Analysis by synthesis of tonal alignment patterns in British English

Saandia Ali

[1] CNRS Laboratoire parole et langage. Aix-Marseille I University. In France

saandiaali@yahoo.fr

Abstract

This paper presents an attempt to investigate tonal alignment patterns in British English via a procedure of analysis by synthesis. This study encompasses both a top down and a bottom up approach, enabling to test different models of tonal alignments (anchor points and modes of alignment) previously described in the literature, and to optimize and evaluate the representation of those models. The whole procedure is reversible so that proposed models can be synthesized and their output compared to the original recordings of the corpus.

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

1. Introduction

The question of how tonal targets are aligned with the segmental string is a crucial one in research on intonation. Tonal alignment has been extensively investigated since the eighties, but to our knowledge there has been no discovery of a universal model of tonal alignment in English or in any other language. The majority of experimental research on the subject have rather enabled us to have a better understanding of the factors influencing tonal alignment phenomena and their nature (articulatory constraints, perceptual ones, segmental context: see Xu 2002 [1], D'Imperio 2002 [2]). As for English, the authors highlight different anchor points or landmarks to account for the variations in timing of tonal targets such as the syllable, the rime, the foot (cf: House 1996 [3], Pierrehumbert 1989 [4]).

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 on the MARSEC corpus. Prosodic forms and prosodic functions are annotated separately following Hirst (2005) [5]. Intonation patterns are annotated automatically using the MOMEL-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. The output of the MOMEL algorithm provides an automatic discrete representation of raw fundamental frequency contour as a sequence of target points. The same implementation codes the target points as a sequence of tonal symbols using the INTSINT alphabet.

A minimal functional annotation in IF (see Hirst 1977 [6]), is also provided through an automatic conversion of the Tonetic Stress Marks (TSMs) into IF. The TSMs were derived from the British school of intonation and enable us to annotate levels of prominence, F0 configurations and boundaries.

In the rest of this paper, the corpus and the procedure of analysis by synthesis are described. Finally different models of tonal alignment in British English are tested and evaluated.

2. Corpus and annotation

The corpus used in this research is extracted from the Aix-MARSEC corpus (see Auran & *al* 2004 [7]) 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 Auran (2004). There are five different speakers (three male and two female) who read short stories either meant for adults or children.

The corpus contains an orthographic transcription, a phonemic one, the annotation of lexical stress and a prosodic transcription in TSMs carried out by Gerry Knowles and Briony Williams. We also added a level of functional annotation using the IF coding system which annotates levels of prominence (accented, nuclear and emphatic) and boundary type (terminal [| *vs* non-terminal [+). Since the corpus was already annotated with the TSMs which represent both formal and functional information, rules were designed to extract functional information from the TSMs and convert these automatically into IF so that the final annotation of the corpus contains the following tiers: an annotation of boundary type ([| *vs.* [+) and sentence type at the level of the intonation units and an annotation of prominences at the level of tonal units following Jassem (1952) [8]. Having adapted IF to the British approach, it was decided to add a level of prominence to IF: the level of rhythmic prominence so that the following types of Tonal Units (TU) could be distinguished : unaccented TU (U); accented TU (A); stressed TU (S); nuclear TU (N); emphatic nucleus (!N) (see figure 1).

At the level of the tonal units, the functional labels (U, S, A, N) were characterised more precisely in keeping with their position within the Intonation Unit (Initial, medial, final, post-nuclear) and the type of Intonation Unit which they belong to (terminal assertion, terminal question, non terminal parenthesis). 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)

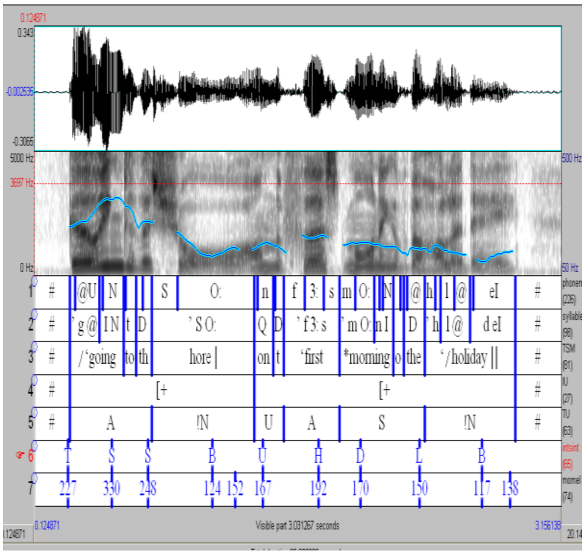
3697 Hz
0 Hz
50 Hz
# @U N S O: n f 3: s m O: N @ h l @ eI #
# 'g @ I N t D 'S O: Q D 'f3: s 'm O: n I D 'h l @ d eI #
# /'going to th hore | on t 'first *morning o the '/holiday || #
# [+ [+ #
# A !N U A S !N #
T S S B U H D L B
227 330 248 124 152 167 192 170 150 117 138
phonem
syllable
TSM
IU
TU
intsint
momel
0.124871
Visible part 3.031267 seconds
3.156138

Figure 1: Different levels of annotation of the MARSEC corpus.

This functional annotation is used as a starting point for the following experiments of optimization of formal representation in terms of F0 configurations and tonal alignment.

3. Testing various models of tonal alignment

Using the functional annotation of the corpus as a starting point, which defined the boundaries of different domains of association of the tones to be tested, the representation of prosodic form was optimized following two steps: First the local variations of F0 were modelled using INTSINT tones at the level of each tone unit; then the alignment of these tones was optimized relative to different anchor points.

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 the initial and medial TU, three for the final ones). Note that the first tone of the IU was necessarily an absolute tone (either T (Top), M (Mid) or B (Bottom)), since a relative tone presupposes that there is a preceding target, this first point was not included in the first TU, however it was aligned with the beginning of the IU.

By means of a Praat script, and for each tonal unit, all combinations of two or three INTSINT tones were tested. There were 8 possible tones at each point; 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 the 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 1)

		Going to the		shore		
Type of unit		[+A1		[+!NF		
Tone number		T1	T2	T1	T2	T3
Possible Tones	M T B	T B H L U D S _	T B H L U D S	T B H L U D S _	T B H L U D S _	T B H L U D S
Tested sequences of tones	M	T T T ... B B Etc.	T B H ... T B Etc.	T T T ... T T Etc.	T T T ... B B Etc.	T B H ... T B Etc.

Table 1. Illustration of the optimization of the coding of INTSINT tones

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; firstly an objective evaluation was carried out using linear correlation and root mean square error (RMSE), for the original hand-corrected MOMEL curve and the curve derived from the model; secondly a subjective evaluation (i.e. auditory) was carried out by comparing the original recordings to the resynthesis of the modelled MOMEL curve.

[+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 2. The best five sequences of tones for each type of nuclear TU in a non terminal IU.

The best thirty 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).

3.2. Optimizing the alignment of the tones

The second step of modelling prosodic form consisted of optimizing the alignment of the best sequences of tones selected in the first experiment. The anchor points which were taken into account in our experiments were the left and right boundaries of the tonal and intonation units, the first and last syllables of the TU, the first and last vowels of the TU, and the distance from the preceding target.

Two modes of alignment were also tested: in the first one, tonal alignment was optimized using absolute distances from the anchor points, in the second timing was allowed to vary as a function of the duration of the segmental anchors. Four different models were tested, the parameters of which are summed up in table 3:

Model 1: alignment with the TU (left and right boundaries)

	T1	T2	T3
1 or M	From 0 to 300ms after StartTU in 7 iterations	From -160 to -10 ms from EndTU in 4 iterations	
F	From 0 to 300ms after StartTU in 7 iterations	From 50 to 300ms from prec in 6 iterations	-10ms from End IU

Model 2: alignment with the TU (in %)

	T1	T2	T3
1 or M	From 10% to 40% of length TU in 7 iterations	From 80% to 95% of length TU in 4 iterations	
F	From 10% to 40% of length TU in 7 iterations	From 80% to 95% of length TU in 4 iterations	-10ms from End IU

Model 3: alignment with the stressed syllable (in %)

	T1	T2	T3
1 or M	From 0 to 150% of first syllable in 7 iterations.	From -160 to -10 ms from End TU in 4 iterations	
F	From 0 to 150% of first syllable in 7 iterations	From -40% to 80% of last syllable in 7 iterations	-10ms from End IU

Model 4: alignment with the stressed vowel (onset of the vowel)

	T1	T2	T3
1 or M	From -100 to 300 from first vowel in 9 iterations	From -160 to -10 ms from End TU in 4 it.	
F	From -100 to 300 from first vowel in 9 iterations	From -100 to 300 from onset of last vowel in 9 iterations	-10 ms from End IU

Table 3. Parameters of the models used to optimize tonal alignment.

T1, T2 and T3 represent the melodic targets, 1, M, F represent the position of the TU within the IU. The iterations correspond to the tested locations for each target. In model 3, the alignment of T1 is allowed to vary from 0 to 150% of the duration of the first syllable in 7 iterations which correspond to five different locations for T1 (alignment at 0% of the syllable, at 25%, 50%, 75%, 100% and 150%).

The alignment of the best thirty sequences of tones (derived from the first step of optimization) was optimized using these parameters and another Praat script. The best alignments and sequences of tones were then selected using linear correlation and RMSE. These results are saved as a basis for the statistical analysis described in the following section and for the subjective evaluation of our models of alignment using resynthesis with Psola.

3.3. Analysis of the results of the models

The best sequences of tones and alignments are saved at the level of each tonal unit. The distance between these points and the different anchor points, is the basis for the first type of statistic analysis. Different locations were tested in our models in order to find out which locations would allow us to obtain the best correlation with the original MOMEL curve. The locations where most best optimal representations are found should be the best landmarks for tonal alignment. We are seeking to discover a satisfactory response to the following questions; which anchor point is the most efficient at accounting for the variations in timing; which landmarks make it easier to predict the alignment of INTSINT tones?

3.4. Statistical evaluation

For each tonal unit, the alignment of two main tones was optimized (T1 and T2), the last one (T3) was aligned at a fixed offset (10 ms from the end of the IU). The best locations for the alignment of T1 in models 2 and 3 (aligned in percentage) can be summed up in figure 2:

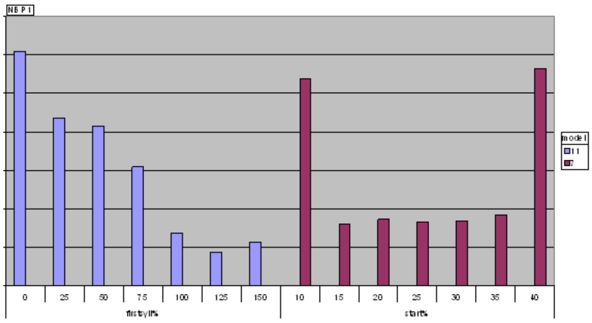

Figure 2. Count of best alignments for each tested location in model 3 (in blue on the left) and model 2 (in red)

In the case of model 3, we can see that the majority of the best sequences of tones and alignments occur within the boundaries of the stressed syllable. Indeed, 3811/4210 T1s are aligned from 0 to 100% of the first syllable, which amounts to 90.5%. On the other hand, model 2's results suggest that more locations should be tested at the level of the TU. Most T1s are aligned at either 10% or 40% of the TU which was the last tested location. A wider variation of T1's alignment could help confirm or modify these results.

The alignment of T2 in model 3 was also consistent with the boundaries of the last syllable of the TU since 90% of the best alignments were found between 20% and 80% of the last syllable. The optimal alignments for model 2 occurred at 80% of the last TU.

The alignment of T1 in models 1 and 4 is optimized in absolute distances from the boundaries of the TU for the first one and the onset of the stressed vowel in the second one:

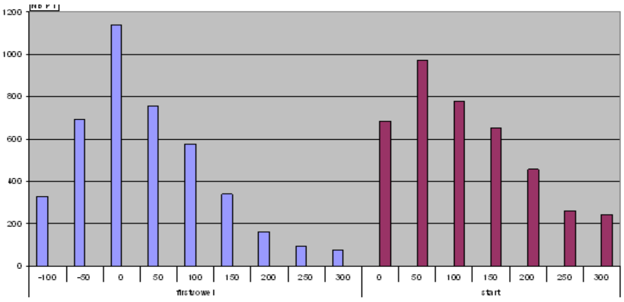

Figure 3. Count of best alignments for each tested location in model 4 (in blue on the left) and model 1 (in red)

The alignment of T1 in model 4 gave the best results from 50 ms before the onset of the vowel to 150 ms after the onset of the vowel. In model 1, the best results were obtained when the alignment of T1 varied from 0 to 200 ms from the beginning of the TU.

The alignment of T2 in model 4 was also better when varying from -50 before the onset of the last vowel of the TU to 150 ms after it. Finally, in model 1, T2 was aligned from 50 to 200 ms after the preceding tonal target.

3.5. Evaluation by analysis by synthesis

The quality of the models is evaluated by calculating the linear correlation of the MOMEL curve issued from the models and the original MOMEL curve from the data, intonation unit by intonation unit. Root Mean Square Error is also used to quantify the difference between the modelled curves and the original one:

*****	correlation	*****			
All		[+		[|	
M3	0.87126	M3	0.87883	M3	0.87528
M4	0.86605	M4	0.87736	M4	0.87419
M1	0.84018	M1	0.85619	M1	0.83868
M2	0.82726	M2	0.82492	M2	0.82966

Table 4. Linear correlation for the models 1 to 4 and for different types of IU (all, non-terminal and terminal)

As can be seen in table 4, Model 3 obtains the best correlations with all types of IU. The lowest values for RMSE are obtained with model 4 followed by model 3.

*****	RMSE	*****			
All		[+		[|	
M4	1.97466	M4	1.84865	M3	2.02415
M3	1.97963	M3	1.88364	M4	2.02988
M1	2.18724	M1	2.06272	M1	2.28558
M2	2.34219	M2	2.24355	M2	2.35985

Table 5. RMSE for all models and different types of IU.

The best models of tonal alignment seem to be those for which the alignment of the tones was allowed to vary within a small domain such as the syllable or the vowel (see figure 4: the MOMEL curve issued from model 3 on an utterance taken from the MARSEC corpus):

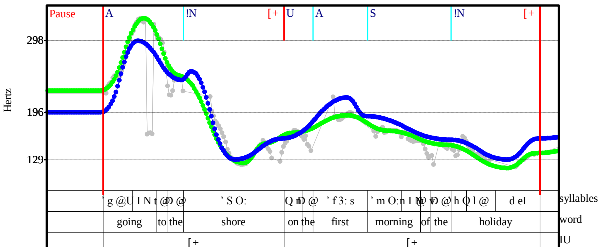

Figure 4. Optimized representation of prosodic form for an utterance in model 3 (tier 3= IU, tier 4= TU, tier 5= combination of correlation and RMSE, tier 6= the best sequences of tones and alignment).

4. Conclusion

In this paper we present a method of investigation of tonal alignment patterns in British English. It is based on the study of a large corpus of British English and a procedure of analysis by synthesis. Four models of tonal alignment based on the results reported in the literature are tested. The duration of the stressed syllable of a TU and the last syllable of a final TU seem to stand out as consistent landmarks in our experiments. It is hoped that further research on the subject will enable us to test more models of tonal alignment and include more prosodic features other than the melodic targets such as tempo and duration for instance.

5. References

[1] Xu, Y. 2002. Articulatory constraints and tonal alignment. In *Proceedings of the 1st International Conference on Speech Prosody,* Aix-en-Provence, France. pp. 91-100.

[2] D'Imperio, Mariapaola. 2002. Language-specific and universal constraints on tonal alignment: The nature of targets and anchors. *In the Proceedings of Speech Prosody 2002.*

[3] House, J, Wichmann, A. 1996. Investigating Peak timing in naturally occurring speech : from segmental constraints to discourse structure. Department of Phonetics and Linguistics. At the university of Central Lankashire.

[4] Pierrehumbert, J.B, Steele. 1989. Categories of tonal alignment in English. *Phonetica 46*: 181-196.

[5] Hirst, D.J. 2005. 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.

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

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

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