Prosodic chunking algorithm for dictation with the use of speech synthesis

Sebastien Le Maguer[1], *Elisabeth Delais-Roussarie*[2], *Nelly Barbot*[1], *Mathieu Avanzi*[2], *Olivier Rosec*[3], *Damien Lolive*[1]

[1]IRISA, Université de Rennes 1, Lannion, France
[2] UMR 7110-Laboratoire de Linguistique Formelle, Université Paris-Diderot, France
[3] VOXYGEN, Lannion, France

Sebastien.Le_maguer@irisa.fr, Elisabeth.roussarie@wanadoo.fr, Nelly.Barbot@irisa.fr, mathieu.avanzi@gmail.com, olivier.rosec@voxygen.fr, Damien.Lolive@irisa.fr

10.21437/SpeechProsody.2014-205

Abstract

The aim of this paper is to present an algorithm that automatically segment a text in prosodic chunks for a dictation by conforming to the rules and procedures used in real settings to dictate a text to primary school children. A better understanding and modeling of these rules and procedures is crucial to develop robust automatic tools that could be used in autonomy by children to improve their spelling skills through dictation with the use of speech synthesis. The different steps used to derive the prosodic chunks from a given text will be explained through concrete examples. The proposal made here relies on the analysis of a corpus of 10 dictations given to children in French and French Canadian elementary schools, and more precisely during their first three years in elementary school (i.e. cycle 2 in the French school system). The phrasing observed in the data is described. It is thus simplified in order to develop an algorithm that automatically generates prosodic chunks from texts.
Index Terms: speech synthesis, prosodic phrasing, automatic parsing.

1. Introduction

The use of software and automatic tools in language teaching offers several advantages among which we may mention the ability to adapt to the learner needs and to provide an environement for him to work in autonomy. Despite these advantages, there are nowadays in France few softwares which are currently used in primary schools to teach reading and writing skills with the use of speech synthesis (e.g. Lectramini[1] and PLATON[2]).

One of the goal of the collaborative ANR research project Phorevox (`http://www.phorevox.fr/`) is to develop this kind of tools, with special emphasis given to the acquisition of writing skills. The software will automatically propose some practice exercises which allows children improving their written skills. Thus, dictations may be given to the children to work on specific grammatical or orthographical aspects.

In order to use automatic procedures and speech synthesis systems for dictation, it is necessary to (i) provide a very intelligible synthesized speech to allow children to hear all the words and sounds to be written, (ii) divide the texts to dictate into chunks that allow accessing all relevant grammatical information, while being of a reasonable size, and (iii) provide a user-friendly typing environment that takes into account the typing speed of the different children. Among these issues, we will focus in this paper on the ability to automatically divide a text to dictate into chunks. To develop an automatic chunking procedure for dictation, we have first analyzed a set of dictations made in primary school. The results of the data observation were used to select and formalize the rules used by the algorithm which generates the dictation.

The paper is organized as follows. In section 2, the rules that are used to derive the prosodic chunks in standard French are briefly presented, and the main characteristics of prosodic phrasing in French are described. The segmentation procedure used by our speech synthesis system are then explained. In section 3, the data and methodology chosen to study the phrasing patterns of dictations in French are presented. The phrasing rules extracted from the observation of the data are listed in section 4. Section 5 explains which pieces of information are taken into account to develop the algorithm that automatically provides a chunking to any text.

2. Background

Studies on French prosody have traditionally pointed out that accentuation, phrasing and intonation are closely intertwined (see, among others, [3], [4] and [5]). In French, the lack of lexical stress causes a syncretism between intonation and accentuation.

In the studies on prosodic phrasing, two or three distinct levels of phrasing above the word are argued for. The lower level, i.e the minor phrase (MiP) – which is also called accentual phrase (see [6] and [7], among others), phonological phrase (see [4] among others) or rhythmic group (see [8] among others) – plays a crucial role. This unit is characterized by the realization of a phrasal stress on its last metrical syllable, which indicates its right edge. In the literature, there is a broad consensus about the definition of this unit: it corresponds minimally to a lexical word and to all the function words that this word governs (see, among others, [4], [6], [7] and [8]). The sentences in (1) are segmented in Minor Phrases as shown in (2).

(1) a. *Les enfants sont venus dans l'après-midi.*
The children came in the afternoon.

b. *Bernard est rentré de son voyage en Asie.*
Bernard came back from his travel to Asia.

(2) a. (Les enfants)$_{MiP}$ (sont venus)$_{MiP}$ (dans l'après-midi)$_{MiP}$.

b. (Bernard)$_{MiP}$ (est rentré)$_{MiP}$ (de son voyage)$_{MiP}$ (en Asie)$_{MiP}$.

In addition to the level of the MaP, two additional levels of phrasing are often referred to: the intermediate phrase or ip (see,

among others, [7] and [9]), which is also called major phrase (see, among others, [10]) or restructured phonological phrase (as in [4]); and the intonational phrase or IP (see, among others, [3], [4], [6] and [7]), which is also called the Breath group. Even if there is no broad consensus on the existence of the ip, this level of phrasing is often requested when the morphosyntactic structure is relatively complex. As to the Intonational phrase or IP, it is the largest prosodic unit in the prosodic hierarchy. It is characterized by a presence of an intonational contour at its right edge, the strongest degree of phrase-final lengthening, and also often followed by a pause. In sequences of clauses, each clause is normally phrased as an independent IP.

In section 4, the prosodic phrasing observed in the data, and the information requested to generate these phrases will be explained in details. It will allow evaluating what differentiate phrasing in standard French from phrasing in dictations.

3. Corpus and methodology

3.1. Corpus

To study prosodic chunking and intonation in dictation, a set of dictations has been gathered. This set consists of dictations that have been given to children enrolled in first to third year elementary school in France and Quebec (Canada). The data come from three different sources:

- Four short dictations come from the website of the Canadian association "Fondation Paul Gerin-Lajoie"[11], in particular, the dictations for first year level (CP in France);
- Four dictations come from the French website La-dictée.fr[12], which offers a wide range of dictations and grammatical exercises to school children;
- Two dictations have been recorded in class situations by some researchers belonging to the Phorevox project.

The choice of the dictations has been done in order to cover the various levels we are interested in for the software development. Moreover, as they come from different sources, some variation may occur in the way to dictate a text, in particular for the repetition of certain sequences. Some teachers repeat each sequence a few times (two or even more), while some others don't. The software will allow the user to configure such repetitions for any given dictation.

3.2. Methodology

The dictations have been annotated by two of the authors. The annotation indicates for each text two distinct types of information: the phrasing obtained (i.e. the way the texts were segmented into chunks during the dictation); and the form of the tonal contours that occur at the end of the various chunks.

The annotation has been achieved by means of a perceptual and instrumental data analysis. The perceptual analysis was done by a careful listening of the data, and allowed determining the chunking and the form of the pitch contours at the end of the various chunks. The acoustic analysis, achieved with the Praat software [13], confirmed what was perceived. Special attention was given to the occurrence of pauses to determine the segmentation in chunks, and to the form of the tonal contours occurring at the end of the prosodic chunks.

4. Data analysis: defining chunking rules

Before describing in details the rules used to derive the prosodic chunks, and the procedures at stake to repeat and introduce new chunks, we want to mention three major features that have been observed in all the analyzed dictations. First, the title and the whole text are said once at a relatively slow rate at the beginning, and then the dictation proper begins. Second, during the dictation proper, the whole sentence is usually repeated once after all its parts have been dictated in separate chunks. Third, punctuation marks are pronounced at the position they occur in the written text for the children to encode them as shown in (3) and (4), where the orthographic transcription of what has been pronounced is given.

(3) *Avec Papa. Point. Je marche dans la nature avec Papa. Point.*
With Daddy. Full stop. I am walking in the country with Daddy. Full stop.

(4) *A l'école, virgule, je travaille toujours avec lui. Point.*
At school, comma, I am working with him. Full stop.

Nevertheless, variation occurs in the pronunciation of the punctuation marks: the pronunciation of commas may be omitted when the sentence as a whole is repeated for the last time, whereas the pronunciation of stops may be done only when the sentence as a whole is produced. Since the latter realizations are not systematic, they will not be taken into account in the elaboration of the dictation algorithm presented in section 5: punctuation marks will always be pronounced in the position where they occur in the written text.

4.1. Phrasing and prosodic structure observed

In the observed data, the chunks used during the dictation proper correspond, in more than 95% of the cases, to minor phrases (see section 2), that is to lexical words preceded by the function words they syntactically govern. In a prepositional phrase, for instance, the noun is always phrased with the preposition and the determinant as in (5a), and, in the same vein, an auxiliary is phrased with the verb as in (5b).

(5) a. *Je marche dans la nature avec Papa* $\rightarrow$ (Je marche)$_{MiP}$ (dans la nature)$_{MiP}$ (avec papa)$_{MiP}$

b. *Il se demande si sa maman a trouvé les bons médicaments* $\rightarrow$ (Il se demande)$_{MiP}$ (si sa maman)$_{MiP}$ (a trouvé)$_{MiP}$ (les bons médicaments)$_{MiP}$

Note however that two Minor Phrases, which are derived from morphosyntactic information, may be restructured in a single one when the size of the phrase is inferior to two syllables. In many cases, for instance the copula *être* is restructured with what follows as shown in (6). The restructuring fails sometimes to apply, in particular when the resulting phrase would be relatively long as in (7).

(6) *Le lac est bleu* $\rightarrow$ [Le lac] [est bleu]

(7) *C'est mon meilleur ami* $\rightarrow$ [C'est] [mon meilleur ami]

As for the prosodic realization, each prosodic chunk is treated as an IP, be it in isolation as in (8) or integrated in a sentence as in (9), when the whole sentence is produced at the beginning or at the end.

(8) *Avec mon ami.* $\rightarrow$ [avec mon ami]$_{IP}$ [point]$_{IP}$

(9) *Je joue dans l'eau avec mon ami* → [Je joue]$_{IP}$ [dans l'eau]$_{IP}$ [avec mon ami]$_{IP}$ [point]$_{IP}$

This means that in dictation as a speaking style any prosodic group that would be a minor phrase in a normal reading style is realized with the major prosodic features of an IP such as an important final lengthening and a presence of a pause. Such a phrasing has been described as completely appropriate in French by [8]. It results from what [8] called the *élasticité prosodique* (i.e. prosodic elasticity), and account for the fact that any MiP could be realized as an IP, without any further restructuring.

The segmentation procedure used to dictate the text could thus be based on a parsing which introduces a major break after the words categorized as nouns, verbs, adjectives or adverbs if they are not modifying the following word. This latter principle should allow phrasing together in the same IP prenominal adjective and noun as in "le petit garcon", modifier adverb and adjective as in "très ennuyeux", or verbal auxiliary and past participle as in "est arrivé".

4.2. Procedures of repetition

Apart from the segmentation and the pronunciation in chunks, new chunks and sentences are introduced by special procedures that will be described in the following subsections. From the observation of the data, it was possible to infer three distinct procedures

4.2.1. Procedure IP by IP

This procedure consists in pronouncing the whole sentence where each MiP is realized as an IP first, and then to produce each IP in isolation (be they repeated or not). When all IPs have been uttered, the whole sentence is pronounced once again. For the sentence in (10), this procedure will lead to the chunking and the realization in (11). Each line break indicates that the chunk is uttered isolated from the preceding ones.

(10) Je joue dans l'eau avec mon ami.

(11) [je joue]$_{IP}$ [dans l'eau]$_{IP}$ [avec mon ami]$_{IP}$ [point]$_{IP}$
[je joue]$_{IP}$
[dans l'eau]$_{IP}$
[avec mon ami]$_{IP}$ [point]$_{IP}$
[je joue]$_{IP}$ [dans l'eau]$_{IP}$ [avec mon ami]$_{IP}$ [point]$_{IP}$

This procedure has mostly been used for the dictations given to the younger children (first year in elementary school).

4.2.2. Procedure by IP chaining

The procedure by IP chaining relies on a dictation of the sentence IP by IP, each IP consisting of what would be a MiP in colloquial speech. When the first IP has been realized once, it is repeated followed by the next IP, then the added IP is also produced in isolation once. For the sentence (12), this procedure will lead to the chunking and pronunciation in (13).

(12) Le chien s'étire sur le tapis.

(13) [Le chien]$_{IP}$
[Le chien s'étire]$_{IP}$
[S'étire]$_{IP}$
[S'étire sur le tapis]$_{IP}$ [Point]$_{IP}$
[Sur le tapis]$_{IP}$ [Point.]$_{IP}$
[Le chien]$_{IP}$ [s'étire]$_{IP}$ [sur le tapis]$_{IP}$ [point]$_{IP}$

Even if this procedure has been used in approximately 25% of the case in our data, it has some serious drawbacks, in particular in case of more complex sentences. In a sentence with a branching NP subject the verb is not uttered with the subject as shown in (14). Such a way of dictating is really error prone, as the subject-verb agreement cannot be interpreted in a straightforward manner.

(14) La porte de la chambre s'ouvre.
[La porte]$_{IP}$
[La porte]$_{IP}$ [de la chambre]$_{IP}$
[De la chambre]$_{IP}$
[De la chambre]$_{IP}$ [s'ouvre]$_{IP}$ [point]$_{IP}$
[S'ouvre]$_{IP}$ [point]$_{IP}$
[La porte]$_{IP}$ [de la chambre]$_{IP}$ [s'ouvre]$_{IP}$ [point]$_{IP}$

4.2.3. Procedure sentence by sentence

The last procedure consists in dictating the text sentence by sentence. Each sentence is pronounced once or twice, depending on its size, at a relatively slow rate. The segmentation in IP should be clearly realized as in (15), where a succession of two sentences is given.

(15) Le lac est bleu. J'aime le lac.
[Le lac]$_{IP}$ [est bleu]$_{IP}$ [point]$_{IP}$
[J'aime]$_{IP}$ [le lac]$_{IP}$ [point]$_{IP}$

In cases of complex or long sentences, the segmentation proceeds clause by clause. Clause refers here to different types of elements:

- Comma clause such as peripheral adjunct followed by a clause as the underlined sequence in (16).
- Subordinated or coordinated clauses as in (17)

(16) A l'école, je travaille toujours avec lui.

(17) Mon père rentre très tard à la maison parce qu'il est musicien.

When such a sub-segmentation is used, the sentence as a whole is repeated once when all parts have been dictated.

5. Adaptation of the rules and procedures for speech synthesis

The achieved segmentation and observed procedures have been used to automatically dictate any text with a speech synthesizer. The implementation of the various elements just described was achieved by generating the chunks, and producing the text while respecting the various features mentioned at the beginning of section 4.

5.1. Chunk generation

To explain how we generate the chunk list from an input text, we are going to take the following French sentence as an example:

(18) *En réalité, c'est un hélicoptère. Avec une cheminée qui crache de la fumée, comme une locomotive à vapeur.*
Actually, this is an helicopter.With a smokestack to expel smoke, like a steam locomotive.

5.1.1. *Main algorithm*

To generate a word chunk list, we first need to determine the syntax tree associated to the dictation text. This is given by the Synapse pos tagger[14]. We suppose that the nodes are corresponding to a syntactic phrase and each leaf is associated to a word.

For a given node N_0, we identify the children of N_0 by $(N_1 \dots N_n)$. Each node N_j represent a chunk of words. So N_j is defined by $N_j = (s_j, ej)$ where s_j is the first word's index and e_j the index of the chunk's last word. The syntax tree has the following properties which implies an order between children: $s_0 = s_1$; $e_0 = e_n$ and $s_j = e_{j-1} + 1$

The goal of the main algorithm is to find the "syntactic group" sequence which could be used as baseline chunks. To achieve this goal, we suppose a user defined parameter w representing the ideal number of words contained in a chunk. Based on this parameter, we define a cost function $C(N_j)$ associated to a node:

$$C(N_j) = |(e_j - s_j + 1) - w|$$

By using this cost function, the idea is to locally determine if, by splitting the current group N_0 into smaller syntactic groups $(N_1 \dots N_n)$, we approach the ideal chunk size or not.

This recursive algorithm distinguishes three cases:

1. if N_0 is a leaf, the recursion is stopped and the chunk is defined by the word contained in N_0,
2. if $(C(N_0) < \sum_{j=1}^{n} C(N_j))$ and N_0 is not a leaf, we consider two cases:
   - if $(N_1 \dots N_n)$ contains only leaves then the chunk is defined by N_0,
   - else we try to recompose the node sequence $(N_1 \dots N_n)$ by eliminating leaves.

     To do this, we try to merge each leaf to the preceding group in an incremental way. We identify by $(N'_1 \dots N'_{n'})$ the obtained node sequence. If $(C(N_0) >= \sum_{j=1}^{n} C(N'_j))$ then we apply the current algorithm by considering $N_0 = N'_j$. In the other case the chunk is defined as N_0 without considering the merging step.
3. in other cases, we apply the current algorithm by considering $N_0 = N_j$ for each N_j in $(N_1, \dots, N_n)$;

By applying this algorithm on the previous example, we achieved the segmentation presented in figure 1(b).

5.1.2. *Post-processing*

The previous stage results in a chunk sequence which is not yet optimal. If we consider the example, we can see that some chunks are composed by only one word ("comme"), other chunks contains punctuation in the middle which is not good in a dictation context ("en réalité, c'est").

In order to improve the consistency of the chunks, we have defined a post-processing stage, using a rule-based approach. Three steps are achieved in this stage: a splitting step, a merging step and finally an annotation step.

The splitting step goal is to isolate punctuations and split large chunk according to part-of-speech information. Punctuations have to be isolated because of their special treatment in dictations. Furthermore, for the moment, we split a chunk into two parts only before a coordinating conjunction. By applying this step, we achieved the segmentation presented in figure 1(c).

The merging step goal is to deal with two constraints. The first one aims to avoid any isolated word or any chunk starting with a non-alphabetical character. The objective of the second rule is to assess a minimum number of syllables in each chunk. As we don't have access to a segmentation in syllables (since we deal with a written text), we made the assumption that number of non-consecutive vowels in the text gives an approximation of the number of syllables. The minimum number of syllables in a chunk is then defined in a parametric way. Consequently, we merge a chunk to the previous one if it contains only one word, if it starts with a non-alphabetical character or if the number of consecutive vowels included in both chunks are inferior to the number selected as parameter. The result of this step is presented in figure 1(d).

5.2. Entire text dictation procedure

Once the text has been segmented into chunks, it is possible to automatically dictate it. To do so, the entire text is first copied while using the chunking automatically generated. In a second stage, each chunk is annotated in such a way as to be pronounced in isolation (followed by pause), the punctuation marks being made explicit. This leads for (18) to the final segmentation and the pronunciation given in Fig. 1(d).The text is then given to the TTS system that can treat it, while using a synthesized voice that has been specifically designed for dictation.

6. Conclusion and perspectives

The paper presents a chunking algorithm that allows segmenting any text into chunks that are comparable to the ones we observed in dictation data. Further research is currently achieved in order to (i) decide which procedure is more appropriate depending on the level of the pupils; an (ii) provide a closer analysis of the intonation contours used at the end of the different non-final IPs.

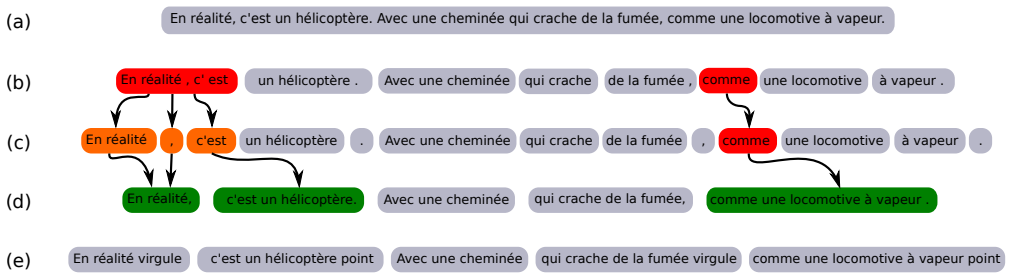

Figure 1: Chunking procedure for (18)

7. Acknowledgement

The work presented here is related to the research project ANR-CONTINT 2011 *PHOREVOX* funded by ANR/CGI.

8. References

[1] "Lectramini," http://www.lectramini.com/.

[2] R. Beaufort and S. Roekhaut, "Automation of dictation exercises. a working combination of call and nlp," *Computational Linguistics in the Netherlands Journal*, vol. 1, pp. 1–20, 2011.

[3] A. Di Cristo, "Intonation in french," *Intonation systems: A survey of twenty languages*, pp. 195–218, 1998.

[4] B. Post, *Tonal and phrasal structures in French intonation*. Thesus, 2000, vol. 34.

[5] P. Martin, *Intonation du français*. A. Colin, 2009.

[6] S.-A. Jun and C. Fougeron, "A phonological model of french intonation," in *Intonation*. Springer, 2000, pp. 209–242.

[7] *Developing a ToBI system for French*. Oxford University Press, Accepted, ch. 3.

[8] S. P. M. Verluyten, *Investigations on French prosodics and metrics*. University Microfilms, 1982.

[9] A. Michelas, "Caractérisation phonétique et phonologique du syntagme intermédiaire en français: de la production à la perception." Ph.D. dissertation, Université de Provence-Aix-Marseille I, 2011.

[10] E. Selkirk, "On derived domains in sentence phonology," *Phonology yearbook*, vol. 3, no. 1986, pp. 371–405, 1986.

[11] F. P. Gerin-Lajoie, "La dictée p.g.l." http://fondationpgl.ca/audio/.

[12] Ladictée, http://www.ladictee.fr/.

[13] P. Boersma and D. Weenink, "Praat (version 5.5)," *Amsterdam: Institute of Phonetic Sciences*, 2012.

[14] Synapse, "Documentation technique: Composant d'étiquetage et lemmatisation," 2011. [Online]. Available: http://www.synapse-fr.com/API/API_Etiquetage_lemmatisation.htm