Pauses and punctuation marks in Brazilian Portuguese read speech

Izabel C. Seara, Fernando S. Pacheco, Rui Seara Jr., Sandra Kafka, Rui Seara, Simone Klein

LINSE – Circuits and Signal Processing Laboratory
Department of Electrical Engineering
Federal University of Santa Catarina
88040-900 – Florianópolis – SC – Brazil
E-mails: {izabels, fernando, ruijr, kafka, seara, klein}@linse.ufsc.br

Mots-clés : pauses, ponctuations, lecture à haute voix.

Keywords: pauses, punctuation marks, read speech.

Résumé Dans cet article, nous avons examiné la relation entre pause et ponctuation (virgule, point et virgule, deux-points). Toutes ces pauses sont internes aux phrases. À l'aide de l'analyse de plusieurs milliers de pauses dans un corpus de presque 17 heures d'enregistrement réalisé par une locutrice professionnelle native du portugais brésilien, nous avons vérifié une proportion importante des pauses hors ponctuations (61,3%). Les données renforcent aussi la présence des structures topique/commentaire dans la lecture à haute voix. Les résultats des durées de pause correspondantes aux ponctuations sont consistants avec les données présentées dans les grammaires.

Abstract In this paper we assess pause effects corresponding to comma, semicolon, colon and the ones that are not related to any punctuation marks, all of them within sentences. Thus, through the analysis of a corpus of approximately 17 hours of recording, carried out by a female professional speaker (native) of the Brazilian Portuguese language, we observe a large proportion of pauses without punctuation (61.3%). Besides, our data reinforce the presence of topic-comment structures in reading. The results here presented with respect to pause and punctuation are consistent with several studies about this theme.

1 Introduction

Several research work about pause (Vannier *et al*., 1999; Marin *et al*., 2002; Campione & Véronis, 2002) have shown the existence of an inaccurate concept about the relationship between pause and punctuation. This concept states that there exists a one-to-one relationship between pause and punctuation; that is, if there is a pause there exists punctuation, no pause, no punctuation. Other works also have shown that several structures, which seem inadequate in reading, reflect usual situations of the spoken language, such as the topic-comment structures (Cagliari, 1992; Pontes, 1987).

The aim of this study is to evaluate situations in which a Brazilian Portuguese speaker inserts pauses in read speech. For such we analyze a Brazilian Portuguese speech corpus and also

investigate the relationship between pause and punctuation marks. The used material consists of recorded read texts by a female professional speaker. This paper is organized as follows. Section 2 presents our speech material and the analysis methodology. Section 3 shows how grammar books have considered a relationship between pause and punctuation. Experimental analyses are carried out in Section 4, investigating the occurrence of pauses, its duration, and the association pause-punctuation. Finally, conclusions and remarks are presented in Section 5.

2 Corpus and Methodology

In the open literature we have found some research works about pause dynamics in natural language which include: (a) a large number of small speech corpora (considering different speakers), according to Campione & Véronis (2002); or (b) a large corpus of a single speaker, according to Marin *et al*. (2002); Vannier *et al*. (1999). By considering that our goal is to obtain a pause model for a Brazilian Portuguese speech synthesis system, we believe that the investigation allowing for a single speaker not only could give rise to the required model, but also to determine some important characteristics (concerning pauses) for the language in question. In this way, we have based our research on a speech corpus with approximately 17 hours of recording, accomplished by a female professional speaker (native) of the Brazilian Portuguese language. This speaker read aloud texts at a normal speaking rate. The data were recorded in a studio and care was taken to avoid environmental noise. The recorded corpus includes different kinds of text, taken from newspapers and magazines, reports, extract of contemporaneous novels, short tales, and academic texts all transcribed and labeled in terms of its morphosyntactic classification. For this task we have used an ad hoc parser conceived for a speech synthesis system (Seara *et al*., 2002). After that, an expert linguist corrected manually the labeling.

Our analysis is restricted to pauses associated with silent intervals. We have not considered filled pauses (related to hesitations) and pauses created by the lengthening of phonetic segments. To investigate the occurrence, position, and duration of pauses we have considered two main classes: pauses with a silent interval larger than 300 ms, termed long pauses, and pauses with a silent interval between 90 and 299 ms, named short pauses. Seara (2000) has shown that silent intervals associated with the stop closure interval have an average duration of 45 ms for the Brazilian Portuguese language. In this way, if we consider a lower limit equal to 90 ms, it is possible to assure that there will always be a silent interval associated with a pause, even if the stop closure occurs within this interval.

In order to study the duration and types of acoustic pauses, the speech data have been processed automatically by a pause detector. This detector retrieves each context with pauses from the corpus. In this way, we obtain a total of 9,985 pauses from the 17 hours of recording of which 3,858 are associated with punctuation marks. From those, 3,633 are accompanying commas; 52, semicolon; and 173, colon. The other 6,127 pauses are not related to punctuation marks.

3 Pause and Punctuation Relationship in Grammar Books

Usually, Brazilian Portuguese teachers complain about mistakes associated with punctuation made by their students. Such mistakes refer to the idea that it is imperative to observe the pauses to assign a comma within a sentence. Thus, we perceive how difficult it is to teach the students that the punctuation marks are based on syntactic structures of the sentences and not on the carried out pauses.

If we observe Brazilian Portuguese grammar books, almost all of them mention pauses when they discuss punctuation marks (Almeida, 1988; Faraco & Moura, 1991, among others). Grammarians state that punctuation marks, such as comma, period, and semicolon are used for pause marking. They also explain that the comma, period, and semicolon represent, respectively, pauses of short, long and intermediate duration. However, they emphasize that the comma should not be used either between the subject and verb or between the verb and object. Thus, we can assume that pauses would not occur in reading aloud in these situations, since the punctuation marks, which indicate such pauses, would not be included. Nevertheless, our speaker inserts pauses in these structures. The following example presents pauses between verb-object and subject-verb, respectively:

> *Os pesquisadores **afirmam** [short pause] **que os resultados** são a primeira evidência de que os **transgênicos** [long pause] **podem** gerar conseqüências*
>
> (*The researchers affirm [short pause] that the results are the first evidence that the transgenic organisms [long pause] may produce consequences.*)

Almeida (1988) states that a comma is never found where no pause is placed. Despite our analysis data we have found 1% of commas with no correspondence to pauses. Thus, it would be more accurate to state that frequently if there is no pause there is no punctuation. Remark that Vannier *et al*. (1999) also have found 4.6% of pauses without punctuation in a French corpus.

To obtain a notion of the insertion dynamics of long and short pauses, mainly for the cases in which there is no association with punctuation marks, we achieve an analysis of the syntactic sequences previously mentioned.

4 Data Analysis and Discussion

To evaluate pauses between both subject-predicate and verb-object (structures in which the comma is forbidden), we divide the data that present such sequences into two groups: subject and verb (or predicate) termed Group 1, and verb and object, Group 2. Results are shown in Tables 1 and 2.

Table 1: Number and occurrence frequency of the Group 1 sequences with and without pauses

Group 1	Without pauses		Long pauses		Short pauses	
	Number of pauses	%	Number of pauses	%	Number of pauses	%
Subject-verb	1658	51.88	459	14.40	1079	33.80
			459/1538	29.84	1079/1538	70.16
			1538 (48.12%)			

According to Tables 1 and 2, we can verify that there exists a larger tendency of occurring pauses between subject and predicate than between verb and object. However, in both cases the prohibition of the grammar books of inserting punctuation in these sequences has not inhibited the presence of long pauses, albeit less frequent than short pauses. Besides the pauses observed in Groups 1 and 2, we also verify that the pauses that are not associated with punctuation precede syntactic boundaries (29.23% correspond to conjunctions, 17.48% prepositions, and 4.13% adverbs). In these cases 81.22% are short pauses.

Table 2: Number and occurrence frequency of the Group 2 sequences with and without pauses

Group 2	Without pauses		Long pauses		Short pauses		Total
	Number of pauses	%	Number of pauses	%	Number of pauses	%	
Verb-object	1489	98.30	11	0.70	15	1	1515
Total	1489 (98.30%)		25 (1.70%)				100%

Examining the data that present pause between subject and verb, we notice that they seem to be associated with topic-comment structures. However, Cagliari (1992) points out that in written texts topic-comment structures should contain comma. As in the written texts there is no comma, we expect that in reading aloud the topic-comment structure has not occurred. In such a structure the speaker divides the statement into two tonal groups. A pause could (or not) be placed between a tonal group and another. Our speaker reinforces such a structure, since her reading presents the two tonal groups and pauses.

Topic-comment structure is a characteristic inherent to the spoken language. In Brazil the spontaneous speech presents a large quantity and diversity of clauses with topic-comment structures (Pontes, 1987). Nevertheless, as the grammar disapproves such structures, they do not appear widely in written language, as occurs in spontaneous speech.

Our data show the presence of topic-comment structures in the text reading, since we notice that 48.12% of the subject-verb sequences are interrupted by pauses. On the other hand, the large tendency of short pauses (70.16%) between subject and verb shows that in a certain way in reading, the grammatical objection inhibits the presence of long pauses. Thus, based on the results here shown we can verify the presence of topic-comment structures in reading, ratifying the data obtained by other authors (Pontes, 1987; Cagliari, 1992; Mollica, 1993).

In Campione & Véronis (2002) a multilingual study about pauses in read speech and its relationship with punctuation is presented. They show that 11.9% of pauses are not associated with punctuation in the French language. The largest percentage of this kind of occurrence is found in the Italian language, in which 33% of pauses are not associated with punctuation. However, Vanier (1999 *apud* Campione & Véronis, 2002) presents a study about the same theme, in which 36% of pauses are not associated with punctuation in the French language. Our data seems to indicate that Brazilian Portuguese is the language that presents a larger occurrence of pauses not associated with punctuation (61.36%), and in general the results have shown that pauses that are not related to punctuation marks are mostly short pauses (see Fig. 1 and Table 3).

The pauses related to punctuation marks (whose plots are also shown in Fig. 1) include: (a) when concerning commas, there is a slight tendency for them to be short; (b) when referring to semicolons and colons there is a strong tendency to be long (with semicolon presenting an intermediate duration) (see Table 3). However, we have not observed distinct clusters in our data, since overlapping occurs between these three classes. The total average duration of the pauses analyzed within the sentences is 224 ms. Data that present punctuation marks which do not correspond to a pause are few (less than 1%).

Data related to the topic-comment structures in the written text do not present continuity violation. This fact means that the speaker produces entire constituents without interrupting them, as described in Strangert (2004). These pauses are inserted before syntactic boundaries. However, in the reading it is possible to perceive a few cases of boundaries in syntactically unmotivated positions. We have found in our corpus less than 1% of cases involving pauses without syntactic

motivation. This occurrence shows the continuity violation defined by Strangert (2004), as can be verified in the following example:

*... dentre **outros** [short pause] **agentes** de doenças.* (*...among other [short pause] disease agents.*)

In this example, the pronoun *outros* and the noun *agentes* form a syntactic constituent (noun phrase), which is interrupted by a short pause, representing a continuity violation.

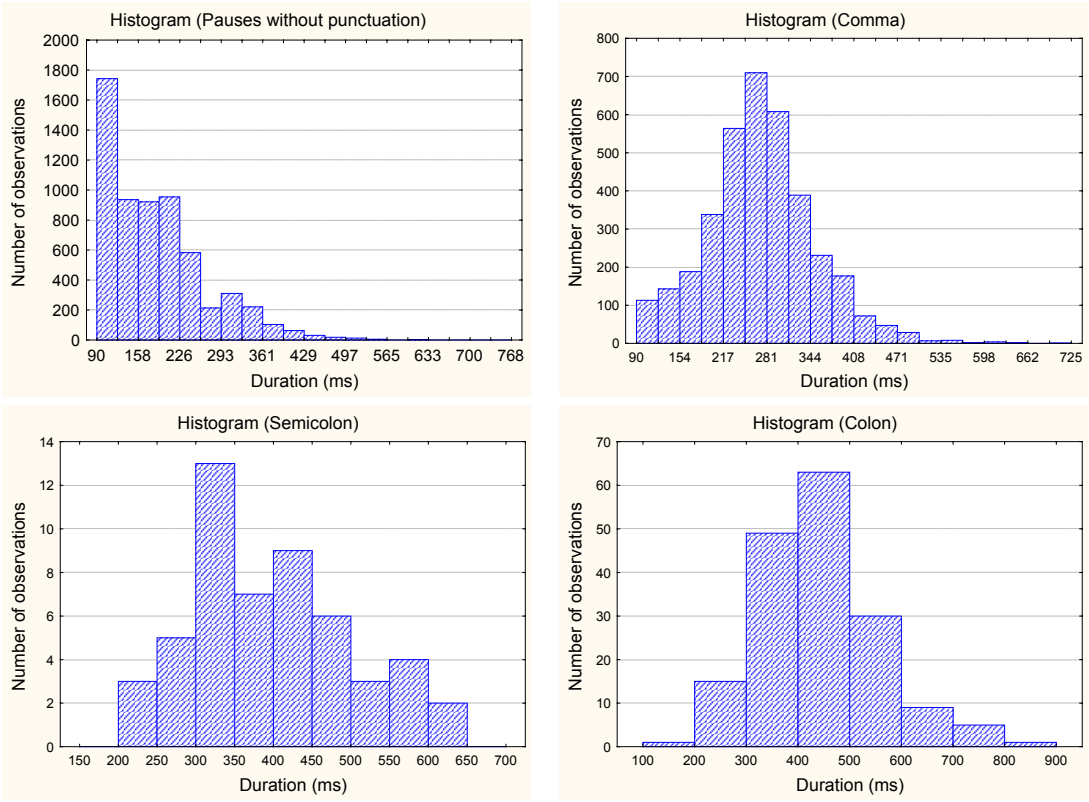

Fig. 1: Duration histogram of pauses: (top left) pauses not associated with punctuation; (top right) associated with comma; (bottom left) with semicolon; (bottom right) with colon.

Table 3: Average duration of pauses for the analyzed data

Pause Classification	Occurrence		Duration (ms)	
	Number	%	Average	Std. Dev.
[long pause]	759	7.60	356	58
[short pause]	5368	53.76	163	54
Comma [long pause]	1194	11.96	357	54
Comma [short pause]	2439	24.43	231	51
Semicolon [long pause]	44	0.44	426	91
Semicolon [short pause]	8	0.08	259	23
Colon [long pause]	157	1.57	478	103
Colon [short pause]	16	0.16	252	32
Total	9985	100		

5 Conclusions and Remarks

In this paper we have discussed the relationship between pauses and punctuation in the Brazilian Portuguese language. We have shown that the claim where there is pause there is also punctuation is not accurate, since we notice a large proportion of pauses without punctuation (61.3%). Also the categorical claim where there is no pause there is no punctuation is not quite accurate, because in 1% of the cases punctuation without pause has occurred. In addition, our data reinforce the presence of topic-comment structures in reading. This fact was not expected due to the objection to that structure. The results here presented with respect to pause and punctuation are consistent with several studies about this theme. However, other analyses using more speakers of Brazilian Portuguese are needed for obtaining a generalization of these findings as inherent to such a language.

References

ALMEIDA N. M. DE (1988), *Methodic Grammar of the Portuguese Language* (in Portuguese), São Paulo, Brazil, Saraiva.

CAGLIARI L. C. (1992), On the importance of the prosody in the description of grammatical events (in Portuguese), In: ILARI, R. (org.) *Grammar of the spoken Portuguese* (in Portuguese), v.2. Campinas (SP), Brazil, Unicamp.

CAMPIONE E., VERONIS J. (2002), Etude des relations entre pauses et pontuations pour la synthèse de la parole à partir de texte. Proceedings of *Traitement Automatique des Langues Natureles* (TALN 2002), Nancy, France, 1-10.

FARACO C. E., MOURA F. M. DE. (1991), *Grammar: phonetic and phonology, morphology, syntax, stylistic* (in Portuguese), São Paulo, Brazil, Ática.

MARIN R., AGUILAR L., CASACUBERTA D. (2002) Placing pauses in read Spanish : a model and an algorithm.*Language Design,* 4, 46-66.

MOLLICA C. (1993). Intervals between the silence and the speech and their representations in written text. *Cadernos de Letras* (in Portuguese), no. 9, Rio de Janeiro, 143-149.

PONTES E. (1987) *The topic in the Brazilian Portuguese* (in Portuguese), São Paulo, Brazil, Pontes.

SEARA I. C. (2000), Analysis acoustic-perceptual of nasality of the vowels in the Brazilian Portuguese. *PhD. Thesis* (in Portuguese), Federal University of Santa Catarina, Florianópolis, Brazil.

SEARA I. C., KAFKA S. G., KLEIN S., SEARA R. (2002), Vowel sound alternation of verbs and nouns of the Portuguese spoken in Brazil for application in text-to-speech synthesis. *Journal of the Brazilian Telecommunication Society* (in Portuguese), vol. 17, no. 1, 79-85.

STRANGERT E. (2004), Speech chunks in conversation: Syntactic and prosodic aspects. Proceedings of *Speech Prosody*, Nara, Japan.

VANNIER G., LACHERET-DUJOUR A., VERGNE J. (1999), Pauses location and duration calculated with syntactic dependencies and textual considerations for T.T.S. system. Proceedings of *XIV International Congress of Phonetics Sciences (ICPhS)*, San Francisco, USA.