Speech Prosody 2016
31 May - 3 Jun 2106, Boston, USA

Using Hierarchical Information Structure for Prosody Prediction in Content-to-Speech Applications

Mónica Domínguez[1], *Mireia Farrús*[1], *Alicia Burga*[1], *Leo Wanner*[2,1]

[1]TALN Group, N-RAS Research Centre
Department of Information and Communication Technologies
Universitat Pompeu Fabra, Barcelona, Spain
[2]Catalan Institute for Research and Advanced Studies (ICREA)
{monica.dominguez|mireia.farrus|alicia.burga|leo.wanner}@upf.edu

Abstract

State-of-the-art prosody modelling in content-to-speech (CTS) applications still uses the same methodology to predict intonation cues as text-to-speech (TTS) applications, namely the analysis of the generated surface sentences with respect to part of speech, syntactic dependency relations and word order. On the other side, several theoretical studies argue that morphology, syntax, and information (or communicative) structure that organizes a given content (semantic or deep-syntactic structure) with respect to the intention of the speaker show a strong correlation with intonation. However, little empirical work based on sufficiently large corpora has been carried out so far to buttress this argumentation. We present empirical evidence for the Information Structure–Prosody correlation using the Wall Street Journal Penn Treebank corpus recorded by native American English speakers. Our experiments reach a prosody prediction accuracy of 80% using the hierarchical information structure from the Meaning-Text Theory, compared to 59% of the baseline.

Index Terms: information structure, thematicity, theme, rheme, prosody, prosodic phrase, ToBI.

1. Introduction

Speech technologies have evolved in a relatively short time span from undertaking mere reading tasks, as, e.g., the well-known MITalk [1], to handling conversations with human interlocutors; cf., e.g., an application in healthcare [2]. This development implies an architectural shift from what is known as *text-to-speech* (TTS) to *content-to-speech* (CTS) [3]. However, the shift has not fully been accomplished yet, especially as far as prosody generation is concerned. In a TTS application, prosody generation is regarded as a final stage module that uses the syntactic structure, part of speech (PoS) tags and word order derived from the input text to predict intonation contours. CTS applications are currently deploying the same end-of-stage prosody modelling as TTS applications, although it has been argued in the literature that: (i) prosody expresses the communicative intention of the speaker [4]; (ii) the communicative intention of the speaker is to a large extent encoded in terms of the *Information Structure* (IS) [5]; (iii) IS is rendered both through syntax and prosody [6]; (iv) in CTS, the IS of a sentence can be derived in a content organization procedure, as done in Natural Language Text Generation (NLTG) [7, 8]. Should this argumentation hold, monotonous and unnatural intonation contours (especially in multiple sentence discourse) inherited from TTS technologies can be avoided (or at least reduced) in CTS applications drawing upon the IS derived automatically using techniques employed in NLTG. Still, there is little work carried out so far to test this argumentation.

In [9], we presented some first qualitative hints that confirm this argumentation. Furthermore, we have shown that the hierarchical IS as put forward in the Meaning-Text Theory [6] correlates with intonation contours to a considerably higher degree than the traditional IS used, e.g., in [10]. In what follows, we provide empirical evidence for the hierarchical IS–Prosody correlation using a selection of the Wall Street Journal corpus recorded by native American English speakers. In addition, we apply this correlation to predict prosody markers using a prosody annotation schema that captures the role of the hierarchical IS to guide the projection of the deep structure of a sentence onto the surface structure.

The remainder of the paper is structured as follows. In the next section, we sketch the theoretical background underlying our work on the IS–Prosody interface and lay out what we consider our theoretical contribution to the problem of prosody modelling. Section 3 describes the set-up of the experiments. Results are presented and discussed in Section 4. And, finally, some conclusions are drawn and future work is briefly summarized in Section 5.

2. Theoretical Background

This section presents the theoretical background relying on the relation between communicative structure and prosody, together with our theoretical contribution towards a more versatile IS – Prosody Interface.

2.1. The Information Structure – Prosody Interface

Information Structure, whose origin goes back to Mathesius [11], and which is also known as *Topic-Focus Articulation* (TFA) [12] in the Prague School [13], and *Communicative Structure* in the Meaning-Text Theory [6], determines the "communicative" segmentation of the meaning of an utterance.

From the perspective of prosody, a number of authors identified a correlation between *theme* (or *topic*, i.e., what the statement is about) and *rheme* (or *focus*, i.e., what the statement says) and characteristic intonational tunes [6, 10, 14–17]. To determine, in their turn, theme/rheme in a statement, it is common to picture the statement as an answer to a hypothetical question, as the following example taken from [10] shows:

(1) *Q: I know what Marcel SOLD to HARRY.*
But what did he GIVE to FRED?

 doi: 10.21437/SpeechProsody.2016-209

A: (Marcel GAVE)$_{Th}$ (a BOOK)$_{Rh}$ (to FRED.)$_{Th}$

In Mel'čuk's Meaning-Text Theory (MTT), the theme/rheme division is covered by the *thematicity*, dimension that is part of a more comprehensive *Communicative Structure* (*CommStr*). Such CommStr is composed of eight distinct dimensions[1] and is first modelled at the semantic level, to be propagated then to the deep-syntactic and surface-syntactic level of the linguistic structure [6].

Compared to the traditional theme/rheme dichotomy, MTT thematicity introduces two key elements that enhance the scope of the theme/rheme span division, namely: (i) the notion of specifier, which sets up the context of the sentence, and (ii) the fact that thematicity is defined over propositions, rather than over sentences. This second element implies that thematicity is *per se* hierarchical: if a proposition is embedded, its thematicity will be embedded as well. Consider an example, taken from our corpus, of the theme(T1)/rheme(R1)/specifier(SP1) distribution over propositions (P1, P2) in the sense of Mel'čuk [6], annotated following the guidelines described in [18]:

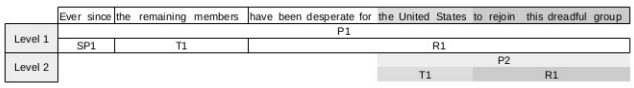
(2) $\{$[*Ever since*]$_{SP1}$, [*the remaining members*]$_{T1}$ [*have been desperate for* $\{$[*the United States*]$_{T1}$[*to rejoin this dreadful group*]$_{R1(P2)}\}_{P2}$]$_{R1}\}_{P1}$

In example (2), the hierarchical thematicity structure is represented as illustrated by Figure 1:

1. at Level 1, P1 contains a theme, rheme and specifier;
2. at Level 2, P2 is embedded into R1(P1) and has a theme and rheme.

That is, spans at Level 2 are hierarchically structured as an embedded thematicity representation.

Figure 1: *Example of hierarchical thematicity*



The relation between information structure and intonation by using ToBI labels has been discussed [19, 20] even before ToBI was agreed as a convention to represent intonation cues [21].

Beckman and Pierrehumbert [19] suggest that the characteristic bitonals for theme and rheme are L*+H and H+L* respectively. Steedman [10] builds upon this assumption and hypothesizes on complete pitch accent/boundary tone patterns, claiming that theme/rheme as in example (3) correlate with intonation patterns as follows:

(3) *Q: I know what Marcel SOLD to HARRY.*
But what did he GIVE to FRED?
A: (Marcel GAVE) (a BOOK) (to FRED.)
L+H LH% H* L L+H* LH%*

Thereupon, a broad characterization of theme and rheme containing rising and falling patterns respectively can be done. However, if we are to deal with complex sentences or even whole texts, we must apply a more versatile model that is able to grasp a multidimensional prosody scheme responding to a communicative correlate. In what follows, we present such a model.

[1]Apart from thematicity, CommStr contains the dimensions of givenness, focalization, perspective, emphasis, presupposedness, unitariness and locutionality.

2.2. Towards a more Versatile IS – Prosody Interface

Domínguez et al. [9] provided some qualitative evidence that a hierarchical thematicity structure that involves a specifier element captures better intonation contours than a flat theme/rheme structure. For further validation of this claim, we have carried out an empirical study of a real monologue discourse corpus read by native speakers. This study confirms [9] and allows us to distill detailed findings concerning the correlation between thematicity at different levels of embeddedness and intonation, namely:

- Main themes (T1$_{L1}$) are characterised by either a rising final tone LH% or a falling final tone HL%.
- Embedded themes (T1$_{L2}$) present two possible intonation contours: L*+H LL% and H* LL%, regardless they are embedded in T1 or R1. If they are embedded in SP1 they show a purely rising contour L*+H LH%.
- Main specifiers (SP1$_{L1}$) tend to include a rising pitch accent L*+H, unless they contain an embedded rheme (R1(SP1)), in which case they are characterized by a flat contour L* LL%.

Table 1 summarizes the most characteristic intonation patterns in main (or Level 1) and embedded (or Level 2) spans (L1 and L2 respectively).

L1		T1		R1		SP1
		L*+H HL%	L*+H LH%	H* LL%		L*+H LL%
L2	T1	L*+H LL%	H* LL%	L*+H LL%	H* LL%	L*+H LH%
	R1	H* LL%		H* LL%		L* LL%

Table 1: *Intonation patterns and their associated hierarchical thematicity aspects*

The hierarchical structure in thematicity (as well as in prosody labeling) is compatible with the labeling of prosodic events at the Prosodic Phrase (PPh) level [17]. Within a PPh, some words are prosodically highlighted, whereas some other words are not so prominent (even if they carry a lexical stress). In this sense, we distinguish between words that are not prosodically marked (*False*) and words that carry a special prosodic label (*True*) at the PPh level. Thereafter, words marked as *False* are annotated as lexically stressed (S) or unstressed (U), whereas words marked as *True* are labeled as pitch accents (PA) or boundary tones (BT). Each one may take one of the possible ToBI labels shown in Table 2.

Prosodic Mark	Prosodic Type	Prosodic Label
True	PA	H*
		L*
		L*+H
	BT	HL%
		LL%
		LH%
False	S	
	U	

Table 2: *Prosodic Annotation Schema*

As our goal in this experiment has been to correlate IS with prosody, the prosody annotation schema does not account for a detailed intonation contour description in the way it is in the case in standard ToBI annotation. Rather, a simplification of prosodic labels is deployed in order to observe prosodic characterisation from a broader perspective. Such hierarchical thematicity–prosody correlation patterns can be readily exploited, e.g., by stochastic transduction algorithms as used

in natural language sentence generation [22], to project deep prosody markers to surface generation of acoustic features.

In the next section, we test this prosody annotation schema in a prediction experiment. The results of the experiment are presented and discussed in Section 4.

3. Experimental Set-up

This section describes the application of the hierarchical thematicity–prosody correlation patterns introduced above. In 3.1, we present the characteristics of the corpus used in our experiment, and in 3.2 we discuss the methodology that draws upon the hierarchical thematicity structure for prosody prediction.

3.1. Data set description

Our data set consists of 109 sentences from the Wall Street Journal corpus [23], selected to cover different sentence typologies regarding length and complexity in syntax and communicative (= information) structure. Thus, the data set contains simple sentences, but also sentences with coordination, subordination, and the combination of both. This varied syntactic composition permits to play with some parameters related to information structure, such as the amount of hierarchical levels of thematicity (up to three in the data set), the presence/absence of each communicative span, their position within the sentence and with respect to each other, as well as their continuity or lack of it.

Native speakers of American English were recruited for a recording session in a professional studio. The session lasted approximately one hour. A total of 15 people (ranging from 20 to 61 years old) were asked to read the corpus, from which 12 were finally included in the experiments, given that three of them exhibited speech disfluences affecting prosody.

Speakers were asked to make a short pause after each sentence, as we have chosen to restrict our experiment to this linguistic unit. Audio files from the recording sessions were segmented into sentences and saved as separate `wav` files. Segmentation and analysis of the audio files were carried out using the Praat software [24]. Interval tiers were automatically created with the division into words. Prosody and hierarchical thematicity structure were annotated manually by expert linguists. Acoustic parameters were extracted for each word interval in order to apply a consistent labelling of prosodic events using not only pitch variations, but also intensity and duration as prosodic markers of saliency at the PPh.

3.2. Methodology

The goal of this experiment has been to assess the prosody prediction capabilities of our hierarchical IS-prosody model as detailed in Section 2.2. We aim to compare our results with a baseline which draws upon traditional textual features, namely, PoS, syntactic dependencies and word order to predict standard ToBI labels. All sentences and speakers from our corpus as detailed in the previous Section 3.1 are included in this experiment.

We implemented the prosody prediction as a supervised classification exercise, with the eight prosodic labels presented in Table 2 as the targeted classes. The goal is to predict the correct prosodic label, given a series of features (including thematicity). Table 3 shows what features are included in this experiment and how many distinct values each feature contains in both the baseline and our IS-prosody model. Each of the syntactic and thematicity features is detailed as follows:

		Distinct Values	
	Feature	**Baseline**	**IS-pros**
Other	Gender	2	
	Word Position	27	
	Total Words	27	
	Number Syllables	11	
Syntax	Function	29	
	PoS	31	
Thematicity	Proposition	—	8
	Embeddedness		3
	L1		6
	L2		4
	L3		3
	Total Spans		8
	Span Position		10
N-gram	ToBI-1	32	9
Class	ToBI	32	9

Table 3: *Features and number of their distinct values used in the prosody pattern prediction experiment.*

- Function: the syntactic function, e.g., subject, direct object.
- PoS: part of speech tag, e.g., noun, verb, adjective, etc.
- Proposition: this feature covers propositions (P2, P3, P4) which are not the main proposition (P1).
- Embeddedness: specifies whether a proposition (other than the main one P1) is embedded or not (subordinated or coordinated); if there is only a main proposition (P1), the instance takes a '0' value.
- L1: first level thematicity, which includes all main spans (T1, R1, SP1 and SP2) as well as split rhemes (R1-1 and R1-2).
- L2: second level thematicity, containing T1, R1 and SP1; if the sentence does not contain L2, the instance takes a '0' value.
- L3: in the third level of hierarchical thematicity, only T1 and R1 occurred in our corpus; if the sentence does not contain L3, the instance takes a '0' value.
- Total Spans: contains the total number of spans in the sentence, including all levels.
- Span Position: the position of the span, distinguishing between initial (A) and final (Z) positions, location of intermediate spans (totalling 7 positions from B to H) and unique spans (U), i.e., sentences that have only one span.
- N-gram: the previous prosodic label taking into account bigrams.
- Class: the actual prosodic label to be predicted changes in the baseline and our model due to the differences in the annotation scheme described in 2.2. The baseline prosodic annotation contains a total of 32 different classes. On the other hand, our proposal implies a simplification of those labels as previously sketched in Table 2.

In order to account for the linear nature of our classification problem, a time series filter has been applied to all features. That is, the preceding prosodic marker (n-1) and all its features are used to predict the intonation pattern n.

For the realization of the classification procedure, we use the machine learning platform Weka [25] and a SimpleCART algorithm deploying a ten-fold cross-validation.

4. Results and Discussion

To compare the performance of the hierarchical IS-based prosody prediction against the baseline, we use accuracy, kappa and relative absolute error. Table 4 shows an accuracy of 80%, kappa of 0.75 and a relative absolute error of 33%. That is, the hierarchical IS-prosody model improves the baseline in 21% accuracy, 23% kappa and 27% error.

	Accuracy	Kappa	Absolute Error
Baseline	59%	0.52	60%
Hierarchical IS	80%	0.75	33%

Table 4: *Prediction results using hierarchical IS vs. baseline*

A closer analysis of the confusion matrices shows significant advances of our proposal compared to the state of the art, especially when predicting boundary tones (BT), which are instrumental for the generation of communicative pauses in long complex sentences containing few punctuation marks.

Figure 2: *Hierarchical IS confusion matrix*

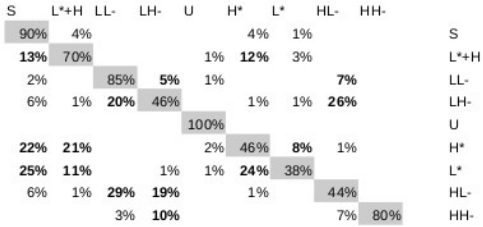


Figure 2 shows the full confusion matrix of intonation pattern prediction based on the hierarchical IS. Correctly classified instances are highlighted in grey. The confusion matrix shows that the majority of errors are made within the same type of prosodic markers (highlighted in bold). In the case of PA in Figure 2, L*+H is mostly confused with H* or S, while it is never confused with a BT.

As far as the classification errors of BT are concerned, they equally occur mostly within the same type of prosodic mark. However, in this case, we must take into account the speaker's choice in making shorter or longer PPh and placing the PA in one word or another. Such differences may lead to one sentence having two different possible prosodic realizations—as the example in Figure 3 shows.

Figure 3: *Example of matching IS–Prosody interface*

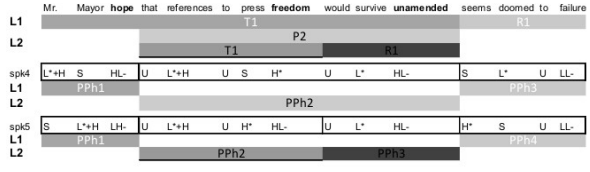


Figure 3 sketches the existence of common prosodic markers that are consistently located (as further explained in [26], which set broad PPh divisions coinciding with different hierarchical levels of the thematicity structure. In this figure, PPh2 of speaker 4 ('spk4') matches the whole Proposition 2 (P2), while speaker 5 ('spk5') splits P2 into two PPh, i.e., PPh2 coincides with T1(P2) and PPh3 with R1(P2). Therefore, we can observe in this example the importance of annotating thematicity over propositions and including embedded thematicity.

5. Conclusions and Future Work

Our experiments provide empirical evidence for the capacity of the prediction of prosody markers for monologue discourse using a hierarchical IS (and, more specifically, a hierarchical thematicity structure). A prediction model based on the hierarchical IS improves state-of-the-art approaches in terms of accuracy in that it achieves 80%), kappa (0.75) and relative absolute error (33%. Especially boundary tones (BT) in complex long texts are predicted well, which renders a more natural speech generation by means of prosody.

Our work also offers a theoretical contribution to intonation patterns related to IS in that it enlarges the scope of state-of-the-art theories by providing empirical evidence on its correlation with prosody. It introduces the idea that prosodic markers can be reliably predicted at a deep level communicative structure. Moreover, our model foresees a hierarchical scaffolding of prosodic events focused on the transition from Prosodic Phrases into Prosodic Words which needs to be explored in future work.

All in all, we believe to have shown that a communicative approach conveyed by hierarchical IS is instrumental for ensuring advanced prosody implementation. Such an improvement is key for CTS technologies as it leads to a versatility of the model, which allows for the adaptation to expressive speech requirements in human-machine interaction. Further qualitative experiments and perception tests will be carried out to measure the overall improvement in a real Content-to-Speech (CTS) application.

6. Acknowledgements

This work is part of a project that has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under the Grant Agreement number H2020-RIA-645012. The second author is partially funded by a grant from the Spanish Ministry of Economy and Competitivity in the framework of the *Juan de la Cierva* fellowship program.

7. References

[1] J. N. Holmes, "From text to speech: The MITalk system," pp. 359–362, 1987.

[2] G. Bierner, "TraumaTalk: content-to-speech generation for decision support at point of care." *Proceedings / AMIA Annual Symposium. AMIA Symposium*, pp. 698–702, 1998.

[3] S. Pan, "Prosody Modeling in Concept-to-Speech Generation," Ph.D. dissertation, 2002.

[4] P. H. Grice, "Further Notes on Logic and Conversation," in *Studies in the Way of Words*, 1989, pp. 41–57.

[5] M. Steedman, "The surface-compositional semantics of english intonation," *Language*, vol. 90, pp. 2–57, 2013.

[6] I. Mel'čuk, *Communicative Organization in Natural Language: The semantic-communicative structure of sentences*, 2001.

[7] L. Wanner, B. Bohnet, and M. Giereth, "Deriving the Communicative Structure in Applied NLG," in *Proceedings of the 9th European Workshop on Natural Language Generation at the Biannual Meeting of the European Chapter of the Association for Computational Linguistics*, 2003, pp. 100–104.

[8] N. Bouayad-Agha, G. Casamayor, S. Mille, and L. Wanner, "Perspective-Oriented Generation of Football Match Summaries: Old Tasks, New Challenges," *ACM Transactions on Speech and Language Processing*, vol. 9, no. 2, 2012.

[9] M. Domínguez, M. Farrús, A. Burga, and L. Wanner, "The Information StructureProsody Language Interface Revisited," in *Proceedings of the 7th International Conference on Speech Prosody (SP2014)*, Dublin, Ireland, 2014, pp. 539–543.

[10] M. Steedman, "Information structure and the syntax-phonology interface," *Linguistic inquiry*, 2000.

[11] V. Mathesius, "Zur Satzperspektive im modernen Englisch," in *Archiv für das Studium der neueren Sprachen und Literaturen, 155*, 1929, pp. 202–210.

[12] P. Sgall, "Functional Sentence Perspective in a generative description of language," *Prague Studies in Mathematical Linguistics*, vol. 2, pp. 203–225, 1967.

[13] F. Daneš, "One instance of Prague School methodology: Functional analysis of utterance and text," *Garvin*, pp. 132–141, 1970.

[14] K. Lambrecht, *Information structure and sentence form: Topic, focus and the mental representations of discourse referents*. Cambridge: Cambridge University Press, 1994.

[15] E. Hajiĉova, B. Partee, and P. Sgall, *Topic-Focus Articulation, Tripartite Structures, and Semantic Content*. Kluwer Academic Publishers, Dordrecht, 1998.

[16] N. Erteschik-Shir, *Information Structure: The Syntax-Discourse Interface*. Oxford University Press, Oxford, 2007.

[17] E. O. Selkirk, *Phonology and Syntax: The relation between sound and structure*, 1984.

[18] B. Bohnet, A. Burga, and L. Wanner, "Towards the Annotation of Penn TreeBank with Information Structure," in *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, Nagoya, Japan, 2013, pp. 1250–1256.

[19] M. Beckman and J. Pierrehumbert, "Intonational structure in Japanese and English," *Phonology Yearbook*, vol. 3, pp. 255–310, 1986.

[20] M. E. Beckman, J. Hirschberg, and S. Shattuck-Hufnagel, "The original ToBI system and the evolution of the ToBI framework," in *Prosodic Typology – The Phonology of Intonation and Phrasing*, S. A. Jun, Ed., 2005.

[21] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "TOBI: A Standard for Labeling English Prosody," in *2nd Internatioanl Conference on Spoken Language Processing (ICSLP 92)*, no. October, 1992, pp. 867–870.

[22] M. Ballesteros, B. Bohnet, S. Mille, and L. Wanner, "Data-driven sentence generation with non-isomorphic trees," in *Proceedings of the Annual Conference of the North American Association for Computational Linguistics – Human Language Technologies (NAACL – HLT)*, 2015.

[23] E. Charniak and E. al., "BLLIP 1987-89 WSJ Corpus Release 1 LDC2000T43." Philadelphia, 2000. [Online]. Available: https://www.cis.upenn.edu/ treebank/

[24] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," 2015. [Online]. Available: retrieved 21 february 2015 from http://www.praat.org/

[25] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.

[26] M. Domínguez, F. Mireia, and L. Wanner, "Combining Acoustic and Linguistic Features in Phrase-Oriented Prosody Prediction," in *International Conference on Speech Prosody (SP2016), Boston, USA*, 2016.