ANALYSIS BY SYNTHESIS OF ENGLISH INTONATION PATTERNS: GENERALISING FROM FORM TO FUNCTION.

Saandia Ali & Daniel Hirst

CNRS, Laboratoire Parole et Langage, Université de Provence
{saandia.ali, daniel.hirst}@lpl.univ-aix.fr

ABSTRACT

This paper presents a general model for the relation between representations of form and function for speech prosody on a multi-lingual basis. It outlines a procedure for analysing prosody by synthesis applied to English intonation patterns, generating formal representations from a minimal representation of prosodic functions and comparing the output with the observed data. This then allows the functional representation to be enriched and tested to see whether it provides a closer fit to the data.

Keywords: English, intonation, analysis by synthesis, prosodic form, prosodic function

1. INTRODUCTION

The central aim of the study of speech prosody is to understand how prosodic functions are related to prosodic forms in different languages. In this presentation we discuss a general model for the relation between representations of form and function for speech prosody on a multi-lingual basis. We outline a procedure for analysing prosody by synthesis, generating formal representations from a minimal representation of prosodic functions [8] and comparing the output with the observed data. This then allows us to enrich the functional representation and test whether the enriched representation provides a closer fit to the data. The procedure is specifically applied here to the intonation patterns of British English, although there is nothing in the model which is specific either to intonation rather than, say rhythm and tempo, or to British English rather than any other language or dialect.

The example presented is a fairly rudimentary mapping from known prosodic functions to prosodic forms within an explicit representation system briefly described in the next section. Both the functional representation and the mapping rules are supplied by hand but the output is in a format which allows comparison with the output of an automatic (or semi-automatic) analysis of prosodic forms. This makes it possible to provide an explicit evaluation metric, illustrated in the final section.

2. LEVELS OF REPRESENTATION

Following [10] and [8], we assume three intermediate levels of representation between the level of the acoustic signal and the level of prosodic function. These are, from the most concrete to the most abstract:

- A phonetic representation
- A surface phonological representation
- An underlying phonological representation

The phonetic representation is provided in our system by the output of the Momel algorithm [10], which provides an automatic discrete representation of a raw fundamental frequency contour as a sequence of target points. A recent implementation of this algorithm [9] was used which requires fewer manual corrections than the original version. The same implementation codes the target points as a sequence of tonal symbols using the INTSINT alphabet.

This symbolic representation provides a convenient format for the formulation of explicit rules for converting functional representations into prosodic forms, which we detail in the next section. Given a sequence of tonal symbols, each provided with a specified temporal alignment, this can be converted back into a sequence of target points, which can in turn be used to produce a smooth, continuous fundamental frequency pattern, using two speaker specific parameters: *key* (in Hz) and *range* (in octaves).

For the underlying phonological representation, [8] argues that the IF (intonation function) annotation proposed there is essentially equivalent to a two-layered phonological structure containing Intonation Units (IUs), each composed of a sequence of Tonal Units (TUs) together with additional features [±terminal] and [±emphatic]. This was the basis for the way in which the different models described below were implemented.

3. MODELS FOR SYNTHESIS

The approach is an incremental one of successive approximation, each time comparing the output with the observed intonation patterns. The aim is ultimately to derive the annotation automatically from the data, but at this stage we provide manual functional annotation of the data in order to bootstrap the annotation procedure and to test the possibility of objective evaluation, which would to some extent avoid the extremely time-consuming process of subjective evaluation (cf [11]).

The material used for the analysis consists of the continuous passages from the Eurom1 corpus [3]. We analysed the 15 passages (each of 5 sentences) read by speaker **fa**. The passages were labelled by hand for the functional representation as described below. Rules were then implemented by means of a script running in the Praat environment [2], converting the IF symbols to a sequence of INTSINT symbols aligned with the signal. These were then converted to Momel targets, interpolated quadratically to provide a pitch-tier for Psola resynthesis, allowing an informal subjective evaluation of the output.

In the rest of this section we present the successive models implemented. In the next section we describe their objective evaluation.

3.1. Model: *none*

The null hypothesis is that intonation is derived from no functional information at all. In order to provide an F0 curve which can be compared with the original we implemented an initial rise from M to H followed by a gradual fall throughout the whole passage to a final B. Figure 1 shows the "model" curve superimposed on the curve derived from hand-corrected Momel pitch targets for one passage of our corpus.

3.2. Model: *IU*

It is obvious that the intonation pattern of the passage illustrated in figure 1 does not consist of one continuous pattern but of a sequence of distinct patterns, which, as a first approximation, we can represent as falls. To account for this we annotated the text with boundary symbols ([ ]) dividing it into Intonation Units (IU). Each IU was then modelled as a sequence [M-T…B] where M and B were aligned with the left and right boundaries of the sequence and T was aligned at a fixed offset (200ms) from the beginning of the unit.

The functional representation was actually implemented as an interval tier in a *Praat* TextGrid, where each interval corresponded to an Intonation Unit, generating the patterns closest to the output of the hand-corrected targets.

Figure 1: F0 curve derived from the INTSINT annotation [MH … B] superimposed on the quadratic spline curve from the hand corrected *Momel* modelling of passage fao30072 from the Eurom1-EN corpus.

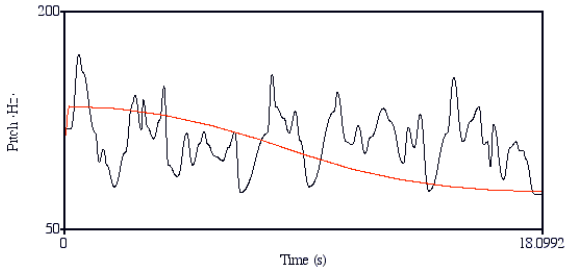
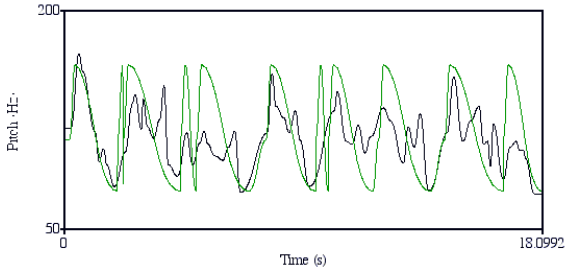


Figure 2: The same f0 pattern as figure 1 with the output of model *IU*.



3.3. Model: *terminal*

The intonation units in model *IU* were all implemented with a falling pattern although many of them exhibit a global fall with a final rise. Functionally, this can be described as a distinction between finality and non-finality (cf Tune 1 and Tune 2 [1] or [±terminal] [6]). This was captured by labelling the IUs with the symbol [+] or [|]. The corresponding INTSINT labels generated were either [M-H…B-B] or [M-H…B-H].

3.4. Model: *accent*

In this model we introduce the functional category *accent*, associated with local pitch movements and annotated ['']. While words in English are lexically marked for stress, not all stresses are accented and sometimes (although rarely and not at all in our corpus) accents can occur on unstressed syllables. The accents were represented on a second interval tier as Tonal Units (TU) following a tradition which dates back to [13].

The INTSINT labels are assigned as follows. M is aligned with the beginning of each IU and either H or B with the end. The label H is aligned with the beginning of the first TU of each IU which is labelled ['] and the label D with subsequent TUs. A tone B is then aligned after the last H or D of the last TU labelled ['].

3.4.1. nucleus

Following [6], we next introduce the label *nucleus,* labelled [*].

Our Intonation Unit now corresponds to the familiar structure:

(prehead) (head (body) nucleus (tail)

as described in the literature on intonation in the British tradition ([15], [4]), where (X) indicates an optional element.

The tones associated with this structure are then:

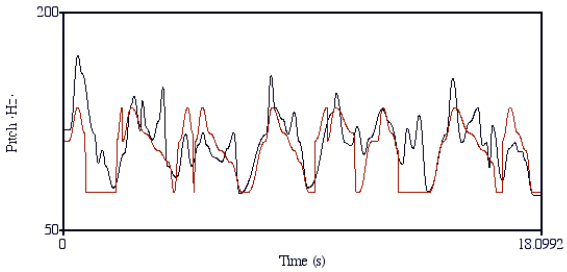
	(prehead)	(head (body…))		nucleus	(tail)
function	[	'	'$_1$	*	'$_1$ |/+
form	M	T	D	D/H	-B B/H

or more compactly:

[M ('T ('D_1)) *H/D '$_1$-B B|/H+

where the choice between H/D depends on whether there is a preceding head and that between B/H on the terminal or non-terminal label of the IU The -B is aligned with the last tonal unit of the intonation unit.

Figure 3: The same f0 pattern as figure 1 with the output of model *accent*.



3.5. Model: *emphasis*

3.5.1. emphatic nucleus

Introducing a label for emphasis [!*] allows a distinction between emphatic and non-emphatic nucleus.

This defines the four tunes described by [1] and is basically the IF notation of [6] and [11].

The basic characteristic of an emphatic nucleus is that it is realised as starting on a higher pitch than preceding accents. This is implemented as a T tone aligned with the TU labelled [!*].

3.5.2. emphatic head

The emphatic terminal or non-terminal nucleus can be pre-signalled in the head [7]. For this we allow emphasis to be specified either on the TU (for the nucleus) or on the IU for an emphatic head, produced as a sequence of falling tones (T+L, H+L H+L…) in a non-terminal IU or as an *upstepping* sequence (B U U…) in a terminal IU.

3.5.3. improper bracketing

Some intonation phenomena can be rather neatly accounted for by allowing improper bracketing of Intonation Units [7]. This means that an Intonation Unit could have an initial boundary ([) but no final boundary (cf the "interrupted glide down" of [14]) or a final boundary but no initial boundary (cf "parenthetical" intonation patterns, "afterthoughts" or "divided fall-rise" patterns).

Our last implementation has the following functional categories:

Intonation units : { [, +, |, +, [|, [+, [!, [!+, [!|}

Tonal unit: { 0, ', *, !* }

This allows us to define the following patterns with associated tonal marks using the same notational conventions as above:

(i) no nucleus, initial:
[M ('T ('D_1))

(ii) neutral, terminal:
[M ('T ('D_1)) *H/D-B B|

(iii) neutral, non-terminal:
[M ('T ('D_1)) *H/D-B H+

(iv) emphatic, terminal
[M ('T ('D_1)) !*T -B B|

(v) emphatic, non-terminal
[M ('T ('D_1)) !*T-B H+]

(v) marked emphatic, terminal
[!T ('B ('U_1)) !*T-B B|

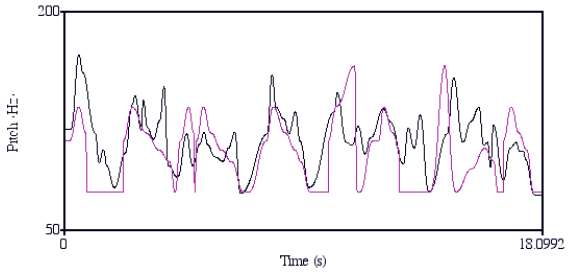
(vi) marked emphatic, non-terminal
[!M ('TL ('HL_1)) !*T-B H+

(vii) no nucleus, terminal
('B ('B)) B|

(viii) no nucleus, non-terminal
('B ('B)) H+

Figure 4 illustrates the output of the model emphasis applied to the same passage ao30072.

Figure 4: The same f0 pattern as figure 1 with the output of the model *emphasis*.



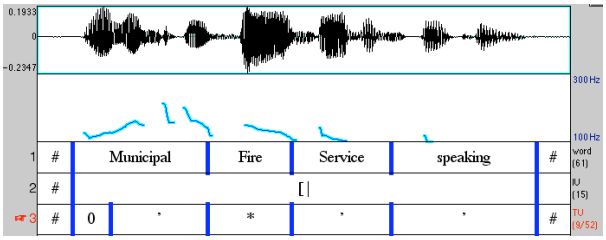
The following is the final functional annotation of one of the passages.

Figure 5: Passage fao30072 with the final functional annotation used in the analysis.

[Mu'nicipal *Fire 'Service 'speaking| # [We're 'trying to lo'cate an e'mergency *caller+ [who 'rang *off+ # [wi'thout 'giving any 'personal *details| # [He ap'peared to 'be on the 'local *network| # [He **c**on'nected on our !*line 'number+ # ['seven six *two+ 'five 'eight *four| # [!We'd ap'preciate im'mediate at'tempts to *trace him| [be'cause he 'sounded *desperate|

The first Intonation Unit from the equivalent in the form of a TextGrid is shown in Figure 6.

Figure 6: A sample extract from the TextGrid of passage fao30072 showing the IU and TU tiers

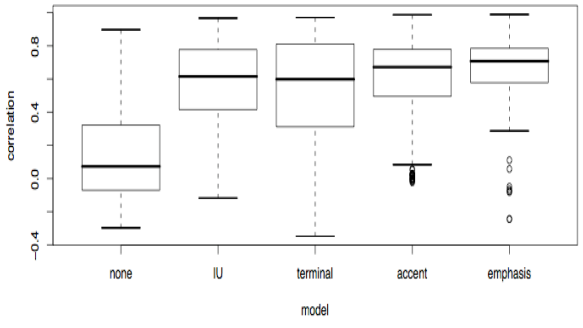


4. EVALUATION

The five models described in the previous section were implemented by means of a Praat script and applied to the fifteen passages of the Eurom1-EN corpus read by speaker **fa**. The raw fundamental frequency of the recordings was modelled with the Momel algorithm and the speaker dependent parameters *key* and *range* were estimated using the INTSINT algorithm. Since there was relatively little difference between the values of key and range for the 15 different passages, the same values (key = 111 Hz, range = 1.1 octaves) were subsequently used for all the passages.

The linear correlation coefficient was then calculated for each Intonation Unit between the f0 curve generated by the output of the model and that generated from the (corrected) Momel targets. The boxplot of the correlation coefficients for each model is given in figure 7.

Figure 7: Boxplot of correlation coefficients for each Intonation Unit between the output of the model and the output from the hand-corrected Momel targets



5. CONCLUSION

Even the best mean correlation obtained in this experiment was not particularly high (r = 0.66) (although several individual values were much higher, often over 0.9) but it should be said that there was no particular attempt at this stage to optimise the parameters of the model. The aim here is rather to demonstrate the analysis by synthesis method, which we feel suggests a great number of possible improvements to the implementation.

The next step will be to use this preliminary system of analysis by synthesis to progress towards fully automatic functional annotation, providing a representation optimised in terms of formal output.

6. REFERENCES

[1] Armstrong, L.E. & Ward I.C. 1926. *A Handbook of English Intonation*. Second edition. Heffer. Cambridge.

[2] Boersma, P. & Weenink, D. 2006. Praat. Doing phonetics by computer. [computer program]. Version 4.5.06 http://www.praat.org/

[3] Chan, D., Fourcin, A., Gibbon, D., Granstrom, B., Huckvale, M., Kokkinakis, G., Kvale, K., Lamel, L., Lindberg, B., Moreno, A., Mouropoulos, Senia, F., Trancoso, I., Veld. C. & Zeiliger, J. 1995. EUROM- A Spoken Language Resource for the EU, in *Proceedings of Eurospeech'95.* (Madrid, Spain, September, 1995). (1), 867-870

[4] Cruttenden, Alan 1986. *Intonation*. Cambridge University Press.

[5] Hirst, D.J. & Di Cristo, A. (eds) 1998. *Intonation Systems. A survey of Twenty Languages.* (Cambridge, Cambridge University Press).

[6] Hirst, D.J. 1977. *Intonative Features. A Syntactic Approach to English Intonation*. (Mouton; La Haye).

[7] Hirst, D.J. 1998. Intonation in British English. in Hirst & Di Cristo (eds) 1998., 56-77.

[8] Hirst, D.J. 2005. Form and function in the representation of speech prosody. *Speech Communication,* 46 (3-4), 334-347.

[9] Hirst, D.J. 2007. A Praat plugin for Momel and INTSINT with improved algorithms for modelling and coding intonation. *ICPhS XVI,* Saabrücken, August 2007.

[10] Hirst, D.J., Di Cristo, A. & Espesser, R. 2000. Levels of representation and levels of analysis for intonation. in M. Horne (ed) *Prosody : Theory and Experiment*. Kluwer Academic Publishers, Dordrecht. 51-87.

[11] Hirst, D.J., Aubergé, V. & Rilliard, A. 1998. Comparison of a subjective and an objective evaluation metric for prosody in text-to-speech synthesis. *Proceedings ESCA/COCOSDA workshop on Speech Synthesis*, Jenolan Caves, Australia. November 1998, 1-4.

[12] Hirst, D.J. & Auran, C. 2005. Analysis by synthesis of speech prosody: the ProZed environment. *Proceedings Interspeech,* September 2005, Lisbon. 3225-3228.

[13] Jassem, W. 1951. *Intonation of Conversational English. (Educated southern English).* Travaux de la société des sciences et des letters de Wroclaw seria A. Nr 45

[14] Kingdon, R. 1958. *The Groundwork of English Intonation*. London, Longman

[15] O'Connor, J.D. & Arnold, G.F. 1961. *Intonation of Colloquial English*. London: Longman (2nd edition, 1973).

[16] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. & Hirschberg, J. 1992. ToBI : a Standard for Labelling English Prosody. *Proceedings ICSLP92*, 2, 867- 870, Banff, Canada.