Adapting prosodic chunking algorithm and synthesis system to specific style: the case of dictation

Elisabeth Delais-Roussarie [1], *Damien Lolive* [2], *Hiyon Yoo* [1], *Nelly Barbot* [2], *Olivier Rosec* [3]

[1] UMR 7110-LLF (Laboratoire de linguistique formelle), Université Paris-Diderot
[2] IRISA, University of Rennes 1, Lannion, France
[3] VOXYGEN, Lannion, France

Elisabeth.Roussarie@wanadoo.fr, Damien.Lolive@irisa.fr, yoo@linguist.univ-paris-diderot.fr, Nelly.Barbot@irisa.fr, olivier.rosec@voxygen.fr

Abstract

In this paper, we present an approach that allows a TTS-system to dictate texts to primary school pupils, while being in conformity with the prosodic features of this speaking style. The approach relies on the elaboration of a preprocessing prosodic module that avoids developing a specific system for a so limited task. The proposal is based on two distinct elements: (i) the results of a preliminary evaluation that allowed getting feedback from potential users; (ii) a corpus study of 10 dictations annotated or uttered by 13 teachers or speech therapists (10 and 3 respectively).

The preliminary evaluation focused on three points: the accuracy of the segmentation procedure, the size of the automatically calculated chunks, and the intelligibility of the synthesized voice. It showed that the chunks were judged too long, and the speaking rate too fast. We thus decided to work on these two issues while analyzing the collected data, and confronting the obtained realizations with the outcome of the speech synthesis system and the chunking algorithm. The results of the analysis lead to propose a module that provides for this speaking style an enriched text that can be treated by the synthesizer to constrain the unit selection and the prosodic realization.

Index Terms: speech synthesis, prosodic phrasing, prosodic parsing, speaking styles.

1. Introduction

Using software and automatic tools in language teaching offers several advantages, among which we may mention the ability to adapt to learners needs and to provide an environment in which the learner is autonomous. Note, however, that educational programs using speech synthesis are not so common, especially for teaching French language to primary school children (e.g. *Lectramini* [1], and PLATON [2]). In order to use speech synthesis systems in language teaching software, several issues have to be taken into account: (i) the synthesized voice has to be highly intelligible, especially for program dedicated to children; (ii) prosody and voice used by the speech synthesis system have to be suitable for reading different text types of speech (instructions, dictations, poems, rhymes, etc.), and (iii) speech rate has to be consistent with the task and the public (e.g.: for dictations made to children).

In a collaborative ANR research project *Phorevox* (http://www.phorevox.fr/), which aims at developing a tool for the acquisition of writing skills in primary school by using speech synthesis, we aimed at developing a procedure that allows the system to dictate texts. To achieve this task, a chunking algorithm and a high quality synthesized voice designed for the task were developed, and the outcomes were evaluated by children and teachers. From this evaluation, it appeared important to improve some points: the chunking algorithm and the speech speed of the synthesized stimuli.

In this paper, we present the approach we used to achieve this improvement task, while avoiding creating a new voice: on the basis of the analysis presented here, we elaborated a preprocessing module that provides, for the dictation style, an enriched text that can be treated by the synthesizer to constrain the unit selection and the prosodic realization. The paper is organized as follows. Section 2 briefly presents the chunking algorithm and the results of the evaluation in order to clearly state which points need to be improved. In section 3, the method used to gather additional data is described. The results of the analysis are then presented and discussed in section 4. Focus is given to two main issues: the chunking procedure, and the improvement of two temporal variables of prosody, e.g. the articulation rate, and duration of pauses.

2. Background

2.1. Parsing algorithm for dictation

During dictations, teachers usually segment texts into chunks that are read at a slow speech rate, and often repeated. To develop the chunking algorithm, a set of dictations produced by teachers in real settings was analyzed so as to understand which linguistic elements are taken into account to build up the dictated chunks. To take into account these elements, the algorithm consists of three steps (for more details, see [3]). First, the text is tagged and parsed with the Synapse Pos Tagger [4]. The nodes corresponding to words and consisting of final leaves in the trees are merged together according to parsing information and minimal size requirements (in terms of number of words per chunk, cf. on that issue [5], [6], [7] and [8] among others). Due to size requirement, the system generates some chunks that are limited to a single word. In a second step, such chunks are merged from left to right to the following chunk, as the verb "*est*" 'is' in (1).

(1) *Médor est le nom de mon chien.* 'Médor is the name of my dog'.

→ [Médor est] [le nom] [de mon chien]

For the merging mechanism, further research is in progress to take as size parameter the number of syllables per chunk, and not only the number of words (see, for the importance of this parameter in French, [7] and [8] among others). Finally, punctuation marks are replaced in the text by a prosodic chunk that overtly utters the punctuation mark as in (2).

(2) *Le soir, Papa ferme la porte*. 'At night, Dad closes the door.'

→ [le soir] [virgule] [Papa] [ferme] [la porte][point]

2.2. Evaluation by potential users

In order to evaluate the outcomes of the chunking algorithm, potential users (7 children enrolled in cycle 2, i.e 6 to 8 years old, and 14 teachers) were asked to participate to an evaluation by means of a written questionnaire (for the use of questionnaire in phonology, see [9] and [10] among others).

The procedure for the evaluation was adapted for each group of users. We thus built two questionnaires in order to gather information regarding the parsing procedure and the task itself. The evaluation test for the pupils was built in two steps. First, children had to perform the dictation task, and then they had to answer orally to questions (mostly yes/no choice) concerning the intelligibility of the synthesized voice and the size of various chunks. The evaluation for the teacher group consisted of an online test where participants could listen to the synthesized stimuli (dictation texts chunked according to the algorithm and uttered by the synthesis system), answer to the questions, and give their opinions for some aspects of the tool. The questionnaire was divided into two parts: the first part concerned the chunking of the stimuli while the second one presented specific questions on speech rate and on the adequacy of the prosodic patterns used in specific sequences.

The answers given by the children reveal that they could really be potential users of the program since they find such a tool useful for dictation exercises. The synthesized voice quality is judged natural, but three children out of seven find that sometimes the way of reading is weird. Most children find difficult to understand some specific words (*algue*, *crabe* and the first name *Lila*). Moreover, most of the children (6 out of 7) find the way of reading too fast for the dictation task. However, none of the pupils have difficulties understanding the meaning of the dictated sentences. As for the chunking, pupils do not seem to be preoccupied by the fact that some chunks, resulting from the merging of isolated words with the previous chunks regardless of the morpho-syntactic structure, were not very accurate. However, all of them find that chunks consisting of a whole sentence such as "*Lila marche sur la plage*" 'Lila is walking on the beach' or "*il a disparu dans le sable*" 'he disappeared in the sand' are too long.

The analysis of the answers given by the adult group reveals that, contrarily to pupils, they are more attached to boundaries related to syntactic constructions. About three-quarters of participants, for instance, consider that preposition such as "*avec*" 'with' should be group with its complement. By and large, teachers considered that dictation should almost always respect syntactic boundaries, arguing that the contrary may induce children to errors. Concerning the size of chunks, the answers given showed that different strategies can be adopted. Thus, 10 out of 14 find the sequence "*là virgule*" too short, but "*et moi*" is unanimously accepted by the participants. For long chunks, answers are less homogenenous: the sequence "*Martin tape sur un tambourin*" 'Martin is hitting on a tambourine' has been considered too long to form a single chunk (13 out of 14) by all the teachers, but "*Lila marche sur la plage*" and "*il a disparu dans le sable*" are accepted as a single chunk (respectively 6 and 5 answers). As far as speech rate is concerned, for all items of the questionnaire, the participants showed a preference for speech rate used in dictations (13 out of 14 for a given sequence), judging the 'normal voice' too fast. Nevertheless, some of them considered that the speech rate, even with the voice designed for dictation, should be slower.

2.3. Summary

To sum up, the results reveal that pupils and teachers differ slightly in the evaluation. Compared to pupils, answers from the teacher group show that syntax seems important in determining boundary placement, at least at the lowest syntactic level (i.e. prepositions should be grouped with the noun phrase they introduce, relative pronouns with the relative clause, determinants with the noun they governs, and so on). As for chunk size, teachers seem to have different strategies, whereas pupils do prefer shorter chunks. Note also that the number of syllables is not the only parameter to take into account; other parameters such as the difficulties in the introduced lexicon play a role. Speech rate during dictation is generally accepted as correct by the teachers, but judged too fast by the pupils.

On the basis of the evaluation, it appears that two points request improvement: chunking itself, in particular the location of certain boundaries, and speech rate. Additional analyses have been thus undertaken to evaluate how the algorithm and the synthesis system could be improved.

3. Methodology

To improve the algorithm and the outcomes of the synthesis system, additional data were gathered and analyzed, special attention being given to the two problematic issues.

3.1. Material

Four distinct types of data sets were gathered:

- A written corpus that consisted of 38 sentences extracted from 10 dictations for cycle 2 pupils (aged 7 to 8 years). The sentences were manually parsed in prosodic chunks by 10 teachers, who were asked to chunk the texts as they would do while doing a dictation to cycle 2 pupils;
- A written corpus which consisted of the same 38 sentences, but automatically parsed by the algorithm;
- An audio corpus composed of the 38 sentences and recorded by three teachers that read aloud the texts as they would do during dictation in cycle 2 classroom;
- An audio corpus consisting of the 38 sentences produced with the synthesis system while using the algorithm for chunking, and the voice specially designed for dictating.

3.2. Methodology

The analysis of the data has been divided in two phases, each of them focusing on a given issue and using designated data sets.

A first phase was achieved by exploring the written corpora, i.e. the chunking provided for the 38 sentences by 10 teachers and by the algorithm. The analysis of these data sets focused on the size of the chunks, and the distribution of the boundaries. By confronting the two sets (manually parsed vs. automatically parsed), recurrent and consistent points of divergences should be revealed, and improvements of the algorithm could be proposed.

The other phase consisted in analyzing temporal variables on the audio data sets, special attention being given to articulation rate and pauses. The 152 sentences (38 sentences

produced by 3 speakers, and 38 stimuli from the synthesis system) were first orthographically transcribed within Praat [11], then aligned in phones, syllables and words with EasyAlign [12]. All alignments were manually verified and corrected by one of the authors.

4. Results

This section presents the results of the analyses achieved on the written and the audio corpora respectively.

4.1. Analysis of the chunking on the written texts

The evaluation of the synthesized dictations showed clearly that the size of the chunks and the location of the boundaries are the two issues that needed to be explored.

4.1.1. Length of the prosodic chunks

Since the pupils mentioned during the evaluation that some chunks were too long (see section 2.2), it was important to calculate and to compare the average size of the chunks for the teacher group and for the algorithm (see Table 1, left part). For both, the lengths of the groups were thus computed.

From these results, we can observe that the mean group length is relatively comparable between what teachers and the algorithm produce. In the same way, the standard deviation of the parameter of group length is also comparable between the two groups. This result puts forward that the algorithm produces chunks with an appropriate length. We now need to check whether the predicted chunk boundaries are accurately distributed or not, in particular in terms of syntax-prosody mapping.

4.1.2. Location of the prosodic boundaries

From analyzing boundary placement in the teachers' data, by taking into account the alignment and rhythmic constraints that comes into play in French (see [7], [8] among others), it appears that positions where everybody agree in realizing a boundary are very limited in terms of syntax-prosody mapping: a boundary is always realized at the right edge of left-peripheral adjuncts (which are often delimited by a comma in written texts), and at the end of a clause in case of coordination (e.g.: after *démodé* 'old-fashioned' in *il est démodé et il pleure* 'it is old-fashioned and it cries'). In all these positions (with the exception of one where a parsing error occurs), the algorithm predicted also a boundary. Between nominal subjects (NP subject) and verbs, the teachers realize also often a boundary, despite some variability which finds its explanation in the managing of the readjustment for rhythmic reasons (see [7], [8], [13] and [14] on rhythmic readjustment in French). In 66 % of the cases, more than 80% of the teachers assigned a boundary between the subject noun phrase (NP) and the verb, and only once was the NP subject phrased together with the verb by more than 50% of the teachers (the NP subject *une voiture* 'a car' was phrased with the verb in *une voiture passe* 'a car is passing by'). By contrast, in around 33 % of the case, the algorithm predicted to phrase together the NP subject and the verb (e.g: *[papa ferme] [la porte]* 'Papa closes the door', *[Maman lit] [sous un parasol]* 'maman is reading under a parasol'). The last point where some discrepancies occur between the teacher group and the algorithm concerns the prosodic behavior of function words. In French, it is usually assumed that function words are metrically weak, and as such phrased with the lexical item from which they depend syntactically (a determinant with the noun, etc., cf. among others [7], [8] and [15]). In the predicted chunks, a function word or a modifier may not be phrased with the item from which it depends syntactically (i.e., the determinant *le* in *[est le][chat du voisin]*' is the neighbor's cat'). The problem in the predicted chunking relies mostly on the fact that the function words are not left in isolation, but grouped with lexical items to which they are not syntactically related. Indeed, the analysis of the chunks assigned by the teacher group reveals that function words may be phrased in isolation in dictation as speaking style.

Table 1. Comparison between teachers' chunks and boundaries, and the algorithm outputs

	Group length		Precision	Recall	F-score
	mean	std			
Teachers	2.67	1.18	0.78	0.79	0.73
Algorithm	3.13	0.90	0.80	0.70	0.72

4.1.3. Interim discussion

The analysis of the data and the comparison between the algorithm and the teacher group reveals that the size of chunks is comparable in both cases. No modification of the algorithm is thus requested to improve this issue.

As shown in Table 1, we have computed the level of agreement between the teachers and the algorithm estimated with three measures: precision, recall and F-score. For precision and recall, the following method was used: for the teacher group, we consider one teacher as the system under test and the others as reference to compute a confusion matrix with values boundary/no boundary. From each confusion matrix, we have computed the values of precision and recall. The same method is applied for the algorithm thus considering the algorithm versus the teachers. As a consequence, for teachers, the numbers presented in the table are mean values of precision, recall and F-score for each teacher.

From this information, it appears that the boundaries predicted by the algorithm are observed in 80% of the cases in the data annotated by the teachers. In addition, in 70% of the cases, the algorithm does not omit a boundary. Because of the variability that exists even between the teachers, it appears that the results provided by the algorithm are comparable to the agreement between teachers. In other words, the algorithm is as precise as the teachers. Now, the analyses of the positions where disagreement between the predictions and the teacher group is larger, reveal that the discrepancy almost always comes from the merging mechanism: as said in section 2.1, a chunk consisted of a single word is always regrouped with the preceding chunk, independently of the morpho-syntactic structure. So verbs are often phrased with the NP subject on its left (e.g. *[papa ferme]*, *[maman lit]*, etc.). This mechanism often explains the phrasing of function words and modifiers discussed in section 4.1.2. It is thus important to improve this procedure in the algorithm. The analyses of the chunking obtained in the teacher group reveal that several options are possible, and need to be tested to choose the optimal solution, even in case of parsing errors in the morpho-syntactic analysis:

- the merging mechanism could be constrained by the morpho-syntax, and thus the syntactic break level between the items. A singleton should thus be phrased with the lexical item to which it is syntactically related to;

- the algorithm could rely only on a distinction between leaners and non-leaners as proposed by [15] to derive prosodic words, and treat any prosodic word as a prosodic chunk without restructuring for metrical reasons. Such a strategy is also used by some teachers to segment texts into prosodic chunks.

4.2. Analysis of the temporal variables of prosody

The temporal variables, in particular the articulation rate (AR) and the pause duration were analyzed on the recorded data set. The aim of this analysis was twofold:

- Evaluating whether the AR observed in the dictations recorded by teachers is very different from what is observed with the synthesized voice;
- Understanding how pauses are used in dictation. This should allow developing a post-processing module that could insert pauses with an accurate duration.

4.2.1. Articulation rate (AR)

The articulation rate has been calculated automatically from the annotated tier segmented at the phone and syllable levels (see section 3.2) by the Praat script *Calculate_tempo* [16]. Table 2 shows the result of this analysis.

The analysis of the data reveals that the AR may vary greatly in the teacher group. Two teachers speak very slowly during dictations (FD and ER in Table 2), but AP (or teacher 3) produces much more speech units (phone and syllable) per second. Even if the number of speakers is restricted, the results show that the AR in the synthesized voice is comparable to the outcome observed in some of the teachers' data, and should be suitable for dictation.

Despite the results we obtained, it is important to recall that some teachers and pupils judged the speech rate too fast during the evaluation. We thus decided to explore the realization of the pauses, since changes in the distribution and the duration of pauses in the synthesized voice may improve the outcome. This study is presented in the following section.

Table 2. AR observed in the recorded data and compared to the AR of the synthesized voice. The AR is given in phone/sec, and syll/sec.

Speakers	Rate (Phone/sec)	Rate (syll/second)
Teacher 1 (FD)	4.17	1.80
Teacher 2 (ER)	4.52	1.98
Teacher 3 (AP)	6.98	3.14
Synth. voice (D)	6.49	2.73

4.2.2. Pauses

A preliminary analysis of the pauses has been achieved on two speakers, FD and ER. This choice is motivated by the fact that the analysis was done to evaluate how pauses could be used to slow down the tempo. It could thus be restricted to the two speakers having the slower tempo. For the 38 sentences extracted from 10 dictations (each dictation consisting of roughly 3 sentences), the pause duration has been calculated automatically with the script *Get_pause_duration* [16] from the annotated tier segmented in words. The mean pause duration and the percentage of pause duration over the entire recording were then calculated for each speaker. The results show that pauses represent 68% of the recording duration for FD, and 63% for ER. As for the mean pause duration (in second), it is 1,907 sec. for FD, and 1, 631 sec for ER.

In addition, pauses were classified according to their position in the sentences (in terms of syntactic constructions), and to the size of the preceding chunk (in terms of number of syllables and duration). The latter parameter does not seem to affect the duration of pauses. By contrast, the location of a pause in the sentence – more precisely from a morpho-syntactic point of view – partly determines its duration: pauses within a lower level syntactic phrase (e.g. between a preposition and what follows in a prepositional phrase, between a prenominal adjective and a noun in a noun phrase) are shorter than the other pauses. Their durations are approximately equal to 0.8 times the mean duration, whereas pauses at the juncture between syntactic phrases (SN subject/ Verb, SN/SN in case of double object constructions) are usually longer than mean pause duration. As for the pauses that occur at the introduction of punctuation mark (see example (2)), the duration of pauses that occur after the elicitation of the punctuation mark are equal to 1.5 to 2 times the mean pause duration, whereas the duration of the pauses occurring before the elicitation is shorter than the mean duration by 0.8.

Finally, a very short pause (almost comparable to a glottal stop insertion) is often realized at the juncture of two words, even in a context of "liaison". This procedure leads to block the liaison. In our data, for instance, the liaison between a preposition and what follows is often blocked (*dans / une maison* 'in a house' instead of *dans_une maison* [dɑ̃zynmɛzɔ̃], or between a determinant and a noun (un artiste 'an artist' is realized as [ɛ̃aʀtist]).

4.2.3. Interim discussion

The analyses of the temporal variable showed that the articulation rate used by the synthesizer is acceptable, even if it is among the faster one. This means that incorporating a duration model to the system or recording a new voice is not necessary.

As for pauses, the analysis reveals that they could be inserted into the text by a preprocessing module, their duration being in part determined from the position of the pause in the sentence. In addition, the insertion of micro-pauses at word boundary should allow slowing down the speech speed, and making the synthesized voice more intelligible, at least for young children doing dictations. These two elements can easily be added by a pre-processing module, before the selection of the speech units by the synthesizer.

5. Conclusion

In this study, we showed that speech synthesis can be used for tasks that request a highly intelligible synthesized voice, i.e. dictations addressed to very young children. Of interest here is the fact that the method used to develop such a task-oriented voice does not rely on creating new voices, which is very costly, or on elaborating very complex prosodic module that would modify the tempo (duration, etc.), but just by processing the text to be treated by the synthesis system by a simple algorithm that derives the prosodic chunk boundaries, inserts the punctuation, indicates pause duration, and blocks some liaison.

6. Acknowledgements

The present is funded by the ANR/CGI (ANR-CONTINT 2011 PHOREVOX project). It is also partly related to the work package “Prosodic phrasing and prosodic hierarchy: a data driven approach” of the Labex Empirical Foundations of Linguistics (ANR-10-LABX-0083).

7. References

[1] http://www.lectramini.com/

[2] Beaufort, R and Roekhaut, S., “Automation of dictation exercises: A working combination of CALL and NLP”, Computational Linguistics in the Netherlands Journal, 1:1-20, 2011.

[3] Le Maguer, S., Delais-Roussarie, E., Barbot, N., Avanzi, M., Rosec, O. and Lolive, D., “Prosodic chunking algorithm for dictation with the use of speech synthesis”, Proceedings of Speech Prosody, 2014.

[4] Synapse, “Documentation technique: Composant d’étiquetage et lemmatisation,” 2011.

[5] Prieto, P., “Syntactic and eurhythmic constraints on phrasing decisions in Catalan”, Studia Linguistica 59(2–3): 194–222, 2005. (Special issue on Boundaries in Intonational Phonology, M. Horne and M. van Oostendorp (eds)).

[6] Frota, S., D’Imperio, M., Elordieta, G., Prieto P., and Vigário. M., “The phonetics and phonology of intonational phrasing in Romance”, in Prieto, P., Mascaró, J. & Solé, M-J. [eds],. Prosodic and Segmental Issues in (Romance) Phonology, 131-153, John Benjamins, 2007.

[7] Martin, P., “Prosodic and Rhythmic Structures in French”, Linguistics, 25: 925-949, 1987.

[8] Delais-Roussarie, E., “Phonological phrasing and accentuation in French”, in M. Nespor & N. Smith (Eds.), Dam phonology: HIL phonology papers II, 1-38, Holland Academic Graphics, 1996.

[9] Honeybone P. Using questionnaires to investigate nonstandard dialects, Department of Linguistics and English Language, University of Edinburgh, 'Using questionnaires to investigate non-standard dialects. 'Northern English and Scots, Phonology and Syntax' website. 2010 http://www.lel.ed.ac.uk/dialects/nesps.html,

[10] Wells J.C. Phonetic research by written questionnaire. In: M. J. Solé, u.a. (Hrsg.): Proc. 15th Int. Congress of Phonetic Sciences, Barcelona, R.4.7:4, 2003

[11] Boersma, P. and Weenink, D., “Praat: doing phonetics by computer (Version 5.5)”, http://www.fon.hum.uva.nl/praat/, 2013.

[12] Goldman, J.-Ph., “EasyAlign: an automatic phonetic alignment tool under Praat”, Proc. of Interspeech, 3233-3236, 2011.

[13] Post, B., “Restructured phonological phrases in French: Evidence from clash resolution”, Linguistics, 37 (1): 41-63, 1999.

[14] Dell, F., “L'accentuation dans les phrases en français”, in Dell, F., Hirst, D. & Vergnaud, J.R [eds], Forme sonore du langage: structure des représentation en phonologie, 65-122, Hermann, Paris, 1986.

[15] Bonami, O., Delais-Roussarie, E., “Metrical Phonology in HPSG”, Proc. of the HPSG 06 Conference, Varna, Bulgaria, CLSI Online publications, 2006.

[16] De Looze, C., Analyse et interprétation de l'empan temporel des variations prosodiques en français et en anglais contemporain, PhD dissertation, Université de Provence, 2010.