To Appear: Proc. Int. Congress Phonetic Sciences, San Francisco, 1999

INTONATION MODELLING IN PROSYNTH: AN INTEGRATED PROSODIC APPROACH TO SPEECH SYNTHESIS

Jill House, Jana Dankovičová, Mark Huckvale
University College London, UK

ABSTRACT

Intonation modelling in ProSynth involves mapping the defining characteristics of an F0 contour on to the constituents of a hierarchical prosodic structure, which constitutes our core linguistic representation. The paper describes the use of a labelled speech database exemplifying selected structures to create a template for a particular pitch pattern in a given context, and the observed systematic structural effects on the alignment and shape of that template. The research confirms the importance of structural domains in determining systematic variation in pitch accent realization. Implemented in XML, our structure integrates intonational, temporal and segmental information to determine coherent parameter values for synthesis.

1. INTRODUCTION

1.1. ProSynth

A fundamental motivating principle in ProSynth [1] is that when timing, intonation and fine acoustic detail are correctly integrated to reflect speech coarticulatory patterns the resulting signal will be perceptually natural, and relatively robust in adverse listening conditions. For the generation of more natural-sounding synthetic speech, we propose a declarative computational model comprising a rich, hierarchical prosodic structure which integrates relevant linguistic, phonetic and acoustic information.

Like other acoustic parameters, F0 must be appropriately represented on our structure, as must linguistic information about the intonation pattern of which the F0 is an exponent. The interpretation of F0 parameters will depend on the structural position of F0 events and their integration with other attributes of that structure. The preliminary F0 modelling described here concentrates on systematic variation in the alignment of F0 phenomena within structural domains.

1.2. F0 alignment

A number of factors influencing F0 alignment have been identified in the literature (see [2, 3] for survey). For a given intonational pitch accent, such factors include (a) the internal composition of the associated accented syllable, together with the intrinsic properties of its onset and rhyme; and (b) characteristics of the foot and accent group over which the pitch accent is realized, such as the number and strength of component syllables, and its position in relation to adjacent feet or accent groups.

The relevant units of structure and their attributes are all represented on our prosodic hierarchy. Mapping knowledge of F0 events on to the structure is the first step in determining the temporal interpretation of F0 contours in synthesis.

2. THE PROSODIC HIERARCHY

Our prosodic hierarchy, building on [4, 5], is a relatively flat, head-driven, and strictly layered structure. Its richness comes from the information stored within structural nodes in the form of attributes and parameter values. The computational representation of this structure is made using the extensible mark-up language (XML). Word-level and syntactic-level information is hyper-linked into the prosodic hierarchy. In this way lexical boundaries and the grammatical functions of words can be used to inform phonetic interpretation without the need to introduce a separate level of “phonological” words.

2.1. Prosodic constituents

For current work, our largest constituent is the Intonational Phrase (IP), the domain for a complete, well-formed intonation contour. IP attributes include discourse factors used to determine the choice of intonation contour.

Each IP comprises one or more Accent Groups (AG), defined as the domain for a pitch accent configuration. The IP’s head is the rightmost AG, site of the intonation “nucleus”. AG attributes include a phonological specification for pitch accent.

AGs in turn contain one or more Feet (F), of which the leftmost is the head of the AG and the domain for the T* of a pitch accent. Feet with the attribute [strong] correspond to rhythmically strong beats.

Each Foot has as its head a strong initial syllable (S) and will include any weak syllables following. The “accented syllable” is thus the leftmost S of the leftmost F of an AG.

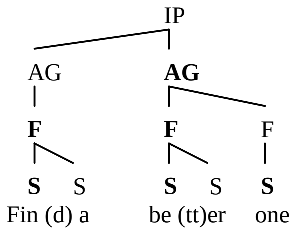

Figure1. Supra-syllabic tree structure for “Find a better one”.

Syllable constituents are Onset and Rhyme, which may in turn branch into Nucleus and Coda. The Rhyme is head of the Syllable, and the Nucleus is head of the Rhyme. Subject to certain constraints, consonants may be ambisyllabic, linked simultaneously to both a Coda and a following Onset.

3. INTONATION MODELLING

3.1. Objectives

We aim to use our structural framework to optimize the representation of a set of pitch patterns which can be used as a predictive model for high-quality synthesis. The preliminary research described here has more limited objectives: (i) for a specific nuclear pitch accent, produced by a single speaker over a range of structural contexts, to identify the important turning-points in the F0 contour, so as to define a template representation for that pitch pattern; (ii) to assess the significance of variability in the alignment of the turning-points in relation to our prosodic constituents.

3.2. Procedure

3.2.1. ProSynth Database. Our analysis is based on a core database of over 450 utterances, recorded by a single male speaker of southern British English. The database includes structures selected for various aspects of modelling within ProSynth [1]. For preliminary intonation modelling, the range of prosodic structures was restricted to a small subset of possible IPs, containing a maximum of 2 AGs. Nuclear AGs all contained a single foot of one or two syllables; accented syllables themselves covered a wide range of sub-syllabic structures. The intonation used for all nuclear AGs was essentially the same: a (low) falling tone (H*L), consistent with an unmarked, utterance-final discourse context.

Database speech files have been exhaustively labelled to identify segmental and prosodic constituent boundaries, using careful hand-correction of an automated procedure. F0 contours, calculated from a simultaneously recorded Laryngograph signal, can be displayed time-aligned with constituent boundaries.

3.2.2. Visual analysis. Phonological accounts of phrase-final H*L pitch accents predict that we should be able to model the contour using an F0 peak (H*) associated with the accented syllable, a valley (L) representing the end of the fall, and a low boundary tone (L%) associated with the right edge of the IP. However, in some contexts the precise location of e.g. H* may be obscured, or there may be competing positions [6, 7]. In constructing our own F0 template, we treated all observed turning-points as potentially important, rather than searching for a single H* candidate. Our visual analysis of all final AGs confirmed the difficulties reported in the literature. The peak F0 was often not so much a sharp peak as a level plateau. For these items, a single turning point to denote the peak was insufficient.

Within the final foot, three major F0 turning points were identified: Peak Onset (PON), Peak Offset (POF) (the beginning of the fall), and Level Onset (LON) (the point from which the low tone spreads till the end of voicing). The timing of these points seemed sensitive to type of Onset and Coda in the accented syllable and to the presence/absence of a post-nuclear syllable (tail) within the final foot. Four categories of Onset and Coda were potentially important for F0 timing: (i) empty, (ii) sonorant, (iii) voiced obstruent and (iv) voiceless.

Level peaks were observed more often in two-syllable (Tail) feet, especially in Tail feet with voiced obstruent Onsets, more variably with sonorant Onsets. POF was found around the middle of the vowel, and PON either at the beginning of the vowel or within a (non-nasal) sonorant Onset. Nasal Onset syllables tended to have a simple but flattened peak mid-vowel; Tail feet with voiceless or empty Onsets typically had simple peaks placed early in the vowel. In Notail feet the simple peak was more common, but level peaks were observed in a subset of cases of sonorant Onsets, where PON was located within the Onset and POF near the beginning of the vowel.

The Level Onset (LON) appeared sensitive to the proximity of the IP boundary and to Coda type: sonorant or other. In Tail feet, where the accented syllable had an obstruent Coda, LON occurred at the end of the vowel; with a sonorant Coda it coincided with the end of the Coda. In Notail Feet, where a sonorant Coda was available to carry the sustained low F0, LON typically occurred at the end of the vowel; obstruent Codas often pushed LON back to mid-vowel position. In a few cases, LON coincided with the IP boundary.

3.2.3. Informal auditory verification. To check the validity of using only PON, POF and LON to model the F0 contour for these phrases, an informal listening experiment was conducted using speech synthesized with the MBROLA system [8]. The durational parameters for synthesis were taken from the natural recording, while the F0 parameters were modelled on the natural F0 or simplified according to the three turning points. Perceptual comparisons between the natural and the simplified contours showed that they were essentially equivalent. However further simplification always gave a poorer result.

3.2.4. Automatic identification of F0 turning points. The statistical analysis of the F0 contour described below is based on a set of essential parameters derived by automatic means from the Laryngograph recording. Firstly, the location of the key syllable components was established using the manual annotations. Then the peak F0 value in the accented syllable was found. The onset (PON) and the offset (POF) of the peak were then found by finding the range of times around the peak where the F0 value was within 4% (approximating to a range for perceptual equality). The schematic representation below illustrates the search for PON and POF.

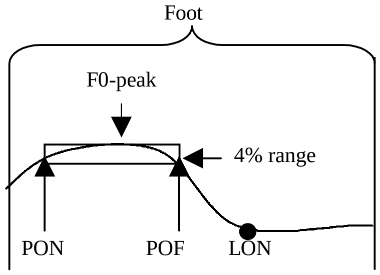

Figure 2. Schematic representation of PON, POF and LON.

To find the onset of the low level tone (LON) a fair estimate of the final F0 value was found from the mean value of the last 50ms of the contour. The onset was then taken to be the earliest point at which the F0 contour dipped 75% down from the peak value to the final tail value. Voiceless regions were ignored.

3.2.5. Statistical analysis. A set of analyses of variance was performed separately on the timing of PON, POF and LON. The timing for each of these was expressed in relation to the (accented) syllable constituent. To allow for differences in syllable duration across observations, the distance was measured between each point and syllable onset, and this distance itself expressed as a proportion of the syllable duration. The factors in the analysis were Onset type, Coda type (categories for both: empty, sonorant, voiced, voiceless) and Foot type (categories: Tail, Notail). A General Linear Model (GLM) was used for the analysis, combined with a post-hoc Tukey test to enable significantly different (p<0.01) pairs of means to be identified. To provide a rough estimate of the relative strength of the individual factors, a measure of the strength of association, η^2 was calculated (the sum of squares for the factor divided by the total sum of squares [9]).

4. SUMMARY OF STATISTICAL RESULTS

Of the 457 utterance-final AGs analysed, 193 had Tail feet, and 264 Notail. Within accented syllables, Onset types were: empty 13, sonorant 284, voiced obstruent 55, voiceless 105, Codas: empty 31, sonorant 161, voiced obstruent 106, voiceless 159.

4.1. Peak Onset (PON)

Overall, two factors were significant:

Onset type: [*F* (3,449) = 72.7, *p*<0.001] $\eta^2 = 0.178$

Foot type: [*F* (1,449) = 54.6, *p*<0.001] $\eta^2 = 0.045$

2-way interaction Onset*Foot type [*F* (3,449) = 9.0, *p*<0.001]

The model based on these significant components accounted for 63% of the variance in the timing of PON. Onset type was a stronger factor than Foot type. A separate analysis of Tail and Notail feet showed that the significant factor for both was Onset (*p*<0.001).

The Tukey test revealed that not all the categories of Onset were responsible for the overall significance. Disregarding empty Onsets (which are not strictly comparable since the Onset constituent is missing altogether), it emerged that in Tail feet, mean PON was significantly later when the Onset was voiceless, and in Notail feet, mean PON was significantly earlier for sonorant Onsets. The other Onset categories were not significantly different from each other.

Figure 3 shows PON as a function of Onset categories within Foot type. Notice that for all Onset categories apart from 'empty' there is a clear tendency for PON to be later in Tail than in Notail feet.

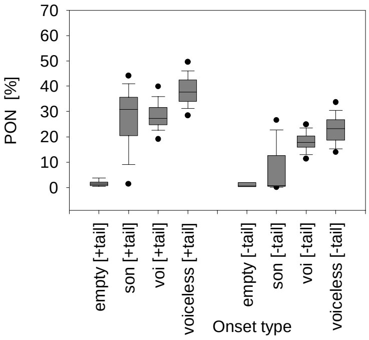

Figure 3. PON as a function of Onset type.

4.2. Peak Offset (POF)

Three factors were significant overall:

Onset type: [*F* (3,435) = 4.6, *p*=0.004] $\eta^2 = 0.010$

Coda type: [*F* (3,435) = 9.2, *p*<0.001] $\eta^2 = 0.017$

Foot type: [*F* (1,435) = 140.2, *p*<0.001] $\eta^2 = 0.100$

2-way interaction Onset*Foot: [*F* (3,435) = 8.8, *p*<0.001]

2-way interaction Coda*Foot: [*F* (3,435) = 4.8, *p*<0.001]

2-way interaction Onset*Coda: [*F* (3,435) = 2.4, *p*=0.014]

The model based on these significant components accounted for 67% of variance in the timing of POF. Foot type was clearly a stronger factor than both Onset and Coda type.

A separate analysis of Tail and Notail feet highlighted differences between the two groups as to the importance of Onset and Coda. In Tail feet both Onset and Coda were significant factors (*p*<0.001), though only empty Onsets were responsible for the significance of Onset type. Of the Codas, only voiceless types were significantly different from the other categories, having earlier POF.

In Notail feet, only Onset type was significant (*p*<0.001). The Tukey test showed that POF in syllables with voiceless Onsets was significantly later than in sonorant (and empty, disregarded) Onsets, but was not significantly different from voiced obstruent Onsets.

Figure 4 shows the POF as a function of Onset categories for each Foot type. Figure 5 shows POF for Coda categories. Notice a clear tendency for POF in both Onset and Coda categories to be later in Tail feet than in Notail feet.

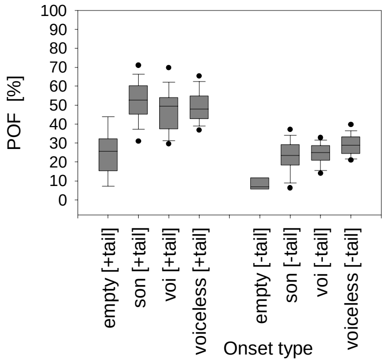

Figure 4. Peak Offset as a function of Onset type.

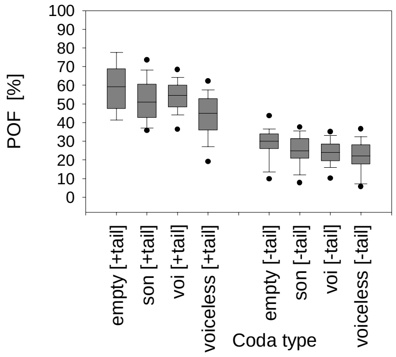

Figure 5. POF as a function of Coda type.

4.3. Duration from LON to foot boundary

A one-way ANOVA was conducted to find out whether Tail and Notail feet differed significantly with respect to the duration from LON to the end of the foot, expressed as a proportion of the whole foot. Our hypothesis was that there would be no difference: that in both Tail and Notail feet there must be a stretch of speech at the end of the foot where the tone stays low.

The hypothesis was confirmed – there was no significant difference between the foot types (*p*>0.05). In both cases the level stretch occupied about 50% of the foot duration.

4.4. Level Onset (LON)

In the visual analysis there appeared to be a relationship between the type of Coda and the timing of LON with respect to the beginning of Coda. In order to test this hypothesis, we made additional calculations of LON in relation to Coda position, to clarify when LON occurred in the vowel and when in the Coda. A difference between foot types with respect to LON was also tested. Otherwise the analysis was the same as for PON and POF.

Overall, two factors were significant:

Coda type: [*F* (3,451) = 29.3, *p*<0.001] $\eta^2 = 0.098$

Foot type: [*F* (3,451) = 277.3, *p*<0.001] $\eta^2 = 0.329$

The model accounted for 46% of variance in LON. Foot type had a much stronger effect on the timing of LON than Coda type.

In Tail feet analysed separately, only Coda type was significant ([*F* (3,188) = 10.9, *p*<0.001]), accounting for 14% of variance in LON. When Coda was empty, LON obviously occurred in the vowel. For other Coda categories, it occurred either in the Coda itself or in the Tail syllable (see positive values in Figure 6).

In Notail feet, both Onset and Coda were significant:

Onset type: [*F* (3,257) = 3.9, *p*<0.001] $\eta^2 = 0.031$
Coda type: [*F* (3,257) = 35.2, *p*<0.001] $\eta^2 = 0.276$

The model accounted for 31.2% of variance in LON. Coda type was clearly a much stronger factor.

For these feet, LON was typically found in the vowel, but considerably earlier when Codas were empty. LON was also significantly earlier before voiced obstruent than sonorant Codas. Figure 6 shows LON results separately for Tail and Notail feet, expressed in relation to the start of the Coda.

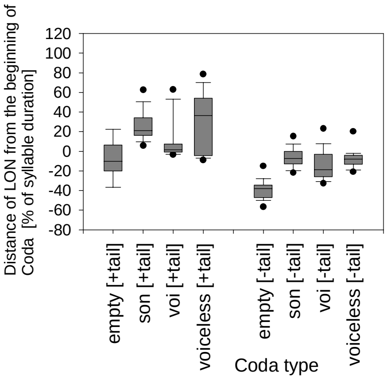

Figure 6. LON as a function of Coda type.

5. DISCUSSION

The systematic variability we have observed in our data is largely consistent with that reported in the literature (e.g. [2, 10]) for both American and British English. In our AGs with two-syllable feet, there was a consistent rightward shift of the contour compared with monosyllabic feet, so that alignment of the pitch fall itself began later in the vowel of the accented syllable. The effects of segmental structure in monosyllabic final AGs is sometimes harder to compare with earlier studies, where the peak values described may sometimes relate to our PON, sometimes to POF. However, where for example [10] reports a systematically earlier peak in sonorant-initial accent groups, this fits in with our findings for PON, whereas their finding that peaks occur later in sonorant-final groups accords with our findings for POF. Our initial informal experiments suggest that where a level peak is observed, both PON and POF are required for a perceptually equivalent modelling in synthesis. Further investigations will include the interaction between peak alignment and variations in the height of the peak, and the alignment of turning points in rising intonation patterns. We would predict significant effects from the same structural factors.

5.1. Phonological implications

Since both simple and level peaks were found as exponents of the same target intonation patterns, our findings suggest that these alternative phonetic interpretations reflect systematic structural differences. In level peaks, it is plausible that POF reflects a truer intonational target than PON, whose value may be masked by microprosodic effects of onset consonants and interpolation from a preceding AG.

Our finding that there was no significant difference in the position of LON relative to the IP boundary over Tail and Notail feet suggests that the pitch accent itself should more properly be associated with the foot than with the accented syllable. Further data are needed to test this hypothesis; firstly, we need to see what happens in three-syllable feet, and secondly, since the feet studied here were coterminous with the final AG, we need to look at AGs containing more than one foot, to see whether foot or AG is the more relevant domain. We would predict that while the AG is the phonological domain for the pitch accent, the leftmost foot within the AG is the domain for its phonetic interpretation. A potential effect from word boundaries within the foot/AG needs investigation.

The perceptual significance of the detailed mapping to structure we propose will be subjected to more stringent evaluation. It is notoriously easy to make subtle changes in the timing and/or frequency of F0 and for the resulting percept to be plausible. Informal experiments with MBROLA have suggested that the percept of finality associated with the natural speech tokens may be altered by shifting the alignment of our turning-points: the further left the position of POF, the more final the percept. More formal perceptual tests are in progress to test this.

5.2. Implications for synthesis

We now understand how to link the defining F0 events of an intonation pattern into our prosodic structure. Attributes (such as number of syllables) of the foot at the head of the AG determine whether alignment of the pattern as a whole is relatively early or late; attributes of the accented syllable, such as the type and duration of its constituents, are used to calculate the positions of the turning-points appropriately. Within our system, the phonetic interpretation of a defined pattern is made declaratively with reference to the structure over which it is realized.

ACKNOWLEDGMENTS

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

REFERENCES

[1] 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.

[2] 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.

[3] Ladd, D.R. 1996. *Intonational Phonology*. Cambridge, CUP

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

[5] 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.

[6] 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.

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

[8]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

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

[10] 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.