Analysis used factors of:

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)

3. Results of the statistical analysis

3.1 Peak onset and offset alignment

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

3.1.1 Peak Onset (Fig. 1)

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

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

NOTAIL (67% variance explained)

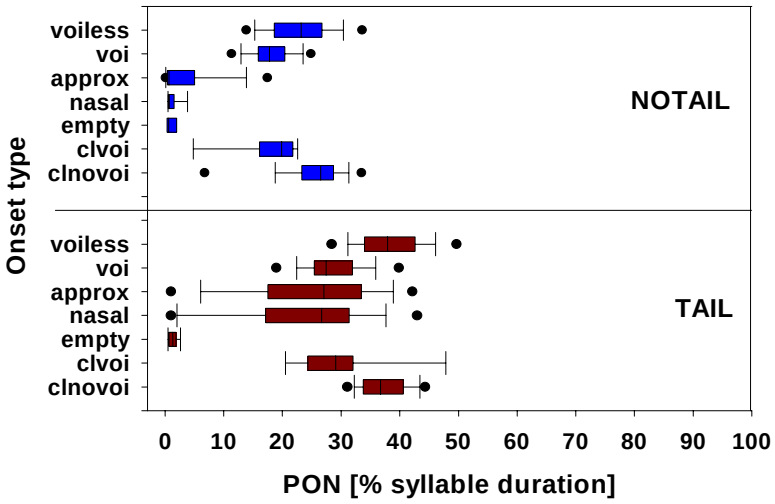
Significant factor ($p < 0.001$)

- 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

Figure 1. *Peak onset as a function of Onset type*

3.1.2 Peak offset (Figures 2 and 3)

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

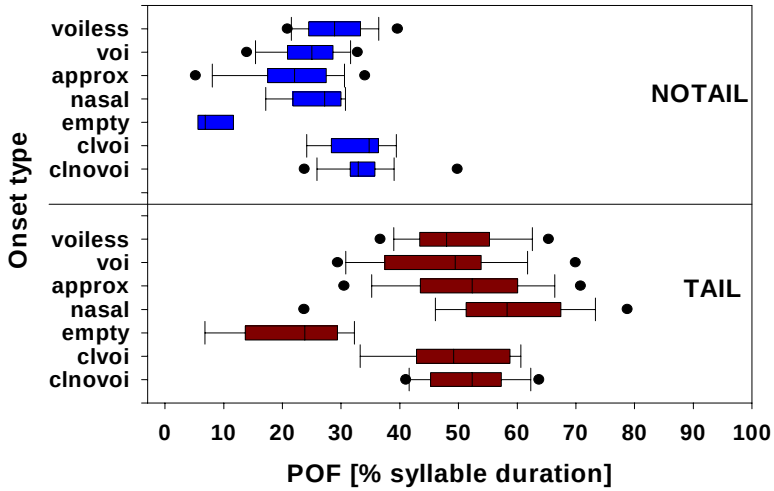


Figure 2. *Peak offset as a function of Onset type*

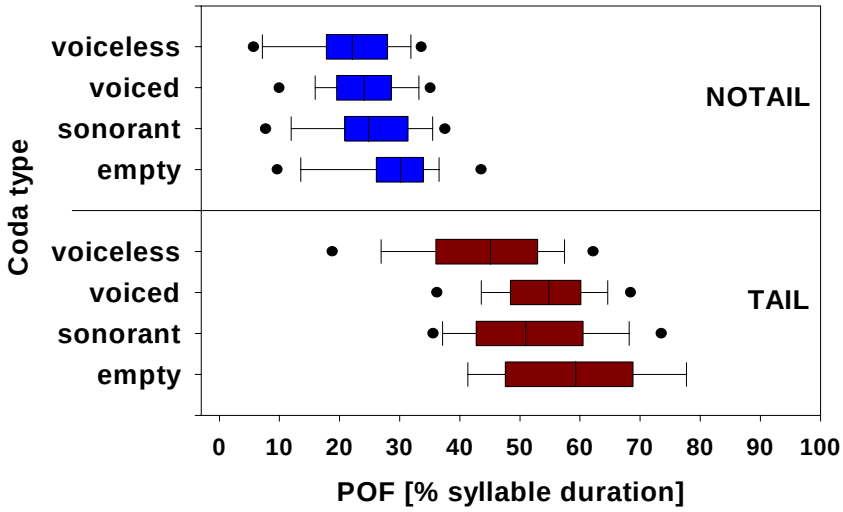


Figure 3. *Peak offset as a function of Coda type*

3.2 Peak onset and offset alignment

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

- identical statistical analysis was carried out for peak onset and offset in relation to Foot duration
- results for NOTAIL Feet are the same as in 3.1.1 and 3.1.2 since Foot = syllable

3.2.1 Peak onset

TAIL (50% variance explained)
Significant factor ($p < 0.001$)

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

3.2.2 Peak offset

TAIL (30% variance explained)
Significant factors ($p < 0.001$)

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

3.3 Peak duration (PON–POF distance) in relation to syllable duration (Fig. 4)

- consistent rightward shift in alignment of both peak onset and offset 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

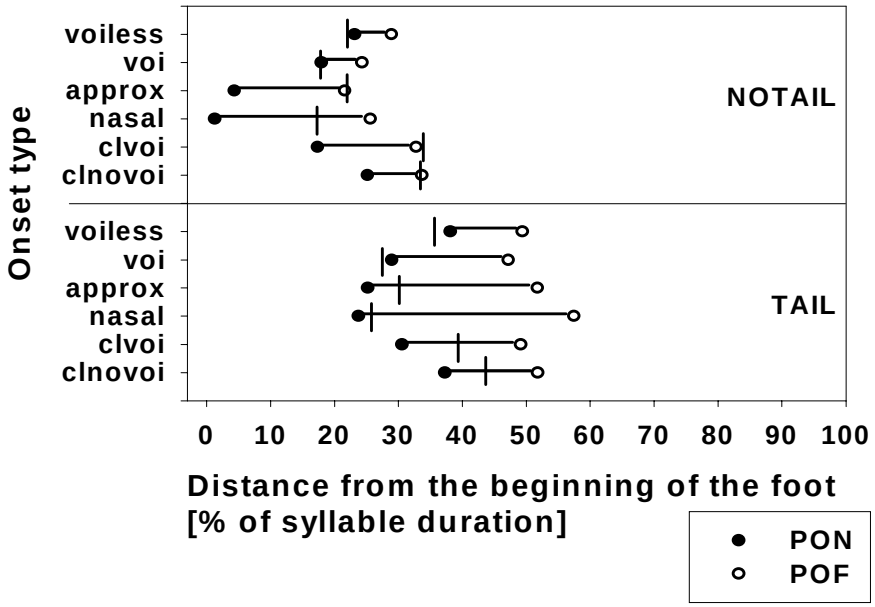


Figure 4. *Mean peak duration as a function of Onset type (related to syllable). Vertical lines = mean values for the beginning of Rhyme.*

3.4 Peak duration (PON–POF distance) in relation to Foot duration (Fig. 5)

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

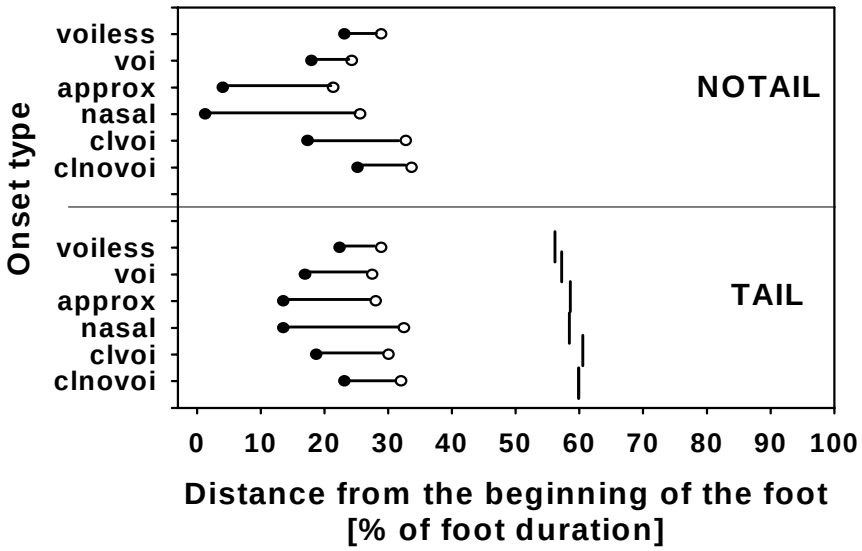


Figure 5. *Mean peak duration as a function of Onset type (related to Foot). Vertical lines = mean values for syllable boundary.*

3.5 Level onset alignment related to Foot duration (Fig. 6)

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

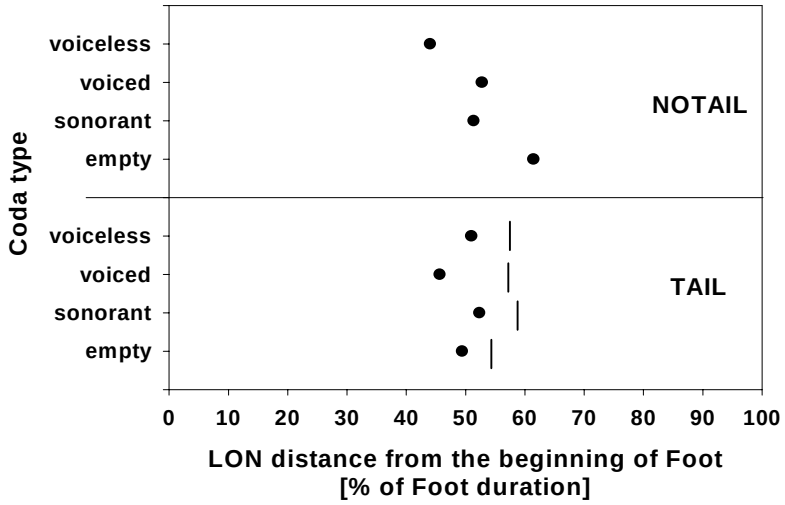


Figure 6. *Level onset as a function of Coda type*

4. Modelling F_0 Turning Points For Synthesis

- temporal alignment for peak onset and offset, and level onset, based on the statistical analysis, is now specified at Foot level on the prosodic hierarchy
- phonetic interpretation is sensitive to the identified structural constraints
- F_0 values for peak onset and offset and level onset are (for now) based on the visual analysis and auditory evaluation using MBROLA

5. 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 F_0 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 F_0 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 F_0 turning points leads to faster comprehension in a task involving true/false judgements.

6. 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 F_0 turning points for pre-nuclear and nuclear AGs templates and (ii) alignment of these templates within the prosodic structure

7. References

Dutoit, T., Pagel, V., Pierret, N., Bataille, F. and van der Vreken, O. 1996. The MBROLA project: towards a set of high-quality speech synthesizers free of use for non-commercial purposes. *Proc. ICSLP'96*, Philadelphia, vol. 3, 1393-1396

Hawkins, S., House, J., Huckvale, M., Local, J. & Ogden, R. 1998. ProSynth: an integrated prosodic approach to device-independent, natural-sounding speech synthesis. *Proc.5th ICSLP'98*, Sydney, 1707-1710.

House, J. and Hawkins, S. 1995. An integrated phonological-phonetic model for text-to-speech synthesis. *Proc. ICPhS XIII*, Stockholm, vol. 2, 326-329.

House, J. and Wichmann, A. 1996. Investigating peak timing in naturally-occurring speech:from segmental constraints to discourse structure. *Speech, Hearing & Language 9*, UCL, 99-117.

Ladd, D.R. 1996. *Intonational Phonology*. Cambridge, CUP

Linton, M. and Gallo, P. S. 1975. *The practical statistician: Simplified handbook of statistics*. Monterey, CA: Brooks/Cole.

Local, J. and Ogden, R. 1997. A model of timing for nonsegmental phonological structure. In van Santen, J., Sproat, R., Olive, J. & Hirschberg, J. (eds.), *Progress in Speech Synthesis*. Springer, New York, 109-122.

Silverman, K. and Pierrehumbert, J. 1990. The timing of prenuclear high accents in English. In Kingston, J. and Beckman, M. (eds.), *Papers in Laboratory Phonology I,* Cambridge. CUP, 72-106.

van Santen, J. & Möbius, B. 1997. Modeling pitch accent curves. In Botinis, A.,Kouroupetroglou, G. and Carayiannis, G. (eds.), *ESCA Workshop on Intonation: Theory, Models and Applications*. Athens, 321-324.

Wichmann, A. and House, J. 1999. Discourse constraints on peak timing in English: experimental evidence. *Proc. ICPhS XIV*, this volume.

8. Acknowledgement

Supported by EPSRC grant no. GR/L52109. The database labelling was undertaken jointly with ProSynth collaborators in York and Cambridge.

9. Further Information

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