Towards Automatic Extraction of Prosodic Patterns for Speech Synthesis

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

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

Abstract

This paper deals with the adaptation of AuToBI annotation for speech synthesis purposes. AuToBI is a tool that automatically determines and classifies the standard ToBI labels for American English. AuToBI annotation is performed word-by-word. However, for speech synthesis applications that use various layers of linguistic annotation (syntax, semantic information and prosody structures) and, in particular, for the detection of the correlation between the information structure and prosody, a labeling of intonation patterns at the intonational phrase level is essential. We present a rule-based procedure for initial AuToBI annotation and its adaptation a phrase-based annotation, avoiding thus a post-processing stage of the extracted labels. To validate our proposal, the outcome of the procedure is compared with manual annotation and with patterns prognosticated by information structure–prosody correlation argued for by main stream theories.

Index Terms: prosody, annotation, ToBI, AuToBI, thematicity, theme, rheme, speech synthesis.

1. Introduction

Prosodic features, such as rhythm, intonation, and stress are instrumental for the naturalness of speech and play thus an important role in the context of the "semantics–syntax–intonation" language interface in all speech-oriented Natural Language Processing (NLP) applications, especially in speech synthesis. The importance of prosody in NLP led linguists and speech technologists establish annotation standards for labeling prosodic events. One of them is ToBI (Tone and Break Indices) [1], a widely used convention thanks to its easy adaptation as a markup language for open-source speech synthesizers such as Festival [2].

The decade following the introduction of the ToBI convention, speech technology experienced an increasing interest in automated prosody labeling, mainly to avoid the time-consuming procedure of manual annotation.[1] As a consequence, a rather exhaustive number of works focused their interest on the automatic detection and annotation of prosodic events in speech; see, among others, [4, 5, 6, 7, 8]. One of the most well-known of them is AuToBI [9] for automatically detecting and classifying ToBI labels for American English. However, AuToBI labels prosody word by word, while what is required for NLP applications is segmentation at the phrase level that is based on a simplified converging model. Word-by-word segmentation is far too detailed to facilitate the connection between the other layers of annotation of the abovementioned "semantics–syntax–intonation" interface, especially when dealing with information structure [10].

In this paper, we discuss a rule-based procedure for the adaptation of AuToBI's word-by-word output to the needs of expressive speech synthesis, with the goal to be able to automatically establish a link between the information structure of an utterance and its prosody structure. The procedure groups AuToBi's word labels into intonational phrases (IPs) and proposes a single intonation pattern for each IP on the grounds of a set of criteria based upon the more detailed word-by-word labeling. Note, however, that we do not aim to address the general problem of phonologic/acoustic recognition of intermediate intonational phrases ('level 3' in ToBI terminology) that still pose a challenge for the state of the art; we merely aim to fit the needs for our research on the "semantics–syntax–intonation" interface.

The corpus that we label in our experiments is also manually annotated with the basic categories of the information structure theme and rheme (referred to as *thematicity*), which allows us to establish the correspondence between the prosodic patterns and themacity structures. This correspondence is used to validate our proposal of automatic prosodic pattern annotation and to contrast our work with the classical work of Steedman [10], which states that theme tends to be associated to the patterns L+H* and LH% (a clear increasing Low-High pattern), while rheme tends to be associated to the patterns H*L and H*LL% (clearly decreasing High-Low)—although both theme and rheme may be associated with other patterns as well.

The paper is structured as follows. Section 2 describes the complete procedure of automatic annotation of phrase-based prosodic patterns, which involves both the AuToBI system and its adaptation to speech synthesis applications. Section 3 presents the annotation results and its validation through themacity structures, before Section 4, finally, summarizes the conclusions we draw from our preliminary work.

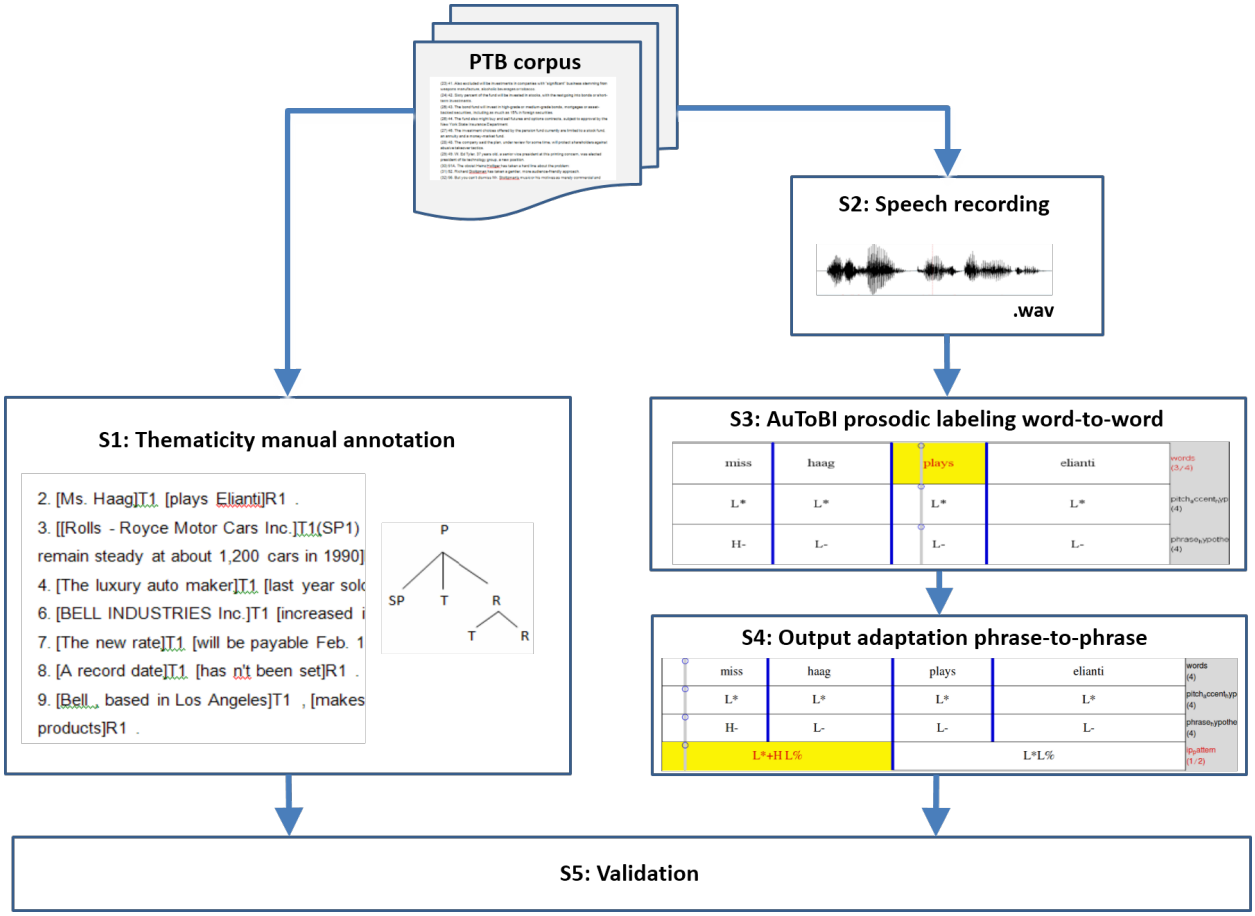
2. Annotating the Information–Prosody Interface

Our annotation procedure consists of five different stages, as shown in Figure 1: (1) thematicity annotation, (2) corpus recording, (3) AuToBI annotation, (4) output adaptation, and (5) validation of the results using manual reference annotations and the outcome of stage (1). In the first stage (S1), a reference corpus is annotated with the information structure (as pointed

[1]Syrdal et al. [3] estimated that experienced labelers could need between 100 and 200 times of the real time speech episode to annotate it.

10.21437/SpeechProsody.2014-210

Figure 1: *The stages of the proposed prosodic pattern annotation procedure and its validation.*



out above, we focus on the thematicity categories theme and rheme). In the second stage (S2), the reading of the corpus (or, as in our case, of a subset of the corpus) by a native speaker of American English is recorded. In the third stage (S3), the recorded speech is automatically labeled with the AuToBI tool. In the fourth (adaptation) stage (S4), the AuToBI word-by-word labels are transformed into IP pattern labels in accordance with our criteria. A final stage (S5) is used to assess the obtained patterns by comparing them with manual annotations and validate them with Steedman's theory [10] on the correlation between prosody and theme/rheme structures.

Next, stages S1 to S4 are described in more detail; the validation stage S5 is presented in a separate section that follows.

2.1. Thematicity annotation stage (S1)

The annotation of thematicity is assumed to be carried out manually over a plain text containing the consecutive sentences. In our experiments, this has been done in a series of blocks of about 40-50 sentences of a fragment of the Wallstreet Journal corpus extracted from the Penn Treebank [11], in accordance with the hierarchical thematicity structure of the Meaning-Text-Theory (MTT) [12].[2] In Figure 1, square brackets mark each communicative span (cf., e.g., '[. . .]T') and parentheses annotate embedded thematicity (cf., e.g., '[. . .]T(R)'). The annotators took into account the context (i.e., the previous sentence), but assumed an interpretation in which none of the elements is focalized or emphasized. The annotation was done by two groups of annotators (two in each group), who discussed in plenum their corresponding annotations to achieve a consensus and to refine the annotation guidelines.

For the validation of the IP pattern annotation procedure, the MTT-oriented annotation has been simplified to match Steedman's theme/rheme structures.

2.2. Speech recording stage (S2)

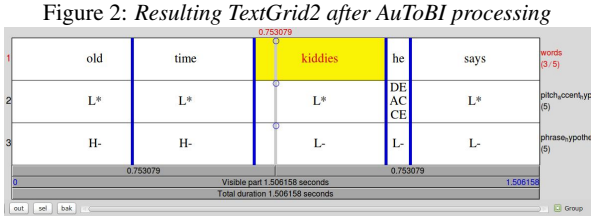
Within the speech recording stage, a subset of the corpus annotated in S1 with thematicity has been recorded. In our experiments, a non-expert native speaker of American English (not involved in the study) was instructed to read a selection of 109 sentences. The sentences contained varied information structure patterns, such that they were prosodically interesting for our study. The recording was done under professional conditions. The sentences were analyzed in order to create a reduction model from the AuToBI output to the IP level to be applied to the entire corpus.

[2]For details about the annotation criteria, see [13].

2.3. Automatic prosodic annotation stage (S3)

This stage consisted in segmenting audio files into words as required by AuToBI. This has been done to automatically process AuToBI's labeling task using Praat [14] and thus generating a TextGrid file for each audio file. Results were saved as TextGrid2 (see Figure 2), which has three interval tiers: the manually segmented word tier and two interval tiers generated automatically by AuToBI, one for the pitch accents and the second for the boundary tones.

Figure 2: *Resulting TextGrid2 after AuToBI processing*



2.4. AuToBI output adaptation stage (S4)

In spite of the fact that AuToBI meant a great step forward in the systematization of prosodic labeling, it has some major constraints for our descriptive approach, as has already been mentioned above.[3] Consequently, the information from AuToBI needs to be manipulated to meet our description requirements for intonational phrases within the information structure framework. For this purpose, we established a limited and manageable inventory of intonation patterns at the phrase level based upon the ToBI annotation convention [16]. We are labeling one main pitch accent (PA) and the boundary tone in each IP. Furthermore, while in the standard ToBI convention [17] four tiers of data are foreseen, namely a tone tier, an orthographic tier, a break tier and a miscellaneous tier, we are only making use of the tone tier, as reduction of detail is prioritized. The collection of patterns we currently use comprehends the items and their possible combinations summarized in Table 1.

Table 1: Patterns set by Pierrehumbert and Hirschberg [16].

pitch accents	L*, H*, L*+H, L+H*, H*+L, H+L*
boundary tones	L%, H%

This limited collection of twelve intonation patterns certainly does not cover all the possible natural tonal realizations of utterances, but it is expected to meet our initial requirements for a model to predict prosody events in speech synthesis applications. The integration of several layers of analysis is assumed to aid to solve the challenge of the prediction of prosody events in speech synthesis applications in that it caters for main pitch accents within the intonational phrase and relevant boundary tones at a clause level.

Our automatic adaptation stage can be envisaged as a loop of three steps over all sentences of the annotated corpus. The steps are: (1) Initial step, (2) Reduction step, and (3) Prerevision step.

[3] Some work has been done by Rosenberg [15] on the incorporation of intonational phrase boundaries into syntactic parsing for automatic summarization, but there is still a great deal of work to be done in this direction.

1. **Initial step.** This step consists in matching the output Textgrid from AuToBI to the sentence annotated in terms of theme/rheme. The result is a txt file (see Figure 3) that contains the following fields:

   - Id number of sentence
   - Chain of words
   - Communicative label
   - Chain of prosodic labels for those words

Figure 3: *Resulting txt matching AuToBI to thematicity labels*

```
Id.S.   Words                   AuToBI

0196    old time kiddies        R1      L*H-L*H-L*L-
0196    he says                 SP1     DEACCENTEDL-L*L-
0196    he                      T1(SP1) DEACCENTEDL-
0196    says                    R1(SP1) L*L-
```

2. **Reduction step.** The greatest part of the prosodic analysis is carried out during this step of the process. The strings of patterns from Step 2 are envisaged from the perspective of the intonational phrase in the pursuit of establishing not only the possible reduction models, but also the communicative and prosodic criteria to segment long utterances into smaller units. These units can help to draw a suitable intonational curve for speech synthesis purposes. As AuToBI does not predict bitonals, our reduction step seeks to predict possible bitonal PAs. The following automatic processing is performed on each pitch accent plus boundary tone (PABT) sequence:

   (a) Total deletion of deaccented items or word chains with a low BT (DL%). These intonation patterns match deaccented words which are disregarded in our IP characterization.

   (b) Substitution of deaccented items with a high BT (DH%) by a bitonal marker H+. High BTs in general may provide information on adjacent word stresses which are relevant in the detection of bitonals when they are followed by a main stress. A sequence of various H+ markers is reduced to a single H+ as it belongs to a sequence of deaccented words. Thus, the resulting single H+ matches a main stress and predicts a bitonal PA.

   (c) Word chains labeled as L*L% in a row can be disregarded for the IP contour definition. Three word-chains with such a label can be reduced to one L*L% IP label as only one word in such a chain will be more salient within the IP.

   (d) Initial L*H%L*L% has been reduced to L*+HL%. In this case, a high BT is turned into a bitonal.

   (e) 3-word combinations of L*H% and L*L% are turned to bitonals with either low or high BTs depending on the pattern chain. For instance, L*H% L*L% L*H% gives H+L*H%.

   The results from this label reduction process are saved into a txt file (see Figure 4) that contains the following fields:

   - Id number of sentence
   - Chain of words

- Communicative label
- Number of words
- Number of IPs
- Proposed ToBI label for each IP

Figure 4: *Resulting txt after reduction model processing*

Id.S.	Words	IS	N.W.	N.IP	PrePattern
0196	old time kiddies	R1	3	1	H+L*L%
0196	he says	SP1	2	1	L*L%
0196	he	T1(SP1)	1	0	0
0196	says	R1(SP1)	1	1	L*L%

3. **Pre-revision step.** After getting a proposed IP label, a Praat file needs to be created in order to revise all the material that has been automatically generated. Therefore, TextGrid3 file merges the existing tiers from TextGrid2 plus three more, namely:

   (a) clauses divided into intonational phrases and with their corresponding communicative labels,

   (b) same intonational phrases containing the proposed ToBI pattern, and

   (c) word divisions as in tier 1 for AuToBI input that will serve to place a pitch accent (PA) into the main stressed word within the IP and BT to be able to detect intermediate IP easily regardless pauses or silence dependency.

   See Figure 5 for illustration.

   Once the Textgrid3 file is generated, the manual process of revisor's validation of the proposed patterns takes place. The manual changes are saved as TextGrid4 (see Figure 6).

Figure 5: *Resulting TextGrid3 including processed tiers*

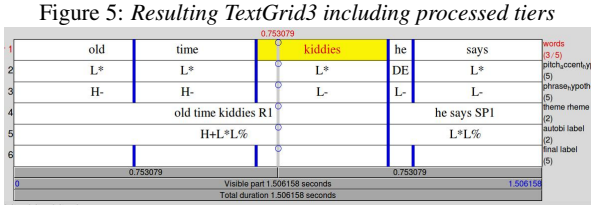
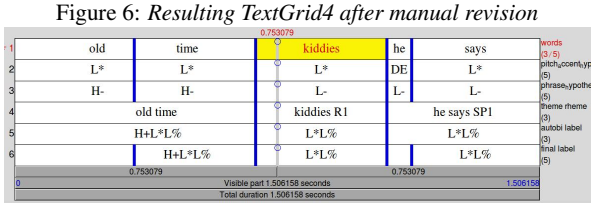


Figure 6: *Resulting TextGrid4 after manual revision*

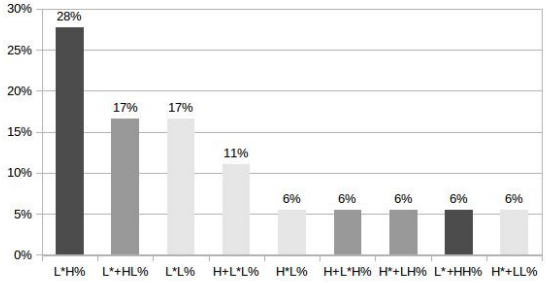


3. Validation stage (S5)

As mentioned above, the validation stage serves to assess the prosodic patterns obtained during the adaptation stage in order to evaluate the efficiency of our model. For this purpose, we first compared the results from our automatic reduction model at the intonational phrase level to a manual annotation. The comparison revealed that the model matches exactly the whole pattern in 58% of the total number of IPs. This includes number of IPs division and exact ToBI pattern assigned. There is a 18% of partially matched patterns (whose match corresponds in all cases to the BT). And the rest 24% of IPs does not match with the manual annotation.

Then, we compared the obtained prosodic patterns with those prognosticated by Steedman's [10] based on sentential theme/rheme structures. Figure 7 shows that themes tend to contain a rising intonation pattern as [10] claims, given that L*H%, L*+H H%, H*+L H% and H+L* H% (highlighted in dark gray) have a final rising intonation and L*+H L% contains a rising PA. These patterns add up to 63%, which proves that our model represents the general characterization made in theoretical approaches on this topic. Hence, the results can be regarded as reliable and we can conclude that apart from the obvious save in time in labeling effort, our reduction model is validated by existing theories on intonation applied to thematicity.

Figure 7: *Theme intonation patterns distribution (%)*



4. Conclusions

We presented a procedure for automatic prosodic pattern annotation that has been shown to be sufficiently reliable for experiments on the "semantics–syntax–intonation" language interface. In our validation, we have shown that our prosodic extraction is in concordance with Steedman's hypothesis for simple sentence structures. However, the great variety of intonation patterns that we found also proves that existing theories on thematicity characterization in prosodic terms, such as Steedman's [10], require a deeper insight from a qualitative perspective using real examples from different contexts, registers and speakers. For this reason, our procedure presented contributes, on the one hand, to the possibility of labeling large corpora with a substantial cut-off on manual revision efforts and minimizes, on the other hand, the risk of obtaining different labels as result of different annotators' subjective viewpoint, as long as all annotators are given a sistematic pattern and are asked to check whether there is an error with the initial output from AuToBI. Annotators can be trained to detect these errors and make more objective decisions when they spot an error than when they are labeling from scratch and therefore, have to make all decisions on their own.

5. Acknowledgements

Parts of this work have been funded by a grant from the European Commission under the contract number FP7-ICT-610411. The second author is partially funded by a grant from the Spanish Ministry of Economy and Competitivity in the framework of the Juan de la Cierva fellowship program (JCI-2012-12272).

6. References

[1] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. and Hirschberg, J., "ToBI: a standard for labelling English prosody". Proceedings of the ICSLP, vol. 2, 867-870, Sydney, Australia, 1992.

[2] Steedman, M., "Using APML to specify intonation". Magicster Project Deliverable 2.5. University of Edinburgh, 2005. Available at http://www.ltg.ed.ac.uk/magicster/deliverables/annex2.5/apml-howto.pdf

[3] Syrdal, A. K., Hirschberg, J., McGory, J. and Beckman, M., "Automatic ToBI prediction and alignment to speed manual labeling of prosody". Speech Communication, 33(1-2): 135–151, 2001.

[4] Noguchi, H., Kiriyama, K., Matsuda, H., Taniguchi, M., Den, Y. and Katagiri, Y., "Automatic labeling of Japanese prosody using j-toBI style description". Proceedings of the Eurospeech, 2259–2262, 1999.

[5] Lee, J.-S., Kim, B. and Lee, G. G., "Automatic corpus-based tone prediction using K-ToBI representation". Proceedings of the Conference on Empirical Methods in Natural Language Processing, 134–142, 2001.

[6] Ananthakrishnan, S. and Narayanan, S. S., "An automatic prosody recognizer using a coupled multi-stream acoustic model and a syntactic-prosodic language model". Proceedings of the ICASSP, 269–272, Philadelphia, PA, 2005.

[7] Ananthakrishnan, S. and Narayanan, S. S., "Automatic prosodic event detection using acoustic, lexical, and syntactic evidence". IEEE Transactions od Audio, Speech, and Language Processing, 16(1): 216–228, 2008.

[8] Rangarajan Sridhar, V. K., Bangalore, S. and Narayanan, S., "Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework". IEEE Transactions on Audio, Speech, and Language Processing, 16(4): 797–811, 2008.

[9] Rosenberg, A., "AutoBI - a tool for automatic toBI annotation". Proceedings of Interspeech, 146–149, 2010.

[10] Steedman, M., "Information structure and the syntax-phonology interface", Linguistic Inquiry, 4(31):649–685, 2000.

[11] Marcus, M.P., Santorini, B. and Marcinkiewicz, M.A., "Building a Large Annotated Corpus of English: The Penn Treebank". Computational Linguistics, 19(2):313–330, 1993.

[12] Mel'čuk, I. A., "Communicative Organization in Natural Language: The semantic-communicative structure of sentences". Benjamins Academic Publishers, Amsterdam, 2001.

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

[14] Boersma, P. and Weenink, D., "Praat: doing phonetics by computer" [Computer program]. Version 5.1.51, retrieved September 2013 from http://www.praat.org/.

[15] Maskey, S., Rosenberg, A. and Hirschberg, J., "Intonational Phrases for Speech Summarization". Interspeech, 2008.

[16] Pierrehumbert, J. and Hirschberg, J. , "The intonational structuring of discourse". Proceedings of the ACL, 136–144, New York, NY, 1986.

[17] Beckman, M. and Hirschberg, J., "The ToBI Annotation Conventions". The Ohio State University, OH, 1999. Available at http://www.ling.ohio-state.edu/ tobi/ame_tobi/annotation_conventions.html