ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Modelling pause duration as a function of contextual length

David Doukhan, Albert Rilliard, Sophie Rosset, Christophe D'Alessandro

LIMSI-CNRS, UPR 3251, 91403 Orsay - France
{doukhan,rilliard,rosset,cda}@limsi.fr

Abstract

Effects of contextual length are known to affect pause durations in neutral speech. The present study investigates these effects on an expressive corpus of read tales in French. Computational models of intra-sentence, and inter-sentence pause durations, as functions of contextual lengths are proposed. These models are aimed at improving Text-To-Speech synthesis systems, and provide clues for synthesizing prosodic instructions above the level of the sentence. They are also aimed to help in the prosodic analysis of pause durations, which may be biased by contextual length effects. We find the phoneme to be the best unit for measuring contextual length. Inter-sentence pause durations were more influenced by the length of the preceding sentences. Intra-sentence pause durations were more influenced by the length of the following pseudo-clauses.

Index Terms: pause duration, prosodic analysis, Text-To-Speech Synthesis

10.21437/Interspeech.2012-203

1. Introduction

Mainstream Text-To-Speech (**TTS**) systems are able to produce natural sounding speech. However, their stereotypic prosodic capabilities result in monotonous speech streams. While this limitation is not crucial for some applications (vocal servers), it is more problematic when considering the synthesis of longer texts (audio books, child-directed robots). The trade-off to be done between sound naturalness and prosodic control [1] is partly responsible for this limitation. Pause duration and localization have been shown to improve the expressiveness of synthesized speech [2]. Predicting pauses in TTS systems is particularly relevant to this aim, since it can be achieved without inducing synthesis artefacts.

Several studies carried on neutral speech, uttered by non-professional speakers, reported correlations between contextual length and pause durations. [3] reported correlations between utterance initiation time and their corresponding length. [4, 5] reported context length effects on pause duration.

This study describes the contextual length effects on pause durations observed on expressive speech. Our observations were done on a corpus of read tales in French [6, 7], uttered by a professional speaker. Models of pause duration as function of contextual length are proposed, aimed at transposing the observations done by [3, 4, 5] to the field of expressive TTS. The models proposed allow predicting the duration of inter-sentence pauses and intra-sentence pauses, providing answers to the necessity of synthesizing prosodic instructions above the level of the sentence. Given that contextual length has consequences on pause duration, it also constitutes a bias when analyzing effects linked to expressive functions of pauses. The second aim of this work is to propose normalization procedures of pause durations, with respect to their sentential context, in order to help measurement of the factors influencing pauses' lengths.

Section 2 presents the properties of the pause distributions found in the tale corpus, and the methodological choices retained for this study. Section 3 present the results obtained using a kernel-based regression algorithm. Section 4 present the results obtained using parametric models of pause duration. Section 5 discusses the relevance of our models for Text-To-Speech synthesis, and prosodic analysis.

2. Material

The GV-LEx read tale corpus [6, 7] was used for the purpose of this study. The corpus contains twelve tales, told in studio conditions by a professional speaker. This recording method allowed to obtain speech material with large prosodic variations, and minimized impact of physiological (breathing) and cognitive (sentence planning) necessities. Phonetic transcription, and phoneme alignment of the speech transcriptions were obtained using the LIMSI semi-automatic software [8, 9]. Pause boundaries (either for silent or breath pauses) were manually checked and corrected. A minimum of 40ms was used as a threshold for pause detection.

Pseudo-clauses were defined as a group of words separated by punctuation marks, e.g.: `One, the brother's wife, was very rich.` Contain three pseudo-clauses: `One` being the first pseudo-clause, `the brother's wife` being the second, and `was very rich` being the third.

Characteristics of the corpus relevant for this study are given in table 1. Large variations were observed in mean and maximal sentence size per tale, highlighting

Table 1: *Global characteristics of the 12 tales of the corpus. Average and Maximal size reported for sentences and breath groups is calculated as their respective number of phonemes.*

	Min	Max	Mean
Tale duration (second)	215	374	299.4
Word count	626	1031	805.3
Phon. count	1995	2914	2374.2
Filled pause count	1	33	11.5
Paragraph count	10	33	19.9
Sentence count	43	122	67.25
Pseudo-clause count	80	215	129.75
Avg Sentence size	22.2	58.4	38.3
Max Sentence size	76	197	117
Speech Rate	5.6	6.6	6.2
Avg Breath group size	12.5	19.7	17.1
Max Breath group size	46	85	55.6

the diversity of tale syntactic content. The recordings conditions, and the professionalism of the speaker, resulted in low filled pauses count, high speech rates (expressed in number of syllables by second, excluding pauses) and long breath group sizes (number of syllables between two pauses).

Table 2 describes pause instantiations, given four mutually exclusive categories: paragraph boundary (**PB**), sentence boundary (**SB**), pseudo-clause boundary (**PCB**), and within pseudo-clause (**PCW**). The number of occurrences reports the number of boundaries found for a given category. For the within pseudo-clause category, it refers to the amount of pause realization not associated to punctuation marks. The realization percentage refers to the amount of boundaries that were eventually associated to a pause realization. Median pause duration was calculated using the realized pause durations.

29% of within sentence punctuation marks (**PCB**) were not associated to a pause realizations. Together with the amount of pause realizations not associated to punctuation signs (**PCW**), we observed 41% of intra-sentence pause realizations not associated to punctuation marks, accounting for a speaker expressive reinterpretation of sentence syntactic structure.

Pause durations were converted to the logarithmic scale, which is more relevant from a perceptive point of view. We used the $\log_{10}$ value of the pause durations expressed in milliseconds ($\log_{10}(10\ ms) = 1$, $\log_{10}(100\ ms) = 2$, $\log_{10}(1000\ ms) = 3$). While pause distributions (see figure 1) show overlaps, and a quite large dispersion, they still mark a hierarchical tendency: $PCW < PCB < SB < PB$. The length of the preceding and following sentences were used as contextual length descriptors for pauses occurring at

Table 2: *Pause distributions in the GV-LEx read tale corpus, given the categories defined in section 2*

	number of observations	realization percentage	median duration (ms)
PCW	376		261.0
PCB	750	70.9	385.7
SB	556	98.2	719.9
PB	239	100.0	1126.8

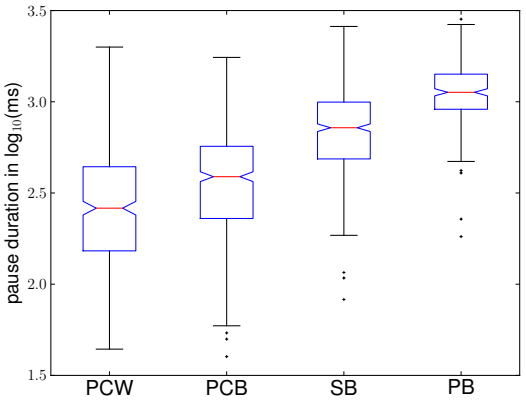

Figure 1: *Variability of pause distributions given the categories defined in table 2.*

sentence boundaries (**SB**). For pseudo-clause boundaries (**PCB**) we considered the length of the preceding and of the following pseudo-clause. Length of sentences and pseudo-clauses was either computed as their respective number of phonemes, characters, syllables, or words.

The pause realization rate associated to pseudo-clause boundaries (**PCB**) instances was 54% when the lengths of the adjacent pseudo-clauses were below 12 phonemes, and 87.4% when both of these lengths were greater than 24 phonemes. This indicates a clear dependence between pseudo-clause boundary contextual length, and the pause realization phenomenon.

In the following experiments, 40ms long pauses were assigned to pseudo-clause and sentence boundaries that were not associated to a pause realization. This choice was aimed to integrate pause deletion phenomenon in the pause duration models proposed.

3. Kernel Regression

The Nadaraya-Watson regression technique [10] was used to model pause durations, leading to equation 1:

$$dur(cont) = \frac{\sum_{i=1}^{n} K(cont, cont_i)\ dur_i}{\sum_{i=1}^{n} K(cont, cont_i)} \qquad (1)$$

with dur representing the pause duration function to infer. The function argument $cont = (lp, lf)$ is a

bidimensional vector representing the contextual length information of the pause duration to predict, with lp being the length of the preceding unit, and lf the length of the following unit. The $cont_i = (lp_i, lf_i)$, and dur_i being respectively the ith contextual length and pause duration observed in the corpus. n stands for the number of observations, for a given pause category (sentence boundary or pseudo-clause boundary).
The kernel function K allows the weighting of the observations, according to their distance to the point to approximate. We used the kernel function K defined in equation 2, with h being the kernel bandwidth, used as a smoothing parameter.

$$K(x, y) = \exp(-h||x - y||^2) \quad (2)$$

Figure 2 and 3 display duration prediction functions inferred from the corpus, using the number of phonemes to calculate the length of the context. For distinctiveness purposes, contextual lengths above the 9th decile were not displayed in these figures.

Figure 2 refers to an approximation of inter-sentence pauses, and was obtained using a Kernel parameter of 0.005, predicting pause durations between 427 and 912 ms at sentence boundaries. Function presented in figure 3 predicts pauses between pseudo-clauses, and uses a Kernel parameter of 0.02. Pause durations reported in this figure lay between 93 and 347ms.

4. Parametric models

While providing optimal fit on the data, kernel regression methods suffer from several limitations. They use opaque knowledge representation mechanisms, similar to black box devices. They may be sensitive to noise when the data is sparse, and hardly allow to constrain the solution space.
Based on the observation of figure 2 and 3, we defined a total of 16 parametric models aimed at calculating pause durations as increasing functions of the preceding and following context length. The models used composition of a few basic functions: exponentiation, logarithm, addition, multiplication, minimum. The compositions were then combined using additive and multiplicative operations. The resulting models include polynomials of order less than 4, as well as polynomials considering the log of the contextual length.
Model parameters were inferred using Matlab `nlinfit` package, which performs non-linear least square regression. To avoid over-fitting, models were benchmarked using 20-fold cross-validation. Several contextual length descriptors were tested: number of phonemes, number of characters, number of syllables, number of words. Best results were systematically obtained using the number of phonemes as the measure of context length. Table 3 reports the average squared error obtained on the testing sets by these models, using the phoneme count as measure of contextual length. The performance of the model returning the constant value satisfying the least-square criterion was also reported, to quantify the contribution of contextual length models in explaining pause duration. The mean of the squared errors obtained on the testing sets, and their respective standard deviation, did not allow distinguishing a particular model as being the more appropriate to the pause duration prediction task. Consequently, we reported maximal and minimal mean squared error obtained by the models. The mean squared error difference between best and worst models was always smaller that the difference between the worst and the constant model. The predictions of sentence boundary pause durations (**SB**) were slightly better than those obtained for pseudo-clause boundary, excluding the pseudo clause boundaries not associated to a pause realization (**PCB***). The efficiency of the model taking into account all pseudo-clause boundaries (**PCB**)

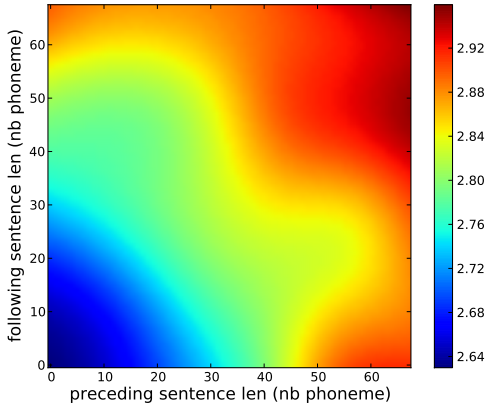

Figure 2: *Inter-sentence pause duration ($\log_{10}(ms)$) as function of preceding and following sentence length*

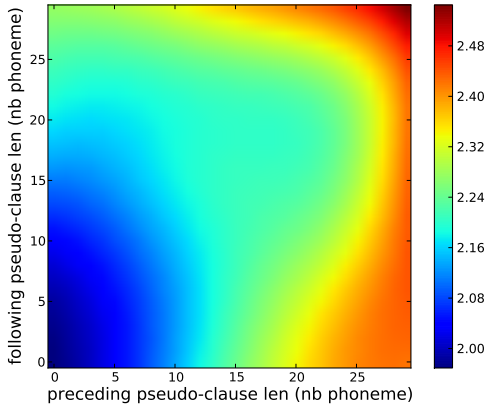

Figure 3: *Intra-sentence pause duration ($\log_{10}(ms)$) as function of preceding and following pseudo-clause length*

Table 3: *Mean squared error obtained by parametric models on the pause duration predictive task related to sentence boundaries (SB), pseudo-clause boundaries (PCB), pseudo-clause boundaries excluding boundaries not instantiated to a pause realization (PCB*)*

task	worst model	best model	constant model
SB	0.069	0.070	0.077
PCB	0.213	0.221	0.244
PCB*	0.071	0.073	0.077

was much lower, with respect to the difficulty of predicting both pause duration and pause deletion phenomena in expressive speech.

The obtained models were then manually inspected. Linear models trend to over-estimate pause duration for extreme contextual lengths (very small, or very large). Logarithmic models trend to underestimate pause duration for extreme contextual length. Polynomial models trend to overfit the data. Equations 3, 4, and 5 reports the parameters found for the simplest duration models benchmarked, respectively obtained on the SB, PCB and PCB* task. lp and lf being the length in number of phonemes of the preceding, and following contextual unit.

$$dur_{SB}(lp, lf) = 0.0026\ lp + 0.0017\ lf + 2.6696 \quad (3)$$

$$dur_{PCB}(lp, lf) = 0.0068\ lp + 0.0129\ lf + 1.9521 \quad (4)$$

$$dur_{PCB*}(lp, lf) = 0.0031\ lp + 0.0038\ lf + 2.4337 \quad (5)$$

5. Conclusion and future work

This study demonstrates contextual length effects on inter-sentential and intra-sentential pause durations, on an expressive read-speech corpus. The findings of this study should ultimately be applied to expressive Text-To-Speech synthesis. Context length information may help to increase the variability of TTS output, resulting in less monotonous speech. It may also be used for normalizing pause durations, with respect to their context length, and observing other phenomena related to expressivity. Similar experiments were done for predicting pause duration associated to paragraph marks, without obtaining convincing results. The validation of the parametric and non-parametric models that were proposed should ultimately be done through perceptual experiments.

Given the mean squared error estimator, no parametric model was found to perform significantly better in the prediction of pause durations. However, some interesting properties were inferred from the observation of the simplest models. Models predicting inter-sentence pause duration (equation 3) were found to give more importance to the length of the preceding sentence, which is consistent with [4] observations. The bigger importance of the length of the following pseudo-clause for predicting intra-sentential pause duration (equations 4 and 5) was not, to our knowledge, reported in the literature. The same conclusions were drawn considering the logarithmic and polynomial models. The significantly lower fit of the models predicting intra-sentence pause durations is explained by the high amount of within-sentence punctuation not instantiated by a pause realization. Expressive speech is known to contain high amount of pausing irregulaties [2]. Improvements of intra-sentencial pause models will require to separate the prediction of pause realization from the prediction of pause duration.

Other descriptors were proposed in the literature to predict pause duration: measures of syntactic complexity [3], surrounding intonational phrases length, and prosodic structure complexity [5]. Future work may include investigating such descriptors, given the TTS inherent constraint that they should be inferable automatically. Several studies (see [11]) reported correlations between speech rate and pauses. Since our material was obtained from a single speaker, pronouncing an average of 6.2 syllables per seconds, the validity of the models proposed is constrained to similar speech rate. Empirical studies should define how to map our observations to different speech rates.

6. Acknowledgements

This work has been funded by the French project GV-LEx (ANR-08-CORD-024 `http://www.gvlex.com`).

7. References

[1] F. Burkhardt and J. Stegmann, "Emotional speech synthesis: Applications, history and possible future," *Proc. ESSV*, 2009.

[2] X. Wang, A. Li, and C. Yuan, "A preliminary study on silent pauses in mandarin expressive speech," in *Speech Prosody*, 2008.

[3] F. Ferreira, "Effects of length and syntactic complexity on initiation times for prepared utterances," *Journal of Memory and Language*, vol. 30, no. 2, pp. 210–233, 1991.

[4] E. Zvonik, "Pausing and the temporal organization of phrases. an experimental study of read speech." Ph.D. dissertation, National University of Ireland, 2004.

[5] J. Krivokapić, "Prosodic planning: Effects of phrasal length and complexity on pause duration," *Journal of phonetics*, vol. 35, no. 2, pp. 162–179, 2007.

[6] D. Doukhan, A. Rilliard, S. Rosset, M. Adda-Decker, and C. d'Alessandro, "Prosodic analysis of a corpus of tales," in *InterSpeech*, 2011, pp. 3129–3132.

[7] D. Doukhan, S. Rosset, A. Rilliard, C. d'Alessandro, and M. Adda-Decker, "Designing french tale corpora for entertaining text to speech synthesis," in *LREC*, 2012.

[8] M. Adda-Decker and L. Lamel, "Pronunciation variants across system configuration, language and speaking style," *Speech Communication*, vol. 29, no. 2-4, pp. 83–98, 1999.

[9] J. Gauvain, G. Adda, M. Adda-Decker, A. Allauzen, V. Gendner, L. Lamel, and H. Schwenk, "Where Are We in Transcribing French Broadcast News?" in *Ninth European Conference on Speech Communication and Technology*. ISCA, 2005.

[10] E. Nadaraya, "On estimating regression," *Teoriya Veroyatnostei i ee Primeneniya*, vol. 9, no. 1, pp. 157–159, 1964.

[11] F. Grosjean and M. Collins, "Breathing, pausing and reading," *Phonetica*, vol. 36, no. 2, pp. 98–114, 1979.