Phrase Break Prediction Using a Finite State Transducer

Antonio Bonafonte, Pablo Daniel Agüero

Department of Signal Theory and Communication
TALP Research Center
Technical University of Catalonia (UPC), Barcelona, Spain

Abstract

This paper presents a method for phrase break prediction using a finite state transducer. In the literature, several algorithms have been proposed using statistical techniques for predicting phrase breaks. Some of these methods rely on linguistic information, such as syllables, words, part-of-speech, accents, etc. Our proposal is a probabilistic finite state transducer to convert part-of-speech tags into phrase break boundaries. The results show that even using only part-of-speech tags the accuracy is high, which is advantageous because of its simplicity. This method can be extended to include more linguistic features.

1. Introduction

In natural language, people organize information into the discourse using different acoustical clues. One of these clues are phrase break boundaries.

Several studies have been performed to uncover the influence of the strenght of phrase break boundaries on certain speech parameters.

One of the parameters analized in the literature is the lengthening of final segment. Several studies concluded that there is a proportional relationship between the strength of the phrase break, and the duration of the final segment [10].

Phrase break boundaries cause changes in the fundamental frequency contour. Strong variations of the fundamental frequency contour are usually due to the presence of a break.

Energy and silence are also closely related to phrase break boundaries. There is a general declination in the energy contour in the neighborhood of a boundary. Some boundaries can have a pause. This pause can be used for breathing or can be motivated by discourse structure. For example, in the sentence *"Are you going to do <pause> that?"*, the pause is used to emphasize the word **that**.

Text-to-speech synthesis systems perform a conversion of a text into voice. In order to produce a speech signal, several steps are performed between the input (text) and the output (voice). In general, the text is converted into phonemes, adding syllable and accent marks, and annotating the part-of-speech of the words. Then, phrase breaks are predicted. All this information is provided to the modules that predict duration and energy of each phoneme, and the fundamental frequency contour of the sentences. In the case of UPC text-to-speech system, the following step is the selection of the speech segments (unit selection) that are concatenated to produce the final speech signal.

The modules that predict the energy, duration and fundamental frequency contour need information about phrase break boundaries in order to perform their task. As mentioned before, these parameters vary according to the presence of a phrase break.

The accuracy of phrase break prediction is important, because a mistake can not only cause a loss of naturalness and errors in the downstream modules, but the meaning of a sentence can be radically altered.

The problem of prediction can be faced with two different approaches.

The knowledge based approach consists of using a group of experts in the area, which provide information to the system in a variety of formats (rules, trees, etc.) to perform the task. The main advantage of this approach is that the rules are hand-written, and the behaviour of the system is under control. The main drawbacks are coverage problems, maintenance and development time.

In the opposite side, data driven approaches require trained people to annotate corpus, and then machine learning techniques are used to extract information and build classifiers. This approach has several advantages compared to the previous approach:

- **Development time**. This approach has a shorter development time, because the annotation of a corpus , in general, requires less time than building a classifier manually. In addition, everyday more corpora are publicly available.
- **Domain and speaker adaptation**. The adaptation to a different domain, speaker or even to other similar language only requires a corpus related to the task to be accomplished, and minor changes in the features used in the classifier (tuning).
- **Extraction of new information**. Machine learning techniques allow to uncover some regularities in the task, that were not previously known.

The speech synthesis system of Universitat Politècnica de Catalunya is designed for multilingual purposes. A short development time and fast adaptation to different domains are required. As a consequence, machine learning techniques are used to perform different tasks in the process of conversion of text into speech.

Several methods have been proposed in the literature to overcome the problem of automatic phrase break prediction.

Hirschberg et al. [5] proposed the use of classification and regression trees to predict phrase break boundaries from text. The features include a context of part-of-speech tags, number of words and syllables counting from the beginning and end of the utterance, total number of words and syllables in the utterance, etc. Kohen et al. [6] improved the previous algorithm adding syntactic information. The main problem of these approaches is that the decision is taken locally for each word. As a consequence, the previous decisions about the existence of a phrase break boundary are not taken into account. Navas et al.

[7] describes a similar system for phrase break boundary prediction applied to Basque language, using a classification and regression tree and morphological and syntactic features.

Black et al. [1] and Sun et al. [9] describe a method for phrase break prediction that combines a statistical model and a classification and regression tree. The tree is used to estimate the probability of a phrase break boundary taking into account a window of part-of-speech tags. The combination of local decision with a language model of phrase break boundaries enables a more reliable prediction.

Sanders et al. [8] proposed several algorithms to predict phrase break boundaries, using a window of three part-of-speech tags to establish the probability of existence of a phrase break boundary in a place. These methods perform exhaustive searchs, making some assumptions in order to increase the speed of the algorithms.

An analysis of the previous approaches show that it is common to calculate the probability of a phrase break boundary taking into account context information (for example, part-of-speech tags) and the history of previous decisions.

In this paper we propose a method for phrase break prediction using a finite state transducer (FST). The FST performs the conversion from part-of-speech tags into phrase break boundaries. The FST models the joint probability of a phrase break according to the context information and previous decisions. The main advantage of this method is its simplicity. The input information is part-of-speech tags, without needing more complex information as in [5].

The main drawback of this method is that it relies on information provided by a part-of-speech tagger. The part-of-speech tagger may have some mistakes. As a consequence, the performace of phrase break prediction is degraded due to error in the input tags.

Section 2 details the proposed algorithm. Section 3 describes a Spanish corpora and how the accuracy is high. However, unfortunately, the Spanish corpus has been tagged automatically using a simple tagger. To validate the experimental results, it will be shown the prediction accuracy taking into account different part-of-speech tagging accuracies. Finally, section 4 shows the conclusions of this work.

2. Algorithm description

The transducer in this work performs a conversion of part-of-speech tags into phrase break boundary tags. In the training step, the transducer is given a sequence of pairs of part-of-speech - phrase break boundary tags:

$$(p_1, b_1)(p_2, b_2)...(p_n, b_n) \tag{1}$$

where p_i is the part-of-speech tag of word w_i, and b_i indicates a phrase break boundary tag (B) or no phrase break boundary tag ($\neg B$) after the word w_i.

The task of the transducer is to find the sequence of phrase break boundary tags that maximize the equation 2.

$$argmax_b P(b/p) = argmax_b \frac{P(b,p)}{P(p)} = argmax_b P(b,p) \tag{2}$$

$P(b,p)$ is the joint probability of a sequence of part-of-speech and phrase break boundary tags. This can be modeled using n-grams, as shown in equation 3.

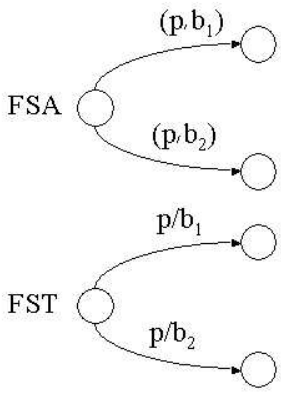

Figure 1: *FSA and FST.*

$$P(b,p) = \prod_{i=1}^{N} P(b_i, p_i / b_{i-k}^{i-1}, p_{i-k}^{i-1}) \tag{3}$$

The language model obtained with the n-grams can be represented as a finite state automata (FSA). Each state represents a history $(b_{i-k}^{i}, p_{i-k}^{i})$ and the arcs contain the conditional probability of an observation given the previous history ($P(b_i, p_i / b_{i-k}^{i-1}, p_{i-k}^{i-1})$). In this way, the joint probability of a sequence of observations can be obtained travelling the finite state automata given the observations, as shown in equation 4.

$P(b,p) = P(b_1, p_1).P(b_2, p_2 / b_1, p_1).$

$$.P(b_3, p_3 / (b_1, p_1)(b_2, p_2))... \tag{4}$$

In this paper, n-grams are estimated using variable length n-grams [2]. The basic idea is that states with history $(w_{t-m}...w_t)$ are candidates to be merged with states $(w_{t-m+1}...w_t)$, in order to obtain more reliable probabilities for longer histories. The criteria employed to take this decision are:

- The states are merged if the number of times that the history $(w_{t-m}...w_t)$ has been observed on the training data is below a threshold.
- The states are merged if the information of distribution $p = p(w/w_{t-m}...w_t)$ is similar to that of distribution $p = p(w/w_{t-m+1}...w_t)$.

The n-grams are smoothed using linear-discounting and back-off.

The FSA is converted into a FST, taking into account that the observation of a part-of-speech p_i produces an output b_i, as shown in Figure 1. Given the inputs p_i, there are several possible paths in the FST that can be travelled with the sequence p. Viterbi decoding is used to obtain the path that maximizes $P(b/p)$. Given the optimal state sequence, it is posible to obtain the phrase break boundary tags (b_i) that correspond to the best path through the FST.

FST's have been used in several tasks, such as phonetic transcription[4] and machine language translation [3]. These tasks are more complex, because in some cases there is a mapping of many-to-many from input to output. In addition, in some cases the output sequence has a different order than the input.

In this approach we decided to use part-of-speech as input due to two reasons:

Text	Input	Output
The	DT	$\neg B$
king	NN	B
is	VBZ	$\neg B$
playing	VBG	$\neg B$
the	DT	$\neg B$
piano	NN	$\neg B$
,	FC	B
while	IN	$\neg B$
the	DT	$\neg B$
queen	NN	$\neg B$
sings	VBZ	$\neg B$
.	FP	B

Table 1: Inputs and outputs of the FST.

- **Reduction of the size of the input space**. The part-of-speech tags are used instead of words. The use of words would cause a need of a huge amount of corpus in order to obtain reliable probability estimations.
- **Relationship between part-of-speech tags and phrase breaks**. Several works in the area have shown that part-of-speech tags are an important source of information to decide the placement of a phrase break boundary [1, 5].

The information given to the transducer is shown in Table 1. The labels of the output are $\neg B$ (no phrase break boundary tag) or B (phrase break boundary tag). The phrase break position is associated to the end of the words. Punctuation marks are considered words, in order to model the phrase break boundaries that are generally present after some punctuation marks.

3. Experiments

3.1. Corpus

The Spanish and English corpora were manually tagged with phrase break boundary tags, without using recorded speech. In this way, the tags were annotated using only the text. In the Spanish corpus, the part-of-speech information is predicted using a part-of-speech tagger (reduced set of PAROLE tags). The English corpus has manually annotated part-of-speech tags (WSJ corpus).

The size of the phrase break boundary corpus in Spanish is 100Kw. and the size of the corpus in English is 50Kw (only a part of WSJ corpus has been tagged with phrase break boundaries).

3.2. Experimental measures

The experiments are performed to obtain quality measures of the algorithm.

A missing or a misplaced phrase break can cause a change in the meaning of the sentence, or a loss of naturalness. In order to measure the quality on these tasks, we use:

- **Precision of phrase break prediction** is the number of phrase break boundary tags that the system predicted correctly. (equation 5).
- **Precision of no phrase break prediction** is the number of no phrase break boundary tags that the system predicted correctly (equation 6).
- **Phrase break prediction recall** is the number of phrase break boundary tags that the system detected correctly from the total number of manually annotated phrase break boundary tags (equation 7).

In order to have a joint measure that balances precision and recall, we use the F-Measure (equation 8).

	Manual B	Manual $\neg B$
Predicted B	t_B	f_B
Predicted $\neg B$	$f_{\neg B}$	$t_{\neg B}$

$$Precision_B = \frac{t_B}{t_B + f_B} \quad (5)$$

$$Precision_{\neg B} = \frac{t_{\neg B}}{t_{\neg B} + f_{\neg B}} \quad (6)$$

$$Recall_B = \frac{t_B}{t_B + f_{\neg B}} \quad (7)$$

$$F = 2\frac{Precision_B Recall_B}{Precision_B + Recall_B} \quad (8)$$

3.3. Experimental results

The experiments are performed using 70% for training and 30% for testing purposes.

The results with the Spanish corpus are shown in Table 2. The results are better than the experiments we performed using [5] using the same corpora (F-measure=75.45). As a consequence, the results in Spanish are high using only part-of-speech tags.

Exp.	$\neg B$ prec.	B prec.	Rec.	F
Spanish	95.88	77.52	75.64	76.57

Table 2: Results of the experiment with the Spanish corpus.

In order to analyze the influence of part-of-speech tagger accuracy in the phrase break boundary F-measure, we performed some experiments with the English corpus.

Table 3 shows the results using part-of-speech taggers for English. The part-of-speech taggers were trained with 10Kw (74% of accuracy) and 100Kw (89% of accuracy). The results show that the performance of the phrase break prediction is degraded because of the inferior accuracy of part-of-speech tags.

The results using manually annotated tags (EngMan) are the best, because of the accuracy of the input information.

Exp.	$\neg B$ prec.	B prec.	Rec.	F
EngMan	94.48	81.33	77.60	79.42
Eng100Kw	94.09	80.09	75.50	77.72
Eng10Kw	93.21	76.88	74.78	75.81

Table 3: Results of the experiments with the English corpus, introducing part-of-speech errors for English phrase break prediction.

3.4. Phrase break prediction accuracy according to manual check

Experimental evaluation of the agreement of two human labelers of phrase break boundaries in written text (no sound information) show differences between them. As a consequence, the

evaluation of phrase break prediction accuracy needs a manual check of the sentences. In some cases, the labels provided by the phrase break prediction algorithm can be valid alternatives to the human labeler version.

The analysis is performed taking into account only the sentences that are fully correct. An error in a part of a sentence can cause a change of the meaning of the entire sentence.

The results of the manual verification show that the accuracy of the phrase break prediction rises to 89% and the no phrase break prediction accuracy rises to 96%.

4. Conclusions

Phrase break prediction is a very important task into a text-to-speech system, because many other tasks require this information to achieve more natural speech synthesis: phone duration, pauses, energy and fundamental frequency contour prediction.

In this paper we proposed a phrase break prediction algorithm using a finite state transducer. Phrase break prediction is performed using part-of-speech tags as inputs. The finite state transducer converts these inputs into phrase break boundaries (output). This approach has the advantage of its simplicity. It does not rely on additional information such as syllables, accents, etc., used in [5]. The main drawback is the dependency of the accuracy of phrase break prediction on the accuracy of part-of-speech tagger.

We expect that the use of additional information will improve the performance of the transducer. However, it will make more complex the estimation of the probabilities of the sequences.

5. Acknowledgements

This work has been partially sponsored by the European Union under grant FP6-506738 (TC-STAR project, http://www.tc-star.org) and the Spanish Government under grant TIC2002-04447-C02 (ALIADO project, http://gps-tsc.upc.es/veu/aliado).

6. References

[1] A. Black and P. Taylor, "Assigning phrase breaks from part-of-speech sequences," *Computer, Speech and Language*, vol. 12, pp. 99–117, 1998.

[2] A. Bonafonte and J. B. Mariño, "Language modeling using x-grams," *Proceedings of ICSLP*, pp. 394–397, 1996.

[3] A. de Gispert and J. B. Mariño, "Using x-grams for speech-to-speech translation," *Proceedings of ICSLP*, pp. 1885–1888, 2002.

[4] L. Galescu and J. F. Allen, "Bi-directional conversion between graphemes and phonemes using a joint n-gram model," *Proceedings of the 4th ISCA workshop on Speech Synthesis*, pp. 103–108, 2001.

[5] 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.

[6] P. Koehn, S. Abney, J. Hirschberg, and M. Collins, "Improving intonational phrasing with syntactic information," *Proceedings of ICASSP*, pp. 1289–1290, 2000.

[7] E. Navas, I. Hernaez, and N. Ezeiza, "Assigning phrase breaks using CART's in Basque TTS," *Proceedings of the 1st International Conference on Speech Prosody*, pp. 527–531, 2002.

[8] E. Sanders and P. Taylor, "Using statistical models to predict phrase boundaries for speech synthesis," *Proceedings of Eurospeech*, pp. 1811–1814, 1995.

[9] X. Sun and T. H. Applebaum, "Intonational phrase break prediction using decision tree and n-gram model," *Proceedings of 7th European Conference on Speech Communication and Technology*, pp. 537–540, 2001.

[10] C. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, and P. Price, "Segmental durations in the vicinity of prosodic phrase boundaries," *Journal of the Acoustical Society of America*, vol. 92, pp. 1707–1717, 1992.