Towards the Adaptation of Prosodic Models for Expressive Text-To-Speech Synthesis

Mathieu Avanzi[1], *George Christodoulides*[2], *Damien Lolive*[3]
Elisabeth Delais-Roussarie[1], *Nelly Barbot*[3]

[1] UMR 7110-LLF (Laboratoire de Linguistique Formelle), Université Paris-Diderot
[2] Centre Valibel, Institute for Language & Communication, University of Louvain, Belgium
[3] IRISA, University of Rennes 1, Lannion, France

mathieu.avanzi@gmail.com, george@mycontent.gr, damien.lolive@irisa.fr
elisabeth.roussarie@wanadoo.fr, nelly.barbot@irisa.fr

Abstract

This paper presents a preliminary study whose main aim is to characterize four distinct speaking styles according to a limited set of prosodic features, including the length of prosodic phrases (AP and IP), the distribution of stressed syllables, pitch register span, the duration of silent pauses, etc. The analysis was performed using semi-automatic procedures on a corpus consisting of 30 minutes of speech per style. The study focuses on four styles, all of which are "overtly addressed to a given audience", but differ as to the nature of the audience (adults vs. children) and the desired impact of the address ("importance of being understood and convincing, or not"). Data analysis reveals that (a) dictation (addressed to children) and political speeches (addressed to adults) are different to the two other speaking styles (reading of novels and fairy tales) with respect to a specific set of prosodic cues; while (b) the speeches addressed to children differ from the ones addressed to adults, with respect to another set of prosodic cues (especially pitch register span). These results have an interesting practical application: refining the design of pre-processing prosodic modules in a text-to-speech system, in order to improve the expressivity of synthesized speech.

Index Terms: accentuation, phrasing, prosody, tempo, dictation, speaking style, pitch register, dictation, read speech.

1. Introduction

In this paper, we study the differences between four speaking styles that are addressed to children (dictation and reading of fairy tales) and adults (reading of novels and political speeches). The main goal is to characterize these four distinct speaking styles according to a limited set of prosodic parameters. These parameters were selected based on two criteria: previous research [1]-[10] that has revealed their relevance as predictors for discriminating speaking styles in French; and the possibility to control them in order to adapt a TTS system to specific audiences. The paper is organized as follows: in section 2, we describe the processing methodology and the corpus used; after a short description of the tools used for analysis, the results are presented in section 3 and discussed in section 4.

2. Methods

2.1. Material

The study focuses on four speaking styles, all of which are read speech "overtly addressed to a given audience": reading of fairy tales (TAL), dictations (DIC), political speeches (POL), and reading of novels (NOV). The four speaking styles differ as to the nature of the audience (adults for POL and NOV, vs. children for DIC and TAL) and the desired impact (importance of being understood and convincing for POL and DIC; less important for NOV and TAL). 30 minutes of speech per style are analyzed. Table 1. details the number of speakers and the exact duration of the samples in our corpus. All participants speak a standard variety of French.

Table 1. *Corpus composition.*

Speaking Style	Nb. of speakers	Nb. of syll.	Nb. of tokens	Duration (sec.)
Tales (TAL)	6F/2M	5942	4189	1065.25
Dictation (DIC)	2F	4175	2918	893.56
Political (POL)	3F/3M	6875	4539	1362.02
Novel (NOV)	2F/2M	7496	5226	1286.97
Total	13F/7M	24488	16872	4607.81

2.2. Data Annotation

The recordings were first orthographically and then transcribed with the usual HMM technique used in forced alignment mode and implemented within EasyAlign [11], a plugin of the Praat software [12]. All alignments were manually verified and corrected by one of the authors by inspecting both spectrogram and waveforms. The orthographic transcription was then annotated with part-of-speech tags using the DisMo software [13]. This allows assigning a phonological status to each word, indicating whether they can be stressed or not (cf. [14]-[16], among other), and then segmenting the data in Phonological Words (henceforth PW).

In addition, prominent syllables were identified by two different ways, once by one of the authors (on the basis of his perceptual judgment only) and by the Analor algorithm [17], which automatically detects prominent syllables on the basis of a reduced set of acoustic parameters. The agreement between the manual and the automatic annotation was statistically measured [18], and found moderate (κ = 0.59) according to [19]. For that reason, a second expert intervened in cases of disagreement between the two annotations and decided the final value of the syllable (+/- prominent). This annotation was entered in a dedicated tier.

Finally, Accentual Phrase (AP) boundaries and Intonation Phrase (IP) boundaries were automatically identified and annotated on two separate tiers. AP boundaries were derived from prominent syllables, and were inserted at the end of any PW whose last metrical syllable is prominent. IP boundaries were inserted after any AP final syllable followed by a silent pause and/or significant lengthening, and associated with a major pitch rise, following the protocol outlined in [20].

2.3. Data Analysis

The dataset was processed by using Praaline [21], a toolkit that interfaces with Praat and runs a cascade of scripts and/or external analysis tools, each of which may add features to an annotation level (e.g. syllables, AP, IP, etc.), storing all annotations in a relational (SQL) database.

For this study, we extracted different prosodic parameters which have been found to play a significant role in studies relating to speaking style discrimination in French [7]-[10]. Regarding accentuation and prosodic phrasing, we focused on **AP Length** (number of syllables per AP), **IP Length** (number of syllables per IP) and **Initial Rise Ratio** (number of prominent PW-initial syllables divided by the total number of PW-initial syllables). Such initial rises have often been described as characteristic of didactic style (see, among others, [22] and [23]). As for temporal variables, we studied **Articulation Rate** (calculated as the mean syllabic duration per IP) and **Silent Pause Duration and Distribution**. We finally evaluated the effects of **Pitch Register** by calculating the difference between the minimum and maximum pitch (f_0) per IP (expressed here in semi-tones, ST). All these prosodic parameters may be of interest to differentiate speeches addressed to children from the other ones. For example, it has been shown that prosodic cues are crucial in "motherese" (see among others [24]).

Data were analyzed by means of Generalized Estimated Equations (GEE) with repeated measures. GEEs are a kind of Generalized Linear Models which are particularly useful to assess significant differences in datasets where the predictors are highly correlated [25]. This is true for the prosodic parameters we chose to study. For example, it has been shown that speech rate (tempo) affects stress (accentuation) and prosodic phrasing: the faster one articulates, the fewer syllables one stresses, the longer the prosodic groups are, the tighter the pitch range is, etc. [26][27][28]. Articulation Rate is also dependent on the size of the constituent it is measured on: in long constituents, syllabic duration tends to be shorter than is short constituents ([29] and [30]). Finally, Bonferroni corrections were systematically applied when examining pairwise comparisons between the levels of a given predictor.

3. Results

3.1. Accentuation and Prosodic Phrasing

For each of the three chosen parameters related to accentuation and prosodic phrasing (i.e. AP length, IP length and Initial Rise Ratio), a GEE model was created, having the prosodic parameter as a dependent variable, and Speaking Style and Local Articulation Rate (henceforth AR) as independent variables. Local AR was calculated as the mean syllabic duration per AP for the models of AP Length and Initial Rise Ratio; per IP for the model of IP Length. It is expressed in ms/syll [31]. Means and standard deviations obtained for each speaking style regarding these 3 prosodic measures are given in Table 2:

First, the statistical analysis reveals that Speaking Style has a significant effect on AP Length (Wald χ^2 (3) = 29.980, $p < .001$). DIC presents shorter AP Length than NOV ($p < .001$), but does not differ from TAL ($p = .094$) and POL. On the other hand, no differences are found between POL, NOV and TAL. The analysis also shows an effect of AR on AP Length (Wald χ^2 (3) = 79.564, $p < .001$), which interacts with speaking style. As can be seen in Figure 1, AR seems to have no effect on AP Length for POL (i.e. the average number of syllables per AP remains constant, and the speaker articulates faster or slower). The effect of AR on AP Length seems to be also slightly different when comparing the three other speaking styles (Figure 1) when predicting the AP Length using the GEE model.

Table 2. *Means and standard deviations for the AP length, IP Length and Initial Rise Ratio, in the 4 speaking styles under study.*

	DIC	POL	TAL	NOV
AP Length (syll/AP)	2.96 (1.25)	3.17 (1.5)	3.16 (1.28)	3.29 (1.34)
IP Length (syll/IP)	4.98 (3.05)	5.66 (3.77)	6.17 (3.68)	7.63 (4.7)
Initial Rise Ratio (%)	63.76 (2.01)	32.05 (1.52)	21.65 (1.41)	22.02 (1.31)

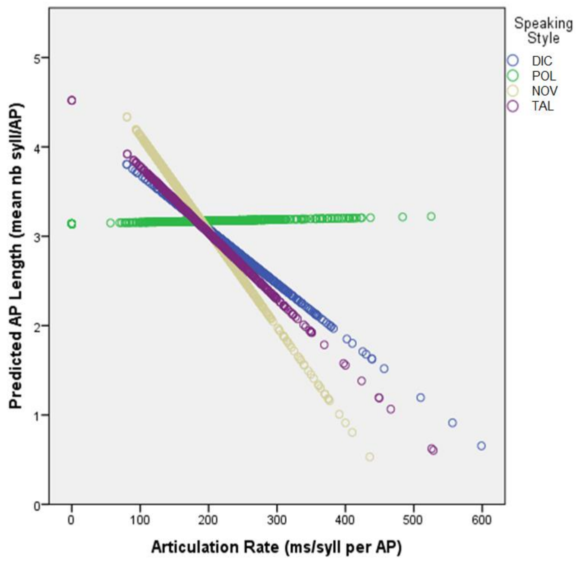


Figure 1: *Predicted AP Length (mean nb of syll/AP) as a function of the Articulation Rate (in ms/syll per AP) and of speaking style (DIC, POL, NOV and TAL).*

Furthermore, the analysis shows that there is a significant effect of Speaking Style on IP Length (Wald χ^2 (3) = 63.343, $p < .001$). But the post-hoc tests reveal a more contrasted situation than what is found for AP. Thus, DIC differs from NOV ($p < .001$) and TAL ($p < .01$), but not from POL. As for POL, it differs only from NOV ($p < .001$). Finally, it is found that NOV differs from TAL ($p < .05$). In summary, NOV presents longer IPs than the three other speaking styles, POL and TAL have a similar IP length and DIC has shorter IPs than NOV and TAL, but not when compared to POL. The analysis also reveals an effect of AR on IP length (Wald χ^2 (3) = 280.162, $p < .001$), and an interaction between AR and Speaking Style predictors (Wald χ^2 (3) = 50.803, $p < .001$). There is a similar effect of AR on IP Length for DIC and POL, which is less important than the observed effect for TAL and NOV.

Finally, the statistical analysis shows that Speaking Style has a significant effect on Initial Rise Rate (Wald χ^2 (3) = 175.223, p < .001). As it can be seen on Table 1, DIC presents a significant higher Initial Rise Ratio than the three other speaking styles (p < .001), which do not manifest any significant differences between them. An effect of AR was also found, showing that the Initial Rise Ratio increases when Articulation Rate decreases (Wald χ^2 (1) = 34.576, p < .001), as one could have hypothesized.

In conclusion, these first results show that AP Length, IP Length and Initial Rise Ratio are robust measures to differentiate some of the speaking styles of our corpus. More importantly, our results show that the differences observed among the speaking styles are not due to differences in tempo.

3.2. Temporal variables

As for temporal variables, we tested the effects of Speaking Style on two prosodic parameters: Articulation Rate and Silent Pause Duration and Distribution. First, a GEE model was run with the Articulation Rate as the dependent variable, Speaking Style and IP Length as independent variables. A global effect of Speaking Style on AR was found (Wald χ^2 (3) = 106.565, p < .001). The post-hoc analysis indicates that there are no significant differences between DIC and POL (in these speaking styles, IPs have an average AR of 222.59 ms/syll and 218.87 respectively), neither between NOV and TAL (IPs have an average AR of 182.36 ms/syll and 193.96 ms/syll respectively). DIC and POL present a longer mean syllabic duration than NOV and TAL, which means that speakers from DIC and POL articulate slower than speakers from NOV and TAL (Figure 3):

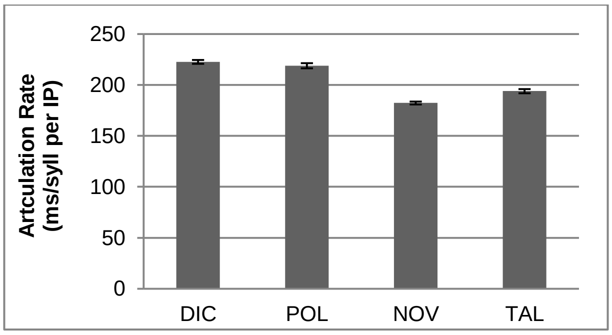


Figure 2: *Articulation Rate per IP (in ms/syll) as a function of speaking style (DIC, POL, NOV and TAL). Error bars are standard error of the mean.*

An effect of IP Length was also found on AR (Wald χ^2 (1) = 307.538, p < .001). Results show that the shorter the prosodic group, the faster AR. The presence of an interaction between IP Length and Speaking Style on AR (Wald χ^2 (3) = 38.378, p < .001) reveals that the effect of IP Length is not the same among the speaking styles, more important for DIC and POL than for the two others.

We modeled the Silent Pause Length as a mixture a log-normal distributions, following the methodology in [34] and [35]:

$$f(x) = \sum_{i=1}^{N} \pi_i \Lambda_i(\mu_i, \sigma_i^2, x)$$

where each component distribution is Gaussian with mean μ and standard deviation σ. Its weight in the mixture model is π and silent pause durations are log-transformed. We identified whether two or three component distributions better model the observed silent pause lengths by using the Bayesian Information Criterion. After selecting the number of component distributions, their parameters are estimated using the Expectation-Maximization algorithm. As it can be seen in Table 3, we observe that TAL and NOV are bi-modal, whereas DIC and POL are tri-modal. We hypothesize that the long pause component distribution in DIC are the pauses the speaker makes to allow for writing time (a very specific characteristic of the dictation speaking style) and that in POL the long pauses component distribution is mainly connected to rhetorical style (cf. for example [32]).

3.3. Pitch Register

Finally, a GEE model was applied with Pitch Register as the dependent variable; Speaking Style, speaker Gender, IP Length and Local AR (calculated as the mean syllabic duration per IP) were the independent variables (Gender was added to take into account the fact that female speakers have been shown to have a wider pitch register than male speakers [33]). As can be seen on Figure 5, DIC and TAL seem to have a wider Pitch Range than POL and NOV (8.33 st and 7.71 st vs 6.34 st and of 5.36 st, namely).

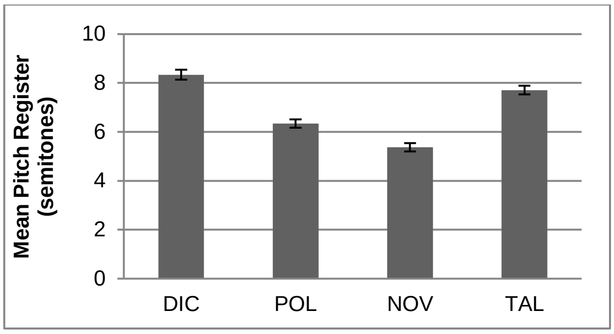


Figure 3: *Mean Pitch Register per IP (in ST) as a function of speaking style (DIC, POL, NOV and TAL). Error bars are standard error of the mean.*

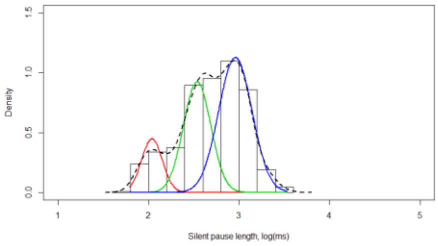
A statistical analysis reveals that there is a significant effect of Speaking Style on Pitch Register (Wald χ^2 (3) = 39.482, p < .001). Post-hoc tests indicate that DIC differs indeed from POL (p < .01) and from NOV (p < .001), but not from TAL. Significant differences are also found between TAL and NOV (p < .01), but not between NOV and POL. Surprisingly, Gender does not appear to have any effect on Pitch Register, nor on AR. Statistics nevertheless show a significant effect of IP length on Pitch Register (Wald χ^2 (1) = 184.600, p < .001), revealing that the longer the IP, the wider pitch register. An interaction between IP Length and Pitch register (Wald χ^2 (3) = 24.342, p < .001) is also found, showing that the effect of IP Length is more important for DIC than for the three others.

4. Discussion

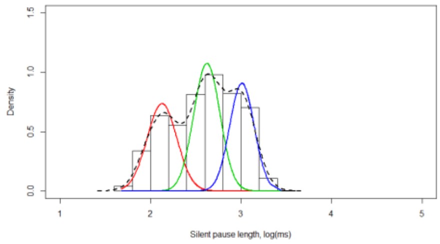
This section summarizes the main findings of our study. Regarding Accentuation and Phrasing, the results indicate that speakers in DIC and POL have a strong tendency to segment their speech flow in smaller prosodic units than speakers in the

Table 3. *Log-normal mixture model of Silent Pause Length (in ms).*

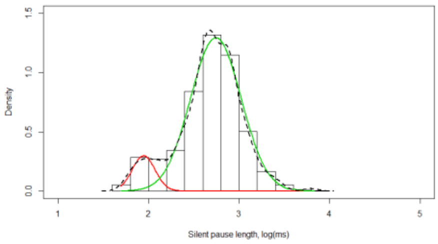
	DIC			POL			TAL		NOV	
π	13%	35%	52%	30%	39%	31%	9%	91%	12%	88%
µ (log ms)	2.038	2.538	2.960	2.127	2.624	3.010	1.946	2.744	2.028	2.782
σ (log ms)	0.115	0.152	0.184	0.165	0.145	0.135	0.127	0.280	0.163	0.263



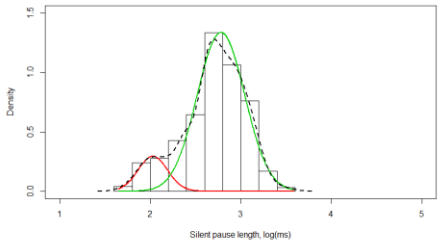
DIC



POL



TAL



NOV

NOV and TAL subsets. The results obtained for NOV and TAL regarding AP and IP length agree with those of previous studies, and confirm that during a dictation or a political speech, speakers tend to align their IP on their AP (on average, an IP is 1.5 APs long in these speaking styles, while it is 2 APs long in NOV and TAL). Nevertheless, it has to be stressed that DIC differs from POL by its higher Initial Rise Ratio, and that POL does not differ significantly from the two other speaking styles regarding this parameter. As for temporal variables, we also find some similitudes between DIC and POL vs. NOV and TAL. Speakers in DIC and POL use three levels of pause, while only two levels are found for NOV and TAL. This is confirmed by the fact that DIC and POL present a slower articulation rate than NOV and TAL. Finally, regarding Pitch Register, we observe some similarities between DIC and TAL on the one hand and some similarities between POL and NOV on the other hand. These results indicate that speech addressed to children tends to exhibit a greater pitch range (and thus greater pitch movement) than speech addressed to adults.

Consequently, if we consider a standard TTS system as producing a speaking style similar to NOV, which may be considered as the closest to typical reading style, the following rules may be employed to adapt the system's prosodic model to one of the speaking styles studied:

- DIC: shorter APs and IPs have to be produced with much more initial rises, more pauses including long pauses and finally a higher pitch register. These rules seem consistent with the fact that this is a didactic style.
- POL: IP Length has to be reduced as in the dictation style, and the number of initial rises increased, but not as much as in dictation. Moreover, this style also shows long pauses (around 1s) that need to be added.
- TAL: the most important parameter is the pitch register which has to be higher with a slightly higher articulation rate. The other parameters are similar in behavior to the ones of NOV, i.e. of current models in TTS.

Finally these rules could be implemented in a TTS synthesis system, and especially a corpus-based one, into a pre-processing prosodic module and also during the selection step of the system.

5. Conclusions

In this paper, we have presented a study of different speaking styles while keeping in mind the application in the TTS synthesis context. Four styles "overtly addressed to a given audience" have been compared: dictation, political speeches, tales and novels. The comparison has been made in terms of some carefully selected prosodic features which have been recognized as robust when distinguishing different speaking styles in French. The results show that significant differences exist between the speaking styles studied in this paper and some adaptations of the TTS prosodic model have to be made to render in an appropriate way these styles. Some rules have been given to explicit what has to be done. Further work will be directed to the integration of these rules into a corpus-based system.

6. Acknowledgements

The work presented here 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] Erickson, D. “Expressive speech: production, perception and application to speech synthesis”, Acoust. Sci. Technol., 26(4):317-325, 2005.

[2] Campbell, N., “Expressive/affective speech synthesis”, in Springer Handbook of Speech Processing, Springer, 505-518, 2008.

[3] Schröder, M., “Emotional Speech Synthesis: A Review”, Proc. of Eurospeech, 2001.

[4] Yamagishi, J., Onishi, K.,Masuko, T and. Kobayashi, T., “Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis”, IEICE Trans. Inf. Syst., 88:502-509, 2005.

[5] Iriondo, I., Socoro, J. C., and Alias, F., “Prosody Modelling of Spanish for Expressive Speech Synthesis” IEEE International Conference on Acoustics, Speech and Signal Processing, 2007.

[6] Rebordao, A. R. F., Shaikh, M. al M., Hirose, K., and Minematsu, N. “How to Improve TTS Systems for Emotional Expressivity”, Proc. of Interspeech, 2009.

[7] Roekhaut, S., Goldman, J.-P. and Simon, A. C., “A Model for Varying Speaking Style in TTS systems”, Proc. of Speech Prosody, 2010.

[8] Obin, N., Lanchantin,P., Lacheret, A. and Rodet, X., “Discrete / Continuous Modelling of Speaking Style in HMM-based Speech Synthesis: Design and Evaluation”, proc. of Interspeech, 2011.

[9] Simon, A.C., Auchlin, A., Avanzi, M. and Goldman, J.-Ph., “Les phonostyles: une description prosodique des styles de parole en français”, in M. Abecassis and G. Ledegen [eds], Les voix des Français. En parlant, en écrivant, 71–88, Peter Lang, Berne, 2010

[10] Goldman, J.-Ph., Pršir, T., Christodoulides, G. and Auchlin A. “Speaking style prosodic variation: an 8-hour 9-style corpus study speaking style”. Proc. of Speech Prosody, 2014.

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

[12] Boersma, P. and Weenink, D. “Praat: doing phonetics by computer (Version 5.5)”. www.praat.org, 2014.

[13] Christodoulides, G., Avanzi, M. and Goldman, J.-P. “DisMo: A Morphosyntactic, Disfluency and Multi-Word Unit Annotator. An Evaluation on a Corpus of French Spontaneous and Read Speech”, Proc. of the 9th LREC, 2014, 3902-3907. www.corpusannotation.org/dismo

[14] Mertens, P.; Goldman; J-P., Wehrli, E. and Gaudinat, A. “La synthèse de l'intonation à partir de structures syntaxiques riches”, Traitement Automatique des Langues 42(1):142-195, 2001.

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

[16] Delais-Roussarie E., “Chapitre XIX : Structuration du flux de parole: constituants prosodiques et structure prosodique”, in: Godard, D., Abeillé, A. & Delaveau, A. [eds], Grande Grammaire du Français, Actes Sud, to app.

[17] Avanzi, M., Obin, N., Lacheret-Dujour, A. and Victorri, B. “Toward a Continuous Modeling of French Prosodic Structure: Using Acoustic Features to Predict Prominence Location and Prominence Degree”. Proc. of Interspeech, 2011.

[18] Cohen, J. “A Coefficient of Agreement for Nominal Scales”, Educational and Psychological Measurement, 20(1):37-46, 1969.

[19] Landis, J. R. and Koch, G. “The Measurement of Observer Agreement for Categorical Data”. Biometrics, 33:159-174, 1977.

[20] Avanzi, M., “Note de recherche sur l’accentuation et le phrasé à la lumière des corpus du français”, Tranel, 58, 2013, 5-24.

[21] Christodoulides, G., “Praaline: integrating tools for speech corpus research”, Proc. of LREC, 2014, 31-34.

[22] Fónagy, I., “L’accent en français : accent probabilitaire”, in Fónagy, I., and Léon, P. [eds], L’accent en français Contemporain, Studia Phonetica 15, 123-233, Paris, Didier. 1980.

[23] Di Cristo, A., “Le cadre accentuel du français contemporain: essai de modélisation”, Langues 2(3): 184-205, and Langues 2(4):258-267, 1999.

[24] Kemler-Nelson, D.G., Hirsh-Pasek, K., Jusczyk, P.W., and Cassidy, K.W., “How the prosodic cues in motherese might assist language learning”, J Child Lang. 16(1):55-68, 1989.

[25] Ghisletta, P. and Spini, D. “An Introduction to Generalized Estimating Equations and an Application to Assess Selectivity Effects in a Longitudinal Study on Very Old Individuals”. Journal of Educational and Behavioral Statistic, 29(4), 421-437, 2004.

[26] Fougeron, C. and Jun, S. A. “Rate Effects on French Intonation: Prosodic Organization and Phonetic Realization”. Journal of Phonetics, 26, 45-69, 1998.

[27] Post, B. “The multi-facetted relation between phrasing and intonation contours in French”. In C. Gabriel & C. Lleó (Eds), Intonational Phrasing in Romance and Germanic: Cross-linguistic and bilingual studies (pp. 43–74). Amsterdam: John Benjamins, 2011.

[28] Avanzi, M., Rousier-Vercruyssen, L., Schwab, S., Gonzalez, S. and Fossard, M., “C-PROM-Task: A New Annotated Dataset for the Study of French Speech Prosody”, Proceedings of TRASP, Aix-en-Provence, 27–30, 2013.

[29] Wioland, F., Prononcer les mots du Français: des sons et des rythmes, Hachette, 1991.

[30] Quené, H., “Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo”. Journal of the Acoustical Society of America, 123, 1104-1113, 2008.

[31] Miller, J.L., Grosjean, F. and Lomato, C., “Articulation rate and its variability in spontaneous speech: A reanalysis and some implications”. Phonetica, 41, 215-225, 1984.

[32] Duez, D. “La fonction symbolique des pauses dans la parole de l'homme politique”, Faits de langue, 13:91-97, 1999.

[33] De Looze, C. Analyse et Interprétation de l'Empan Temporel des Variations Prosodiques en Français et en Anglais, PhD Thesis, Aix/Marseille, 2010.

[34] Goldman, J.-Ph., François, T., Roekhaut, S., Simon, A. C., “Étude statistique de la durée pausale dans différents styles de parole”, Actes des 28èmes Journées d'Étude sur la Parole (JEP), Association Francophone de la Communication Parlée, Mons, Belgium, 25-28 May, 2010.

[35] Little, D., Oehmen, R., Dunn, J., Hird, K., Kirsner, K., “Fluency Profiling System: An automated system for analyzing the temporal properties of speech”, Behavior Research Methods 45(1): 191–202, 2012.