



A Thematicity-based Prosody Enrichment Tool for CTS

Mónica Domínguez¹, Mireia Farrús¹, Leo Wanner^{2,1}

¹Universitat Pompeu Fabra, Barcelona, Spain

²Catalan Institute for Research and Advanced Studies (ICREA), Barcelona, Spain

monica.dominguez@upf.edu, mireia.farrus@upf.edu, leo.wanner@upf.edu

Abstract

This paper presents a demonstration of a stochastic prosody tool for enrichment of synthesized speech using SSML prosody tags applied over hierarchical thematicity spans in the context of a CTS application. The motivation for using hierarchical thematicity is exemplified, together with the capabilities of the module to generate a variety of SSML prosody tags within a controlled range of values depending on the input thematicity label.

Index Terms: prosody, information structure, thematicity, CTS, TTS, SSML.

1. Background

Proposals to consider extended *Information* (or *Communicative*) *Structure* in Natural Language Text Generation for the definition of the transition between conceptual and syntactic structures on the one side [1] and advances on the empirical study of the so-called *Information Structure–Prosody Interface* [2, 3] on the other side, put the development of communicatively-oriented prosody derivation for concept-to-speech (CTS) applications back on the research agenda.

Traditional Information Structure (IS) interpretations assume a bipartite division of a sentence into “what is being talked about” (aka ‘theme’, ‘topic’ or ‘given information’) and “what is being said” (aka ‘rheme’ or ‘new information’). Several works in the context of speech synthesis (see, e.g., [2, 3, 4]) draw upon this flat binary theme–rheme division to establish a deterministic correlation between theme–rheme and rising–falling intonation patterns in text-to-speech (TTS) applications. However, such methodology has several drawbacks: (i) it fails to describe longer sentences with complex syntactic structures; (ii) it ignores other prosodic elements, such as rhythm and intensity (that also relate to information and prosody structure [5, 6]); and (iii) it presupposes a fully deterministic mapping between intonation labels and acoustic parameters. These drawbacks make this methodology also insufficient when it comes to find remedies for monotonous prosody in synthesized speech.

Mel’čuk [7] proposes an alternative interpretation of the thematicity. Firstly, he introduces, apart from theme and rheme, a third element, namely ‘specifier’ (which sets the utterance’s context), and, secondly, he defines thematicity over *propositions* instead of sentences – which implies that thematicity can be embedded (i.e., a theme or a rheme can contain another theme/rheme/specifier division) and, thus, be hierarchical. In what follows, we present a stochastic hierarchical thematicity-based tool for enrichment of prosody using SSML prosody tags [8]. The tool, which is presented as part of a whole CTS system, has the following characteristics:

- it generates a varied range of SSML prosody tags: three prosodic attributes (pitch, loudness and speech rate) are used to introduce a range of prosodic variations in each

sentence;

- it is communicatively oriented: the selection of spans is based upon hierarchical thematicity structure (annotated following the guidelines outlined in [9]);
- it uses relative attribute values from the empirical analysis of a corpus of read speech: these values correspond to mean values of z-scores for each thematicity span converted to percentages. Values are randomly varied over a range of $\pm 5\%$;
- it automatically converts thematicity spans into SSML prosody contour tags.

The demonstration material (code and sample sentences annotated with thematicity) is available under a GNU General Public License v.3 from our repository¹.

2. From thematicity to SSML prosody control tags

The prosody tool presented in this demonstration automatically transforms a text file annotated with thematicity into an SSML format output for prosody enrichment of synthesized speech. For the demonstration, three optional attributes, overall pitch, speech rate and volume with relative values are chosen for conversion to the SSML *prosody tag*.

2.1. The hierarchical thematicity

In order to show the contrast between the default synthesis (raw text without any SSML tag) and our prosody tool output, a number of sentences, which cover a typical range of thematicity partitions at different levels, are selected to exemplify:

1. Level 1 thematicity:
 - (a) Theme–Rheme (L1-T1, L1-R1): [*The luxury auto maker*]T1 [*last year sold 1,214 cars in the U.S.*]R1
 - (b) Long Theme (more than 8 words, L1-T1>8): [*For its employees to sign up for the options*]T1, [*a college also must approve the plan.*]R1
 - (c) Theme–Rheme–Specifier (L1-T1, L1-R1, L1-SP1): [*Men who have played hard all their lives*]T1 [*aren’t about to change their habits*]R1, [*he says.*]SP1
2. Level 2 embeddedness:
 - (a) Embedded thematicity in theme and rheme spans (L2-T1, L2-R1): [*For its employees*]T1(T1) [*to sign up for the options*]R1(T1), [*a college*]T1(R1) [*also must approve the plan.*]R1(R1)
 - (b) Embedded proposition in theme span (L2-P2): [*Men { who have played hard all their lives }*]P2]T1 [*aren’t about to change their habits*]R1, [*he says.*]SP1

¹<http://github.com/TalnUPF/thematicitytoSSML>

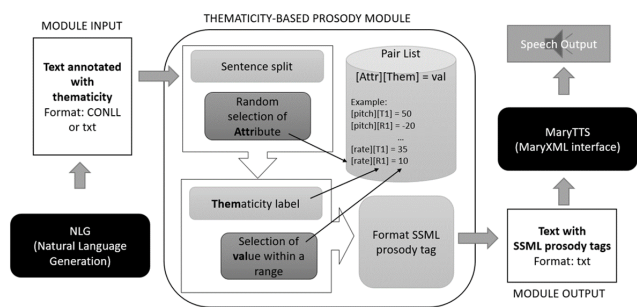


Figure 1: *Thematicity-based prosody enrichment pipeline*

Despite we show the capabilities of our module with a limited set of examples, the tool is scalable to the whole possible range of hierarchical thematicity spans.

2.2. The prosody enrichment tool

The pipeline of our tool within a CTS application, which uses MaryTTS, is displayed in Figure 1. It can take as input an annotated file with thematicity in *txt* or *CONLL* format (the later being generated by the NLG module) to begin the processing. The core part of the tool is a pair list of attributes and thematicity spans with their corresponding values. This parameter–span list of values is built using a data-driven approach from a corpus of read speech, annotated with thematicity (see [10] for further reference). In the first place, the module splits the text into sentences and assigns each sentence a random prosody attribute (either pitch, volume or speech rate in this demonstration). Then, a query to the pair list that contains percentages for each thematicity span² is performed to assign the value of the selected attribute. So as to attain more variability, the final value of the prosody attribute is randomly assigned within a range of $\pm 5\%$ from the pair list entry value for each span. A result SSML output for our example sentences is shown below:

1. `<prosody pitch="+30%">The luxury auto maker</prosody><prosody pitch="+10%">last year sold 1,214 cars in the U.S..</prosody>`
2. `<prosody rate="+15%">For its employees</prosody><prosody rate="+5%">to sign up for the options,</prosody><prosody rate="+10%">a college</prosody><prosody rate="-10%">also must approve the plan.</prosody>`
3. `<prosody volume="+25%">Men who have played hard all their lives</prosody><prosody volume="-5%">aren't about to change their habits,</prosody><prosody volume="-10%">he says.</prosody>`

3. Conclusions and future work

Derivation of communicative prosodic contours is instrumental for CTS applications, but also useful for TTS applications. In this work, previous empirical findings are used to develop a stochastic tool for generation of SSML prosody control tags, based on hierarchical thematicity. Thematicity-based prosody enrichment has several advantages: communicative segments, which are instrumental for reading comprehension [11], are used; three prosodic elements are generated that contribute

²Percentages are computed from the mean z-score values for each thematicity span in our corpus.

to prosodic variability and, thus, avoid monotony in speech synthesis; and an empirical derivation of attribute values for prosody control tags guarantees the appropriateness of the generated output in terms of thematicity spans.

The implementation of the presented prosody enrichment tool is scalable in various ways: its output can be tested in any TTS application, given that it complies with the SSML format procedures, and other SSML control tags can be included, given that empirical data are extracted to derive them.

4. Acknowledgements

This work is part of the KRISTINA project, which has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under the Grant Agreement number H2020-RIA-645012. It has been also partly supported by the Spanish Ministry of Economy and Competitiveness under the María de Maeztu Units of Excellence Programme (MDM-2015-0502). The second author is partially funded by the *Ramón y Cajal* program.

5. References

- [1] 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.
- [2] M. Steedman, "Information structure and the syntax-phonology interface," in *Linguistic inquiry*. Cambridge, Massachusetts: The MIT Press, vol. 31, no. 4, pp. 649–689.
- [3] I. Kruijff-Korbayová, S. Ericsson, K. J. Rodríguez, and E. Karagjosova, "Producing Contextually Appropriate Intonation in an Information-State Based Dialogue System," in *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2003, pp. 227–234.
- [4] M. Haji-Abdolhosseini, "A Constraint-Based Approach to Information Structure and Prosody Correspondence," in *Michigan State University, East Lansing*, S. Muller, Ed. CSLI Publications, 2003, pp. 143–162.
- [5] S. Calhoun, "The centrality of metrical structure in signalling information structure: A probabilistic perspective," *Language*, vol. 1, no. 86, pp. 1–42.
- [6] C. Tseng, "Intensity in relation to prosody organization," in *International Symposium on Chinese Spoken Language Processing*. IEEE, 2004, pp. 217–220.
- [7] I. A. Mel'čuk, *Communicative Organization in Natural Language: The semantic-communicative structure of sentences*. Amsterdam, Philadelphia: Benjamins, 2001.
- [8] P. Taylor and A. Isard, "SSML: A Markup Language for Speech Synthesis," *Speech Communication*, vol. 21, no. 1-2, pp. 123–133, 1997.
- [9] 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.
- [10] M. Domínguez, M. Farrús, A. Burga, and L. Wanner, "Using hierarchical information structure for prosody prediction in content-to-speech applications," in *Proceedings of the 8th International Conference on Speech Prosody*, Boston, USA, 2016, pp. 1019–1023.
- [11] D. Meurers, R. Ziai, N. Ott, and J. Kopp, "Evaluating answers to reading comprehension questions in context: Results for german and the role of information structure," in *Proceedings of the TextInfer 2011 Workshop on Textual Entailment*, ser. TIWTE '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 1–9.