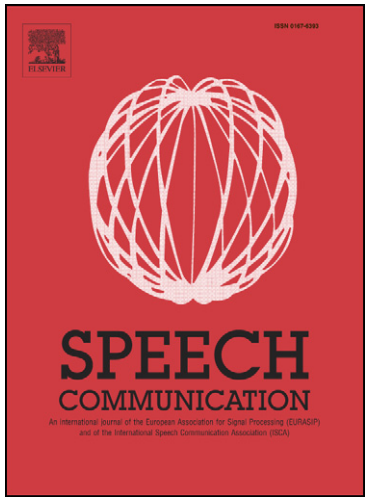
Integrating phrasing and intonation modelling using syntactic and morphosyntactic information

Francisco Campillo Di'Az, Jan van Santen, Eduardo Rodri'Guez Banga



Integrating phrasing and intonation modelling using syntactic and morphosyntactic information

Francisco Campillo Dı ´az, Jan van Santen, Eduardo Rodrı ´guez Banga

PII: S0167-6393(09)00009-0
DOI: 10.1016/j.specom.2009.01.007
Reference: SPECOM 1778

To appear in: *Speech Communication*

Received Date: 26 June 2008
Revised Date: 12 November 2008
Accepted Date: 23 January 2009

Please cite this article as: Dı ´az, F.C., van Santen, J., Banga, E.R., Integrating phrasing and intonation modelling using syntactic and morphosyntactic information, *Speech Communication* (2009), doi: 10.1016/j.specom.2009.01.007

Integrating phrasing and intonation modelling using syntactic and morphosyntactic information

Francisco Campillo Díaz, Jan van Santen, Eduardo Rodríguez Banga [a,b,a]

[a] *Dpto. Teoría de la Señal y Comunicaciones. ETSI Telecomunicación. Universidad de Vigo. Campus Universitario. 36200. Vigo. SPAIN*
e-mail: campillo@gts.tsc.uvigo.es; erbanga@tsc.uvigo.es

[b] *Center for Spoken Language Understanding*
OGI School of Science & Engineering, Oregon Health & Science University
20000 NW Walker Rd. Beaverton, OR. 97006, USA
e-mail: vansanten@ogi.cslu.edu

Corresponding author: Francisco Campillo.
Address:
Dpto. Teoría de la Señal y Comunicaciones. ETSI Telecomunicación.
Universidad de Vigo. Campus Universitario. 36200. Vigo. SPAIN
e-mail: campillo@tsc.uvigo.es
Phone: +34 986 812683 **Fax:** +34 986 812116

Abstract

This paper focuses on the relationship between intonation and syntactic and morphosyntactic information. Although intonation and syntax are both related to dependencies between the different parts of a sentence, and are therefore related to meaning, the precise influence of grammar on intonation is not clear. We describe a novel method that uses syntactic and part-of-speech features in the framework of corpus-based intonation modelling, and which integrates part of the phrasing algorithm in the unit selection stage. Subjective tests confirm an improvement in the quality of the resulting synthetic intonation: 75% of the sentences synthesised with the new intonation model were considered to be better or much better than the sentences synthesised using the old model, while only 7.5% of sentences were rated as worse or much worse.

Key words: intonation modelling, unit selection, corpus based, syntax, POS, phrasing

1 Introduction

One current reason for unnaturalness in synthetic intonation is the fact that the influence of syntax and semantics on the intonation generation is not accounted for (Prevost and Steedman (1993)). Most current intonation models only make use of features such as number of syllables, accent distribution and type of sentence, all of which are unrelated to the meaning of the sentence.

Intonation modelling is acknowledged to be one of the most relevant stages in speech synthesis, since languages use intonation variations to mark important parts of the discourse and dependencies between the different phrases (Pierrehumbert and Hirschberg (1990)). One of the functions of intonation is to divide up sentences into sequences of chunks or phrases (Ladd (1996)). Called prosodic structure, this plays a determining role both in naturalness and intelligibility (Ostendorf and Veilleux (1994)).

Syntax is the part of grammar that dictates how to coordinate and join words together to compose sentences and express concepts, while morphosyntax is the part of grammar that integrates morphology and syntax. From a practical point of view, morphosyntax refers to the function of each word in a sentence (a noun, for example), while syntax refers to the way in which words are put together to form constituents (a noun phrase, for example [1]).

[1] We will use the term syntagma in this paper to refer to a syntactic phrase, in order to avoid confusion with a prosodic phrase

The relation between syntax and intonation has been widely discussed in the literature. Although syntax is known to reflect dependencies among the different constituents of a sentence, it is not clear how it relates to intonation. Prosodic phrases cannot simply be predicted from syntactic structure, since prosodic boundaries do not always coincide with syntactic boundaries (Abney (1992), Ostendorf and Veilleux (1994), D'Imperio et al. (2003)). Moreover, sometimes it is the intonation itself which is used to clarify the correct syntax from among several alternatives (Steedman (1990)).

Text-to-speech systems traditionally estimate the prosodic structure of sentences by means of some kind of phrasing algorithm, implemented prior to the prosody generation stage (Black et al. (1999), Möebius (1999), Hernáez et al. (2001), Campillo and Banga (2006)). These phrasing algorithms make use of syntactic and/or morphosyntactic features to decide the best place to insert a phrase boundary within a sentence (Taylor and Black (1998), Koehn et al. (2000)). Given the prosodic structure, most intonation models do not use syntactic or morphosyntactic features to generate target intonation contours (van Santen and Möbius (1999), Garrido (1996), Campillo and Banga (2006), Navas (2003)). In this work we propose a novel approach that integrates part of the phrasing algorithm into an intonation unit selection model. Syntactic and morphosyntactic information–which we will refer to as grammatical information–is used to decide the best place to insert a prosodic boundary and the emphasis or strength of the accented syllables. Different prosodic structures are considered, and the best one is chosen according to the intonation units available in the prosodic corpus.

The article is outlined as follows: Section 2 summarises the intonation hierarchy that will be assumed for the purposes of this paper; Section 3 describes the corpus used in the study, and Section 4 gives a general overview of the key features of an intonation unit selection model; Section 5 first describes the experiments that were carried out to ascertain the influence of syntactic and morphosyntactic context on the shape of the contours around the accented syllable, and then follows up with the corresponding results and discussion; Sections 6 and 7 describe different phrasing algorithms and the new approach that combines phrasing and unit selection and Section 8 describes the cost functions for the unit selection stage; Section 9 presents a subjective test that demonstrates how the new method represents an improvement, and finally, Section 10 is dedicated to our overall conclusions and future lines of research.

2 Intonation hierarchy

As mentioned in the introduction, it is generally assumed that one of the functions of intonation is to divide up sentences into phrases (Ladd (1996)).

Although the literature refers to several linguistic theories on intonation structure (see Beckman and Pierrehumbert (1986) and Ladd (1986), for English, or Navarro (1977) for Spanish), many of them share a hierarchical structure with several levels that interact in different domains. The intonation group (IG)–or intonational phrase, in the terminology of Beckman and Pierrehumbert (1986)–is the most widely accepted unit of description in intonation (Garrido (1996)). It can be defined as a coherent intonation structure with no important prosodic boundaries within it (Escudero (2002)). These boundaries are hard to identify consistently (Ladd (1996)) since they vary from a clear pause accompanied by a local f_0 fall or rise to a subtle local pitch change. Thus, another unit commonly used is the phonic group (PG), defined in Navarro (1977) as the part of the discourse between two consecutive pauses. As noted in Garrido (1996) there is a tendency to confuse these two units, although their boundaries do not always coincide: a PG boundary is always an IG boundary, but the reverse does not always hold true.

In this research we include these two groups. We will also refer to the accent group (AG) as a sequence of non-accented words ending in an accented word. The intonation hierarchy will thus consist of a sentence containing a sequence of PGs, with each PG containing a sequence of IGs, and with each IG containing a sequence of AGs. Since every PG boundary implies an IG boundary, we will use the term "intonation break" (IB) to refer to IG boundaries not associated with a pause.[2] Moreover, given that this study is limited to the sentence domain, when talking about PG boundaries, we will only refer to internal pauses, rather than sentence final boundaries. Figure 1 shows the relation between the different intonation levels considered in this study.

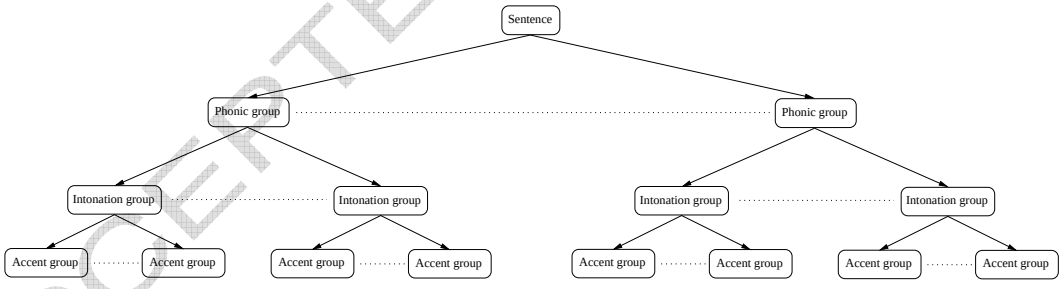


Fig. 1. Intonation hierarchy

This hierarchical model of PGs and AGS (and excluding the IGs) has already been successfully applied to Galician and Spanish in Campillo and Banga (2006). Consequently, another purpose of this article will be to study the suitability of this intonation hierarchy for Galician and to find a method to include the intonation group level in the intonation model.

[2] In this paper we will draw a distinction between major phrasing, associated with the estimation of PG boundaries, and minor phrasing, dedicated to the insertion of intonation breaks.

3 Corpus description

A Galician corpus was employed, consisting of recordings by two male professional speakers, identified as speakers 1 and 2, and with an average fundamental frequency, respectively, of around 88 Hz and around 110 Hz. It consists of two well differentiated sub-corpora: 807 sentences manually designed by an expert linguist to be a good sample of the basic prosodic structures of the Galician language (basic), and 559 sentences automatically selected to include more complex syntactic structures (complex).

The work presented here is limited by the characteristics of this corpus, which consistes of isolated sentences recorded in a neutral style. Consequently, factors related to discourse structure–such as controlling the focus of the sentence or reflecting the intentions of the speaker (see Grosz and Sidner (1986) or Pierrehumbert and Hirschberg (1990), for example)–are beyond the scope of this work.

This corpus is organised in terms of sentences, PGs, IGs and AGs, as described in Section 2. Table 1 shows the number of sentences, PGs, IGs, AGs and words for the speaker 1 corpus. The mean value of occurrences for the next level up (for example, the average number of IGs in a PG) are given in brackets.

Table 1
Corpus statistics: Mean values for the next level up in the hierarchy given in brackets

	Sentences	PG	IG	AG	Words
Basic	807	1440 (1.8)	2302 (1.6)	4496 (2.0)	6755 (1.5)
Complex	559	756 (1.4)	1288 (1.7)	2557 (2.0)	3696 (1.4)

Labelling was carried out in three steps. First of all, the accented words were automatically labelled since, in Galician–unlike in languages such as English–the accent is mainly determined by the part-of-speech (POS) tag of the word. Secondly, PG and AG boundaries were aligned to segmental boundaries that had previously been computed by forced alignment and then manually revised by a group of expert linguists. Finally, the IGs were manually labelled by an expert (one of the authors). An automated method for IG labelling will be the subject of future research.

As a result of this labelling, the following four boundary types will be possible at the beginning and end of every AG:

- PG boundary: a pause in the discourse, corresponding to silence in the waveform.
- IB: an IG boundary that does not coincide with a PG boundary.
- Comma IB: input text commas rendered spontaneously as an IB, with no pause.

- Non-breaking boundary: the absence of any boundary.

4 Overview of intonation unit selection systems

In this study we will mainly refer to the intonation model described by Campillo and Banga (2006). Following the same principles used for acoustic unit selection systems (Black and Campbell (1995)), it is assumed that a synthetic contour is indistinguishable from a natural one as long as it is created by a concatenation of basic intonation units applied in contexts similar to the contexts they were originally extracted from. In Campillo and Banga (2006) the basic unit used for concatenation purposes was the AG as defined in Section 2. In this way, an input sentence is divided up into a sequence of target AGs, which are parameterised according to a set of features such as duration, number of syllables, type of sentence, position of the AG within the PG and types of boundary surrounding the AG. All possible sequences of candidate groups are considered and the best one according to a cost function is chosen by means of a Viterbi search, as illustrated in Figure 2. The square boxes represent the sequence of desired feature vectors, while the round boxes represent the available candidate units. Shaded and unshaded candidates reflect the presence of units with different boundary conditions for the same target AG. We will come back to this figure in Section 7.

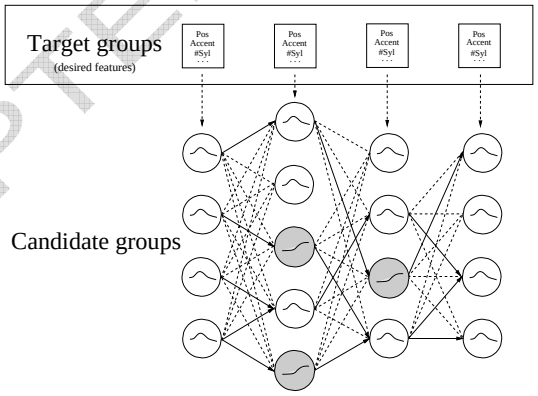


Fig. 2. Intonation unit selection (Campillo et al. (2008))

The total cost of a sequence of candidate units for a sentence with N target accent groups is computed using Equation (1) (Hunt and Black (1996)), combining two cost functions: a target cost function (C_{tar}) which measures the similarity between the target unit t_i and a candidate unit c_i, and a concatenation cost function (C_{con}) which measures the distortion associated with the concatenation of two candidate units, c_{i-1} and c_i, corresponding to adjacent

target units in the unit selection stage.

$$C = \sum_{i=1}^{N} C_{tar}(\mathbf{t}_i, \mathbf{c}_i) + \sum_{i=2}^{N} C_{con}(\mathbf{c}_{i-1}, \mathbf{c}_i) \quad (1)$$

Not every candidate unit is suitable for a given target unit. For example, an ellipsis typically shows a final flat contour which would not be appropriate for a declarative sentence, given that there would be a change in meaning. Therefore, candidate units are classified into different clusters, taking into account several features that are considered to be crucial. In this case, the corpus is clustered into 48 subsets on the basis of the sentence type (declarative, interrogative, exclamatory or ellipsis), the position of the accented syllable within the AG (ultimate, penultimate or antepenultimate) and the position of the AG within the PG (initial, final, intermediate or initial-final). This first hard-decision, applied to every target unit before the unit selection, is depicted in Figure 3.

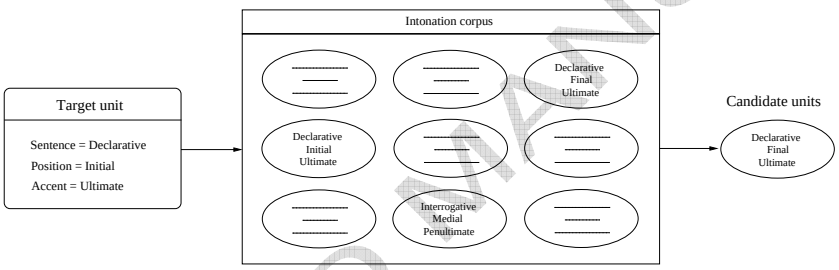


Fig. 3. Accent group clustering: selecting suitable candidates for a target accent group in a declarative sentence, final position in the phonic group, and with the accent on the last syllable

5 Experiments to assess the influence of grammatical context

Although there is no general agreement about the influence of syntactic and morphosyntactic information, its application to intonation modelling is obviously not new, as is evidenced by the many different approaches described in the literature. For example, for speech synthesis Taylor (2000) uses phonological trees which include the syntactic structure of the candidate and target units. A top-down search algorithm designed to pick candidates up in the database tree promotes the use of the longest available units with regard to the tree structure. Therefore, in choosing candidates with the same local syntactic structure, intonation modelling is implicit to the method. [3] A different approach is presented in Raux and Black (2003), who propose an intonation

[3] However, there is also a parallel prosody estimation for the cases when the phonological match is not good enough.

unit selection model with the segment f0 contour as the basic unit for concatenation, with the estimated POS of the word containing the segment and the neighbouring words used as some of the features to decide suitable sets of candidate units. Finally, most intonation models use syntactic and morphosyntactic information only indirectly, taking as input, rather, the prosodic structure of the sentence as given by some external phrasing algorithm that generally makes use of these features (Taylor and Black (1998), Koehn et al. (2000)). Once accent distribution and phrasing are predicted, these models do not use grammatical features to estimate target intonation contours.

Although grammatical information has proven to be useful in all these approaches, there are some questions that remain to be answered. For example, copying intonation contours from sentences with exactly the desired grammatical features may produce very good results, but perhaps it would be better to discern which grammatical contexts correspond to genuinely different intonation contours, since this would allow a more efficient use of the intonation database. Moreover, separating the phrasing algorithm and the intonation model might not be the best choice, especially in the case of a corpus-based intonation system.

In order to try to shed some light on this problem, we conducted two experiments. Firstly, we studied the influence of syntactic information on the insertion of IBs, and secondly, we assessed the influence of grammatical features on the shape of the intonation contour around the accented syllable. The following features were considered:

- POS label: morphosyntactic label of the accented word in the AG.
- Current syntagma: syntactic constituent to which the AG belongs.
- Next syntagma: syntactic constituent following the current syntagma.

We will henceforth refer to these three grammatical features as "grammatical context". As an example, Figure 4 shows the grammatical context for the AG "O rapaz" (*The boy*) in the sentence "O rapaz que viviu" (*The boy who lived*).

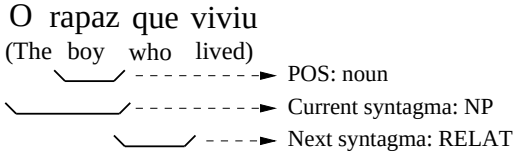


Fig. 4. Grammatical context of the "O rapaz" (*The boy*) accent group

Table 2 shows the classification considered according to the POS label. POS tagging was obtained with an ngram-based tagger (Méndez et al. (2003)), already integrated in the speech synthesiser. Regarding syntactic information, shallow parsing was chosen, which was easily computed by means of a set of linguistic rules from the morphosyntactic analysis, as it does not include the dependencies between the different components. The following syntagmas were

considered: noun phrase (NP), verb phrase (VP), adjective phrase (AdjP), prepositional phrase (PP), adverb phrase (AdvP) and other (OtherP), plus the conjunctions and relative pronoun included in the POS classification. Figure 5 is an example of the syntactic and morphosyntactic analysis of the sentence "A caída da casa de Usher foi escrita por Poe" (*The Fall of the House of Usher was written by Poe*).

Table 2
POS labels

Tag	Meaning	Example
Noun	Noun	Ese **coche** é rápido (That **car** is fast)
Adj	Adjective	O libro era **interesante** (The book was **interesting**)
Verb	Verb	A ela **encántanlle** os libros (She **loves** books)
Pron	Pronoun	**Ela** estará alí (**She** will be there)
Adv	Adverb	Era **demasiado** longa (It was **too** long)
Int	Interrogative pronoun	¿**Onde** estás? (**Where** are you?)
Exc	Exclamative pronoun	¡**Que** demo! (**What** the hell!)
RELAT	Relative pronoun	O rapaz **que** veu (The boy **who** came)
CopCON	Copulative conjunction	Ti **e** mais eu (You **and** me)
ConCON	Contrastive conjunction	Triste **pero** certo (Sad **but** true)
DISJ	Disjunction	Un **ou** o outro (One **or** the other)
SubCON	Subordinate conjunction	Fíxate **que** é un exemplo (Note **that** this is an example)
Other	Other	**1984**

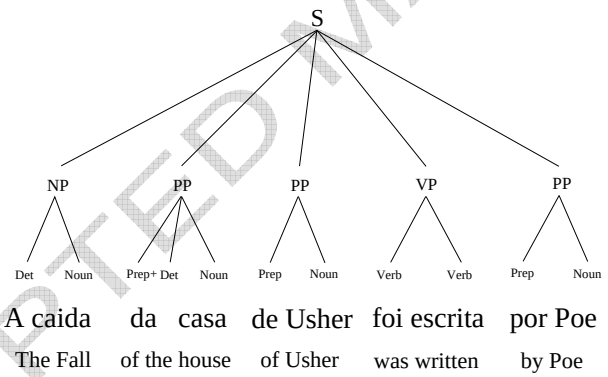


Fig. 5. Example of syntactic analysis and POS tagging

5.1 Experiment 1

In the first experiment, the goal of which was to find a way to integrate the phrasing algorithm into the unit selection stage, we studied the influence of syntactic information on the insertion of IBs. We computed the percentage of IB occurrence before the different types of syntagmas in the corpus described in Section 3, as rendered by the two speakers. Cases where the syntagma was preceded by a punctuation mark were discarded from the statistics since we were only interested in the behaviour of the speaker in the absence of punctuation marks.

5.2 Results of experiment 1

Table 3 shows the results for this experiment. As can be seen, some syntagmas seem more likely to be preceded by an IB than others. Compare, for example, the disjunction (*or*, for example) at 94.7%, and the adverb phrase (*highly*, for example) at 23.3%. A similar analysis was carried out to check if the POS tag of the word succeeding the IB would produce similar results, but in some cases there were clear differences. For example, when the tag was a noun, the percentage was only 1.3%, much lower than the 16.1% obtained with the NP. Likewise, the probability of the occurrence of an IB before the determiner of an NP can change drastically depending on whether the determiner is preceded by a preposition. Although this phenomenon can also be modelled with ngrams of POS labels (Taylor and Black (1998)), this trivial rule-based syntactic parsing is more than adequately efficient in regard to integrating phrasing and unit selection, as will be shown in Section 8.1.

Table 3
Percentage of IBs before different syntagma types (speaker 1)

Syntagma	Total	Preceding IBs	%
VP	1136	271	21.9
DISJ	94	89	94.7
NP	1881	303	16.1
RELAT	36	31	86.1
PP	1174	378	32.2
CopCON	128	97	75.8
ConCON	0	0	0
AdvP	437	102	23.3
SubCON	114	73	64.0
AdjP	475	42	8.8
OTHER	112	48	42.9

5.3 Experiment 2

In the second experiment, we wanted to distinguish which contexts were actually different from the intonation point of view. Considering different syntactic and morphosyntactic conditions we studied the variations in the shape of the contour around the accented syllable clustering the AG database as a function of the grammatical context.

The contour around the accent was modelled in terms of two factors. The first

factor is the slope factor (Campillo and Banga (2006)), defined as the slope of the imaginary line joining the f_0 value in the middle of the nucleus of the accented syllable and the value at the end of the AG (Figure 6, top), which is related to the hierarchical interpretation of the discourse. For example, a rise at the end of an IG usually suggests more information coming about a topic, whereas a fall at the end of an IG is suggestive of an utterance completing a topic (Pierrehumbert and Hirschberg (1990)). The second factor is an estimation of the height of the AG, Δ_{f0}, computed as the difference between the f_0 value in the middle of the nucleus of the accented syllable and that of the preceding nucleus after a simple estimate of the phrase curve of the AG is subtracted (Campillo et al. (2006a)) (Figure 6, bottom), which gives an idea of the emphasis of the AG [4] . These two factors were measured as a function of the three grammatical features comprising the grammatical context, that is, the POS label of the accented word in the AG, the current syntagma, and the next syntagma.

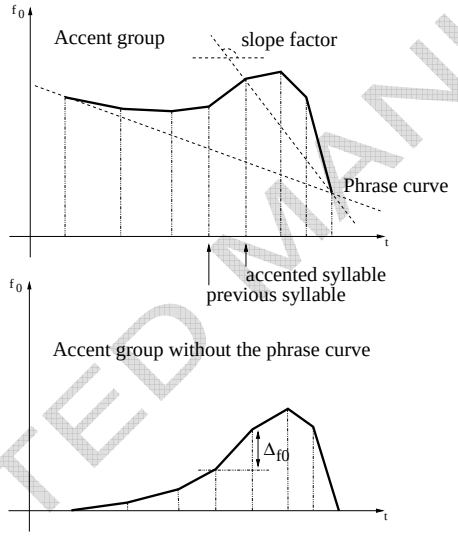


Fig. 6. Accent model: estimating the slope factor (top) and Δ_{f0} (bottom)

Finally, we hypothesise that the influence of grammatical information might be different depending on the type of boundary at the end of the AG. In order to disaggregate this influence, the four boundary conditions mentioned in Section 3 were considered: PG boundary (pause), [5] IB, non-breaking boundary, and comma IB. Summing up, in the second experiment we classified the AG into twelve groups (four boundary conditions by three grammatical features), and compared their average Δ_{f0} and slope factor. Note that, however, these

[4] Although this accent model does not capture the shape of the f_0 contour at the beginning of the AG, this part of the contour seems to have only meaningful variations when the AG is preceded by a prosodic boundary. In our system we consider this behaviour in the concatenation cost (see Section 8.2)

[5] Remember that in this analysis we only take into account pauses other than those at the end of sentences

classes were not exclusive. In the example given in Figure 4, the AG "O rapaz" (*The boy*) would be included in the classes $\{Boundary = no\,break,\ POS = Noun\}$, $\{Boundary = nobreak, current syntagma = NP\}$ and $\{Boundary = no\,break,\ next\ syntagma = RELAT\}$.

5.4 Results of experiment 2

As a consequence of the definition of Δ_{f0}, the AGs with an accented syllable at the beginning of the group were excluded from the statistics. Similarly, AGs with the accent on the last syllable were discarded from the slope factor data. Given the large number of tables in this experiment, only the most relevant results for speaker 1 will be shown, unless otherwise indicated, as similar tendencies were found for the other speaker.

The most interesting result was that the POS label, current syntagma and next syntagma seemed to have different relevance in the four boundary conditions, which confirms our hypothesis of the end of Section 5.3 to study the influence of the grammatical context in those different situations. Table 4 shows the influence of the feature next syntagma when an AG is followed by an IB, while Tables 5 and 6 show, respectively, the p-values from a t-test for the means of the Δ_{f0} and the slope factor in a pairwise comparison of the different clusters, being the null hypothesis that the means of the compared clusters are equal.

Table 4
Influence of the next syntagma. Δ_{f0} and slope factor before an intonation break (speaker 1)

	Δ_{f0}			Slope factor		
Next syntagma	No	Mean	Std Dev	No	Mean	Std Dev
VP	242	14.11	8.60	220	-23.76	55.75
DISJ	80	35.35	12.63	68	-202.81	133.68
NP	252	11.94	11.94	220	-50.51	60.21
RELAT	27	9.30	9.53	29	-22.08	44.26
PP	352	12.32	8.19	303	-52.61	56.87
CopCON	73	13.71	8.82	79	-41.55	54.29
AdvP	86	11.03	6.93	73	-47.46	48.35
SubCON	41	8.13	7.31	53	-23.24	49.82
OtherP	30	13.89	6.95	34	-69.13	54.41
AdjP	37	12.65	8.83	32	-50.10	68.36

As can be seen, there are noticeable differences between some syntagmas. For example, for a disjunction, the mean Δ_{f0} is 35.35 Hz whereas for a relative pronoun the mean is 9.30. Similarly, the slope factor is -23.76 Hz/s for verb phrases, while it is -50.10 Hz/s for adjective phrases. As shown in Table 5, although not all the differences are statistically significant, most are; for example, DISJ and all the other categories, SubCON and NP, VP and NP, etc. These results prove the existence of grammatical contexts with differences

Table 5

Influence of the next syntagma. Intonation break: p-value from a t-test for the mean of the Δ_{f0} of the different classes (speaker 1)

*****	DISJ	NP	RELAT	PP	CopCON	AdvP	SubCON	OtherP	AdjP
VP	0.000	0.007	0.018	0.012	(0.73)	0.001	0.000	(0.88)	(0.35)
DISJ		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
NP			(0.18)	(0.60)	(0.14)	(0.34)	0.004	(0.17)	(0.65)
RELAT				(0.12)	0.042	(0.39)	(0.59)	0.045	(0.16)
PP					(0.22)	(0.14)	0.001	(0.25)	(0.83)
CopCON						0.037	0.000	(0.91)	(0.55)
AdvP							0.037	(0.06)	(0.33)
SubCON								0.001	0.017
OtherP									(0.52)

Table 6

Influence of the next syntagma. Intonation break: p-value from a t-test for the mean of the slope factor of the different classes (speaker 1)

*****	DISJ	NP	RELAT	PP	CopCON	AdvP	SubCON	OtherP	AdjP
VP	0.000	0.000	(0.85)	0.000	0.014	0.001	(0.95)	0.000	0.044
DISJ		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
NP			0.003	(0.69)	(0.22)	(0.66)	0.001	(0.07)	(0.97)
RELAT				0.001	(0.06)	0.014	(0.91)	0.000	(0.06)
PP					(0.11)	(0.43)	0.000	(0.09)	(0.84)
CopCON						(0.48)	0.048	0.014	(0.53)
AdvP							0.007	0.047	(0.84)
SubCON								0.000	(0.06)
OtherP									(0.21)

that are not only statistically significant but perceptually noticeable.[6] This implies that AGs extracted from some contexts should never be used in other target contexts (for example, in the presence of an IB, an AG followed by a disjunction would not be a suitable candidate for a target AG followed by a subordinate conjunction). For comparison purposes, Tables 7 and 8 show the same information for speaker 2. Note that, although the results are not exactly the same, differences among classes are also noticeable.

Table 9 shows the influence of the current syntagma on the Δ_{f0} and the slope factor when there is an IB. Compared with Table 4 it can be seen that values are more regular and that standard deviations are generally larger, which suggests that the current syntagma is a worse classifier than the next syntagma for distinguishing among classes when there is an IB. Table 10 summarises the p-values from the t-test for the means of the Δ_{f0}. Similarly, Table 11 depicts the influence of the POS label when the AG is followed by an IB (it seems, in fact, to be very similar to the current syntagma case).

[6] It is not the purpose of this paper to find an exact perceptual threshold for these differences, although it would be a very interesting line of research in itself. The subjective test in Section 9 will clarify the relevance of including these results in the intonation model.

Table 7

Influence of the next syntagma. Δ_{f0} and slope factor before an intonation break (speaker 2)

	Δ_{f0}			Slope factor		
Next syntagma	No	Mean	Std Dev	No	Mean	Std Dev
VP	212	20.38	9.17	188	-53.77	52.36
DISJ	71	27.35	13.78	59	-44.42	74.83
NP	203	12.95	9.49	179	-38.80	60.99
RELAT	21	15.40	10.57	24	-53.84	39.71
PP	294	17.33	10.71	263	-53.36	56.20
CopCON	72	17.62	11.28	86	-30.58	62.42
AdvP	64	15.69	9.60	62	-52.71	56.86
SubCON	43	11.29	10.56	52	-38.56	58.70
OtherP	19	13.87	8.37	23	-62.46	47.21
AdjP	32	14.66	9.28	23	-41.59	51.57

Table 8

Influence of the next syntagma. Intonation break: p-value from a t-test for the mean of the Δ_{f0} of the different classes (speaker 2)

*****	DISJ	NP	RELAT	PP	CopCON	AdvP	SubCON	OtherP	AdjP
VP	0.000	0.000	0.048	0.001	(0.06)	0.001	0.000	0.004	0.002
DISJ		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
NP			(0.32)	0.000	0.002	0.048	(0.34)	(0.66)	(0.34)
RELAT				(0.43)	(0.41)	(0.91)	(0.15)	(0.61)	(0.80)
PP					(0.85)	(0.23)	0.001	(0.10)	(0.14)
CopCON						(0.29)	0.003	(0.12)	(0.17)
AdvP							0.031	(0.42)	(0.61)
SubCON								(0.31)	(0.15)
OtherP									(0.75)

Table 9

Influence of the current syntagma. Δ_{f0} and slope factor before an intonation break (speaker 1)

	Δ_{f0}			Slope factor		
Current syntagma	No	Mean	Std Dev	No	Mean	Std Dev
VP	321	12.52	9.40	259	-49.83	60.30
DISJ	0	0.00	0.00	0	0.00	0.00
NP	394	14.54	10.93	382	-56.58	80.72
RELAT	0	0.00	0.00	0	0.00	0.00
PP	322	14.80	10.91	276	-53.67	87.54
CopCON	0	0.00	0.00	0	0.00	0.00
AdvP	52	14.51	9.87	51	-44.82	60.34
SubCON	0	0	0.00	0	0.00	0.00
OtherP	9	15.10	8.24	10	-65.05	61.22
AdjP	118	13.24	11.06	131	-47.89	63.47

Table 12 shows the influence of the next syntagma in the absence of a boundary. Obviously the Δ_{f0} means are smaller than in the case of the IB, since there is no boundary at the end of the AG. This, to some extent, validates the accent model used in this work. Similarly to the influence of the current

Table 10

Influence of the current syntagma. Intonation break: p-value from a t-test for the mean of the Δ_{f0} of the different classes (speaker 1)

*****	DISJ	NP	RELAT	PP	CopCON	AdvP	SubCON	OtherP	AdjP
VP	N-DAT	0.08	N-DAT	0.005	N-DAT	(0.18)	N-DAT	(0.38)	(0.53)
NP			N-DAT	(0.75)	N-DAT	(0.99)	N-DAT	(0.84)	(0.26)
PP					N-DAT	(0.85)	N-DAT	(0.92)	(0.19)
AdvP							N-DAT	(0.85)	(0.46)
OtherP									(0.54)

Table 11

Influence of the POS label. Δ_{f0} and slope factor before an intonation break (speaker 1)

	Δ_{f0}			Slope factor		
Next syntagma	No	Mean	Std Dev	No	Mean	Std Dev
Noun	673	14.77	11.07	602	-55.65	84.16
Adj	134	13.59	10.84	146	-49.18	62.97
Verb	316	12.43	9.40	253	-50.16	60.62
Pron	30	11.68	6.92	33	-43.11	75.43
Adv	57	14.83	9.74	57	-43.86	58.73
Other	7	11.85	6.88	4	-51.12	27.83

syntagma in the presence of an IB, the next syntagma does not seem to be relevant when there is no boundary. The p-values from the t-test for the Δ_{f0} means are presented in Table 14.

Table 12

Influence of the next syntagma. Δ_{f0} and slope factor in the absence of a boundary (speaker 1)

	Δ_{f0}			Slope factor		
Next syntagma	No	Mean	Std Dev	No	Mean	Std Dev
VP	134	8.70	8.86	126	-63.91	82.57
DISJ	4	9.07	5.38	2	-51.03	14.41
NP	737	8.75	7.87	806	-42.77	62.25
RELAT	4	3.77	9.06	4	-31.43	17.33
PP	626	10.80	8.03	549	-52.00	53.24
CopCON	15	6.85	6.98	17	-27.63	34.88
AdvP	143	8.68	8.19	140	-55.46	56.64
SubCON	9	7.30	8.98	10	-30.65	49.38
OtherP	24	10.09	8.47	24	-50.84	48.61
AdjP	267	9.97	8.19	226	-49.64	54.21

Table 14 shows the influence of the POS label when there is no boundary after the AG, while Table 15 summarises the p-values from the t-test of the means of the Δ_{f0}. In this case the differences are statistically significant between Adv, OTHER and the rest of the classes. This is not a surprising result, as nouns, adjectives and verbs usually convey the most important information in the sentence.

Regarding the slope factor, it is also remarkable that similar means were ob-

Table 13

Influence of the next syntagma. Absence of a boundary: p-value from a t-test for the mean of the Δ_{f0} of the different classes (speaker 1)

*****	DISJ	NP	RELAT	PP	CopCON	AdvP	SubCON	OtherP	AdjP
VP	(0.90)	(0.95)	(0.36)	0.012	(0.36)	(0.99)	(0.66)	(0.47)	(0.16)
DISJ		(0.91)	(0.36)	(0.57)	(0.52)	(0.90)	(0.67)	(0.76)	(0.76)
NP			(0.35)	0.000	(0.32)	(0.93)	(0.64)	(0.45)	0.035
RELAT				(0.22)	(0.56)	(0.36)	(0.54)	(0.26)	(0.26)
PP				0.048	0.006	(0.28)	(0.69)	(0.16)	(0.62)
CopCON					(0.36)	(0.90)	(0.20)	(0.11)	(0.98)
AdvP						(0.66)	(0.46)	(0.13)	(0.81)
SubCON							(0.43)	(0.40)	(0.94)
OtherP								(0.95)	(0.69)
AdjP									(0.69)

Table 14

Influence of the POS label. Δ_{f0} and slope factor in the absence of a boundary (speaker 1)

	Δ_{f0}			Slope factor		
Next syntagma	No	Mean	Std Dev	No	Mean	Std Dev
Noun	678	11.44	8.27	614	-53.12	54.79
Adj	146	9.95	8.26	151	-51.67	54.36
Verb	819	9.14	7.85	649	-49.22	53.31
Pron	40	9.42	7.13	40	-49.76	68.78
Adv	104	5.39	7.15	74	-44.60	57.85
Other	25	5.71	9.57	28	-22.11	80.86

Table 15

Influence of the POS label. Absence of a boundary: p-value from a t-test for the mean of the Δ_{f0} of the different classes (speaker 1)

*****	Adj	Verb	Pron	Adv	Other
Noun	(0.05)	0.000	(0.09)	0.000	0.007
Adj		(0.27)	(0.69)	0.000	0.045
Verb			(0.81)	0.000	(0.09)
Pron				0.003	(0.10)
Adv					(0.63)

tained for most of the classes. This result holds for the influence of the current and next syntagma in the absence of a boundary, seeming to imply that in this situation the slope factor, that is, the direction of the intonation contour after the accent, does not convey discriminative information.

5.5 Discussion

Regarding the first experiment, Table 3 reflects noticeable differences between certain syntagmas, even on comparing the various types of conjunctions (see, for example, CopCON at 75.8% and SubCON at 64.0%). There are other fea-

tures that can affect the insertion of an IB, but since the corpus described in Section 3 was designed to be a balanced sample of the most frequent syntactic structures in Galician, we conclude that the next syntagma clearly has a bearing on the insertion of IBs in the discourse.

The results of the second experiment show some evidence of the effect of the syntactic and morphosyntactic structure on the shape of the intonation contours and, hence, on the information they convey. The most important result here is the varying influence of the grammatical context on the four boundary conditions. In summary, the most relevant features for each were:

- IB: next syntagma.
- PG boundary: both current and next syntagma.
- Comma IB: next syntagma. Regarding the current syntagma, there are also noticeable differences but they are not statistically significant, probably as a consequence of the lack of data in our corpus for this boundary type.
- Non-breaking boundary: POS label. There are also small consistent differences with respect to the next syntagma, but they are hardly perceptible.

In Ladd (1988) the declination reset is studied for English in sentences of the form "A, and B, but C" and "A, but B, and C", it being concluded that the amount of reset depends on the boundary strength and not on the presence of "and" and "but". Even though our study focuses on a different language, it seems reasonable to think that in Galician the reset also depends on the boundary strength, which is directly related to the meaning of the sentence. However, since current technology provides a very limited understanding of sentences, the rule-based syntactic analysis used here seems to be a good approach to modelling boundary strength.

6 Phrasing algorithms

The results described in Section 5.2 show some evidence of the influence of grammatical information on the intonation contours as a function of boundary type. However, in order to make this distinction between the target boundaries in the unit selection stage, an external phrasing algorithm would be needed.

In the literature there are several approaches, most of them dedicated to major phrasing. Taylor and Black (1998) describe a generic algorithm for inserting different types of boundaries, based on using a Markov model where each state represents one type of boundary and emits probabilities of POS tags. The method is simple and has the advantage of taking information that can easily be obtained at synthesis time as input; however, the study is mainly focused on major phrasing assignment. Hirschberg and Prieto (1996) describe

a different approach for major phrasing, based on using decision trees generated automatically from manually labelled text corpora. They include features such as a POS window of four words around the current word and counts of the number of words/syllables in an utterance and of the distance in words from the beginning and end of an utterance. Ostendorf and Veilleux (1994) describe a model applied to a hierarchy containing sentences, PGs and IGs, where each unit is represented in terms of the probability of the sequence of sub-units comprising it. Decision trees embedded in this hierarchical structure are employed, with questions arising from syntactic and morphosyntactic information.

The use of syntactic information is somehow controversial. Although not used by Taylor and Black (1998), these authors claim that there is a limit to how well a model like theirs can perform with only POS information, as certain decisions about phrase break assignment can only be reasonably made with syntactic information. Similarly, the major phrasing model in Hirschberg and Prieto (1996) is improved on by Koehn et al. (2000) with the use of syntactic information. However, this information was not found to be useful when included in the hierarchical model developed by Ostendorf and Veilleux (1994), although it was useful when applied to the non hierarchical classification tree these authors built for comparison purposes.

7 Integrating unit selection and minor phrasing

A very interesting observation was made in Ostendorf and Veilleux (1994): several different prosodic parses may all be allowed in one sentence without altering naturalness or meaning. Their model takes advantage of this variability considering all the possible prosodic parses and selecting the most likely one.

As mentioned in the Introduction, the prosodic structure of the sentence is an input to traditional intonation models. This is not an optimal approach in the case of corpus-based intonation models, where synthetic contours are generated by concatenation of natural contours extracted from a limited corpus. As a result, the final quality of the synthetic contour will depend on finding a good match between the available candidate units and the desired features, with prosodic structure as a very important feature. Therefore, in a similar way to Ostendorf and Veilleux (1994), corpus-based intonation models can also benefit from considering different prosodic structures for each input sentence. In the particular case of corpus-based intonation modelling, unit selection and phrasing should not be addressed as independent problems.

Another flaw in most phrasing algorithms is that they try to predict intonation

boundaries leaving out the intonation contour itself. Some models (Ostendorf and Veilleux (1994)) include the constituent length or the number of content words as factors that are obviously related to the prosodic structure: the longer the sentence, the more likely a phrase break as a consequence of the dynamic range and declination of the intonation contour, which will need a reset at some point. However, our conjecture is that the decision to include an IB depends not only on the sentence length but also on the f_0 value at each point. The lower this value, the more likely the insertion of an IB. Moreover, since this reset is related to the Δ_{f0} and slope factor results given in Section 5.4 the grammatical context should be taken into account every time an IB is inserted. Note that the actual $f0$ values cannot be taken into account when the stages of phrasing and intonation modelling are separated.

Our approach was as follows:

- The major and minor phrasing algorithms are implemented as different stages.
- Major phrasing is accomplished by means of a decision tree, with factors such as the distance in syllables from the last pause and the distance in syllables to the next pause, and a POS window of three places to the left and right of the current word. Although there may be more sophisticated approaches, but for our purposes the key point is that the pause distribution is an input to the intonation model.
- As mentioned in Section 4, not every candidate AG is suitable for a given target unit, which implies that a first decision has to be taken on the available units, as depicted in Figure 3. The key point in the new approach is thus to organise the AGs according to their position within the PG. Candidate AGs followed by an IB are therefore included in the intermediate or initial clusters.
- At selection time, for every initial and intermediate target group, the corresponding candidate AGs can be followed either by IBs or no boundaries, since the PG boundary is the pause. This situation is depicted in Figure 2, where shaded candidate AGs represent groups followed by an IB. Similarly to Ostendorf and Veilleux (1994), by considering every sequence of candidate AGs, we also take into account all the possible prosodic structures, as long as there are available candidate AGs with different types of boundary. Figure 7 shows an example of two synthetic intonation contours with different prosodic structure for the sentence "Non sabía se saír ou quedar na casa" (*He didn't know whether to go out or stay at home*). Note the presence of an IB around 0.7 s in contour 1.
- A new sub-cost is introduced in the target cost that takes into account the results of experiment 1 (see Section 5.2): as a function of the next syntagma in the target AG, candidate AGs are weighted appropriately depending on whether or not they are followed by an IB (Campillo et al. (2008)).
- For each candidate AG, the results obtained in experiment 2 (see Sec-

tion 5.4) are applied according to its actual and not target boundary. For example, if the intermediate candidate AG is not followed by an IB, morphosyntactic information is used (see Table 14); otherwise, syntactic information is considered (Table 4).

- The actual $f0$ values are taken into account, since $f0$ continuity is weighted heavily in the concatenation cost (Section 8.2 below).

Hence, on considering different prosodic structures, the minor phrasing algorithm is integrated in the intonation unit selection module. The best combination of prosodic structure and intonation units is chosen, and IBs are introduced in the most appropriate places according to the cost functions and taking into account their boundary strengths. Note that modelling the boundary strength is a fundamental point in this approach. Inserting a strong IB in the wrong place would be unnatural and could even change the meaning of the sentence. The conjecture of the authors is, however, that soft IBs (small values of Δ_{f0}) in the right places in accordance with syntactic information would improve the quality of synthetic contours. The subjective test described in Section 9 below supports this hypothesis.

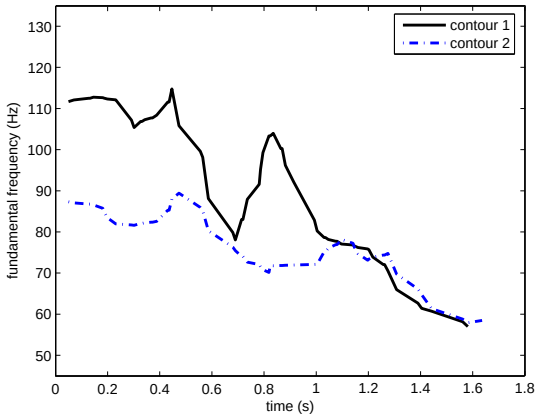


Fig. 7. Two contours with different phrasing for the sentence "Non sabía se saír ou quedar na casa" (*He didn't know whether to go out or stay at home*) (Campillo et al. (2008))

8 New cost functions

In this section we will address the topic of the design of the cost functions. Some features have already been explained in Campillo and Banga (2006), and the reader is referred there for a more detailed description.

8.1 Target cost

The following sub-costs from the model described in Campillo and Banga (2006) were maintained:

- Number of syllables.
- Duration.
- Position of the PG in the sentence.

Other sub-costs such as the position of the IG in the PG and the final slope of the candidate AG were discarded based on informal listening tests. Finally, two further sub-costs were added:

- Syntactic and morphosyntactic cost.
- IB probability.

Adding the syntactic and morphosyntactic cost information to a unit selection system is extremely easy:

```
if (candidate→final–boundary == Intonation–break)
    if (target→next–syntagma != candidate→next–syntagma)
        cost += 1; // just an example
```

Taking into account the candidate final boundary, candidate AGs with a grammatical context found to be significantly different from the target context are penalised, in a series of *if...then* statements like the example above.

In relation to IB probability, since conjunctions, disjunctions and relative pronouns seem to have a close relationship with the insertion of IBs (see Table 3) when the target unit's next syntagma belongs to one of these classes, the candidate groups not followed by one of them are penalised, and vice versa. Note that we are not directly forcing the insertion of an IB, but the use of an AG that is likely to be followed by an IB.

8.2 Concatenation cost

Only three sub-costs are considered:

- Natural continuity. If the candidate AGs are adjacent in the original recording, the concatenation is assumed to be natural and so the cost is zero. This sub-cost is overridden when the candidate AGs are separated by a pause, since the f_0 reset depends on features such as the number of AGs in the

PG after the pause (Campillo et al. (2006b)) which can be different in the target and candidate sentences.

- f_0 continuity. Discontinuities in the concatenation point are heavily penalised. In the presence of a pause, this sub-cost takes into account the f_0 reset, which is computed by means of a neural network with input parameters such as the number of AGs within the PG after the pause, the duration of the pause and the position of the PG within the sentence (see Campillo et al. (2006b)).
- Boundary continuity. Since the IB insertion depends on the candidate AG, in the Viterbi search the concatenation of groups with different boundary types is plausible. It would be possible to avoid this situation, but it was preferred to penalise these cases heavily, and leave them as a fallback mechanism in case there was no suitable candidate AG. Note that this sub-cost was applied basically to IBs, since AGs ending in a pause were always followed by an AG after a pause as a consequence of the clustering depicted in Figure 3.

8.3 Training the weights

Tuning the weights of the different sub-costs is always a problem in a unit selection system. In this case, the syntactic and morphosyntactic sub-cost mentioned in Section 8.1 was generated manually according from data for speaker 1, taking into account differences that were both perceptually the most important and statistically significant. The resulting value (just one weight), was tuned with informal listening tests.

This approach is obviously far from optimal, so we are currently working on an automated method for creating the syntactic and morphosyntactic sub-cost taking into account statistically significant differences from a recorded corpus. This method would automatically learn the characteristics of a given speaker, since, according to our data, there might be important differences from one speaker to another (see, for example, Tables 4 and 7). Moreover, the algorithm could also easily be applied to other languages, since prosodic structure seems to be language-dependent (D'Imperio et al. (2003)).

9 Subjective test of the new intonation model

Testing the influence of new features in a unit selection system can be a difficult task. The relationship between the synthetic contours of the old version and the new features may seem random, given that they are not being considered in the unit search. Therefore, comparing the old version with the new one can be misleading, which makes the process of selecting sentences automatically

for the subjective tests quite difficult.

In this case we applied the following procedure: 1000 sentences extracted from a newspaper (independent from the prosodic corpus) were synthesised with both versions, and 100 of the sentences were chosen to look for large differences in the mean standard error between the intonation contours. Sentences that were too short were discarded. A total of 60 sentences were used to tune the new cost functions. The remaining 40 sentences were used for a pairwise comparison test between the old version (system A) and the new version (system B). Note that the intonation contours of the sentences corresponding to System B may not be the ones resulting from the first 1000 sentences selected, given that the cost functions have been tuned and the chosen contours may therefore be completely different.

The group of listeners for this test was composed of 26 people from the academic world, both with and without experience in intonation modelling. Each listener was presented with a random subset of 20 sentences, in general quite complex (about 25 words on average). The order of the systems was also randomised for each sentence. The listeners were asked to score each pair on a five-point scale, as shown in Table 16. As a verification stage, a random sentence was synthesised with only one of the two systems (and duplicated), in order to exclude any listener not scoring it as equal. Note that there is no specific test for the minor phrasing algorithm. Since our conjecture is that its performance depends not only on the selection of the right places for inserting IBs, but also on the strength of the IBs, we assume that the previous subjective test with long sentences already reflects its behaviour.

1	A version much better than B version
2	A version better than B version
3	Equal
4	B version better than A version
5	B version much better than A version

Table 16
Scoring for the pairwise comparison

Table 17
Results of the subjective test (99% confidence interval)

Rating	Confidence interval	
Much worse	0.008	0.045
Worse	0.113	0.199
Equal	0.119	0.205
Better	0.327	0.441
Much better	0.237	0.343

A one-sample proportion test was conducted to study the influence of the new method. The proportions of **much worse**, **worse** and **equal** were grouped, and the hull hypothesis was that the new approach did not improve the quality of the synthetic intonation. A p-value of 2.87×10^{-14} was obtained, which allows us to reject the null hypothesis. In addition, Table 17 shows the 99% confidence intervals for each choice in the ranking. The quality improvement is clear, since the proportion of **worse** and **much worse** is even lower than the proportion of **much better** alone. On a sentence basis, 75% of them were scored as **better** or **much better** by a majority of the listeners (considering the number of **better** and **much better** for a sentence was larger than the number of **equal**, **worse** and **much worse**), while only 7.5% were scored as **worse** or **much worse**.

10 Conclusions and future work

This paper describes a novel approach to including syntactic and morphosyntactic information in intonation modelling. These features are used to decide not only the appropriate places to insert IBs, but also boundary strength and the emphasis on words in the absence of boundaries.

In contrast with Taylor (2000), we promote the use of AGs with the right syntax/morphosyntax only wherever we find this information to be important. Note that always using the exact grammatical context would not be very efficient, as it would require a larger prosodic corpus, with good coverage of every syntactic/morphosyntactic structure. However, according to the results in Section 5.4, since some grammatical structures do not seem to yield noticeable differences, such a complete prosodic corpus would not even be necessary.

Both POS and syntactic features seem to be important. According to our results, POS labels are related to the emphasis of the AGs in the absence of boundaries, while syntactic information seems to be related not just to the appropriate places to insert IBs, but also to the Δ_{f0} and slope factor of the AGs before them.

Multiple prosodic structures are taken into account as long as there are candidate AGs with and without IBs. This has several advantages. Firstly, the method adds variability to the synthetic intonation. Secondly, in the special case of corpus-based intonation modelling, considering more than one prosodic structure is very important since the database is finite and a restriction to only one alternative could eliminate other prosodic structures that might yield better results. Thirdly, the fact of deciding the appropriate places for inserting IBs according to the available groups in the database helps to mimic the individual characteristics of the speakers.

Another advantage of this algorithm is its simplicity. POS tagging is resolved by means of an ngram-based algorithm that is computationally inexpensive, while syntactic parsing is rule-based. A more complete syntactic parsing would probably yield better results, but the current technology does not seem to be sufficiently advanced. Regarding the unit selection stage, adding new information is very easy (as demonstrated in Section 8); furthermore, it does not slow down the whole synthesis process, which is particularly important in a unit selection system. The only drawback could be the requirement for labelling IG boundaries in the corpus (Hirschberg and Prieto (1996)). However, this task should always be performed on a prosodic corpus, independently of the intonation model, and so it does not imply any extra work.

One of the aims of this work was to study the suitability of the intonation hierarchy model described in Section 2 (sentences, PGs and IGs) when applied to the Galician language. The subjective test in Section 9 seems to confirm the appropriateness both of this model and the new method used for the implementation.

This study was carried out in the special framework of intonation unit selection, but it seems that most of the results in Section 5.4 can be applied to any intonation model. Although integrating the minor phrasing algorithm in the intonation generation stage would probably be more difficult with other models, POS and syntactic information could be used for the strength of the accents in the different boundary conditions.

This work can serve as the basis for a number of future lines of research. As mentioned before, we are currently working on automated methods for generating the target cost function for a given speaker, taking the information described in Section 5 as input. It would also be interesting to consider clustering techniques for differentiating grammatical contexts, since the classification used here was based on linguistic knowledge and a data-based technique could produce better results. Another point would be to let the unit selection algorithm decide the best places for the PG breaks; however, in this case an external model signalling likely places would probably be needed to achieve a satisfactory performance.

Finally, it is important to note that integrating the minor phrasing algorithm into the unit selection stage does not exclude the possibility of using an external algorithm. The unit selection module could easily be modified both to force the insertion of IBs in the places indicated by a previous model, and decide for itself in other cases.

11 Acknowledgements

The work reported here was carried out while the first author was a visiting post-doctoral researcher at the Center for Spoken Language Understanding, with funding from the Xunta de Galicia "Isidro Parga Pondal" research programme and PGIDIT05TIC32202-PR programme, and with support from NSF grant 0205731, "ITR: Prosody Generation for Child Oriented Speech Synthesis" (PI Jan van Santen), and MEC under the project TEC2006-13694-C03-03. Thanks also to Paul Hosom and Raychel Moldover, for their comments and suggestions on the paper.

References

Abney, S. (1992). Prosodic structure, performance structure and phrase structure. In *Proceedings, Speech and Natural Language Workshop*, pages 425–428, San Mateo, CA.

Beckman, M. and Pierrehumbert, J. (1986). Intonational structure in Japanese and English. In *Phonology Yearbook 3*, pages 15–70.

Black, A. and Campbell, N. (1995). Optimising selection of units from speech databases for concatenative synthesis. In *Proceedings of Eurospeech*, volume 1, pages 581–584, Madrid, Spain.

Black, A., Taylor, P., and Caley, R. (1999). *The Festival Speech Synthesis System: system documentation 1.4 for Festival version 1.4.0.*

Campillo, F. and Banga, E. R. (2006). A method for combining intonation modelling and speech unit selection in corpus-based speech synthesis systems. *Speech Communication*, 48:941–956.

Campillo, F., Mishra, T., van Santen, J., and Banga, E. R. (2006a). A method for avoiding f0 discontinuities in a concatenative intonation model. In *Jornadas de Tecnologías del Habla*, pages 167–170, Zaragoza.

Campillo, F., van Santen, J., and Banga, E. R. (2006b). A model for the f0 reset in corpus-based intonation approaches. In *Proceedings of ICSLP*, pages 2362–2365, Pittsburgh.

Campillo, F., van Santen, J., and Banga, E. R. (2008). Combining phrasing and unit selection in intonation modelling. *Electronics Letters*, 44(7):501–503.

D'Imperio, M., Elordieta, G., Frota, S., Prieto, P., and Vigário, M. (2003). Intonational phrasing in Romance: The role of syntactic and prosodic structure. In Sónia Frota, M. V. and Freitas, M. J., editors, *Prosodies*, Berlin.

Escudero, D. (2002). *Modelado estadístico de entonación con funciones de Bézier: Aplicaciones a la conversión texto-voz en Español*. PhD thesis, Universidad de Valladolid, España.

Garrido, J. M. (1996). *Modelling Spanish intonation for text-to-speech applications*. PhD thesis, Facultad de Lletres, Universitat de Barcelona, España.

Grosz, B. and Sidner, C. (1986). Attentions, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Hernáez, I., Navas, E., Murugarren, J., and Etxebarría, B. (2001). Description of the AhoTTS conversion system for the Basque language. In *Proceedings of the 4th ISCA Tutorial & Research Workshop on speech Synthesis*, pages 151–154, Edimburgo.

Hirschberg, J. and Prieto, P. (1996). Training intonational phrasing rules automatically for English and Spanish text-to-speech. *Speech Communication*, 18:281–290.

Hunt, A. and Black, A. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *Proceedings of ICASSP*, volume 1, pages 373–376.

Koehn, P., Abney, S., Hirschberg, J., and Collins, M. (2000). Improving intonational phrasing with syntactic information. In *Proceedings of IEEE International Conference on Acoustics, Speech , and Signal Processing*, volume 3, pages 1289–1290, Istambul.

Ladd, R. (1986). Intonational phrasing: The case for recursive prosodic structure. In *Phonology Yearbook 3*, pages 311–340.

Ladd, R. (1988). Declination 'reset' and the hierarchical organization of utterances. *JASA*, 84:530–544.

Ladd, R. (1996). *Intonational phonology*. Cambridge University Press.

Méndez, F., Campillo, F., Banga, E. R., and Rei, E. F. (2003). Análisis morfológico estadístico en lengua gallega. *Procesamiento del lenguaje natural*, 31:159–166.

Möebius, B. (1999). The Bell Labs German text-to-speech system. *Computer Speech and Language*, 13:319–358.

Navarro, T. (1977). *Manual de pronunciación española*. Consejo Superior de Investigaciones Científicas, Madrid, 19 edition.

Navas, E. (2003). *Modelado prosódico del Euskera Batúa para conversión de texto a habla*. PhD thesis, Universidad del País Vasco.

Ostendorf, M. and Veilleux, N. (1994). A hierarchical stochastic model for automatic prediction of prosodic boundary location. *Computational Linguistics*, 20(1):27–54.

Pierrehumbert, J. and Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. *Intentions in Communication*, pages 271–311.

Prevost, S. and Steedman, M. (1993). Generating contextually appropriate intonation. In *Proceedings of the Sixth Conference of the European Chapter of ACL*, pages 332–340, Utrecht.

Raux, A. and Black, A. (2003). A unit selection approach to f0 modeling and its application to emphasis. In *ASRU*, St Thomas, US Virgin Islands.

Steedman, M. (1990). Structure and intonation in spoken language understanding. In *Meeting of the Association for Computational Linguistics*,

pages 9–16.

Taylor, P. and Black, A. (1998). Assigning phrase breaks from part-of-speech sequences. *Computer Speech and Language*, 12:99–117.

Taylor, P. (2000). Concept-to-speech synthesis by phonological structure matching.

van Santen, J. and Möbius, B. (1999). A quantitative model of f0 generation and alignment. In Botinis, editor, *Intonation Analysis, Modelling and Technology*, chapter 12, pages 269–288. Kluwer Academic Publishers, Netherlands.