9th International Conference on Speech Prosody 2018
13-16 June 2018, Poznań, Poland

Thematicity-based Prosody Enrichment for Text-to-Speech Applications

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

[1]Universitat Pompeu Fabra, Barcelona, Spain
[2]Catalan Institue for Research and Advanced Studies (ICREA), Barcelona, Spain
monica.dominguez@upf.edu, mireia.farrus@upf.edu, leo.wanner@upf.edu

Abstract

Theoretical studies on the information structure–prosody interface argue that the content packaged in terms of theme and rheme correlates with the intonation of the corresponding sentence as regards to rising and falling patterns (L*+H LH% and H* LL% respectively). When such a correspondence is used to derive prosody in text-to-speech applications, it is often the case that ToBI labels are statically mapped to acoustic parameters. Such an approach is insufficient to solve the problem of monotonous synthetic voices for two reasons: it is repetitive with respect to prosody enrichment, and a binary flat theme-rheme representation does not serve to describe properly long complex sentences. In this paper, we introduce a methodology for a more versatile thematicity-based prosody enrichment based on: (i) a hierarchical tripartite thematicity model as proposed in the Meaning–Text Theory, and (ii) a corpus-based approach for the automatic extraction of acoustic parameters (fundamental frequency, breaks and speech rate) that are mapped to a varied range of prosody control tags of the synthesized speech. Such a prosody enrichment has shown to provide higher results in a perception test when implemented in a TTS system.
Keywords: prosody, information structure, theme, rheme, TTS, SSML.

1. Introduction

The consolidation of technologies that enable human users to have a conversation with a virtual assistant suggests to advance further and tackle the complex subtleties of human spoken communication. One of these subtleties is natural prosody, which plays a role in communication in that it expresses the communicative intention and the emotional state of the speaker. This role has been largely left aside in speech technologies, partially due to its intrinsic difficulty to be formalized in a language model that can be used in computational tasks. Such a model should ideally involve a definition of how communicative intention is translated into a formal structure and how this structure is then projected onto prosody generation. In linguistics, the *Information Structure* (IS) is considered to accommodate for the communication intention of the speaker [1, 2]. Existing studies focused so far under the heading of *Information Structure–Prosody Interface* mainly on the correlation between prosody and one of the dimensions of the IS, namely *theme–rheme* [3, 4].

Several works in the context of speech synthesis (see, e.g., [5, 6, 7]) draw upon a theme (i.e., "what is being talked about")–rheme (i.e., "what is being said") division. Such a binary flat division is used to establish a one-to-one correspondence between theme–rheme spans and rising–falling intonation patterns in terms of ToBI labels. Some text-to-speech (TTS) applications generate prosodic contours based upon this correspondence. However, this methodology has several drawbacks: (i) it fails to describe adequately longer sentences with complex syntactic and communicative structures (thus, we see that rheme consists *de facto* of the whole factual statement); (ii) it ignores other prosodic elements, such as rhythm and intensity (that also relate to information and prosody structure [8, 9]); and (iii) it presupposes a one-to-one mapping between ToBI labels and associated acoustic parameters (primarily, variations of fundamental frequency). These drawbacks make this type of approach insufficient when it comes to find remedies for monotonous prosody in synthesized speech.

In what follows, we propose a methodology to derive thematicity–based prosody enrichment using a formal representation of thematicity described by Mel'čuk [10] as one dimension of communicative (aka information) structure[1]. Mel'čuk's representation of thematicity includes theme, rheme and a third span, the *specifier*, which sets the context of the sentence. Moreover, thematicity is defined over propositions what allows embeddedness within spans. Thus, thematicity is hierarchical in nature.

Our methodology is based on corpus analysis of the IS–prosody interface using the Mel'čukian representation of hierarchical thematicity. With respect to prosody representation, an automatic computation of mean prosodic parameters is carried out. The analysis of the distribution of three normalized acoustic parameters, namely, fundamental frequency (F0), intensity, and speech rate over hierarchical thematicity spans in our corpus is used to derive a varied range of prosodic modifications. The methodology is tested in an implementation setting using SSML prosody tags in a TTS application, which is evaluated with a perception test.

The remainder of paper is structured as follows. The next section describes the motivation and background of the presented approach and summarizes the works that are related to ours. Section 3 presents our methodological approach, outlining how synthetic prosody is enriched by thematicity information departing from corpus analysis. Section 4 presents an implementation of such a thematicity-based prosody enrichment of the synthetic speech. The evaluation is introduced in Section 5. Finally, conclusions and future work are presented in Section 6.

2. Motivation and Background

Different linguistic schools have long stated that Information Structure (IS), and, in particular, the dichotomy referred to as *theme–rheme* [11], *given–new* [12], or *topic–focus* [13] is related to intonation.[2] Moreover, it has been claimed that when prosody reflects thematicity structure, comprehension of the

[1]Communicative structure includes eight different dimensions, namely: thematicity, givenness, focalization, perspective, emphasis, presupposedness, unitariness and locutionality.

[2]In our work, we use the first denotation, i.e., *theme–rheme* or *thematicity*.

10.21437/SpeechProsody.2018-119

message is positively affected [14]. Recent empirical studies in different languages provide evidence of such an improvement; cf., e.g., [15] for German and [16] for Catalan. Therefore, there is reason to assume that a conversational application that considers, on the one hand, content packaging in terms of thematicity and, on the other hand, the relation between thematicity and prosody will lead to more natural conversation settings.

Some attempts to include thematicity in TTS applications were made in the past. Consider, for instance, Steedman's work [5] on the correlation of theme and rheme with rising and falling intonation patterns based on a question–answer setting. An implementation on the grounds of Steedman's characterization was tested in concept-to-speech (CTS) applications [6, 7, 17]. However, a derivation of thematicity that is fully dependent on a question-answer setting is limited to a specific conversational environment and, thus, fails to deal with the problem of monotony of TTS applications.

In contrast to IS models that propose a partition of sentences into a theme and a rheme, Mel'čuk [10] argues in the context of the Meaning–Text Theory for a tripartite hierarchical division ('theme', 'rheme', and 'specifier' – the element which sets the utterance's context) within propositions that further permits embeddedness of communicative spans; consider (1) for illustration of hierarchical thematicity (annotated following the guidelines established in [18]) of the sentence *Ever since, the remaining members have been desperate for the United States to rejoin this dreadful group*. A total of five partitions are identified, including three spans at level 1, a specifier (SP1), theme (T1) and rheme (R1), and two embedded spans at level 2 in the rheme, a theme (T1(R1)) and a rheme (R1(R1))[3].

(1) [*Ever since,*]SP1 [*the remaining members*]T1 [*have been desperate* [*for the United States*]T1(R1) [*to rejoin this dreadful group.*]R1(R1)]R1

A hierarchical thematicity structure of this kind has been shown to correlate better with ToBI labels than binary flat thematicity [19, 20]. However, such a correlation still does not solve the problem of a one–to-one mapping between a specific intonation label (e.g., H*) to a static acoustic parameter (e.g., an increase of 50% in fundamental frequency). In what follows, we propose a methodology to derive a more varied range of prosodic cues based on the analysis of the available corpus of read speech annotated with hierarchical thematicity and prosody; see [21]. Furthermore, we optimize a preliminary realization (see [22]) that converts hierarchical thematicity labels into SSML prosody tags.

3. Methodology

According to Xu's review of the methodologies in the field of speech prosody [23], theoretical studies on the syntax-pragmatics-prosody interfaces use an "analysis by introspection" as methodological approach. He argues that this approach is imprecise as the assignment of prosody "by intuition" is inevitably unreliable.

The present paper proposes an approach that tests a formal representation of IS in speech synthesis experiments. This work envisages the study of the information structure–prosody interface from a methodological perspective based on an empirical approach in two experimental setups: corpus analysis of human speech and speech synthesis experiments. The proposed methodology aims to achieve the following goals:

- to allow scalability to other communicative dimensions, registers and languages;
- to analyze the thematicity–prosody correspondence in human speech using a corpus-driven approach;
- to explore the advantages and limitations of a corpus-driven thematicity-based prosody enrichment in a TTS application.

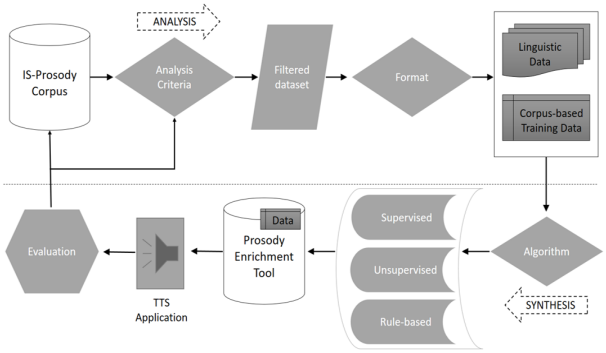

Figure 1: *From analysis to synthesis for prosody generation.*

Such a methodology addresses two main research issues in this field: (i) the lack of empirical analysis of the Information Structure–prosody correspondence; and (ii) testing of the integration of the Information Structure–prosody interface in computational settings. Figure 1 summarizes the proposed working flow.Prosody enrichment in TTS applications consists in applying specific modifications to the default synthesized speech specifying a certain parametric variation for a particular word or group of words. Several XML-based markup languages are used to encode these modifications. The *Speech Synthesis Markup Language* (SSML) [24] is the most well-known convention. SSML establishes a control sequence defined in terms of attributes[4] that is mapped onto the acoustic signal and is, thus, parametric in nature. These attributes can be absolute (e.g., '+50 Hz' for increasing a specific amount of hertz (Hz) in F0) and relative (e.g., '-20%' for decreasing a percentage in F0) within the range $\pm 100\%$.

Our working corpus contains absolute and relative prosodic parameters that were automatically extracted using an extension of Praat for feature annotation [25], which was developed based on the Praat software [26]. These parameters include mean and standard deviations of F0, intensity and speech rate as well as duration at different segments coinciding with thematicity spans; see [21] for further details on corpus compilation. Normalized z-scores relative to the whole sentence are used in our implementation. Z-scores represent positive and negative deviations relative to the sentence mean value of each parameter. Such deviations are used to analyze whether different parametric distributions occur between hierarchical thematicity spans using the average values across speakers. On the other hand, the normalized prosodic parameters across hierarchical thematicity are converted to relative SSML attribute values. Consequently, the acoustic data extracted from the corpus analysis stage is fed

[3] As more than one thematicity span may exist within the same proposition, abbreviations include a number (e.g., 'SP1') that indicates the number of occurrences at each level (e.g., 'SP2' would be the second specifier in a specific thematicity level).

[4] The SSML *prosody tags* allow control of six optional attributes: overall pitch, pitch contour, pitch range, speech rate, duration, and volume.

into the synthesis stage for the thematicity-based prosody enrichment.

4. Implementation

The proposed methodology has been tested in an implementation setting using the corpus in [21] and the prosody enrichment tool presented in [22]. The working dataset contains a total of 109 sentences extracted from the Wall Street Journal and read by twelve native speakers of American English, that is, a total of 1,308 speech samples are analyzed to infer the acoustic data. As acoustic parameters, relative values of F0, intensity and speech rate are used in the implementation. In what follows, we present results from the analysis and synthesis pipelines in Section 4.1 and 4.2 respectively.

4.1. Prosodic parameters distribution in hierarchical thematicity

The analysis of the distribution across speakers of average z-scores for intensity, F0 and speech rate (abbreviated as z_int, z_F0 and z_sr, respectively) is presented in two figures grouped in level 1 (L1) (cf., Table 1) and level 2 (L2) thematicity (cf., Table 2).

	z_int	z_F0	z_sr
T1	**0.16**	**0.47**	**0.95**
R1	-0.04	-0.17	**0.30**
SP1	-0.07	**0.15**	**1.76**

Table 1: *Distribution of prosodic parameters in L1 thematicity.*

Table 1 shows that there are distinct distributions of parameters between theme (T1), rheme (R1) and specifier (SP1) spans within the L1 thematicity spans. T1 displays positive deviations (highlighted in bold) in all z_int, z_F0 and z_sr parameters; R1 has negative z_int and z_F0 and positive (but lower than T1) z_sr; and SP1 is characterized by negative z_int (like R1), positive (but lower than T1) z_F0 and positive (higher than T1) z_sr.

	z_int	z_F0	z_sr
T1(T1)	**0.27**	**0.57**	**0.23**
T1(R1)	**0.12**	**0.21**	**1.30**
T1(SP1)	-0.13	-0.07	**2.25**
R1(T1)	**0.27**	**0.80**	**0.89**
R1(R1)	-0.23	-0.41	**0.48**
R1(SP1)	-0.21	-0.21	**1.11**
SP1(SP1)	**0.25**	-0.06	**3.13**

Table 2: *Distribution of prosodic parameters in L2 thematicity.*

Table 2 shows the average normalized parameters for embedded thematicity spans in three sections for embedded themes, rhemes and specifiers respectively. The L1 spans where L2 thematicity is embedded are represented in rows. Thus, L2 specifiers are embedded only in a L1 specifier, i.e., SP1(SP1), in our corpus, even though it is possible to find L2 specifiers initially in any other span.

The analysis of average z_scores across speakers shows a distinct distribution pattern of prosodic parameters extracted from each thematicity span. If we compare these results to previous studies on the IS–prosody interface that related themes with rising and rhemes with falling F0 contours, the distribution of z_F0 values across our corpus also supports the argument that themes involve a higher z_F0. This argument is further extended to the concept of embeddedness: spans that are embedded in themes (both T1(T1) and R1(T1)) have higher values of z_F0 than other embedded spans. Differences in the distribution of values observed in thematicity elements are exploited in Section 4.2 in terms of a corpus-driven approach for the derivation of a thematicity-based generation of SSML prosody control tags.

4.2. Thematicity–based prosody enrichment

The open source software MaryTTS[5] [27] was used for the implementation of thematicity-based prosody enrichment. The default synthesized speech output has been enriched using MaryXML prosody specifications[6], which follow the SSML recommendation[7].

The SSML *prosody tags* allow control of six optional attributes (overall pitch, pitch contour, pitch range, speech rate, duration and volume). These attributes can be modified independently or in combination. For our implementation, overall pitch and speech rate were chosen individually and in combination. Absolute (e.g., '+50Hz' for increasing a specific amount of hertz (Hz) in F0) and relative values can be used to applied the modification. We use relative values within a range of $\pm$100% because the values are extracted from the corpus for the normalized deviation with respect to the sentence of each prosodic parameter (F0, intensity and speech rate). An example of a SSML prosody tag for modification of two prosodic elements is presented below:

Example (1)

<prosody rate="-10%" pitch="+20%">*text to be modified* </prosody>

Moreover, the SSML *boundary tag* that controls the introduction of pauses at a specific location was also used after each thematicity span. The duration of the break is specified in milliseconds (ms). An example of SSML boundary tag is introduced below:

Example (2)

Text before the break <boundary duration="100"/>*text after the break.*

Thematicity	z_int	'volume'	z_F0	'pitch'	z_sr	'rate'
T1	0.15	15	0.50	50	0.95	35
R1	-0.05	-5	-0.20	-20	0.30	10
SP1	-0.10	-10	0.15	15	1.00	35
R1(T1)	0.30	30	0.50	25	0.25	15
R1(R1)	-0.25	-25	-0.40	-15	0.50	25

Table 3: *Conversion from z-scores to SSML attribute values.*

A set of sentences is selected for the implementation and assessment of the thematicity-based prosody enrichment in MaryTTS using a statistical voice. The distribution of prosodic parameters in the whole corpus (described in the previous section) is mapped onto the values that the attribute of the SSML tag will take. Table 3 presents the characterization of thematicity for the spans selected for the implementation[8]. Some values (especially those for speech rate) were scaled to an appropriate percentage, because previous testing using SSML prosody tags showed an undesired distortion when a very high attribute value was inserted. For instance, if an increase of 95% in speech rate

[5] Available at `http://mary.dfki.de/`

[6] `http://mary.dfki.de/documentation/maryxml/index.html`

[7] `https://www.w3.org/TR/speech-synthesis/`

[8] Figures are round up to the closest half tenth.

was specified, the resulting speech would sound far too quickly with an associated F0 increase, and consequently unnatural and, sometimes, unintelligible.

5. Evaluation

For the prosody enrichment evaluation, a selection of the following six annotated sentences has been made:

1. [*The luxury auto maker*]T1 [*last year sold 1,214 cars in the U.S.*]R1 .
2. [[*For its employees*]T1(T1) [*to sign up for the options*]R1(T1)]T1, [[*a college*]T1(R1) [*also must approve the plan*]R1(R1)]R1 .
3. [*Mr. Mayors hope* {[*that*]SP1(P2) [*references to press freedom*]T1(P2) [*would survive unamended*]R1(P2)}P2]T1 [*seems doomed to failure*]R1 .
4. [[*The researchers*]T1(SP1) [*said*]R1(SP1)]SP1 [*they*]T1 [*have isolated a plant gene* {[*that*]T1(P2) [*prevents the production of pollen*]R1(P2)}P2]R1 .
5. [{[*When*]SP1(P2) [*he*]T1(P2) [*sent letters* {[*offering 1,250 retired major leaguers*]T1(P4) [*the chance of another season*]R1(P4)}P4]R1(P2)}P2]T1 , [{[*730*]T1(P3) [*responded*]R1(P3)}P3]R1 .
6. [*Men* {[*who*]T1(P2) [*have played hard all their lives*]R1(P2)}P2]T1 [*arent about to change their habits*]R1, [[*he*]T1(SP1) [*says*]]R1(SP1) .

Each sentence was synthesized using five different prosody contours (one by default and four modifications using SSML tags)[9]. Thus, thirty synthesized speech samples have been evaluated in the perception test taken by thirty participants, i.e., a total amount of 900 answers have been considered in the evaluation. Prosody modifications included F0, speech rate and breaks after thematicity spans in isolation and in combination. Intensity was excluded from the evaluation as it did not yield the expected effect according to the authors' criteria. The test consisted in two parts: (i) a *Mean Opinion Score* (MOS) test rating at a Likert scale from 1 to 5 rating level of expressiveness[10]; and (ii) a pairwise comparison where the most expressive sentence is chosen between the pairs baseline–F0, combination–break and F0–combination.

	baseline	F0	speech rate	break	combination
sent 1	2.03	**2.67**	2.30	**2.53**	2.43
sent 2	2.97	3.07	3.10	2.83	3.00
sent 3	2.93	2.83	*2.40*	*2.37*	2.90
sent 4	2.83	2.67	2.90	2.70	2.60
sent 5	3.03	2.87	2.73	3.00	3.06
sent 6	2.40	2.67	**3.17**	2.73	2.17
Average	2.70	2.79	2.77	2.69	2.61

Table 4: *MOS test results.*

Results from the MOS test are displayed in Table 4. T-tests were performed to observe the level of significance with a confidence of 95%. Bold figures represent statistically significant improvements over the baseline (t-test, $p < 0.05$). Figures in italic represent those results in which the lower value is also statistically significant (t-test, $p < 0.05$) with respect to the baseline. In overall, F0 and speech rate modifications are rated higher than the default voice. Looking at each specific sentence, thematicity-based modifications tend to be rated higher. Sentences 1 and 6 show statistically significant differences with respect to the baseline for F0 and break modifications (in sentence 1) and for speech rate in sentence 6. Sentence 3 shows statistically significant worse results for modifications concerning speech rate and break compared to the baseline.

In most of the sentences, F0, break and speech rate enrichments are perceived as more expressive than the baseline. All in all, this buttresses the idea that communicative spans are important for generation of expressive synthesized speech, and that a variety of prosodic cues contributes to signaling the IS–prosody correspondence. However, as four out of six sentences do not show statistically significant differences compared to the baseline, results should be interpreted with caution. One plausible explanation to this is the fact that the default prosody sometimes contains unnatural prosody, in some cases artifacts, that cannot be overridden by SSML tags. Consequently, the perception of human listeners might be negatively affected in both cases.

6. Conclusions and future work

Given the relevant role of the Information Structure–Prosody Interface in human communication, it seems reasonable that next generation virtual assistants face new challenges in adopting communicatively-oriented models. Current speech technologies have been oblivious to advances in theoretical fields studying this correlation, basically due to the lack of a formal representation of the communicative (or information) structure and limited capabilities of prosody enrichment standards to achieve variability in implementation settings.

The present study provides a methodology for a more versatile integration of the IS–prosody interface in TTS applications. Such methodology contributes in several aspects to the state-of-the-art: (i) a formal description of hierarchical thematicity is used; (ii) the prosodic representation is automatically computed and tested in a TTS application; and (iii) a derivation of prosody enrichment is done empirically from a corpus of read speech. All in all, this study pivots the transition from theoretical work on the IS–prosody interface to the integration of a corpus-driven prosody enrichment to achieve more expressive synthesized speech.

One limitation of the current study is that it only considers relative acoustic parameters over rather large text segments. Key aspects of prosody modeling, like F0 contour generation in terms of prominence and phrasing remain to be looked into in future work.

7. Acknowledgements

This work is part of the KRISTINA project, which has received funding from the *European Unions 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 Maria de Maeztu Units of Excellence Programme (MDM-2015-0502). The second author is partially funded by the Spanish Ministry of Economy and Competitivity through the *Ramón y Cajal* program.

[9] Samples will be made available via GitHub upon acceptance

[10] Defined as *effectively conveying meaning*.

8. References

[1] W. L. Chafe, *Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. Chicago and London: University of Chicago Press, 1994.

[2] M. Strube and U. Hahn, "Functional centering: Grounding referential coherence in information structure," *Computational Linguistics*, vol. 25, no. 3, pp. 309–344, September 1999.

[3] L. Mereu, Ed., *Information Structure and its Interfaces*. Berlin, Boston: De Gruyer Mouton, 2009.

[4] D. Büring, *Intonation and Meaning*. Oxford: Oxford University Press, 2016.

[5] M. Steedman, "Information structure and the syntax-phonology interface," *Linguistic inquiry*, vol. 31, no. 4, pp. 649–689, Fall 2000.

[6] I. Kruijff-Korbayová, S. Ericsson, K. J. Rodríguez, and E. Karagrjosova, "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.

[7] M. Haji-Abdolhosseini and S. Müller, "Constraint-based approach to information structure and prosody correspondence," in *Proceedings of the 10th International Conference on Head-Driven Phrase Structure Grammar*. CSLI Publications, 2003, pp. 143–162.

[8] S. Calhoun, "The centrality of metrical structure in signalling information structure: A probabilistic perspective," *Language*, vol. 1, no. 86, pp. 1–42, 2010.

[9] C. Tseng, "Intensity in relation to prosody organization," in *International Symposium on Chinese Spoken Language Processing*. IEEE, 2004, pp. 217–220.

[10] I. A. Mel'čuk, *Communicative Organization in Natural Language: The semantic-communicative structure of sentences*. Amsterdam, Philadephia: Benjamins, 2001.

[11] M. Halliday, "Notes on Transitivity and Theme in English, Parts 1-3," *Journal of Linguistics*, vol. 3, no. 1, pp. 37–81, 1967.

[12] R. Schwarzschild, "Givenness, avoidf and other constraints on the placement of accent," *Natural Language Semantics*, vol. 7, no. 1, pp. 141–177, 1999.

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

[14] H. H. Clark and S. E. Haviland, "Comprehension and the given-new contract," *Discourse production and comprehension. Discourse processes: Advances in research and theory*, vol. 1, pp. 1–40, 1977.

[15] 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.

[16] M. Vanrell, I. Mascaró, F. Torres-Tamarit, and P. Prieto, "Intonation as an encoder of speaker certainty: Information and confirmation yes-no questions in catalan," *Language and Speech*, vol. 56, no. 2, pp. 163–190, 2013.

[17] F. Kügler, B. Smolibocki, and M. Stede, "Evaluation of information structure in speech synthesis : The case of product recommender systems perception," in *ITG Conference on Speech Communication, IEEE*, 2012, pp. 26–29.

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

[20] ——, "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.

[21] M. Domínguez, M. Farrús, and L. Wanner, "Compilation of corpora for the study of the information structure–prosody interface," in *Accepted for publication at the 11th edition of the Language Resources and Evaluation Conference, LREC 2018*, Miyazaki, Japan, 2018.

[22] ——, "A thematicity-based prosody enrichment tool for cts," in *Proceedings of INTERSPEECH; show and tell demonstrations*, Stockholm, Sweden, 2017, pp. 3421–2.

[23] Y. Xu, "Speech prosody: A methodological review," *Journal of Speech Sciences*, vol. 1, no. 1, pp. 85–115, 2011.

[24] P. Taylor and A. Isard, "Ssml: A speech synthesis markup language," *Speech Communication*, vol. 21, no. 1-2, pp. 123–133, February 1997.

[25] M. Domínguez, I. Latorre, M. Farrús, J. Codina, and L. Wanner, "Praat on the web: An upgrade of praat for semi-automatic speech annotation," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, Osaka, Japan, 2016, pp. 218–222.

[26] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program], http://www.praat.org/, version 6.0.14, downloaded January 2016."

[27] M. Schröder and J. Trouvain, "The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching," *International Journal of Speech Technology*, vol. 6, no. 4, pp. 365–377, 2003. [Online]. Available: http://mary.dfki.de