Assigning Phrase Breaks Using CARTs for Basque TTS

*Eva Navas, Inmaculada Hernáez & Nerea Ezeiza**

Department of Electronic and Telecommunications
*Department of Computer Languages and Systems
University of the Basque Country
{eva; inma}@bips.bi.ehu.es ; *jibecran@si.ehu.es

Abstract

This paper presents a prosodic phrasing method for the Basque language, to improve naturalness in text to speech synthesis. Binary classification trees are trained with morphological and syntactic information to predict locations of breaks. Overall score achieved by the prediction tree is 92.53%, which compares positively with the results published for other languages.

1. Introduction

Assigning appropriate phrase breaks is basic for the naturalness of synthetic speech, and even for its intelligibility, mainly in long sentences. Unfortunately, this is a difficult task: the decision of placing phrase breaks in natural speech depends on many factors, like context, speech rate and necessity of breathing. Breaks missing when necessary or inserted in incorrect places make the TTS (text to speech) system sound unnatural and boring. Besides, traditionally phrase breaks are used by a number of modules of TTS systems, such as duration module, grapheme to phoneme module and intonation module.

There are systems that place breaks by rule, taking into account a function/content word classification [1], and other systems use diverse methods of statistical analysis to insert breaks [2][3]. As Basque is an agglutinative language very few function word exist, so traditional rules are not applicable in our case. In this study, a classification tree trained with morphological and syntactic information, is used to predict breaks.

The paper layout is as follows: firstly, in Section 2 the data base used in this study is described, then Section 3 details the part-of-speech and syntactic tagging of the data base. In Section 4 features for the statistical analysis of the data are proposed and the tree is described. The tree-based model for prosodic parsing is evaluated in Section 5 and finally section 6 discusses the results.

2. Data

In the early experiments an already available oral database was used. This database was composed by four articles taken from *Campusa* magazine and read in a quiet environment by a native Basque male speaker. Unfortunately, the database was small and the number of breaks was not enough for the reliable prediction of breaks. The research performed using this data base was valuable as starting point for the work described in this paper.

Due to the enormous effort needed to record and annotate a new speech data base, a textual data base was used. Advantages of textual databases have been described in [4]. It was called *Internet* and was formed by 17 texts with diverse topics taken from different Internet sites. Table 1 shows the characteristics of both databases.

Table 1: Main characteristics of databases used

	Campusa	Internet
Type	oral & textual	textual
Recording length	10’	--
Text quantity	7K	38.1K
Nº of words	899	4332
Nº of sentences	49	366
Nº orthographic breaks	93	605
Nº non-orthographic breaks	162	665

2.1. Database labeling

The *Internet* database was annotated by a native Basque speaker who labeled the spaces between each pair of words as a break, when she considered the boundary to sound natural at this point, and as a non-break otherwise. For this labeling task every orthographic sign was considered as indicating a break. No further classification of breaks was made, because the more categories are used, the fewer occurrences of each class there will be in the corpus and for the results to be reliable, a large quantity of examples of each class is needed.

The distribution of the resulting prosodic phrases lengths measured in syllables is shown in Fig. 1, where the mean number of syllables in prosodic phrases is also displayed.

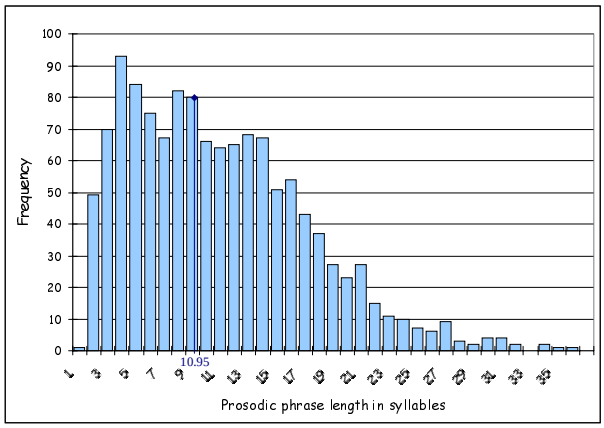

Figure 1: Distribution of prosodic phrase length in Internet database.

The shortest prosodic phrase had only one word (1 syllable) and the longest one has 13 words (36 syllables), while the mean length of the prosodic phrases measured in syllables is 11.

3. Part of speech tagging

Traditionally syntactic and morphological information has proven to be very useful in the prediction of pauses for several languages [5], so the databases used in this study were morphological and syntactically annotated, using some tools developed by the IXA group [6]. For the part-of-speech tagging, the morphological analyzer for Basque MORFEUS [7] has been used. This program provides a set of 15 main tags. Most of those main tags have two more levels of subcategorization, according to grammatical considerations. The number of different tags supplied by MORFEUS was too large for our purposes, especially taking into account the small size of the corpus. So, to avoid sparse data problem the full tag set had to be reduced. Finally considered tags are shown in Table 2.

As Table 2 shows, content word classes were not subcategorized and only the main tag was used for these cases. There were only two function word classes in the tag set: determinants (DET) and conjunctions (LOT). The former did not need any subclassification, but the latter was subclassified attending to the second level of categorization, because of the conclusions from previous studies performed using the *Campusa* database. No manual correction of these labels was done.

Table 2: Part-of-speech tags

Label	Description
ADB	Adverb
ADI	Main verb
ADJ	Adjective
ADL	Auxiliary verb
ADT	Synthetic verb
BEREIZ	Special punctuation mark
DET	Determiner
IOR	Pronoun
ITJ	Interjection
IZE	Noun
LOT_JNT	Sentence connector
LOT_LOK	Conjunction
LOT_MEN	Subordinating conjunction
PUNT	Punctuation mark

Syntactic annotation of the data was also performed by means of an automatic annotation tool that is still under development. However, having some kind of syntactic information even with errors is better than not having it at all. This tool groups words into phrases, and classifies them according to their syntactic function. The main errors contained in these syntactic labels were manually corrected: this way, ambiguities between direct objects and subjects were eliminated and some important words (mainly conjunctions) that had been left unlabelled, were assigned a correct value.

4. Statistical analysis

Statistical analysis of the data was made using CARTs [8]. Binary classification trees were trained to predict the presence or absence of a break after each word of the training data.

The corpus was divided into training data (256 sentences, having 3482 inter-word spaces without break and 1140 with break) and test data (111 sentences, having 379 non-breaks and 130 breaks), and the statistics of both sets were calculated to prove that the division made was acceptable.

4.1. Prediction information

To predict the location of the breaks, the following information was provided to the tree:

- Part of speech of the words included in a five word window centered at the word under study.
- Syntactic function of the current word and the surrounding two words.
- Indication of whether next word belongs to the same phrase as the current word: words corresponding to the same phrase are not likely to be separated by a prosodic break.
- Number of words and syllables from last break and number of words and syllables left to the next punctuation sign. Positions very close to a break seem to be less likely positions for a new break. For training, these data are known; for prediction, the tree is applied from left to right to the input text, so the last break is the last one predicted by the tree.
- Length of the current sentence, measured in syllables, to test the hypothesis that longer sentences are uttered with more breaks.

4.2. Prediction tree

Two different types of errors can be distinguished when predicting location of breaks:

- Insertion of a break when it was not present in the reference data. This seems to be the worst type of error, because phrases that were undivided in the original data and thus had a strong relation between them, are divided.
- Deletion of a break present in the reference data. This error does not look as serious as the first one, because the resulting texts do not sound very unnatural with the eliminated breaks.

The first tree built to predict location of breaks was trained considering that both errors had the same importance to calculate the minimum prediction error. After this experiment, another tree was trained assigning 33% more cost to the error of inserting a non-existing break than to the deletion of a break.

The first split of the tree inserts a break if the current label is a punctuation sign. This was expected because the labeling of the corpus inserted a break after every punctuation sign. Then the part of speech of the following word is evaluated and in case of being an adjective, adverb, synthetic verb, subordinating conjunction or punctuation sign, the break is not inserted. If the next word's part of speech is not included in this list, then the evaluation of whether the next word belongs to same syntactic group as the current word is performed. In case both words belong to the same phrase no break is inserted and then. If both words appertain to the same phrase, the number of syllables from the last break is checked:

if it is greater than 9 and the number of syllables left to the next punctuation sign is greater than 5, a break is inserted.

5. Results

The evaluation of the performance of the trees is not easy. The overall score achieved by the tree is calculated as the number of inter-word spaces correctly classified divided by the total number of inter-word spaces.

This datum has to be interpreted carefully, because it depends on the proportion of breaks in the original corpus. In the test data corpus 74.46% of the inter-word spaces were labeled as non-breaks, so an algorithm that does not place any break at all would achieve almost 75% of performance without doing anything. To avoid this problem another way of computing the score of the tree, the kappa statistic, has been proposed. This measure was first suggested for linguistic classification tasks by Carletta [9] and has since been used by others [10] to avoid the dependency of the score on the proportion of non-breaks in the text.

The kappa statistic is calculated as indicated by Equation 1.

$$\kappa = \frac{\Pr(A) - \Pr(E)}{1 - \Pr(E)} \quad (1)$$

where Pr(A) is the overall score attained by the tree and Pr(E) is the proportion of non-breaks in the data.

In expression (1), overall score achieved by the tree is compared with the probability of having a non-break label in the data, eliminating the dependency on the structure of the data. If the algorithm does not insert any break, the value of the kappa statistic will be 0. If the method predicts every inter-word space correctly, κ=1. Values lower than 0 indicate that the breaks placed by the algorithm are in the wrong places, so it is better not to use it.

Table 3 shows the results obtained by the tree having equal cost for insertion and deletion errors when applied to test data. Partial score indicates the proportion of breaks (or non-breaks) over the total number of breaks (or non-breaks) in the reference data. Overall score achieved by this tree is 92.53%. The value of the kappa statistic in this case is 0.71.

Table 3: Performance of the 1st prediction tree

	Predicted non-break	Predicted break	Partial score
Non-break	351	28	92.61%
Break	10	120	92.31%

The results obtained by the tree having higher cost for insertion errors than for deletion errors are shown in Table 4. Overall score achieved by this tree is also 92.53%, but the distribution of errors is different. In this case, there are more deletions and fewer insertions than in the former case. The value of the kappa statistic in this case is also 0.71.

Table 4: Performance of the 2nd prediction tree

	Predicted non-break	Predicted break	Partial score
Non-break	371	8	97.89%
Break	30	100	76.92%

The data presented in both tables correspond to the use of the tree in training mode, i. e., with the values of the number of syllables and words from the last break calculated with the correct positions of the breaks.

6. Discussion

The scores achieved by the trees trained to predict break locations compare positively with the results obtained for other languages:

- Mexican Spanish: [4] achieved a overall score of 94.2% but testing the algorithm over training data.
- English: [10] has an overall score of 90% (κ=0.5) in the best of the proposed usable methods; [5] attains the 91.5% (κ=0.53) using part of speech sequences and a Markov model to give the most likely sequence of phrase breaks and [2] gets 90.8% (κ=0.59) using memory based learning.
- Korean: [11] achieved an overall score of 77.0% (κ=0.56), [12] of 84.9% (κ=0.62) with a method based in CARTs and [13] of 85.5% (κ=0.64).
- Japanese: [14] attained an overall score of 89.8% through the training of a stochastic context free grammar.

These numbers serve to compare different algorithms, but have to be interpreted carefully. All the errors contribute evenly to the calculation of scores, but they do not have the same seriousness. The errors made by the trees have been classified by an expert in two categories: errors to avoid and permissible errors. The distribution of both categories among the insertion and deletion errors is shown in Fig. 2. As it has been previously commented, insertion errors tend to be more serious than deletion errors, and in 84% of the cases are classified in the category of errors to avoid, while deletion errors are acceptable more or less in the same proportion.

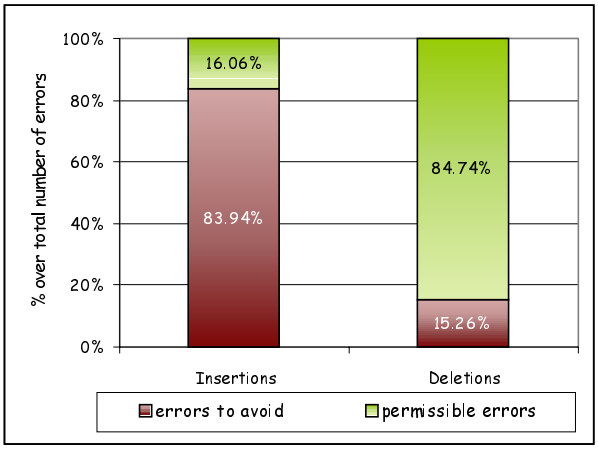

Figure 2: Distribution of seriousness of errors for each type of error.

Both trees built in this study have the same overall score, but they differ in the proportion of insertion and deletion errors they make. Fig. 3 shows the number of errors of each category made by each tree, when predicting the breaks of test data. The total number of errors made does not equal the numbers presented in Tables 3 and 4, because this time, the

number of syllables and words from the last break is calculated based on the locations of breaks predicted by the decision tree. Thus, now the results are prone to be more erroneous, because of the propagation of errors. Looking at the number of serious errors, the first tree has worse behavior than the second one. Hence the second one should be used for the prediction of prosodic breaks.

Figure 3: Classification of errors of each tree.

An analysis of the origin of the errors belonging to the worst category has been made to improve the method. The distribution of the causes of those errors is shown in Table 5.

Table 5: Causes of the "errors to avoid" of the 2nd tree

Nº of errors	Cause
7	Error in syntactic annotation
3	Error due to calculate variables using the breaks already predicted by the tree
1	Error in the syllabification algorithm
1	Error in the part-of-speech tag
1	Break incorrectly labeled in reference data
1	Number of syllables from last break = 10

Looking more closely to the 14 "errors to avoid" made by the 2nd tree, half of them are due to errors in syntactic annotation. This proves the importance of having a good syntactic analysis of the texts for the accurate prediction of the location of breaks. Three errors were due to the propagation of errors resulting from using the breaks predicted by the tree to calculate the variables needed for the prediction. Each of the remaining four errors had a different origin: one was due to an error in the syllabification algorithm that did not consider a diphthong, another one was produced by an error in part-of-speech tagging, another one was the result of an incorrect decision of the tree in a prosodic phrase that had a number of syllables just in the border of decision (10 syllables) and the last one was due to an incorrect label in the reference data.

In spite of the small size of the corpus and the errors present in the labels used, the features chosen to predict prosodic breaks as well as the statistical method selected, have proven to be very valuable, as the achieved results show.

7. Acknowledgements

The authors greatly appreciate the help of Patricia Pérez who worked on the prediction of prosodic breaks from the beginning of the study.

We also acknowledge the finantial support of the University of the Basque Country and the Ministry of Science and Technology (grants UPV147.345-TA066/98 and TIC2000-1005-C03-03).

8. References

[1] Karn, H.,1996. Design and evaluation of a phonological phrase parser for Spanish text-to-speech. *Proceedings of the 4th International Conference on Spoken Language Processing*, Philadelphia, vol. 3,pp. 1696-1699.

[2] Busser, B.; Daelemans, W.; van den Bosch, A., 2001. Predicting phrase breaks with memory-based learning. *4th ISCA Tutorial and Research Workshop on Speech Synthesis*, Edimburgh.

[3] Hirschberg, J., 1991. Using text analysis to predict intonational boundaries. *Second European Conference on Speech Communication and Technology*, Genoa.

[4] Hirschberg, J.; Prieto, P.,1996. Training intonational phrasing rules automatically for English and Spanish text-to-speech. *Speech Communication*, Vol. 18, 281-290.

[5] Black, A.W.; Taylor, P., 1997. Assigning phrase breaks from part-of-speech sequences, *Proceedings of Eurospeech'97*, Rhodes, pp. 995-998.

[6] http://ixa.si.ehu.es

[7] Ezeiza N.; Aduriz I.; Alegria I.; Arriola J.M.; Urizar R., 1998. Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages. *COLING-ACL'98*, Montreal.

[8] Breiman, L.; Friedman, J.H.; Olsen, R.A.; Stone, C. J., 1984. Classification and Regression Trees. *Chapman & Hall*.

[9] Carletta, J. C., 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2), 249-254.

[10] Sanders, E., 1995. Using probabilistic methods to predict phrase boundaries for a text-to-speech system. *Master's thesis*, University of Nijmegen.

[11] Yeon-Jun K.; Yung-Hwan O., 1999. Prosodic Phrasing In Korean; Determine Governor, and Then Split or Not. *Proceedings of Eurospeech'99*, Budapest, pp. 539-542.

[12] Sangho L.; Yung-Hwan O., 1999. Tree-based modeling of prosodic phrasing and segmental duration for Korean TTS systems. *Speech Communication* , vol. 28, pp. 283-300.

[13] Byeongchang K,; Geunbae L., 2000. Decision-Tree based Error Correction for Statistical Phrase Break Prediction in Korean. *The 18th International Conference on Computational Linguistics*.

[14] Fujio S.; Sagisaka Y.; Higuchi N., 1997. Prediction of Major Phrase Boundary Location and Pause Insertion Using a Stochastic Context-free Grammar. In *Computing Prosody: Computational Models for Processing Spontaneous Speech*. Springer, New York.