INTERSPEECH 2009 BRIGHTON

Developing an Automatic Functional Annotation System for British English Intonation

Saandia Ali and Daniel Hirst

CNRS Laboratoire parole et langage. Aix-Marseille I University. France

saandiaali@yahoo.fr, daniel.hirst@lpl-aix.fr

Abstract

One of the fundamental aims of prosodic analysis is to provide a reliable means of extracting functional information (what prosody contributes to meaning) directly from prosodic form (i.e. what prosody is – in this case intonation). This paper addresses the development of an automatic functional annotation system for British English. It is based on the study of a large corpus of British English and a procedure of analysis by synthesis, enabling to test and enrich different models of English intonation on the one hand and work towards an automatic version of the annotation process on the other.

Index terms: phonetic and phonology**,** speech analysis and representation, prosody modeling and generation

10.21437/Interspeech.2009-628

1. Introduction

The question of how to annotate prosodic phenomena is a vital one when one wants to investigate or analyse intonation. Prosodic annotation systems were developed and widely used to research intonation in English (see ToBI [1] and the Tonetic Stress Marks (TSMs) [2]). However, recent evaluations of these systems [3], show that they fail to answer the growing needs to cover bigger corpora at lower costs in terms of time and money.

Hirst [4] has argued that one of the main obstacles to the development of these systems has been the lack of separation between the annotation of prosodic functions and prosodic forms. While prosodic forms can be accurately modelled by algorithms today, prosodic functions remain within the field of human interpretation. “Machines are good at labelling forms, humans at interpreting the message” [4]. Hence, our drive to develop an automatic functional annotation system for British English.

Such an attempt was successfully achieved recently, in Finnish (see [5]). It was suggested that the manual functional annotation of a small corpus could be used as a bootstrap for such an attempt, and as a starting point for high quality multilingual speech synthesis.

In an earlier study [6], we also developed a procedure of analysis by synthesis generating formal representation from a minimal functional representation. The representation of prosodic forms was optimized starting from these functional labels using the INTSINT coding system. The sequences of INTSINT tones were then converted into phonetic representations by means of the Momel algorithm so that the output could be compared directly to the original recordings. The quality of fit of the model was measured by linear correlation with hand corrected modelled fundamental frequency curves, intonation unit by intonation unit. The parameters of this model were optimized on an extract of the Eurom1 corpus for which the functional annotation was carried out manually using the IF annotation system in a subsequent study [7].

This paper presents the developments of this approach based on a large corpus of expressive speech which already contains a prosodic transcription in TSMs.

The functional annotation was provided automatically through a conversion of the TSMs into the Intonation Function (IF) annotation system on the Marsec corpus which will be presented in the next section. This will be followed by the presentation of the automatic procedure of optimization of the representation of prosodic forms, starting from the functional annotation. Finally, our first attempts at making the functional annotation process automatic, will be discussed.

2. Corpus and annotation

The corpus is used in this research is extracted from the Aix-Marsec corpus [8] which contains five and a half hours of continuous speech. The extract used for our experiments consists of 47 minutes of continuous speech, which can be qualified as “authentic” speech following [8]. There are five different speakers (two female) who read different short stories meant for adults or children. The speakers were professional actors providing us a corpus of expressive speech which seemed a good starting point to enrich a functional annotation system.

The corpus contains an orthographic and phonemic transcription, the annotation of lexical stress, and a prosodic transcription (TSMs) which was made by Gerry Knowles and Briony Williams:

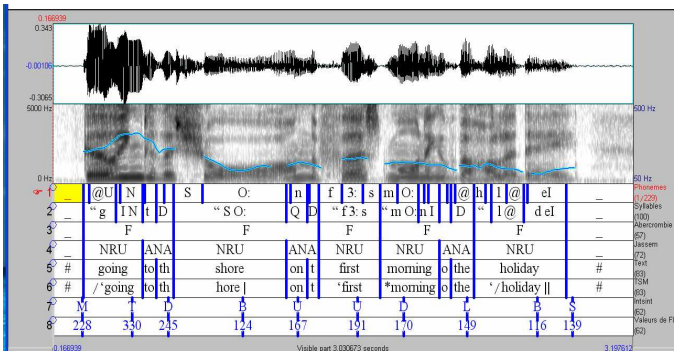

Figure 1: *Different levels of annotation of the Aix-Marsec database, illustrated by a Praat TextGrid.*

Starting from the top, the first tier shows the phonemes, the second tier the syllables, the third the stress feet, the fourth the narrow rhythmic units and anacrusis, the fifth the words, the sixth the TSMs, the seventh the INTSINT tones and the eighth the MOMEL targets.

Intonation was annotated automatically using the Momel and INTSINT algorithms. The annotation of prosodic forms consists of two levels of representation: a phonetic representation with Momel pitch targets and a surface phonological representation in INTSINT tones which can be compared to a phonetic alphabet for intonation transcription [9]. The annotation of prosodic function was provided automatically through a conversion of the TSMs into IF annotation.

2.1. Conversion of TSMs into IF: towards an enriched version of IF

The TSMs represent both prosodic forms (F0 configurations) and functions (prominences and boundaries) so that our goal was to extract functional information from the TSMs and design rules to convert this information into the IF system.

The IF system [10, 4] distinguishes two types of boundary at the level of the intonation unit (IU): terminal boundaries marked "[|" and non-terminal boundaries marked "[+". It also represents four levels of prominence at the level of the tonal unit (TU) [11]: unaccented, accented ['], nuclear [°] and emphatic [!°]. The emphatic label can also be assigned to the level of the intonation unit in order to describe emphatic heads, following the British tradition of intonation analysis.

The TSMs distinguish two types of boundary (major || and minor |) and four levels of prominence, although the definition of these is slightly different, which leads to some adaptations to the IF system. As a first approximation, the minor intonation unit boundary was equated to the non terminal boundary and the major boundary to the terminal boundary.

For the levels of prominence, three problems were addressed: the levels of prominence taken into account preceding the nucleus in IF, the identification of the nucleus from the TSMs and the emphatic function. A particularity of the British approach, encoded in the TSMs, is the distinction between stressed and accented syllables. Stressed syllables represent a rhythmic prominence whereas accented syllables which are made prominent also by pitch. Adapting IF to this functional approach, it was decided to add a level of prominence to IF: that of rhythmic prominence.

The IF coding was implemented using alphabetic symbols so that the following types of TUs were distinguished: unaccented (U), stressed (S), accented (A), nuclear (N), and emphatic nuclei (!N).

The final pitch accent in each IU was then automatically identified as the nucleus.

In the British approach, a distinction is made between "low" and "high" contours. In the case of "high falls", the starting point of the fall is said to be higher than the pitch target which associated with the previous syllable. This phonetic difference is interpreted in this conversion task, as a functional difference between emphatic and non-emphatic pitch contours. Pitch contours preceded by the symbol ">" in the TSMs, (signalling a considerable widening of pitch register at that point), were also annotated as emphatic.

Finally, the annotation of boundary types and thus sentence type was developed on the basis of the analysis of the content of the corpus. The types of utterances labeled were emphatic utterances ([!| or [!+), questions ([?| or [?+), parentheticals (| or +), and interrupted utterances ([ ).

A Praat script was used to convert the TSMs automatically into the modified IF labels using two tiers: one for the annotation of boundaries and sentence-type at the level of intonation units and one for the annotation of levels of prominence at the level of the tone units. The result of this conversion was then double checked and corrected manually when necessary.

TSM	IF
* (stressed but unaccented)	S
Low fall (‘), low rise(,), low fall-rise(\’), low level(_), high level (~)	A or N (if the last pitch accent)
High fall(\), high rise (/), high fall-rise (‘/), >+ pitch accent	A or !N (if the last pitch accent)
Minor boundary (|)	+
Major boundary (||)	|

Table1. *IF equivalents for TSMs*

An example of the conversion of TSMs into IF annotation for an utterance is shown in table 2 below:

Word	#	And I	Needed to	find some	books	#	for my	course at	College	#
TSM	#	~an d I	_needed to	find some	`books |	#	for my	`course at	`college ||	#
IU	#	[+				#	[|			#
TU	#	A	A	S	N	#	U	A	N	#

Table 2. *TSMs and IF annotation for two sections of speech.*

The utterance consists of two intonation units: the first ends with a minor boundary in the TSMs which corresponds to a non-terminal boundary in IF; the second ends with a major boundary which corresponds to a terminal boundary in IF. The first IU contains four TUs, the first of which is labeled "A" (for accented) since the TSMs signal the presence of a high pitch accent; the second is also labeled "A" since it was marked as a low level pitch accent. The third is unmarked in the TSMs and labelled "S" in IF. Finally, the last pitch accent which is a low falling tone is labeled as a nucleus (N) in IF. The second IU contains an unstressed TU (U) followed by two accented TUs the last of which is labeled "N".

2.2. Final modifications of IF

At the level of the tonal units, the functional labels (U, S, A, N) were characterized more precisely in keeping with their position within the Intonation Unit (Initial, medial, final, postnuclear) and the type of Intonation Unit they belong to. (A terminal assertion, a terminal question, a non terminal parenthetical). A nuclear tonal unit (N) in a non terminal IU can be annotated as follows:

- [+(!) N1 (the first TU of a non terminal IU which can be emphatic)
- [+ (!)NM (the second or third TU)
- [+ (!)NF (the last TU of the IU)

In the rest of this paper we show how this functional annotation was used as a basis for the modeling of the mapping between the representations of prosodic forms and functions on the Marsec corpus.

3. Modeling the representation of prosodic form starting from prosodic function

Starting from the functional annotation of the corpus, the representation of the local variations of F0 and their timing, were optimized using the INTSINT coding system. (The register was calculated automatically with the INTSINT algorithm at this stage).

3.1. Optimizing the coding of INTSINT tones

Each tonal unit within a specific IU was modelled using two or three INTSINT tones depending on the position of the TU (two tones for initial and medial TUs, three for final ones).

By means of a Praat script, and for each tonal unit, all of the combinations of two or three INTSINT tones were tested except that the first tone of the IU was necessarily an absolute tone, either T, M or B, since a relative tone presupposes that there is a preceding target. For the other tones there were 8 possible tones at each point and we also included the possibility of no tone at all (_) at each point except the last. This gave a total of 648 (=9*9*8) possible sequences of tones for a final TU. The tonal targets were aligned at a fixed offset from the left and right boundaries of each TU, for this first step of optimization (see table 3).

[+N1	[+!N1	[+NM	[+!NM	[+NF	[+!NF	[+N1F	[+!N1F
HL	HL	_D	TL	SLU	HBU	HLS	TBU
UD	UL	_L	UD	SBU	UBU	HLD	HLD
HD	UD	DL	HL	DBU	TBH	TLD	HLS
_L	HD	DD	UL	SDS	HBH	ULS	TLS
DL	TL	SL	TD	SLS	ULU	ULD	ULS

Table 3. *The five best sequences of tones for each type of nuclear TU in a non terminal IU.*

The sequences of INTSINT tones were then automatically converted into Momel targets so that the output of the model could be evaluated and compared to the original recordings of the Marsec corpus in two ways. First an objective evaluation was carried out using linear correlation and RMSE for the original hand-corrected Momel curve and the curve generated by the model. Next, a subjective evaluation was carried out by comparing the original recordings to the resynthesis of the modelled Momel curve. As illustrated in table 3, the best sequences of INTSINT tones were then saved and used as a basis for the second step of optimization of the representation of prosodic form (i.e. the optimization of the alignment of the tonal targets). The ten best sequences were used for this step.

3.2. Optimizing the alignment of INTSINT tones

The second step of modeling of prosodic form consists in optimizing the alignment of the best sequences of tones selected in the first experiment.

In this approach, we use the procedure of analysis by synthesis as a testing ground for different models of tonal alignment so that different anchor points can be tested and the alignment of the tonal targets relative to these anchor points is optimized for our corpus. Two modes of alignment were also tested: in the first, tonal alignment was optimized using absolute distances from the anchor points, in the second one timing was allowed to vary as a function of the duration of the segmental anchors.

The anchor points which were taken into account in our experiment were the left and right boundaries of the tonal unit, the onset of the stressed syllable, the onset of the stressed vowel, the last syllable of final tonal units and the preceding target. In order to illustrate the procedure of optimization, we focus here on one model taking into account the first and last syllables of TUs as well as their left and right boundaries and finally using both modes of alignment.

The parameters of optimization are specific to the position of the TU within the IU as is detailed below:

- TU1 (the first TU of an IU) and TUM (medial TUs): T1's alignment varies from 0 to 150% of the first syllable of the TU in 7 iterations. T2 is aligned from 160 to 10 ms before the right boundary of the TU in 4 iterations.
- TUF (last TU of an IU): T1= from 0 to 150% of the first syllable of the TU in 7 iterations. T2= from minus 40% to +80% of the last syllable of the TU in 7 iterations. T3= 10 ms before the right boundary of the TU.

The alignment of the ten best sequences of tones identified from the first step of optimization was optimized following these parameters. The best alignments and sequences of tones were then selected using linear correlation and RMSE (see figure 2). The alignment of t1 in all TUs (initial, medial and final) was consistent with the boundaries of the first syllable since 90% of the best correlations were obtained with an alignment varying from 50% to 100% of the first syllable. The alignment of t2 (in final TUs) gave the best results when varying from 0 to 70% of the last syllable which stood out as a consistent anchor point as well.

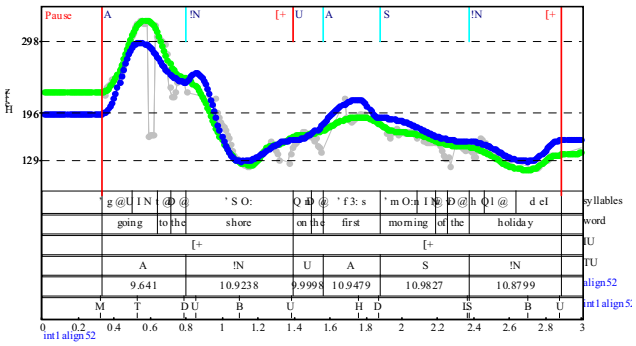

Figure2. *Optimized representation of prosodic form for an utterance (tier3= IU, tier 4= TUs, tier 5= combination of correlation and RMSE, tier 6= best sequences of tones and alignment)*

Using these optimized representations of prosodic forms (sequences of tones and their alignment), our next attempt is to extract the functional information directly from prosodic form.

3.3. Modeling the representation of prosodic function: from form to function

Three types of functional information extraction can be carried out on the basis of our experiments: first the recognition of sentence-type, next the detection of levels of prominence and finally the detection of intonation unit boundaries. Only the two first types of automatic detections will be described here.

3.3.1. Recognition of sentence-type and types of prominence

The functional elements which were taken into account for this experiment were the boundaries of intonation and tonal units. The joint recognition of sentence-type and prominence level was carried out by means of the following automatic procedure implemented as a Perl script.

Knowing the boundaries of a given intonation unit as well as the boundaries of its tonal units, several hypotheses can be made: If the IU which is taken into account is assumed to be a terminal one ([|), then all the best forms associated with the TUs: [|S1, [|A1, [|NM, etc., will be tested at the level of

each tonal unit, until the best global score at the level of the intonation unit is reached. The same procedure is then repeated for each hypothesized sentence-type ([+, [?+, etc.). Finally the type of intonation unit and sequence of TUs which result in the best score is selected.

3.3.2. Results of the recognition of sentence-type

For both experiments, the functional labels of the corpus are compared to the automatic recognition of functional information using a confusion matrix.

	Detected IU									
IU	**[**	**[!|**	**[!+**	**[?|**	**[?+**	**[|**	**[+**	**|**	**+**	**Total**
[	**42**				4			7	16	69
[!|		**167**	43	41		110	32		2	395
[!+		24	**203**	17	8	36	108	1		397
[?|		13	4	**10**		7	3			37
[?+	2		15	3	**49**	6				75
[|	18	514	264	274	35	**1034**	432		20	2591
[+	36	178	263	162	104	334	**566**	1	14	1658
|	46							**34**	33	113
+	17			2				11	**73**	103

Table 4. *Confusion matrix of predicted types of Intonation Units (columns) and the original ones (rows).*

Focussing on the recognition of emphatic and unemphatic IUs ([!| vs [|, [!+ vs [+), the following comments can be made: 167 cases of [!| out of 395 were recognized (= 42%). 110 of the other IUs are identified as [| (terminal but unemphatic) so that 70% of terminal IUs are identified, regardless of whether they are emphatic or not.

203 cases of [!+ IUs out of 397 were automatically recognized (= 51.1 %.) Adding the number of IUs identified as simply non terminal ([+), the results rise to 70%.

This suggests that the representation of the emphatic function at the level of the intonation unit needs to be further enriched or that more emphatic data needs to be modelled in order to improve the automatic identification of emphatic utterances.

3.3.3. Results of the detection of prominences

	Detect ed TUs					
TUs	**!N**	**A**	**N**	**S**	**U**	**Total**
!N	**251**	41	258	100		650
A	39	**493**	86	401		1019
N	168	10	**320**	143		641
S	62	224	194	**629**		1109
U					**895**	895
	520	768	858	1273	895	4314

Table5. *Confusion matrix of the predicted prominences and the ones which were extracted from the TSMs.*

Again, the representation of the emphatic function would need to be further enriched at the level of the nucleus. Indeed 39% of the emphatic nuclei and 50% of the unmarked nuclei are automatically recognized. Taking into account all types of nuclei regardless of the emphatic label, the result rises to 997/1291 (= 77%).

Moreover, 57% of the stressed TUs (S) are correctly recognized as opposed to 48, 4% of the accented TUs (A). This leads to questioning the number of levels of prominence which need to be taken into account in order to model and predict prosodic form.

4. Conclusion

In this paper, we present the developments of a functional annotation system for British English intonation. This was based on the automatic conversion of the TSMs into IF annotation on a large corpus of British English. The mapping between the representation of prosodic forms and prosodic functions was then optimized on this corpus, via a procedure of analysis by synthesis. The results of the first steps towards an automatic labelling of prosodic functions were then presented. It was shown that such a procedure enables to test different theories of intonation and can help reconsidering fundamental components of an intonation model, such as the representation of emphasis or the levels of prominence.

5. References

[1] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J. ToBI: a Standard for Labeling English Prosody. *Proceedings ICSLP 92*, 2, 867- 870, Banff, Canada, 1992.

[2] Knowles, G. Annotating large speech corpora: building on the experience of MARSEC. *Journal of Linguistics,* 13, 87-98, 1994.

[3] Wightman, C. ToBI or not ToBI? In *Proceedings of the First International Conference on Speech Prosody*, Aix en Provence, April 2002.

[4] Hirst, D.J. Form and function in the representation of speech prosody. In K.Hirose, D.J.Hirst & Y.Sagisaka (eds) *Quantitative prosody modeling for natural speech description and generation (=Speech Communication* 46 (3-4)), 334-347, 2005.

[5] Vainio, M.; Hirst, D.J., Suni, A. & De Looze, C. Using functional prosodic annotation for high quality multilingual, multidialectal and multistyle speech synthesis. 1*3th International Conference on Speech and Computer SPECOM 2009*, Saint-Petersburg, Russia, June 2009.

[6] Ali, S. & Hirst, D.J. Analysis by synthesis of English intonation patterns: Generalising from form to function. in *Proceedings of ICPHS 2007*, Saarbrücken, 2007.

[7] Hirst, D.J. & Ali, S. Optimizing the automatic functional annotation of English intonation. In *Proceedings of 4th International Conference on Speech Prosody.* Campinas, Brazil, pp. 127-130, 2008.

[8] Auran, C., Bouzon, C. & Hirst, D.J. The Aix-MARSEC project: an Evolutive database of spoken English. In *Proceedings of the Second International Conference on Speech Prosody*, Nara, Japan, 561-564, 2004.

[9] Hirst, D.J. Intonation in British English. In D.J Hirst and A. DiCristo : *Intonation Systems : a survey of twenty languages,* Cambridge University Press, Cambridge,1998.

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

[11] Jassem, W. *Intonation in Conversational English*. Warsaw, Polish Academy of Science, 1952.