Fourth ISCA ITRW on
Speech Synthesis (SSW-4))
Perthshire, Scotland
August 29 - September 1, 2001

# Prosodic Phrasing: Machine and Human Evaluation

*M. Céu Viana*, Luís C. Oliveira**, Ana I. Mata***,*

*CLUL, **INESC-ID/IST, ***FLUL/CLUL
Rua Alves Redol 9, 1000 Lisboa, Portugal
mcv@clul.ul.pt, lco@inesc-id.pt, aim@mail.doc.fl.ul.pt
http://www.l2f.inesc-id.pt

## Abstract

In this paper we describe a set of experiments aiming at building and evaluating a new phrasing module for European Portuguese Text-to-Speech Synthesis, using Classification and Regression Tree (CART) techniques on hand-labeled texts. Using the assessment criteria of matching boundary predictions against a reference example of phrased sentences, the best solution found up to now achieves an overall performance of 91.9%, with 86.3% of breaks correctly assigned and 4.3% of false insertions. Although in absolute terms such scores may be considered surprisingly good considering the size of the training set, the total number of exact matches at the sentence level is much lower. This suggested a more formal experiment to test the acceptability of the predicted phrasing in the judgment of human evaluators. The experiment involved 90 participants that were asked to grade both the predicted and reference phrasing, and to also express their opinion on where should the breaks be placed. The results showed that, as expected, there is a large variability among the subjects in the acceptance of a specific partitioning. However the performance of the automatic assignment procedure is better rated by human evaluators.

## 1. Introduction

Linguistic theory posits a hierarchy of nested prosodic phrasing levels above the word, which are domains for sandhi rules and manifest themselves more or less directly in the speech signal in terms of F0, duration patterns, location of pauses, etc. (e.g. [1], [2], [3], [4], [5], [6]).

Although the number and designation of phrasing levels differ from author to author and different hypothesis have been presented concerning the number of levels needed to account for tonal as well as durational patterns, there is a common agreement that prosodic structures are not fully congruent with syntactic ones. They are generally flatter and cannot be predicted using syntactic information only: semantics and discourse structure, as well as rhythmic constraints, do play an important role [7].

Previous attempts have been done, using more or less elaborated parsing strategies, to incorporate such types of information in TTS-systems [8]. Many rule-based or statistically based TTS-systems, however, achieved satisfactory results with much simpler phrasing algorithms (e.g. [9], [10], [11], [12]).

In order to build a phrasing module for a new version of the DIXI system [13] in the festival framework, we closely followed [12]. In the line of [10] and [11] work, Classification and Regression Tree (CART) techniques [14] were used and all the experiments were performed on hand-annotated text and not on hand-labeled recorded speech. As shown in [11], text-based methods considerably speed up the process of building new phrasing modules or updating existing ones, and the resulting decision trees may reach equivalent or even slightly higher cross-validation scores.

Models based on self-learning procedures have some well-known advantages over rule-based ones. They can be easily retrained and tested as more and more annotated speech materials become available or its quality is improved. They may be used also as an efficient method to determine which variables are linguistically meaningful and which is their relative weight. They are thus particularly useful in the case of languages like European Portuguese, for which large annotated speech corpora are still under construction and neither the most relevant features for the marking of different prosodic events nor the ways they are interrelated with each other are well known. In this sense, among the available self-learning procedures, CART techniques appear as the most natural choice, since the resulting trees are easier to read and can be manually modified.

## 2. Data and methods

In all the experiments described below, a selection of 35 written texts was used, covering a variety of genres: excerpts from school textbooks, novels, press articles and interviews, letters, cooking recipes, traditional rimes and popular jokes. These were collected in the scope of another CLUL's project to exemplify the heterogeneity of language uses for teaching purposes. This heterogeneity is a desirable feature for the development of a TTS-system supposed to deal with unrestricted text. An additional reason concerning future work motivated, however, the current selection. All the texts were read by two professional speakers (a mail and a female) and recorded with the same protocol used for other speech corpora collected by the group.

Two of the authors, both European Portuguese (EP) native speakers, hand-labeled half of the selected written materials for prosodic boundaries, and checked the other half. Three types of junctures were considered, which roughly correspond to ToBI [15] levels 1, 3 and 4 (no break, minor break and major break, respectively). Breaks were assigned paying attention to punctuation and in accordance to what the annotators believed to be appropriate intonational boundary locations in a slow but fluent and smooth oral reading. Since we currently wanted to predict only if there is or not a break at a given boundary, levels 3 and 4 were collapsed into a single category.

The working corpus has 11167 word boundaries, 3037 (27%) of which were marked has breaks. This portioning gave a phrase length distribution plotted in figure 1 with an average length of 3.8 words per phrase. The part-of-speech tags were also annotated, as well as other kinds of information (sentence length in words and different types of distance from the boundary to previous and following punctuation marks). Dis-

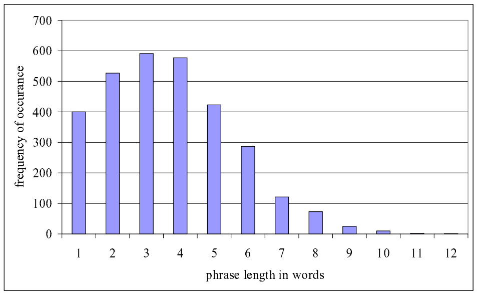

Figure 1: Distribution of phrase lengths in the reference data.

tance measures were automatically extracted from raw text and POS tags were obtained with palavroso, an EP morphological analyzer developed by INESC, whose output has been checked for ambiguities and manually corrected.

For statistical modeling, we used the Edinburgh Speech Tools version of CART procedures. Its output can easily be used in the festival framework in which the new version of the DIXI system [13] is currently being developed.

As most the phrasing methods considered in the following experiments were also tested for English by [12], we employ the same performance measures used in that work.

Given the limited size of the corpus, this was randomly divided in order to allow for five-fold cross-validation estimates using 80% of the data for training and the remaining 20% for testing.

## 3. Experiments and Results

All the experiments described in the following paragraphs aimed at building and testing the performance of some current phrasing methods on European Portuguese data. They all rely on information that can be directly extracted from text and account for a progression in complexity most useful during the process of development of TTS-systems for new languages. They can be used at different stages and in accordance with the availability of the necessary linguistic resources. As they have all been tested for English by [12], the results of this work were used as reference.

### 3.1. Experiment 1: punctuation only

Punctuation marks are an important source of prosodic information, indicating namely how an sentence must be phrased. In our reference corpus the punctuation marks account for more than a half of the total number of breaks. A system using just this information would have an average performance of 61.1% of correct breaks and a total of 89.4% of correctly classified junctures. There were no false insertions and the 10.6% of error correspond to a failure in predicting a break contemplated in the reference corpus. These results are similar to the ones reported for English [12]. As often pointed out, this error rate may result in real bad performances when sentences are relatively long and have little or no punctuation.

### 3.2. Experiment 2: punctuation plus content/function word distinction

Another method with acceptable results reported for English [9], is to insert a break not only on punctuation marks but also after any content word followed by a function word.

Our experiments showed that the use of this strategy increased considerably the number of breaks correctly placed but with an increased rate of false insertions and consequently reducing the number of correctly predicted boundaries. The number of correct breaks changed from 61.1% to 85.1% but the rate of false insertions increased from zero to 16.8%. The rate of failure in predicting a break was reduced from 10.6% to 4.0%.

Since this method always groups together sequences of content words, sentences are often phrased in unacceptable ways. The verb, for instance, is separated from its first complement every time it begins with a function word and is always grouped with a preceding NP, ending in a noun or an adjective. Although in some cases such groupings may be considered acceptable or even good, most of the time they correspond both to rhythmic and syntactically ill structures.

### 3.3. Experiment 3: Punctuation plus part-of-speech information.

The next step was to consider the part-of-speech (POS) information where two questions had to be answered: how many different tags should we consider? and how many words should be included in the analysis window? Those variables needed to be investigated and optimized. In order to do so, the original set of 260 tags produced by palavroso was reduced to 42 by removing all nominal and verbal inflexion marks. Several exploratory experiments were conducted in order to have an insight on the behavior of this variables. The window size was varied from 3 to 5 words and two sets of tags were used, one with 36 categories and the other with 11. The best results were obtained with 36 categories and with the longest window.

In these exploratory experiments it was also taken into account the distance measures: the number of words from the boundary to previous and following punctuation marks. The results showed that the final tree almost did not take this measures into consideration. The major decision factor was the location of the punctuation marks and the POS tags of the words. In the following experiments the distance measures were no longer used.

After the exploratory experiments, we tried to optimize the tag set using a greedy-type algorithm. The initial 42 different labels were reduced to 41 by merging two tags into one. A CART was than trained and tested on the resulting data and the performance was recorded. We than repeated the procedure for all other possible combinations of two labels in the original set. The combination producing the best result in terms of correctly placed breaks was selected for the next step of the algorithm. The procedure was stopped when the tag set reached a size of 4 labels.

In figure 2 the best rate of correctly placed breaks is plotted for each tag set size. Given these results a tag set of 12 labels was selected, with 86.3% of correctly placed breaks, 91.9% of correctly predicted junctures and 4.3% of false insertions (3.8% of missing breaks).

## 4. Evaluation

So far, we have evaluated the performance of a system for automatic phrasing by matching its results with the reference test

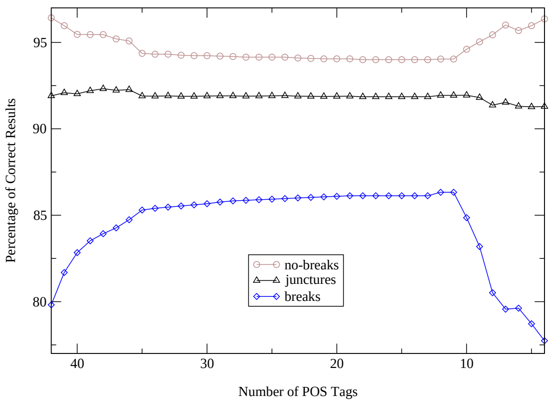

Figure 2: Best rate of correctly placed breaks and the corresponding percentage of correct junctures and no-breaks for each tag set size.

set. This measure can be deceptive: the automatically assigned boundaries can be different from the reference but still acceptable.

An evaluation procedure was developed to assess the ability of the automatic system to locate prosodic breaks and also to get some ideas of how a group of native speakers evaluates possible partitions of a sentence. Given the expected variability of the possible phrasings, the evaluation procedure was designed in such a way that a significant number of evaluators could easily be recruited. The questionnaire was restricted to take around half an hour and to be carried over the Internet.

### 4.1. Evaluation Tool

The evaluation tool developed for this test requires that the evaluator has an Internet access and a web browser such as Netscape or Internet Explorer. The tool runs on an HTTP ("HyperText Transfer Protocol") server and uses the "Common Gateway Interface" (CGI) to generate forms for the evaluator to fill.

The evaluators were recruited by e-mail announcing the URL (Uniform Resource Locator) address of the test. The snow ball recruiting was also attempted by asking the evaluators to spread the address of the test.

The test design had three main objectives:

1. To evaluate the acceptability of the prosodic breaks assigned by the automatic system.
2. To evaluate the opinion of the evaluators on the reference phrasing used for training the automatic system.
3. To evaluate the variability of the evaluators in the task of segmenting a sentence in a limited number of prosodic phrases.

The two first objectives can be reached with same test, where the evaluator is asked to rate a sentence phrasing. The sentence text is marked with the locations of the prosodic breaks, for example:

```
    Na Madeira/ haverá chuva/ passando a
aguaceiros.
    (In Madeira/ it will rain/ followed by
heavy showers.)
```

For the rating, a scale of three values was chosen and it was presented to the evaluators in the following form:

**G:** Good, I could read it this way

| task0 | task1 | task2 |
|---|---|---|
| set0r | set1r | set2r |
| set1a | set2a | set0a |
| set2m | set0m | set1m |

| task3 | task4 | task5 |
|---|---|---|
| set3r | set4r | set5r |
| set4a | set5a | set3a |
| set5m | set3m | set4m |

| task6 | task7 | task8 |
|---|---|---|
| set6r | set7r | set8r |
| set7a | set8a | set6a |
| set8m | set6m | set7m |

Table 1: Each evaluator's task included 3 sets of 10 phrases for the evaluation of the automatic phrasing (a), the reference phrasing (r) and to mark prosodic boundaries (m).

**A:** Acceptable, I would not read it this way but it could be a possible reading.

**U:** Unacceptable, it does not seem to be a natural reading of the sentence.

To carry through the third objective, the sentence was presented to the evaluator with buttons between the words. The evaluator was asked to place breaks where he believed to be an appropriate location. An informal test showed that the evaluators had different sensitivity to the break level: some marked only the major breaks while others produced a larger number of phrases. To solve this problem the evaluators were forced to mark a number of breaks in a specified range, between the number of breaks of the automatic and reference phrasing. This first solution was found too restrictive because in some cases both phrasings had the same number of breaks. The final solution was to allow one break less than the previously calculated minimum.

The first tests showed that the evaluation of each sentence took one minute on average, which limited the number of sentences per evaluator to 30. Since we needed 90 sentences in the test, a strategy had to be devised to allocate each evaluator sentences. The sentences were first randomly split into 9 sets of 10 sentences (set0 set8). To accomplish the three intended experiences, 3 versions of each phrase were produced: one with the break marks produced by the automatic system (a), another one with the reference phrasing (r) and another one with just the sentence text to be marked by the evaluator (m). These 270 sentences were than distributed by 9 tasks of 30 phrases each in accordance with table 1. The sentences of each task were randomly ordered to prevent the evaluator to identify the automatic and reference phrasing.

### 4.2. Test Sentences

The 90 sentences used in the evaluation were select from the the full corpus, one in every five sentences, but restricted to be longer than 7 words. The sentence length range to a maximum 65 words, with an average size of 19 words. The reference phrasing of this set had between 2 and 16 prosodic breaks in each sentence including the final break. In this set, each sentence had, in average, 4 phrases.

### 4.3. Test Description

The test was carried out between March 27th and April 3rd, 2001. The evaluators were asked to participate through e-mail messages sent to researchers of our laboratories (INESC-ID and CLUL), to Professors and students of our Universities (IST and FLUL) as well as to some personal contacts.

Each evaluator had to select a username and was then asked for his full name. After the identification procedure, one of the nine tasks was assigned to him. Thereafter it was possible for the evaluator to stop the test at any moment and continue later, by giving his username.

The 30 sentences of the test were presented in separate pages for the evaluator to perform the required task. He then had to submit his answer to receive another sentence. After submission the answer could not be changed. The evaluator was asked to grade the phrasing of 20 sentences as Good, Acceptable or Unacceptable. For the 10 remaining sentences the evaluator had to mark the location of the prosodic breaks that he would introduce in a slow but fluent reading.

After April 3rd, the evaluators could visit the same URL of the test to compare their answers with those given by other evaluators. The goal was to show them that there was no "right" answer and thus the importance of a large number of participants. We hope that this will increase their willingness to be involved in future tests.

We received a total of 105 evaluator registrations, of which 91 completed the test. So, each task was carried through by at least 10 different evaluators.

### 4.4. Evaluation at the Sentence Level

The results of the automatic phrasing performance, previously presented, accounted for the number of boundaries correctly or incorrectly located. The selected evaluation method allows us to study the performance of the system at the sentence level. The performance results computed this way are more demanding for longer sentences because a single phrase boundary in an unacceptable place is enough to make the sentence unacceptable, even if the remaining boundaries are acceptable. This is clearly shown by the fact that in the 90 sentences of this test only 20 have the same phrasing as the reference partitioning: with this criteria we would have only 22% of correct phrasing. However, the fact of that the automatic phrasing differs from reference phrasing does not mean that it is unacceptable.

#### 4.4.1. Evaluators variability

As expected, the subjectivity of the phrasing evaluation produced large differences in the evaluator's judgments. For the 20 sentences with identical automatic and reference phrasing, we have for each sentence the judgment of 20 evaluators and the results are presented in table 2. For instance, for the sentence 443, 11 evaluators considered the phrasing as good, while 4 found it unacceptable but 18 of the 20 evaluators made that same phrasing when asked to insert breaks in that sentence. It is also noticeable that the longer sentences have a larger unacceptability ratio.

Using these findings, it was decided that a phrasing could only be rated as unacceptable if more than half of the evaluators considered it as such. Using the same criteria, it can only be considered a good phrasing if it was classified as such by more than 50% of the evaluators.

Another reason for the variability of the judgments is the large number of possible phrasings for long sentences. The number of possible break locations tends to grow with the length of the sentence. Figure 3 shows the histogram of the average number of phrasing patterns assigned by the evaluators for the same sentence as a function of the number of words. For sentences longer than 29 words, the 10 evaluators made 6 or more different phrasing patterns.

An interesting result of the test is to compare the phrasing patterns assigned by the test subjects with the automatic and reference phrasing. Table 3 shows the percentage of agreement in the phrasing performed by the test subjects and the evaluation of that phrasing. Even when 90% of the evaluators agreed on a phrasing, 10% found it unacceptable. This confirms the requirement of a majority of more than 50% to find a sentence phrasing unacceptable.

| sentence | | evaluation | | | |
|---|---|---|---|---|---|
| id | length | Agree | Good | Accept. | Unaccept. |
| 8 | 11 | 90% | 85% | 10% | 5% |
| 18 | 11 | 90% | 40% | 40% | 20% |
| 43 | 19 | 0% | 55% | 40% | 5% |
| 53 | 30 | 0% | 30% | 40% | 30% |
| 103 | 23 | 40% | 40% | 45% | 15% |
| 118 | 8 | 70% | 80% | 30% | 5% |
| 133 | 13 | 50% | 85% | 10% | 5% |
| 173 | 8 | 90% | 85% | 5% | 10% |
| 323 | 23 | 10% | 55% | 30% | 15% |
| 378 | 17 | 10% | 40% | 45% | 15% |
| 383 | 31 | 10% | 30% | 35% | 35% |
| 418 | 9 | 80% | 95% | 5% | 0% |
| 438 | 12 | 80% | 90% | 10% | 0% |
| 443 | 11 | 90% | 55% | 25% | 20% |
| 458 | 16 | 80% | 80% | 10% | 10% |
| 473 | 9 | 90% | 80% | 20% | 0% |
| 488 | 9 | 80% | 75% | 25% | 0% |
| 538 | 11 | 100% | 80% | 20% | 0% |
| 543 | 16 | 60% | 100% | 0% | 0% |
| 548 | 34 | 0% | 20% | 35% | 45% |

Table 2: The evaluation results of the 20 sentences for which the automatic phrasing matched the reference.

| Phrasing | Evaluation | | |
|---|---|---|---|
| Agreement | Good | Accept. | Unaccept. |
| 100% | 80% | 20% | 0% |
| 90% | 71% | 19% | 10% |
| 80% | 80% | 17% | 3% |
| 70% | 67% | 23% | 10% |
| 60% | 73% | 15% | 12% |
| 50% | 68% | 17% | 15% |
| 40% | 54% | 30% | 16% |
| 30% | 38% | 38% | 23% |
| 20% | 43% | 26% | 31% |
| 10% | 36% | 39% | 25% |

Table 3: Percentage of evaluators that agreed on the same phrasing and the grading of that phrasing. Even when 90% of the evaluators agreed on a phrasing, 10% found it unacceptable.

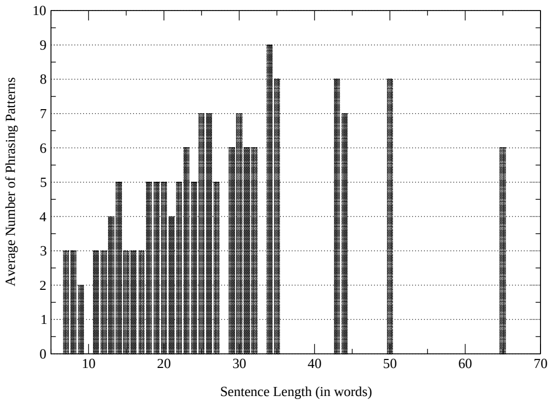

Figure 3: Average number of phrasing patterns assigned by the 10 evaluators for each sentence length.

### 4.4.2. *Automatic phrasing results*

Using the criteria that we have just defined, the evaluators found that the automatic phrasing of 20 of the 90 (22%) sentences were unacceptable. These includes the two longest sentences and a sentence that the evaluators considered unacceptable even in the reference phrasing. Of the remaining sentences, 40 (44%) were considered to have an acceptable phrasing and 30 (33%) a good one. As reference, it can be pointed out that for 40 sentences (44%) there was at least one evaluator that performed the same phrasing of the sentence as the automatic procedure did.

### 4.4.3. *Reference phrasing results*

Applying the same criteria to the judgment made by the evaluators on the reference phrasing, the phrasing of 6 (7%) sentences was considered unacceptable, 31 (34%) good and 53 (59%) acceptable. In this case, the number of sentences for which at least one evaluator made the same phrasing went up to 50 (56%).

### 4.5. Break Level Results

According to [16]: ”There are many reasonable places to pause in long sentence, but few where it is critical not to pause”. That is, the errors in the assignment of prosodic breaks cannot be found only by matching the breaks in the reference phrasing, there are surely other acceptable locations. The problem is to find the breaks that were assigned to places where it is critical not to pause.

We would add to the previous statement that there are also places where it is almost mandatory to pause. If a phrase boundary is missing at that location the phrasing can become unnatural.

Using these basic principles, the evaluation was made using 3 performance measures:

**correct break:** at least one evaluator placed a break at that same location;

**false insertion:** none of the evaluators placed a break at that same location;

**missing break:** there should be a break at that location because more than 2/3 of the evaluators agreed on breaking there.

The evaluation results gave that in the 1715 word boundaries of the 90 sentences, the automatic system inserted 389 breaks (22,7%) giving an average phrase length of 4.4 words. The reference phrasing located 448 breaks (26,1%) with an average length of 3.8 words per phrase, while the evaluators introduced on average 370 breaks (21,6%) with an average phrase length of 4.6 words.

| Errors | False Insertions | | Boundary Deletion | |
|---|---|---|---|---|
| | automatic | reference | automatic | reference |
| 0 | 74% | 84% | 69% | 94% |
| 1 | 22% | 14% | 29% | 6% |
| 2 | 3% | 1% | 2% | 0% |

Table 4: Average boundary errors in sentences.

Of the 389 breaks assigned by the automatic system, 26 (6.7%) were considered false insertions because none of the evaluators placed a phrase break at those locations. On the other hand, the system failed to place a break in 30 locations where more than 2/3 of the evaluators agreed. Considering all possible break locations the system assigned 1.5% of incorrect breaks and failed to introduce 1.7%.

Performing a similar analysis on the reference phrasing, 15 (3.3%) breaks were considered wrong and 5 breaks were missing. Considering all possible locations for the breaks the evaluators did not agree on 0.9% of the assigned breaks and would have added more 0.3% of breaks.

Of the 20 phrases considered unacceptable by the evaluators, 8 were due to missing breaks, 6 incorrectly assigned breaks and the remaining 6 for both reasons.

Table 4 shows the number of sentences without false insertions and deletions and with 1 or 2 errors due to badly located or missing breaks.

## 5. Conclusions

We have described a set of experiments for building and evaluating a new phrasing module for European Portuguese on hand annotated text and using CART techniques.

Results confirm the efficiency of this procedure for acquiring phrasing rules for a new language and for testing the relative weight of different variables. They compare well with the ones obtained for English by [12] which also used information directly obtained from text but on a much larger data set.

To validate our results an evaluation was performed using the judgment of human evaluators. We have also asked the subjects to perform themselves the task of assigning prosodic breaks.

One of the difficulties on the analysis of the results was to deal with the large variability among the evaluators. Criteria had to be defined to summarize the results of the different evaluators: a majority of opinions had to be expressed for the judgment to be accepted.

The full analysis of the results is not yet completed but they clearly show that automatic assignment of prosodic breaks performs much better in the judgment of the human evaluators than predicted by the comparisation with the reference phrasing. For sentence level results, only 22% of the sentences matched the reference phrasing but the evaluators found that 78% of the sentences were acceptable.

When analyzing at the break level the results also show differences between the performance using the reference breaks and those assigned by the evaluators. In the first case we found 4.3% of false break assignment and 1.5% in the second case. Regarding the missing breaks, the system missed 3.8% breaks when compared with the reference corpus, but only 1.7% if compared with the phrasing performed by the subjects.

As future work we plan to analyze the results system for more than one phrasing level. This way we will try to verify if the evaluators were more consistent in assigning the major breaks.

## 6. Acknowledgements

The research reported here has been conducted within the DIXI+ project, supported by the Portuguese Foundation for Science and Technology (FCT), Praxis XXI program. We would like to thank Isabel Mascarenhas who hand-corrected most of the part-of-speech tagging. We also would like to thank all all those who freely gave some of their time to participate in the evaluation experiment.

## 7. References

[1] J. Pierrehumbert, *The Phonology and Phonetics of English Intonation,* Ph.D. thesis, MIT, Boston, 1980.

[2] Janet Pierrehumbert and Mary Beckman, *Japanese Tone Structure,* MIT Press, Cambridge, Mass., 1988.

[3] Elisabeth Selkirk, *Phonology and Syntax,* MIT Press, 1984.

[4] Elisabeth Selkirk, “On derived domains in sentence prosody,” 1986.

[5] M. Nespor and I. Vogel, *Prosodic phonology,* Foris Publications, 1986.

[6] Robert Ladd, *Intonational phonology,* Cambridge University Press, 1996.

[7] J. P. Gee and F. Grosjean, “Performance structures: A psycholinguistic and linguistic appraisal,” *Cognitive Psychology*, , no. 15, pp. 411–458, 1983.

[8] J. Bachenko and E. Fitzpatrick, “A computational grammar of discourse-neutral prosodic phrasing in english.,” *Computational Linguistics,*, vol. 16, no. 3, pp. 155–170, 1990.

[9] K. Silverman, *The Sructure and Processing of Fundamental Frequency Contours,* Ph.D. thesis, University of Cambridge, 1987.

[10] M. Q. Wang and J. Hirschberg, “Automatic classification of intonational phrase boundaries,” *Computer Speech and Language*, vol. 6, pp. 175–196, 1992.

[11] J. Hirschberg and P. Prieto, “Training intonational phrasing rules automatically for english and spanish text-to-speech,” *Speech Communication*, vol. 18, pp. 281–290, 1996.

[12] Alan W. Black and Paul Taylor, “Assigning phrase breaks from part-of-speech sequences,” *Computer Speech and Language*, vol. 12, pp. 99 – 117, 1998.

[13] L. C. Oliveira, M. C. Viana, and I. M. Trancoso, “DIXI – Portuguese text-to-speech system,” in *Eurospeech*, Genoa, Sept. 1991, pp. 1239–1242, http://www.speech.inesc.pt/bib/Oliveira91a.ps.

[14] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees,* Wadsworth & Brooks, Pacific Grove CA, 1984.

[15] K. Silverman, M. Beckman, J. Petrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, “Tobi: A standard for labelling english prosody,” in *Proc. ICSLP’92*, 1992, pp. 867–870.

[16] X. Huang, A. Acero, and H. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development,* Prentice Hall, 2001.