Pauses and respiratory markers of the structure of book reading

Gérard Bailly, Cécilia Gouvernayre

Pauses and respiratory markers of the structure of book reading

Gérard Bailly & Cécilia Gouvernayre

GIPSA-lab, UMR 5216 CNRS/INPG/U. Stendhal /UJF, Grenoble -France

gerard.bailly@gipsa-lab.grenoble-inp.fr

Abstract

The automatic reading of books by text-to-speech synthesizers requires not only the adequate encoding of the many levels of information and discourse structures in the acoustic signals but also the proper patterns of breathing, so that to pace information and organize discourse at an ecological rhythm.

We analyze here the locations and durations of near 4,000 pauses produced by voice donor who has read several audiobooks, freely available on the web. Since the voice was recorded by a close microphone, we also characterized the acoustic markers of inhalation and show that the delay between end of phonation and air intake can be considered as an additional marker of thematic continuity between the two adjacent speech chunks that complements well-documented prosodic cues such as the preboundary tone and lengthening or the pause duration.

Index Terms: prosody, pause, respiration, prediction of pause locations and durations

1. Introduction

Most of the generative models of prosody for text-to-speech synthesis are designed and trained using large sets of sentences or short paragraphs. More recently, the development of storytelling engines [1] and the use of very large corpora such as audiobooks [2-3] for maximizing the coverage of context and building speaker-specific reading strategies motivate a renewed interest in the intricate patterning of speech and respiration. In other words, we paraphraze our colleague Daniel Hirst:

"Speech = text + prosody + respiration"

We obviously need to coordinate speech and breathing to plan and structure our discourse as well as refill the air volume of the lungs. Central pattern generators (CPGs) – assemblies of neurons located in the pons and medulla – are responsible for the generation of rhythmical movements such as locomotion, mastication, swallowing or breathing [4] and these CPGs are modulated by higher centers of the brain – especially the inferio-lateral region of the sensorimotor cortex – and sensory receptors as well as coordinated each other. Atypical patterns of coordination can impact fluency [5].

We indirectly analyze the coordination of speech and breathing by characterizing (a) the patterns of silence and phonation and (b) the acoustic markers of inhalation during silent pauses when observed.

2. State of the art

The patterns of respirations when reading are very typical: slow expirations follow rapid or even partial inhalations [6-7]. If the durations of respiratory pauses are loosely correlated to surrounding phrase lengths [8-9], subjective and silent (with or without inhalation) pauses during speech tend to occur at phrase or sentence boundaries [10-11]. Their positions and durations are strongly influenced by linguistic organization, in particular grammar [11]. Their characteristics combine with linguistic (such as syntactic structure or lexical choices), prosodic cues before or at the boundary (prepausal lengthening, boundary tones) to signal links between previous speech chunks with the upcoming flow of speech. This listener-oriented planning combines with speaker-oriented constraints, since pauses also serve to text understanding and speech planning [12]. Most authors however agree on the preponderance of listener-oriented strategies on speaker-oriented constraints [13].

Prediction of pause locations and durations [14-19] often combine models of prosodic phrasing and pausing. These models use phonotactic (typically the number of syllables of diverse adjacent units) and syntactic (usually a local window of parts-of-speech (POS) surrounding each candidate location or more complex syntactic analysis) information as input of decision and regression trees. The baseline location model that assigns pauses to punctuations is rather difficult to beat [19-20]. For the prediction of pause durations, rather low correlations with observed data are often obtained: 0.4 [21] to 0.6 [18] are typically reached for sentence-internal pauses. Few quantitative evaluation can be found for entire texts [22].

3. Pause analysis

Our corpus consists of the first 15 chapters of the original French version of "Around the World in 80 Days" by Jules Verne, read by Damien Genevois[1]. The total audio size is 2:35 hours. The signal has been automatically aligned with the phonetic transcription of the text tagged with POS information (25616 words) by our French TTS. All phonetic labels and tags have been checked and corrected by hand when necessary. Textual information – such as the structure of paragraphs and punctuations – was preserved in the labeling.

In absence of physiological signals for helping us to parse the audio signals into breath groups, we detect inspiratory loci based on listening. This task is eased by the fact that the speaker used a close microphone. Following Wang et al [23], minimum pause duration for pause with inhalation is 250 msec. Whenever possible, the presence of inhalation noise during silent pauses was labeled since breath sounds as well as noises produced by mouthing are known to contribute to the structuring of dialogs and benefit to synthesis quality [24].

3.1. Number and distribution of pauses

The average phonation rate is 5.67 syllables/s. We distinguish between 4 types of pauses (see Table 1):

[1] www.litteratureaudio.com/livre-audio-gratuit-mp3/jules-verne-le-tour-du-monde-en-80-jours.html

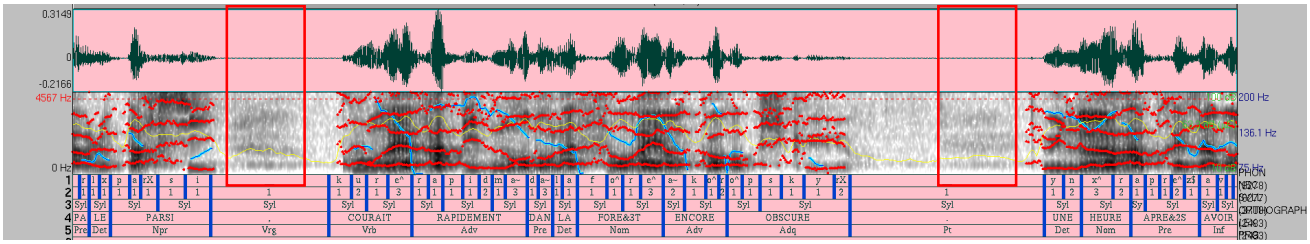

Figure 1: Examples of a close (left) vs. a delayed (right) breath noise, respectively associated with a sentence-internal (associated here with a comma) vs. a sentence-final pause (associated here with a full stop). Red rectangles enlightened their locations in the spectrogram.

- Syntactic pauses (S): short pauses (<200ms) produced with no inhalation
- Sentence-internal pauses (SI): mid and long pauses associated with commas, colons or major punctuations followed by a verb (i.e. relating turns) or located at major syntactic boundaries (i.e. at the edge of following phrases beginning with a verb, a preposition or a coordination as shown in Figure 2). We note SIp and SIb the pauses respectively associated with punctuations and non-marked boundaries.
- Sentence-final pauses (SF)
- End-of-paragraph pauses (EP): a paragraph is cued by two carriage returns in the original text.

A total of 3772 SI, SF and EP pauses have been produced. Average number of syllables between these pauses is 7 (see Figure 3).

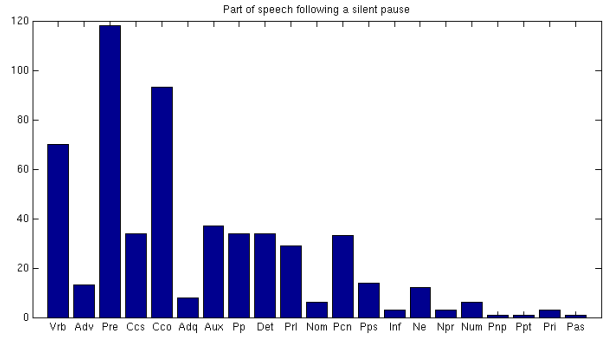

Figure 2. Distribution of sentence-internal – but not elicited by punctuations – pauses according to the POS of the next word. Coordinations, prepositions and verbs feature half of the occurrences.

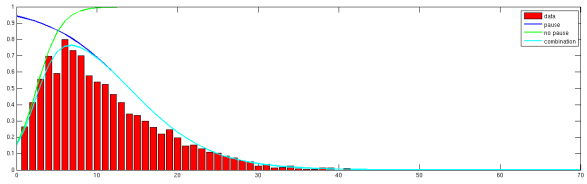

Figure 3. Distribution of the number of syllables between non syntactic pauses. This histogram is fitted by two distributions: (1) probability to have a pause after x consecutive syllables without pause (green curve) and (2) probability to have no pause after x consecutive syllables without pause (blue curve). This is used in the prediction of pause locations (section 4).

3.2. Pause durations

We actually compute the amount of final lengthening by adding to pause duration the length of the preceding rime and following onset (i.e. from the onset of a vowel to the next): this corresponds to the Inter P-Center Group (IPCG) we promoted in our analysis of speech rhythm [25].

As already evidenced by several authors [10, 26], the overall probability density function (pdf) of the pause duration is multimodal and reflects the importance of the syntactic breaks: the four underlying modes are clearly lognormal and explain nicely the observed modes (see Figure 4). The mean durations of the modes are given in Table 1. They display the "quantal" effects we already mentioned in [25]: average IPCGs that include a pause equal to 2.07, 3.92, 4.95 & 8.08 syllables, i.e. close to integer multiples of the mandible cycle. Note that this phenomenon could be specific to the speech style and French but this deserves considerations in further research.

Table 1. Realizations of pauses according to positions.

Position	Realized	Non-realized	Mean IPCG (ms)
S	807		366±159
SIp	1521	1025	692±172
SIb	559		
SF	912	92	874±293
EP	780	7	1426±425

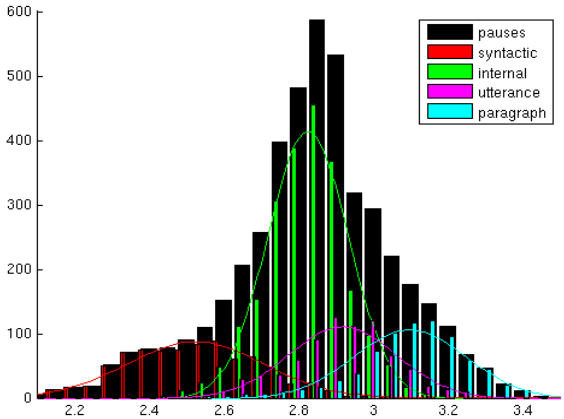

Figure 4: Distributions of log10(durations of IPCG with pauses) according to their positions in the text. From the shortest to the longest ones: syntactic, sentence-internal, between-sentences and between paragraphs pause (see text for explanations).

3.3. Inhalation phasing

We labeled breath sounds when an obvious acoustic trace can be detected on the spectrogram: 69.7% of the mid and long pauses include such one or more audible breath sounds. The average duration of a breath is 355±145 ms (see next figure for the pdf of breath durations). Despite the fact that the breath durations are weakly correlated with pause durations (R=0.65), the pdfs according to pause position are much more overlapping than pdfs for pause durations (cf. Figure 5). Noise typically ends close to phonation onset but the delay between the end of the previous phrase and noise onset is however multimodal (see Figure 5). While this delay is very short for SI, both SF and EP often exhibit large delays: in the first cases, breathing noise act as a pause filler and signals – often together with a rising tone – that the upcoming phrase completes the sentence while, in the second case, the absence of close inhalation – often together with a falling tone and a large pre-boundary lengthening – signals the end of the theme (cf. Figure 5). Note thus that large delays are not systematically associated with long EP: close and delayed breath noises really cue thematic continuity/discontinuity independently of pause durations.

Note that the instances of thematic continuities that seem to be contradicted by late breath noises – very few: only 20 junctures between paragraphs are miss-classified according to breath position – are cued by other features, such as open syntactic phrases or non-final prosodic tones.

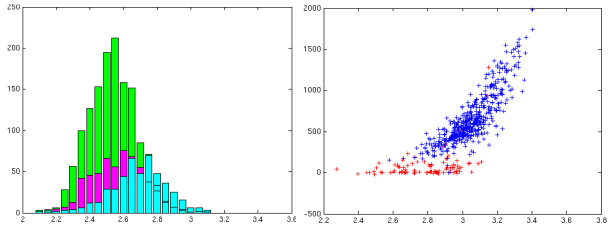

Figure 5: Breathing noises. Left: distribution of log10(noise durations); right: delays (ms) between pause onsets and noise onsets as a function of log10(pause durations) for EP. EP with thematic breaks are figured with dark dots: they exhibit larger delays than EP with thematic continuity.

4. Prediction of pause locations

Syntax-based prediction. We used Wapiti [27] to predict locations from previous, current and following POS (signaling also sentence- and paragraph-final punctuations). The optimal window encloses at least one previous and two following POS. We got similar results with decision trees using the Matlab implementation of C4.5 with a 10-fold cross validation. Accuracy is close to .95 and specificity close to .98 for a F-score equal to 0.8. This confirms data from Figure 2 showing that pauses are placed at major syntactic breaks cued by the incoming clause. Recall is however lower and close to .79: prediction based only on syntax tends to underestimate the number of pauses.

Note that the accuracy, recall and F-score degrade respectively at .93, .71 and .70 when commas are withdrawn (POS provided by syntactic analysis being kept the same).

Adding phonotactic constraints. Wapiti provides priors associated with the predicted values. A Markov decision process (MDP) further links states *s* that figure the 2^n patterns of n successive words – each followed or not with a pause – with two actions *e*: generate or not a pause after the word. Transition probabilities are computed from distributions of Figure 3 that compute the probability of emitting a pause or not after x number of syllables with no pause. Emission probabilities – also called rewards – equal to Wapiti priors. Dynamic–time warping operating on log(pb) is used to compute the cumulative function of transitions and emissions. Contributions of transition probabilities are weighted by α (see its influence on pause patterns in Figure 6).

$$D_{s,t} = \max_{\text{sp,e}=\{0,1\}} (D_{sp,t-1} + \log(pp_t(e)) + \propto . \log(pt_t(sp,e))$$

Where $pp_t(e)$ is the prior for hypothesis e at time t and $pt_t(sp,e)$ is the probability of having the pattern [sp e] at time t. Using n=5 and α=0.25, we are able to keep accuracy at 0.95 recall at 0.97 and F-score at 0.8, while increasing recall to 0.83. When commas are withdrawn, the MDP maintains accuracy at .93 while improving the Syntax-based prediction with a Recall and F-score at respectively 0.71 and .72.

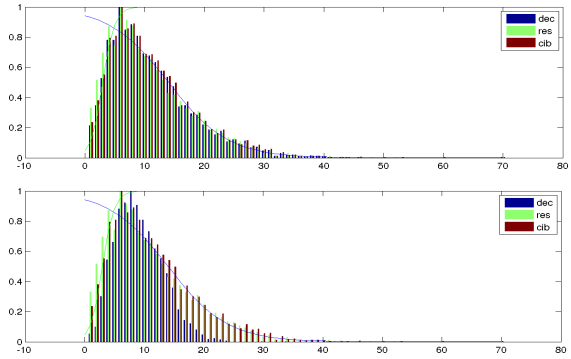

Figure 6: Influence of the weighting of the transition probabilities in the Markov decision process that assigns pauses according to both POS tags (green) and phonotactic information (blue). Original distribution is figured in brown. Top: the optimal weighting; Bottom: a large weight forces inter-pause intervals to be closer to the mean interval.

5. Prediction of pause durations

We used regression trees using Matlab implementation of C4.5 with a 10-fold cross validation. Overall correlation with original pause durations is 0.72 (cf. Figure 7).

We examined if phonotactic information may increase this correlation within each leaf. We used the six following predictors: number of syllables to the next/previous pause location; same for the next/previous sentence-final pause and same for the next/previous paragraph-final pause. The correlation slightly increases to 0.76. The only significant interaction is for SF pauses that are weakly influenced by the number of syllables of the preceding sentence. The coefficient is three times larger than all other weights. This corroborates results obtained by Kentner on isolated sentences [28]: "the effect of phrase length on pausing […] is found to be distinctly stronger for long phrases preceding the pause than for long upcoming phrases." (p. 2637).

6. Conclusions

In this paper we examined the relationship between locations and durations of pauses and their functions. In our data, we evidenced quantal effects in pause planning that tends to align

onsets of nuclei with an isochronous syllabic clock. This is surely dependent of speaker and language but deserves further attention in subsequent research or revisit of previous results.

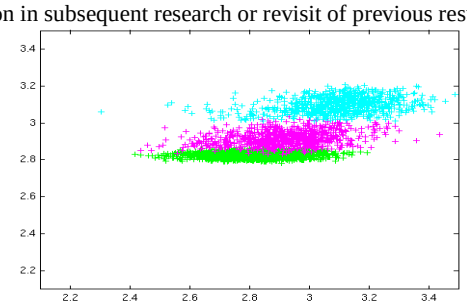

Figure 7: Predicted vs. original pause durations. Multilinear regression with phonotactic predictors within each pause type.

We also show that breathing noise may be used by speakers to cue syntactic/semantic/thematic cohesion between adjacent speech units. Breath noises can be used as fillers to signal close upcoming complementary information to what has just been said while unfilled silence may inform the listener that cognitive processes are engaged by the speaker and should be triggered by the listener to process fresh information. Text-to-speech systems would certainly benefit of a more detailed modeling and synthesis of breath noises and their timings according to the information structure of the discourse. We have suggested that early breaths may signal thematic continuity but the spectrum of motivations may be larger.

We are planning to investigate these issues with synchronous recordings of acoustic and respiratory movements in order to sort out the impact of physiological and communicative needs on the coordination of respiration and phonation. Finally perception tests should be performed to question the subjective impact of the various constraints and strategies used by that particular speaker to breathe and encode information structure.

7. Acknowledgements

We thank volunteer voice donors of the numerous associations that promote reading and give free access to literature.

8. References

[1] Theune, M., K. Meijs, D.K.J. Heylen, and R.J.F. Ordelman, *Generating expressive speech for storytelling applications.* IEEE Transactions on Audio, Speech and Language Processing, 2006. **14**(4): p. 1137-1144.

[2] Prahallad, K., E.V. Raghavendra, and A.W. Black. *Learning speaker-specific phrase breaks for text-to-speech systems*. in *Speech Synthesis Workshop*. 2010. Kyoto, Japan. p. 162-166.

[3] Prahallad, K. and A.W. Black. *Handling large audio files in audio books for building synthetic voices*. in *Speech Synthesis Workshop*. 2010. Kyoto, Japan. p. 148-153.

[4] Lund, J.P. and A. Koltab, *Brainstem circuits that control mastication: Do they have anything to say during speech?* Journal of Communication Disorders, 2006. **39**(5): p. 381–390.

[5] Denny, M. and A. Smith, *Respiratory control in stuttering speakers. Evidence from respiratory high-frequency oscillations.* Journal of Speech, Language, and Hearing Research, 2000. **43**: p. 1024-1037.

[6] Conrad, B. and P. Schönle, *Speech and respiration.* Archiv Für Psychiatrie Und Nervenkrankheiten, 1979. **226**: p. 251-268.

[7] McFarland, D.H., *Respiratory markers of conversational interaction* Journal of Speech, Language, and Hearing Research, 2001. **44**: p. 128-143.

[8] Zvonik, E. and F. Cummins. *The effect of surrounding phrase lengths on pause duration*. in *EuroSpeech*. 2003. Geneva, CH. p. 777-780.

[9] Whalen, D.H. and J.M. Kinsella-Shaw, *Exploring the relationship of inspiration duration to utterance duration.* Phonetica, 1997. **54**: p. 138-152.

[10] Campione, E. and J. Véronis. *A large-scale multilingual study of silent pause duration*. in *Speech Prosody*. 2002. Aix-en-Provence, France. p. 199-202.

[11] Winkworth, A.L., P.J. Davis, E. Ellis, and R.D. Adams, *Variability and consistency in speech breathing during reading: Lung Volumes, speech Intensity, and linguistic factors.* Journal of Speech and Hearing Research, 1994. **37**: p. 535-556.

[12] Goldman-Eisler, F., *Pauses, slauses, sentences.* Language and Speech, 1972. **15**: p. 103-113.

[13] Breen, M., D. Watson, and E. Gibson, *Intonational phrasing is constrained by meaning not balance.* Language and Cognitive Processes, 2011. **26**(10): p. 1532-1562.

[14] Keri, V., S.C. Pammi, and K. Prahallad. *Pause prediction from lexical and syntax information*. in *Proceedings of International Conference on Natural Language Processing (ICON)*. 2007. Hyderabad, India.

[15] Yu, J. and J. Tao. *The pause duration prediction for Mandarin text-to-speech system*. in *IEEE International Conference on Natural Language Processing and Knowledge Engineering*. 2005. Wuhan - China. p. 204 - 208.

[16] Zvonik, E. and F. Cummins. *Pause duration and variability in read texts*. in *International Conference on Speech and language Processing (ICSLP)*. 2002. Denver, Colorado. p. 1109-1112.

[17] Vannier, G., A. Lacheret-Dujour, and J. Vergne. *Pauses location and duration calculated with syntactic dependencies and textual considerations for TTS system*. in *ICPhS*. 1999. San Francisco, CA.

[18] Pfitzinger, H.R. and U.D. Reichel. *Text-based and signal-based prediction of break indices and pause durations*. in *Speech Prosody*. 2006. Dresden, Germany. p. 133-136.

[19] Burrows, T., P. Jackson, K. Knill, and D. Sityaev. *Combining models of prosodic phrasing and pausing*. in *Interspeech*. 2005. Lisbon - Portugal. p. 1829-1832.

[20] Marín, R., L. Aguilar, and D. Casacuberta, *Placing pauses in read spoken Spanish: a model and an algorithm.* Language Design : Journal of Theoretical and Experimental Linguistics, 2002. **4**: p. 49-66.

[21] Apel, J., F. Neubarth, H. Pirker, and H. Trost. *Have a break! Modelling pauses in German speech*. in *Konferenz zur Verarbeitung natürlicher Sprache (Konvens)*. 2004. Vienna, Austria.

[22] Parlikar, A. and A.W. Black. *A grammar based approach to style specific phrase prediction*. in *Interspeech*. 2011. Florence. p. 2149-2152.

[23] Wang, Y.-T., J.R. Green, I.S.B. Nip, R.D. Kent, J.F. Kent, and C. Ullman, *Accuracy of perceptually based and acoustically based inspiratory loci in reading.* Behavior Research Methods, Instruments & Computers, 2010. **43**(3): p. 791-797.

[24] Whalen, D.H., C.E. Hoequist, and S.M. Sheffert, *The effects of breath sounds on the perception of synthetic speech.* The Journal of the Acoustical Society of America, 1995. **97**(5): p. 3147-3153.

[25] Barbosa, P. and G. Bailly, *Characterisation of rhythmic patterns for text-to-speech synthesis.* Speech Communication, 1994. **15**: p. 127-137.

[26] Goldman, J.-P., T. François, S. Roekhaut, and A.C. Simon. *Étude statistique de la durée pausale dans différents styles de parole*. in *Journées d'Etude sur la Parole (JEP)*. 2010. Mons, Belgique.

[27] Lavergne, T., O. Cappé, and F. Yvon. *Practical very large scale CRFs*. in *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*. 2010. Uppsala, Sweden. p. 504-513.

[28] Kentner, G. *Length, ordering preference and intonational phrasing: Evidence from pauses*. in *Interspeech*. 2007. Antwerp, Belgium. p. 2637-2640.