Using Prosody to Classify Discourse Relations

Janine Kleinhans\(^1\), Mireia Farrús\(^2\), 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

\textit{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) \textit{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-
Transcripts as well as audio and video files are available on TED’s website. The transcripts we used for our corpus were re-
charged against the meaning of the message. The two text spans are also called Elementary Discourse Units (EDUs). To
struct 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 elabora-
tion, background, attribution, condition, contrast, explanation and enablement. 2. The authors extracted features from the tar-
get segments only, while we additionally observe the difference between two segments to consider prosodic changes between
two consecutive units. 3. Presumably, the authors included rela-
tions 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, Entertain-
ment, Design) talks1 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 top-
ics, ranging from technology and design to science, culture, or academic topics. Each talk lasts about 15 minutes. The cor-
pus comprises 1156 speakers of English with different accents. Transcripts as well as audio and video files are available on TED’s website. The transcripts we used for our corpus were re-
trieved from talks held before 2014, and the transcriptions were created by volunteers and include punctuation and paragraph

Figure 1: Example of a sentence with two discourse relations Attribution and Condition. For our work, we only regard rela-
tions 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. Tak-
ing 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 FastNLPParser2 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 im-
poved with equal numbers of instances [18]. We only con-
sidered 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, stan-
dard deviation, maximum, minimum, medium, slope, and range
were obtained. The features were normalized in order to elim-
nate the differences between interlocutors. The values were

1http://www.ted.com

2https://github.com/clulab/processors
Table 1: Discourse relations with examples partially extracted from the TED corpus.

<table>
<thead>
<tr>
<th>Relation</th>
<th>Description with examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elaboration</td>
<td>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).</td>
</tr>
<tr>
<td>Background</td>
<td>S gives important information for the reader to comprehend N. It started in 1908 (N), when the Wright brothers flew in Paris (S).</td>
</tr>
<tr>
<td>Attribution</td>
<td>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)?</td>
</tr>
<tr>
<td>Condition</td>
<td>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).</td>
</tr>
<tr>
<td>Contrast</td>
<td>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).</td>
</tr>
<tr>
<td>Explanation</td>
<td>S explains a situation presented in N. My staff was extremely upset (N), because they had already written a story about my speech (S).</td>
</tr>
<tr>
<td>Enablement</td>
<td>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).</td>
</tr>
</tbody>
</table>

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 358 cases, but was classified 528 times as attribution. Few relations were confused with enablement and condition.

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

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) intra- (2) intersegmental, and (3) intra- and intersegmental environments.

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

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

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

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, rhyme, 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 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.

3https://catalog.ldc.upenn.edu/LDC2002T07
8. References


