PAUSES LOCATION AND DURATION CALCULATED WITH SYNTACTIC DEPENDENCIES AND TEXTUAL CONSIDERATIONS FOR T.T.S. SYSTEM

Gérald Vannier[†], Anne Lacheret-Dujour[‡], Jacques Vergne[†]
[†]*GREYC University of Caen, France,* [‡]*ELSAP University of Caen, France*

ABSTRACT

Punctuation is essential in predicting pauses in text-to-speech systems, but it is not sufficient. When reading, silences may be produced without any graphemic indication in the text. In this communication, we offer to introduce a method to calculate pauses occurences and durations according to first, textual constraints and second, syntactic dependencies. We present the role of textual hierarchy for pauses durations and the importance of the syntactic relations from automatic syntactic analysis which allow to indicate the pauses locations and durations inside sentences. The results of the reported work are implemented in a retailed French text-to-speech system.

INTRODUCTION

While punctuation is essential in predicting pauses in text-to-speech system, it is not sufficient: when reading, silences may be produced without any graphemic indication in the text, therefore syntactic cues are also determinant. First, we set out some observations on text corpora and we propose a hierarchical model for the prediction of the durations of pauses which are marked by punctuations (section 2.1). In other words, the processing of marked pauses prooves the reality of a textual prosody. Second, we will focus on pauses within sentences, and we will explain our method for silence automatic generation (section 2.2). Our work hypothesis are:

(i) Textual marks allow us to produce text prosody.

(ii) Within a sentence, pauses location and duration are linked to syntactic dependency: further are syntactic units in relation more important are pauses duration.

(iii) We have to consider first pauses through syntactic approach, before using some phonotactic indications.

Finally, we will present our temporal rules for French based on the analysis of two texts read twice (section 3).

The results of the reported work[1] are implemented in an operational French text-to-speech system, but the background concepts are more general.

1. CORPORA

Our goal was to build a text-to-speech synthesis system for the blind. Therefore, the prosodic study presented here is exclusively based on the analysis of a particular pragmatic task: the reading of French texts at a normal rate. We will observe corpora to bring out some general properties which could be implemented, and tested in perceptual contexts.

1.1 CHOOSEN TEXTS

Two texts, each of about 500 words have been read by one Parisian native speaker (25 years old), no familiar with the problems of speech synthesis and not informed of our goals. The first text is an extract of a contemporary novel (corpus CB), the other is a press article from the newspaper "le Monde" (corpus CM). Both texts have been chosen according to our task (no phonotactic, morphologic nor syntactic constraints *a priori*). Each text was read twice (two weeks separated each reading) in order to test the phonetic consistency (particularly regarding with speech rate) and to ponder over our results taking intra-speaker variability into acount. Speech signals were digitally recorded in a soundproof booth (sampling frequency 16 KHz).

1.2 DATA BASE COMPOSITION

Prosodic events have been manually extracted using Unice (software from Limsi) and Phonedit (www.sqlab.com) software environments. Our phonetic analysis was focused on (i) syllabic segmentation, (ii) segments length extraction and (iii) pauses measure and (iv) choice of relevant melodic target values.

(i) The phonetic syllables were defined in terms of length and various melodic configurations. The manual segmentation of each text in a string of syllables has been processed according to acoustic and perceptual cues.

(ii) For each syllable, a raw length has been calculated independently of the phonematic events which constitute the syllable; knowing that general tendencies which emerge will have to be precise according to the segmental influence on prosody [3, 8].

(iii) Pauses measurements consist on silence (no signal) detection, based on visual and perceptual approach.

(iv) The extraction of melodic targets had to take the detection errors into account.

2. ANALYSIS

In our corpora, about 430 pauses occur with a relative stable distribution and duration. The residual variation highlights the impossibility of developping a deterministic system of pauses duration prediction which copies exactly human strategies: we must make choices according to constraints on different levels, textual, pragmatic, syntactic and phonotactic.

2.1 A HIERARCHICAL TEXTUAL MODEL FOR THE PREDICTION OF PAUSES DURATION

In available French text-to-speech systems, pauses prediction is systematically based on punctuation markers and phonotactic criteria to generate different distributions and pauses length [2]. Variation of length is often used to differentiate between pauses after a punctuation marker inside a sentence and pauses which mark a full stop. This minimalist processing can be performed taking enunciative structuration of a text (types of paragraphs, location of a paragraph in a texte, etc.) also into acccount. Regarding phonotactic criteria, they are used to fix a temporal interval (in temporal unit or number of syllables) between two

pauses. Of course, this temporal interval must respect syntactic constraints (eg. a pause is forbidden between a determiner and a noun). In fact, punctuations in a text are not resricted to full stop and comma: all non alphanumeric caracters which give informations about textual structure are concerned (Table 1).

.	?	…	"	!	(	)	,	:	-	_	¶

Table 1: punctuations in corpora

The analysis of the link between punctuations reported in Table 1, and pauses distribution, leads to the following points:

(i) 4,6% of these punctuations are not instanciated by a pause. This punctuations occur in short sentences (eg: *Suzanne vous attend, elle aussi*) or when they are used in nominal contrastive structures (eg: *service national pour les femmes, grossesse pour les hommes*). In the first example, the enunciative operation of extraction (*elle aussi* corresponds to a given information moved at the end of a sentence [8]) is instanciated by specific modulations of the melodic curve. Melodic variations are also used to indicate the contrastive structure of the second example. Therefore, the adjunction of a pause would be redundant for the understanding of both structures.

(ii) The graphemic punctuation can be phonetically instanciated by a right transfer (eg: *ce qui lui plaisait chez Fouquet, c'était # qu'il ne semblait pas…*), here again, enunciative constraints explain the location of the pause.

(iii) Quotation marks are not necessarily instanciated by a pause (eg: *la femme accomplit son "devoir national"# en étant…*). For all the occurrences of pair found in the corpora, at least one of each pair is instanciated.

In other cases, full stops and paragraphs marks are stable cues for pauses prediction. On the other hand, a great number of pauses do not actualise each kind of punctuation. There, distribution must be analysed.

The raw values are supposed to be speaker dependent, we aim to bring out reading strategy by means of relative considerations. In Fig. 1, pauses are distinguished in inner sentences pauses (lozenges), pauses between sentences (triangles) and pauses between paragraphs (circles).

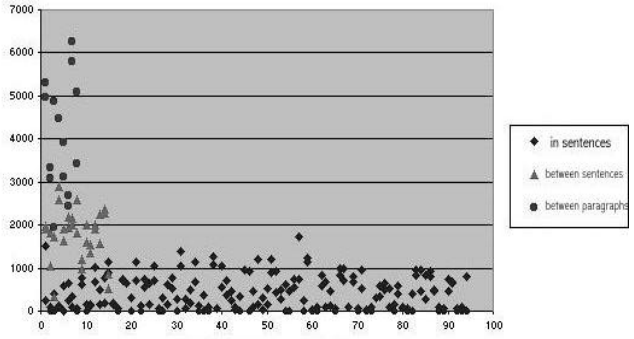


Figure 1: Pauses distribution in CM

We are interested in the distribution of pauses along the vertical axis which corresponds to duration axis (horizontal axis is the numbers of occurrences). Three areas stand out which correspond to the three categories of pauses: inner sentence pauses with duration smaller than 1500, pauses between sentences stretches between 1500 and 3000, pauses between paragraphs over 2500 (3000).

There is a textual organization of prosody which is reflected by pauses durations: "Word < Sentence < Paragraph" is found in the distribution "Pauses duration between words < Pauses duration between sentences < Pauses duration between paragraphs".

We calculate averages of durations of this different pauses categories, as shown in Table 2.

Corpus	***CB***	***CM***
In sentences	505	418
Between sentences	1550	1771
Between paragraphs	2755	4072

Table 2: Duration averages of pauses

Note the distance in paragraph pauses averages in *CM* and *CB*. Two considerations can explain this difference: the different speech rate of the speaker and the really different style of the two texts. Only a specific study could decide.

The differences of averages between the three paradigms are important. They confirm the hierarchical organization: hierarchy of syntactic units, hierarchy of frontiers between these units, hierarchy of pauses duration. The higher the frontier in the hierarchy, the more important the duration of the associated pause.

We have also two remarks for the Fig. 1. First, the area boundaries are not as accurate as expected. We can see some overlaps: some pauses between sentences have durations near the ones of paragraphs pauses. However, paragraphs pauses durations are always greater than durations of pauses in sentences. Second, there is a great dispersion among values of each area. We tried to explain this phenomenon with different experiments, in particular, we compared the durations of pauses between paragraphs with the length of the last sentences (in number of words or syntactic groups) or with the length of the paragraphs (in number of sentences). We failed in isolating a relevant parameter.

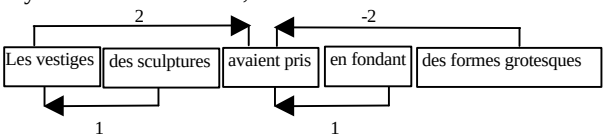
2.2 PAUSES IN SENTENCES AND DEPENDENCE RELATIONS

Our syntactic analysis is based on the bakground of Vergne's parser, implemented at the French university of Caen (www.info.unicaen.fr/~jvergne). Contrary to usual approaches, it is not based on expliciting the syntactic structures but it tries to modelize the process of local deduction propagation. The analysis is processed in two steps: in the first, we assign syntactic categories and segment sentences in syntactic groups, the second consists in linking these syntactic groups.

To introduce our view on syntactic analysis, we have to define the syntactic group. It is an homogeneous syntactic constituent that can be assimilated to a non-recursive phrase. A first level of analysis allows a segmentation in syntactic groups that we divide in verbal (< >) and nominal groups ({ }):

{Il} <s'aperçut que> {son tableau} {à tiroirs} <faisait> {peu de place} {au visage} {de la femme}.

For more details on tagging we can refer to [4, 9]. We do not explain here the linking process, but we discuss and explain the results of this level. The illustration below, a result of our analysis on a real sentence, will be commented.



Arrows show the groups which are linked, and the direction of the links. The values associated are the number of groups to jump from the first group to the linked group. Then, the link between two consecutive groups carried the value 1 or -1 ("*les vestiges*" and "*des sculptures*"). In the example, we have to glance through 2 groups between "*Les sculptures*" and "*avaient pris*", and the value for the link will be 2 or -2 according to the direction. This value is independent of the nature of the relation. We choose, in an arbitrary way, the reading/writing direction as positive direction.

Inside sentences, we observed many pauses without punctuation. In this paragraph, we focused on this kind of pause, through the results of the syntactic analysis, especially through the values of the links between the syntactic groups.

Lot of works highlight a correlation in different languages, between the strength of the syntactic boundaries and the pauses duration. So we propose, first, to define a scale of syntactic boundaries strength, in the context of our analysis. The syntactic links, then the dependence relations, will be associated with the boundaries strength. Some recent studies, on German, seem to confirm our views [1]. In the example above, we give linking associated values:

Les vestiges **|*-1*|** *des sculptures* **|*2*|** *avaient pris* **|*-1*|** *en fondant* **|*-2*|** *des formes grotesques.*

The boundaries strength is calculated from the linking values. The indicated strength is the absolute value of relevant links. We consider the strength of a boundary only as potential candidate for pause occurence. **A relevant link is a link which is on the group immediatly before or after a pause, and which crosses over the pause.** The implied hypothesis is: the longer the link, the stronger the generated prosodic mark. The contiguous relation means a weak prosodic mark. We intend to confirm this hypothesis.

We use this method to associate a pause duration with a boundary strength. In the next sentence, from *CB*, a boundary strength value follows each pause occurence:

*Dans un brouillard confus,**#P1** |+2| Quentin entrevit une manière de tableau,**#P2** |-1| plus chaud qu'une allégorie**#P3** |-3| et plein de bruits de foules,**#P4** |-6| où un père et son fils trinquaient à l'envi,**#P5** |-3| soudés par le même secret;*

The *#Pi* indicates the occurence and the number of each pause. The |X| are boundary values. The first value |+2| marks the link between "*Dans … confus*" and "*entrevit*", similarly "*où*" is distant of 6 groups to "*une manière"*.

Thus we could associate a pause duration with some syntactic knowledge, but we do not take punctuation into account. We said the importance of comma, so we have to re-introduce the punctuation. A comma strengthens a boundary; **the strength of boundary with a comma, will be incremented of two units (experimentaly determinated value).** Let us go back to our example:

*Dans un brouillard confus,**#P1** |+2|+2 Quentin entrevit une manière de tableau,**#P2** |-1| +2 plus chaud qu'une allégorie**#P3** |-3| et plein de bruits de foules,**#P4** |-6|+2 où un père et son fils trinquaient à l'envi,**#P5** |-3|+2 soudés par le même secret;*

We associate the value 2+2 to pause P1, 1+2 to the pause P2, etc… The P3 strength does not change because there is no punctuation. So, as described, for all corpus, we calculate boundaries strengths and note all associated pauses to obtain the graphic Fig. 2 (for *CM*):

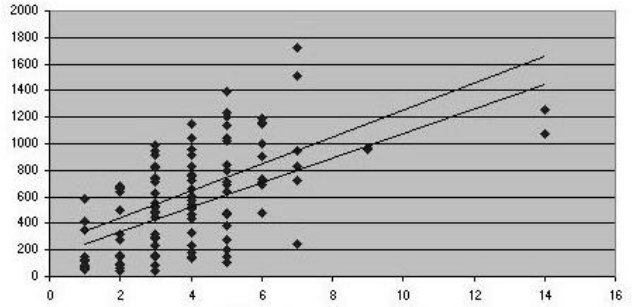


Figure 2: Pauses durations and boundaries strengths for CM1 and CM2

This figure represents the pauses durations (vertical axis) according to boundaries strengths (horizontal axis). We join tendancy curves for each corpus. These curves seem parallel, then the results comparable, even if durations are more important in one version. The graphic shows too that the stronger the boundaries, the longer the pauses. In other words, the further apart two linked groups are, the more important the associated pause is. The correlation seems to follow a linear progression.

Hence, in TTS context, we can predict pauses duration on syntactic cues, and we have now to predict the occurence of these pauses. It brings out this question: "Where do we have inserted pauses without punctuation? " or "Are syntactic relations good cues to predict pauses occurences?".

Our system predicts pause occurences from cues such as syntactic relations. Each not-contiguous relation will engender a pause. If we take again the first example, we will generate two pauses: *Les vestiges des sculptures # avaient pris en fondant # des formes grotesques.*

The process is the reverse of the analysis process first described. This method gives a prediction rate of about 57 % (*CB*) and 70% (*CM*) compared to stable pauses. We introduce the concept of stable pauses for the pauses that appear in both read versions of our corpus. Considering unstable pauses, our predicting rates reach 65% for *CB* and 45% for *CM*.

This result, considered roughly, could seem weak. However, the rate of stable pauses is only about 40% of total pauses in *CB* and 70% in *CM*. There is a great variation in pauses duration but in number of generated pauses too. A third reading of our corpus would decrease these rates. The great

variation constitutes a strong difficulty for observations, automatic processing and validation of results. So we have to practice some perceptual tests. We have not made formal tests yet, but some encouraging feedbacks have been provided from end users.

3. APPLICATION

A closer comparison with real productions underlines some weakness in our system. It does not produce pauses when reader do. For example, it does not generate a pause between syntactic groups in contiguous relation; but we can observe some pauses under these conditions (*eg: Quentin # avait maintenant…*). On the other hand, a long sentence may be made up of syntactic groups only in contiguous relations ("*Ils se turent presque jusqu'au jardin de l'hotel où l'autre lacha Fouquet."*). In this case, our system does not produce any pause, but our reader did. The only cues for the segmentation are enonciative or phonotactic, we do not implement solutions yet for these problems.

Sometimes our pauses generation process produces pauses which could be debatable (remarks from users). It cut short sentences like: "*il voit le chien # dans le jardin*". So, phonotactic constraints, such as a limit number of syllables (7 in our system) have been included to avoid this type of segmentations.

3.1 HIERARCHICAL TEXTUAL PROSODIC MODEL

Our observations show the necessity to modelize text structure; so we consider texts as lists of paragraphs (title is a paragraph), paragraphs as lists of sentences, etc… We would wonder if a superior modeling is necessary (chapters or pages for example).

At the moment, we need this hierarchical text structure to predict pauses duration: between sentences pauses duration < between paragraphs pauses duration < between words pauses duration.

The variations of sentences and paragraphs pauses are a real difficulty. We will have to consider the perceptual contribution for this phenomenon. But, our intermediate solution consists in a controled random process. We propose to generate pauses inside a time interval. Intervals are choosen from observations and respect the textual hierarchy (Fig. 1). This method could be improved further, with the introduction of relevant parameters to calculate these intervals.

3.2 SYNTACTIC PREDICTING MODEL

For pauses inside sentences, the durations respect the hierarchical organization, but the variation is not processed with a controled random. Our approach, in generating durations based on boundary strengths, gives different values enough to generate various durations.

First, each not-contiguous relation engenders a pause. As described, durations are evaluated from relation values and possibly enforced by punctuations marks. We follow a linear relation between durations and frontier strength.

$D = a \times | \text{Strenth of frontier} | + b$

The parameters a and b are constant for a given speech rate, they can be evaluated from the tendancy curves (Fig. 2). Our corpora indicate about 90 for parameter a and 150 for parameter b, in slow rate. That produces pauses about 300 ms for a strength which correponds to a comma (value 2). We have also to limit pauses durations within interval which corresponds to pauses within sentences, as in Fig. 1.

This process of pause generation has to be completed with phonotactic processing. We fix a limit of 7 syllables; if the interval between two related syntactic groups (not contiguous) is less than 7 syllables, the pause is not generated. So we can say that syntactic analysis predicts some pauses in a text, and that pauses are actualized according to phontotactic contraints.

We make the hypothesis that the different durations of pauses calculated with these methods and the prediction of pauses occurrences help the listener to rebuild the syntactic and textual structure and more to decode utterances.

4. CONCLUSION

We have shown that, in TTS context, punctuations are important cues for pauses prediction. Some punctuation considerations lead us to propose a duration hierarchy for pauses, which reflects the textual hierarchy (or boundary hierarchy).

Then, an introduction with our syntactic system gives informations to evaluate boundaries strengths inside sentences. This approach allows us to compare pauses durations and syntactic values. This comparison shows a linear relation between the parameters.

We implemented some proposed solutions, and it appears to be necessary to complete our model with phonotactic considerations. This implementation is based on syntactic concepts and textual representation, which are quite universal concepts: and can be exported to other languages.

NOTES

1. This work was financed by a FEDER (a european financing), a local company and a national association of blinds (Club Micro-Son).

REFERENCES

[1] Alter K., Matiasek J., Steinhauer K., Pirker H., Friederici H.D. (1998), "Exploiting Syntactic Dependencies for German Prosody: Evidence from Speech Production and Perception", KonVeNs'98, Bonn, Germany.

[2] Beaugendre F., Lacheret-Dujour A. (1993), "Automatic generation of French intonation based on a perceptual study and morpho-syntactic information", in proceedings of European Conference On Speech Communication and Technology (Eurospeech 93), Berlin, Germany, vol. 2.

[3] Di Cristo A., Di Cristo P., Véronis J. (1998), "Optimisation d'un modèle prosodique pour la synthèse par règles à partir du texte en français", XXIIèmes Journées d'Etudes sur la Parole, Martigny, Switzland, pp.135-137.

[4] Giguet E. (1998), "Méthode pour l'analyse automatique de structures formelles sur documents multilingues", Thèse de Doctorat d'informatique de l'Université de Caen.

[5] Grosjean F., Deschamps A. (1975), "Analyse contrastive des variables temporelles de l'anglais et du français: vitesse de parole et variables composantes, phénomènes d'hésitation", Phonetica, 31, pp. 144-184.

[6] Hirst D.J., Di Cristo A. (1993), "Intonation Systems. A Survey of Twenty Languages", Cambridge University Press.

[7] Lacheret A., Beaugendre F. (1999), "La prosodie du français", édition du CNRS, Paris.

[8] Rossi M., Di Cristo A., Hirst D., Martin P., Nishinuma Y. (1981), "L'intonation, de l'acoustique à la sémantique", Klincksieck Press, Paris.

[9] Vergne J. (1998), "Entre arbre de dépendance et ordre linéaire, les deux processus de transformation: linéarisation et reconstruction de l'arbre", Cahiers de Grammaire n° 23, Toulouse, France.