INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Using Prosody to Classify Discourse Relations

Janine Kleinhans[1], *Mireia Farrús*[1], *Agustín Gravano*[2,3],
Juan Manuel Pérez[2,3], *Catherine Lai*[4], *Leo Wanner*[1,5]

[1]TALN Research Group, DTIC, Universitat Pompeu Fabra, Barcelona, Spain
[2]Departamento de Computación, FCEyN, Universidad de Buenos Aires, Argentina
[3]Instituto de Investigación en Ciencias de la Computación, CONICET-UBA, Buenos Aires, Argentina
[4]School of Informatics, University of Edinburgh, Edinburgh, UK
[5]Catalan Institute for Research and Advanced Studies, Barcelona, Spain

janine.kleinhans@upf.edu, mireia.farrus@upf.edu, agustin.gravano@gmail.com,
jmperez@dc.uba.ar, clai@inf.ed.ac.uk, leo.wanner@upf.edu

Abstract

This work aims to explore the correlation between the discourse structure of a spoken monologue and its prosody by predicting discourse relations from different prosodic attributes. For this purpose, a corpus of semi-spontaneous monologues in English has been automatically annotated according to the Rhetorical Structure Theory, which models coherence in text via rhetorical relations. From corresponding audio files, prosodic features such as pitch, intensity, and speech rate have been extracted from different contexts of a relation. Supervised classification tasks using Support Vector Machines have been performed to find relationships between prosodic features and rhetorical relations. Preliminary results show that intensity combined with other features extracted from intra- and intersegmental environments is the feature with the highest predictability for a discourse relation. The prediction of rhetorical relations from prosodic features and their combinations is straightforwardly applicable to several tasks such as speech understanding or generation. Moreover, the knowledge of how rhetorical relations should be marked in terms of prosody will serve as a basis to improve speech synthesis applications and make voices sound more natural and expressive.

Index Terms: prosody, discourse structure, RST, speech synthesis, support vector machines

1. Introduction

The quality (and thus, to a major degree the expressiveness) of synthesized speech is judged, amongst other things, by its similarity to the human voice. Human-like speech synthesis should be able to account for different emotions, speaking styles, and also for different discourse relations in a spoken text. Consider for instance the following sentence: (1) *I think* (2) *if the weather is nice* (3) *we can eat outside*. The first segment signals that the upcoming ones represent an attribution while the second segment imposes a condition on the third one. This study attempts to show that those two discourse relations, *attribution* and *condition*, differ in terms of prosody, i.e. pitch (F0), intensity, and speech rate, in human speech. Currently, synthesizers do not take into account this prosodic diversity between different text structures. However, we assume that its implementation in a speech synthesis system would enhance its performance in terms of naturalness and expressiveness.

Several studies have examined the relationship between prosody and discourse markers as explicit indicators of discourse structure [1, 2] or even to find direct relationships between prosody and specific discourse relations [3]. In the current work, we aim to find such relationships by using a larger set of discourse relations and prosodic features retrieved from different contexts within a phrase. We use Rhetorical Structure Theory (RST) [4], which describes the organizational structure of text by defining relations that hold between two spans of text and thus helps to explain coherence in an utterance. RST discourse relations are used to annotate semi-spontaneous monologues. F0, intensity, and speech rate are extracted from corresponding audio files, and supervised classification experiments using Support Vector Machines (SVM) are performed to analyze to what extent the relations can be predicted from their corresponding prosodic features.

The structure of this paper unfolds as follows: in Section 2 we present some related work on investigating the relationship between prosody and discourse structure. In Section 3 we present the experimental setup and the obtained results are shown in Section 4. The discussion of the results and the conclusions are drawn in Sections 5 and 6, respectively.

2. Prosody and Discourse Structure

Several studies have investigated the connection between prosodic information and discourse structure by looking at certain cue words, which are often referred to as *discourse markers* (DMs) as they are regarded as explicit indicators of discourse structure [1]. The marker *now*, for instance, may signal the return to a previous topic, while *but* can indicate contrasting information. DMs are considered to be prosodically independent words separated from their surrounding context by pauses and/or intonation breaks [5]. They relate the segment they introduce to a prior segment [6] and are cohesion building devices in conversations [7]. DMs and their prosodic realization have been investigated in the context of the disambiguation between their sentential (1) and discourse (2) use, e.g. *Now* (2) *now* (1) *that we have all been welcomed here it's time to get on with the business of the conference* [2]. That study shows that, when occurring initially in a larger phrase, the first *now* functioning as a DM is usually deaccented or has a L* accent. By contrast, the second *now* is a temporal marker bearing a L* or a complex accent.

Studying such markers and their prosodic realization can help provide straightforward insights about the relationship between prosody and discourse structure, but looking only at DMs is problematic in two ways. First, it is highly controversial which terms belong to the class of DMs. While some re-

http://dx.doi.org/10.21437/Interspeech.2017-710

searchers include e.g. interjections (*oh*), pause markers (*well*), or phrases (*y'know*) [8, 9], others exclude them from the list of DMs [6]. Second, while the taxonomy of cue phrases can often be directly mapped onto discourse structure [10], it has been shown that discourse relations often exist without the presence of DMs [11]. Moreover, it can be seen in the example above that some DMs belong to certain classes such as adverbs; thus, they can easily be confused as an indicator of discourse structure in places where they are not.

According to RST [4], discourse relations usually consist of two different text spans, a nucleus (N), that conveys the main message and is the more central part of the message, and a satellite (S), which is less central, potentially incomprehensible without N, and could be substituted by another text span without changing the meaning of the message. The two text spans are also called *Elementary Discourse Units* (EDUs). To construct an RST relation, two EDUs of text are related together and can then be connected with another EDU, forming a new relation. This process is repeated iteratively until a whole discourse tree is built from the text. Different relation sets have been proposed. For this study, we use part of the original RST relation set by [4].

In an investigation by [3] on the identification of discourse relations with prosodic features (and not lexical cues), the five relations *contrast, elaboration, summary, question*, and *cause* have been retrieved from texts in the ICSI corpus [12], which contains 75 transcripts of native and non-native speakers participating in meetings. Binary and multi-class classifications were performed with supervised and unsupervised methods, using 75 features including F0, F0 variance, intensity, speech rate, pause, and duration. Their multi-class classification revealed an accuracy which was only slightly better than the baseline.

The current study differs from the previously mentioned work in several ways. 1. While only two of the relations are the same, our relation set differs from the set that is used by [3] regarding the number and types of relations, as we use *elaboration, background, attribution, condition, contrast, explanation* and *enablement*. 2. The authors extracted features from the target segments only, while we additionally observe the difference between two segments to consider prosodic changes between two consecutive units. 3. Presumably, the authors included relations of all levels in the discourse tree, while we only regard relations that are not parent of another relation themselves. This way we ensure that prosodic features only belong to the current segment and are not part of other units. 4. Our work consists of a large-scale study using a corpus of 1564 transcripts that have been annotated automatically with an RST parser.

3. Experimental Setup

3.1. Data

The annotated corpus of 1564 TED (Technology, Entertainment, Design) talks[1] is a set of conference talks, which have been held worldwide in more than 110 languages under the slogan *Ideas worth spreading*. It includes a broad variety of topics, ranging from technology and design to science, culture, or academic topics. Each talk lasts about 15 minutes. The corpus comprises 1156 speakers of English with different accents. Transcripts as well as audio and video files are available on TEDs website. The transcripts we used for our corpus were retrieved from talks held before 2014, and the transcriptions were created by volunteers and include punctuation and paragraph breaks. The data set includes 151820 sentences and 20953 paragraphs, with an average of seven sentences per paragraph [13]. The talks are considered semi-spontaneous speech as they are prepared in advance. Since they are well structured and presented in an engaging way, we assume to be able to recognize a large set of discourse relations as well as prosodic properties that reflect those relations.

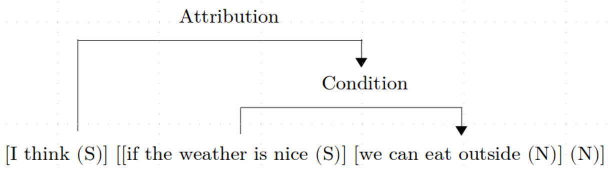
Attribution
Condition
[I think (S)] [[if the weather is nice (S)] [we can eat outside (N)] (N)]

Figure 1: *Example of a sentence with two discourse relations Attribution and Condition. For our work, we only regard relations where the two EDUs are not parents of other segments.*

3.2. RST Annotations

Prior to parsing all transcripts to obtain RST annotations, meta-information about file, page, line number, or surrounding noise like laughter or applause were removed from the files. Taking into account the complexity and considerable time to create manual annotations, it was decided to use an automatic RST parser. Within three of the most recent and commonly known parsers [14, 15, 16], we chose Surdeanu's FastNLPParser[2] to obtain a complete RST discourse tree for each transcript. In his own comparison of its output with manual annotations, an f-measure of 0.55 was rendered, using the same relation set of 18 labels as [17] and [14].

The number of relations used in our analysis was limited to seven (see Table 1) based on several criteria. We discarded eleven relations of the original relation set from the analysis with less than 2000 instances. Furthermore, as the number of instances is highly imbalanced across the whole corpus (e.g. 66710 *elaborations* vs. 2380 *explanations*), we limited the number of instances to the smallest common number (2380). This way, multi-class classification could be achieved and improved with equal numbers of instances [18]. We only considered Nucleus-Satellite relations, not multi-nuclear relations (e.g. *joint*) where no relation holds between the different nuclei, except from the relation *contrast*, as contrasting information is likely to be reflected prosodically in spoken language.

Relations were only considered if they consisted of two text spans that are the leaves of the discourse tree, i.e. they are neighbors and not the parent of other text spans. Hence, in the sentence in Figure 1, the *condition* relation would be included, while *attribution* being parent of another relation would not. This decision was taken as we aim to find direct correlates between a segment and its own prosodic features, not features that belong to its daughter-segments.

3.3. Prosodic Features

Prosodic features based on RST annotations and audio files were extracted using Praat. For each EDU, F0, intensity, and speech rate, together with aggregate statistics like mean, standard deviation, maximum, minimum, medium, slope, and range were obtained. The features were normalized in order to eliminate the differences between interlocutors. The values were

[1] http://www.ted.com

[2] https://github.com/clulab/processors

Table 1: *Discourse relations with examples partially extracted from the TED corpus.*

Relation	**Description with examples**
Elaboration	S gives additional detail about a situation presented in N. *I want to thank all of you for the many nice comments (N) about what I had to say the other night (S).*
Background	S gives important information for the reader to comprehend N. *It started in 1908 (N), when the Wright brothers flew in Paris (S).*
Attribution	S is an Attribution when it is used for reporting direct or indirect speech or to express feelings, thoughts, or hopes, that are stated in N. *I thought to myself (S), what in the world could be wrong (N)?*
Condition	The realization of an action or a situation present in N depends on the situation in S. *If you have invested money with managers (S), don't ever complain about quarterly CEO management (N).*
Contrast	A multinuclear relation consisting of two nuclei which present two situations that differ in one or more respects and can be compared by the reader. *It sounds like a little things to (N), but I looked in the rearview mirror and it just hit me (N).*
Explanation	S explains a situation presented in N. *My staff was extremely upset (N), because they had already written a story about my speech (S).*
Enablement	S increases the readers ability to perform the action presented in N. *They do not have permission (S) to do what needs to be done (N).*

converted to semitones that stand in relation to the mean F0 value of a speaker, represented in Hz. We also retrieved difference features which capture F0 and intensity change between the last word of the first EDU and the first word of the second EDU, as well as between the first and the last word of a current EDU. We are interested in detecting how prosody changes between two EDUs and across each EDU separately in a rhetorical relation. While pause durations between two EDUs were included, we did not look at any contextual features in surrounding words.

3.4. Classification

We performed experiments with several settings: using (1) F0 only (2) intensity only (3) speech rate only, (4) F0 and speech rate, (5) intensity and speech rate, and (6) F0, intensity and speech rate. Furthermore, these features were retrieved from three different environments: (i) *intrasegmental features* that occur within an EDU, which can be absolute features such *mean.normF0* or difference features like *slope.normI* (ii) *intersegmental features* that capture the difference between two EDUs, such as *ndiff.mean.normF0* describing the difference in normalized F0 between the last word of the current and the first word of the next EDU, and (iii) a combination of intra- and intersegmental features (Figure 2). We used Accuracy, Precision, and Recall as evaluation metrics. The data was standardized and randomized before doing cross-validation with 10 folds. We provide results based on supervised classification methods using the SVM algorithm LibSVM C-SVC approach with a RBF kernel, setting the cost parameter to $C = 1$ and $\gamma = 0$ in Weka [19].

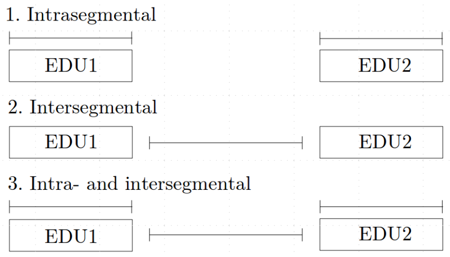

Figure 2: *Intrasegmental features reflect absolute and difference features within a unit, intersegmental features reflect differences between two units, and 3. is a combination of both.*

4. Results

Table 2 shows the results for multi-class classification with seven target classes and different prosodic features and their combinations extracted from three different segmental environments. When considering only one feature at a time, intensity is the feature that provides the highest accuracy of 29.76% and a baseline of 14.29%. This is only the case if intensity features are taken from both intra- and intersegmental environments. We tested whether adding features to intensity improved its performance, and found that this is the case when combined with speech rate (Mann-Whitney $U = 19.5, n_1 = n_2 = 10, P < 0.05$) while it is not significantly different when combined with F0. If intensity is measured only between two units, accuracy is only marginally better than chance (16.14%), and if measured from within the segments, it is not considerably higher (19.98%). Surprisingly, considering F0 alone gives in all three conditions almost the same performance as speech rate, or in the second setting even less.

The confusion matrix for all three features combined from intra- and intersegmental environments shows that *attribution*, *condition*, *enablement*, and *explanation* are discourse relations that are the easiest to classify and exhibit f-measures ranging from 0.32 to 0.41, while *elaboration* has only been correctly identified in 375 of 2380 cases with an f-measure of 0.18. In fact, *elaboration* has been classified in more cases (483) as an *attribution* than as an *elaboration*. Similarly, *background* has been correctly identified in only 558 cases, but was classified 528 times as *attribution*. Few relations were confused with *enablement* and *condition*.

5. Discussion

Our goal was to predict discourse relations from a set of features involving F0, intensity, and speech rate. We found that some relations could be classified with a higher accuracy than others. Using all prosodic features from intra- and intersegmental environments, *attribution*, *condition*, *enablement*, and *explanation* were less often confused with other relations than *elaboration* or *background*. One possible reason for this is that less confusable relations exhibit certain prosodic patterns that are more specific to them than for other relations. An *attribution*, e.g., is often introduced with *he said* or *I think*, and an *explanation* with *because*, and *condition* is marked with *if*, while *elaborations* do not show a characteristic set of cue words.

Another possible explanation is that the picture might be blurred by the diverging recognition rates of the different rela-

Table 2: *LibSVM Classification results for prosodic characteristics retrieved from (1) intrasegmental (2) intersegmental, and (3) intra- and intersegmental environments.*

Features	Accuracy (%)	F1	Precision	Recall
1. Intrasegmental				
F0	21.04	0.18	0.2	0.2
I	19.98	0.19	0.2	0.2
SR	20.01	0.19	0.2	0.2
F0+I	22.07	0.2	0.21	0.3
F0+SR	22.20	0.2	0.21	0.22
I+SR	21.64	0.21	0.21	0.22
F0+I+SR	22.77	0.2	0.21	0.22
2. Intersegmental				
F0	16.84	0.15	0.17	0.17
I	16.14	0.12	0.17	0.16
SR	20.01	0.19	0.2	0.2
F0+I	15.03	0.3	0.3	0.3
F0+SR	17.08	0.1	0.16	0.15
I+SR	17.02	0.13	0.18	0.17
F0+I+SR	15.03	0.1	0.16	0.15
3. Intra- and intersegmental				
F0	22.41	0.21	0.22	0.22
I	**29.76**	0.29	0.3	0.3
SR	20.01	0.19	0.2	0.2
F0+I	**30.24**	0.3	0.3	0.3
F0+SR	23.42	0.3	0.31	0.31
I+SR	**31.12**	0.31	0.31	0.31
F0+I+SR	**30.84**	0.31	0.31	0.31

Table 3: *Confusion matrix for F0, intensity and SR retrieved from intra- and intersegmental environments.*

Classified as →	a	b	c	d	e	f	g	F1
a = attribution	**946**	335	258	223	138	276	204	**0.32**
b = background	528	**558**	257	264	188	315	270	0.25
c = condition	472	281	**713**	263	187	248	215	**0.32**
d = contrast	373	252	256	**601**	197	447	254	0.27
e = enablement	319	225	165	207	**901**	331	232	**0.41**
f = explanation	390	139	149	275	189	**1043**	195	**0.39**
g = elaboration	483	326	269	303	262	362	**375**	0.18

tions by the parser. An *elaboration* is more difficult to be recognized due to the lack of typical DMs, such that the false positives introduce considerable noise into the prosodic pattern of *elaboration*. On the other side, relations such as *attribution* or *explanation* are often introduced by specific cue words, which is why the high number of true positive classified relations might be the result of correct identification by the parser and thus their prosodic patterns are more consistent than for *elaborations*.

We furthermore discovered that features extracted from intra- or intersegmental environments alone do not provide as good results as when they are combined. This suggests that for best classification, we need information both from within and between segments. This makes sense if we think of two relations with equal intrasegmental features but different intersegmental features. For example, two relations can share the same mean F0 value per unit, but they may be distinct with regards to the difference feature between two units, if the difference is measured between the last word of a current EDU and the first word of the next EDU. In this case, absolute features alone would be insufficient for classification. Likewise, two relations can differ in mean features within a unit, but share the same feature values between two units. In both cases, considering inter- and intrasegmental features together would be justified.

It has been shown that some features result in higher accuracy than others. Intensity alone is the best feature for classification and when combined with speech rate, F0, or both, results are higher. However, it should be noted that the audio files were retrieved from TED talks where people usually speak to a large audience and speakers might play more with intensity than with other features to stress something. In the future, another experiment should be made with a corpus of conversations or monologues that take place under moderate intensity conditions to compare the results. This would be especially important for the implementation of results in speech synthesis applications. It should also be mentioned that high accuracy values above the baseline are difficult to achieve as numerous factors are involved in prosodic variation. For instance, speech rate and differences in F0 and intensity have been shown to mark paragraph boundaries [13], pitch accent, duration, and pitch range correlate with negative emotions in speech [20], rising intonation signals uncertainty and surprise [21], and prosody also correlates with information structure elements such as theme, rheme, and specifier [22].

6. Conclusions and Future Work

In this work, we were able to find correlations between discourse structure and prosody. Discourse relations, particularly *attribution*, *explanation*, and *enablement* can best be predicted when looking at prosodic features from both within and between two segments. One limitation of this work is the automatic annotation with a discourse parser. Automatic parsing was necessary due to the large number of transcripts. However, a revision of randomly chosen annotations revealed that the parser performed well in the presence of certain cue words which indicated a discourse relation, but often set incorrect discourse relation tags or EDU segmentation boundaries. For this reason, in a next step, the RST Discourse Treebank[3] containing 385 Wall Street Journal news articles with manually created RST annotations will be used as a gold standard from which audio files will be recorded. Prosodic features will be retrieved from these files, and together with RST annotations the same classification tasks will be performed as in the current work to compare the validity of the automatic discourse parser. We will also include Neural Networks for classification. A next step would then be to do prosodic analyses of single features per relation. Finally, an implementation of results in a Text-to-Speech system by creating tags of discourse relations and a following perception study will round up the work.

7. 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 645012. The second author is partially funded by the Spanish Ministry of Economy, Industry and Competitiveness through the Ramón y Cajal program. The third and fourth authors are partially funded by ANPCYT PICT 2014-1561, and the Air Force Office of Scientific Research, Air Force Material Command, USAF under Award No. FA9550-15-1-0055.

[3] https://catalog.ldc.upenn.edu/LDC2002T07

8. References

[1] J. Hirschberg and D. Litman, "Now let's talk about now: Identifying cue phrases intonationally," in *Proceedings of the 25th annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 1987, pp. 163–171.

[2] J. Hirschberg, D. Litman, J. B. Pierrehumbert, and G. Ward, "Intonation and the intentional structure of discourse," *Proceedings of the 10th international joint conference on Artificial intelligence-Volume 2*, vol. 1, pp. 636–639, 1987.

[3] G. Murray, S. Renals, and M. Taboada, "Prosodic correlates of rhetorical relations," *Proceedings of the HLT-NAACL 2006 Workshop on Analyzing Conversations in Text and Speech*, no. June, pp. 1–7, 2006.

[4] W. C. Mann and S. A. Thompson, "Rhetorical structure theory: Toward a functional theory of text organization," pp. 243–281, 1988.

[5] A. Zwicky, "Clitics and Particles," *Language*, vol. 61, no. 2, pp. 283–305, 1985.

[6] B. Fraser, "What are discourse markers?" *Journal of Pragmatics*, vol. 31, pp. 931–952, 1999.

[7] M. M. Louwerse and H. Mitchell, "Towards a taxonomy of a set of discourse markers in dialog: a theoretical and computational linguistic account," *Discourse Processes*, vol. 35, no. 1, pp. 199–239, 2003.

[8] D. Schiffrin, *Discourse Markers*. Cambridge University Press, 1988.

[9] C. C. Fries, *The structure of English: An introduction to the construction of English sentences*. Longman, 1973.

[10] A. Knott and R. Dale, "Using linguistic phenomena to motivate a set of coherence relations," *Discourse Processes*, vol. 18, pp. 35–62, 1994.

[11] M. Taboada, "Discourse markers as signals (or not) of rhetorical relations," *Journal of Pragmatics*, vol. 38, pp. 567–592, 2006.

[12] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke *et al.*, "The icsi meeting corpus. acoustics, speech, and signal processing, 2003," in *Proceedings.(ICASSP03). 2003 IEEE International Conference on*, vol. 1, 2003.

[13] M. Farrús, C. Lai, and J. D. Moore, "Paragraph-based Prosodic Cues for Speech Synthesis Applications," in *Proceedings of the 8th International Conference on Speech Prosody (SP 2016)*, 2016.

[14] V. W. Feng and G. Hirst, "A Linear-Time Bottom-Up Discourse Parser with Constraints and Post-Editing," *Acl*, pp. 511–521, 2014.

[15] M. Heilman and K. Sagae, "Fast Rhetorical Structure Theory Discourse Parsing," *arXiv preprint arXiv:1505.02425*, 2015. [Online]. Available: http://arxiv.org/abs/1505.02425

[16] M. Surdeanu, T. Hicks, and M. A. Valenzuela-Escárcega, "Two Practical Rhetorical Structure Theory Parsers," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 2015, pp. 1–5.

[17] H. Hernault, H. Prendinger, D. a. DuVerle, and M. Ishizuka, "HILDA: A discourse parser using Support Vector Machine classification," *Dialogue & Discourse*, vol. 1, no. 3, pp. 1–33, 2010.

[18] Y. Liu, N. V. Chawla, M. P. Harper, E. Shriberg, and A. Stolcke, "A study in machine learning from imbalanced data for sentence boundary detection in speech," *Computer Speech & Language*, vol. 20, no. 4, pp. 469–494, 2006.

[19] I. Witten, E. Frank, M. Hall, and C. Pal, *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*, ser. The Morgan Kaufmann Series in Data Management Systems, Fourth Edition. Elsevier Science, 2016.

[20] A. Gravano, S. Benus, J. Hirschberg, S. Mitchell, and I. Vovsha, "Classification of Discourse Functions of Affirmative Words in Spoken Dialogue," *Interspeech*, pp. 1613–1616, 2007.

[21] C. Lai, "What do you mean, you're uncertain?: The interpretation of cue words and rising intonation in dialogue," *Interspeech*, pp. 1–4, 2010.

[22] 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.