Speaking style prosodic variation: an 8-hour 9-style corpus study

Jean-Philippe Goldman[1], Tea Pršir[1,2], George Christodoulides[2], Antoine Auchlin[1]

[1]Département de Linguistique, Université de Genève
[2]Institut Langage & Communication, Université de Louvain

jean-philippe.goldman@unige.ch, tea.prsir@unige.ch,
george@mycontent.gr, antoine.auchlin@unige.ch

Abstract

This paper presents the results of a prosodic and phonostylistic analysis based on C-PhonoGenre, an 8-hour-long spoken French corpus, consisting of 9 speaking situations and (on average) 10 speakers per situation. The corpus was automatically segmented at the phonetic, syllabic and word levels (EasyAlign), and in larger pause-separated units. Part-of-speech annotation (DisMo) and prominent syllable detection (ProsoProm) was added automatically. The corpus was also manually annotated at the syllabic level for stylistic variants, such as post-tonic schwas, liaisons, elisions, disfluencies, audible breaths and noises. Acoustic analyses (ProsoReport, DurationAnalyser) provide more than 100 micro- and macro-prosodic measures, which we correlate with the phonostylistic features and the linguistic annotation. This analysis results in a contrastive, fine-grained *prosometric* description of phono-stylistic and situational variation, over 4 situational, gradual dimensions: audience, media, preparation, and interactivity. Further statistical analysis was carried out to explore the discriminative and explanatory power of combinations of prosodic measures.

Index Terms: situational variation, prosody, classification of speaking styles

1. Introduction

General knowledge of language includes that of its variants, as demonstrated by the study of *genres* [1, 2, 3], and phono-stylistic variation is such an area [4, 5, 6]. Recent research in prosody focuses on phonostylistic situational variation in large corpora, departing from previous binary oppositions (*e.g.* read vs. spontaneous speech), or one-dimensional characterisations of style (formal vs. informal).

This paper presents a selection of global and contrasted results of an on-going research project on situation-dependent speaking styles, or *phonogenres*. It applies the semi-automated methodology of corpus description introduced in previous work [7, 8]. It analyses speaking situations by features [9, 10, 11], reduced to four main dimensions: audience, media, preparation, and interactivity; each dimension has 3 different states (see Table 1). For example, ***audience** = 1* indicates that the speaker is physically present before an audience, while ***media** = 1* indicates speech directed to an individual or a small group, yet in front of a microphone or camera (*indirect* audience). ***Preparation** = 1* indicates semi-prepared speech, situated between spontaneous and read speech. In the case of parliamentary debates, a *question* is prepared, while the *answer* is semi-prepared. ***Interactivity** = 1* indicates that the main speaker may be interrupted. For example we distinguish between dialogue interaction and broadcast sports commentaries. The corpus under study, C-PhonoGenre, is approximately 8 hours-long, and covers 9 different genres, four of which are further subdivided into *sub-genres* (based on their situational specificities).

Results show that phonogenres and sub-genres can be distinguished and characterised by the relation between situational and prosodic dimensions. Various "hidden" or unpredicted influences of situational properties on prosody also emerge. An indirect, but equally important result of this study is a corpus processing methodology, based on the co-ordinated application of several semi-automatic tools.

2. Data

C-PhonoGenre contains data from 8 speaking styles: instructional speech [DIDA]; spontaneous narration [NARR]; speeches during "Question Time" at the French parliament [PARL]; sermons [RELG]; radio press reviews [RPRW]; three kinds of sports commentary [SPOR]: rugby, basketball and football; presidential New Year's wishes [WISH] and weather forecasts [WFOR]. The average sample duration per speaker is 5:30 min.

Table 1. *Situational features by PhonoGenre*

PhonoGenre		Audience	Media	Preparation	Interaction
DIDA	Radio	1	2	2	2
	TV	0	2	2	0
	Lecture	2	0	1	0
NARR	Narration	1	0	0	2
PARL	Question	2	1	2	1
	Answer	2	1	1	1
READ	Reading	0	0	2	0
RELG	Internet mass	0	1	2	0
	Sermon on TV	2	1	2	0
RPRW	Radio press review	0	2	2	0
SPOR	Basket	0	2	0	0
	Rugby/football	1	2	0	2
WFOR	Weather forecast	0	2	2	0
WISH	Pres. New Year	0	1	2	0

For this study, we compiled a corpus including the eight speaking styles of C-PhonoGenre and the "reading" style [READ] from C-PROM-PFC [12]. Table 2 shows number of samples, syllables, words and total speech time per genre. Although [SPOR] and [RELG] contain less than 10 speakers per genre, the total amount of data renders these genres comparable with the others. On the other hand, while 10 different speakers are included under the "weather forecasts" genre [WFOR], the total speech time is 9 minutes, i.e. less than a minute per speaker.

The corpus contains recording of both female and male speakers, originating from 3 different French-speaking areas: Metropolitan France, Belgium and Switzerland [13]. We do not present findings regarding regional variation here, but the information is present in the corpus metadata and can be used for further study. Regional variation may partly explain the observed intra-genre, inter-speaker dispersion.

10.21437/SpeechProsody.2014-9

Table 2. *Number of recordings, syllables and words, and duration by phonogenre*

PhonoGenre	Num. samples	Duration (min)	Num. syllables	Num. words
DIDA	17	100	26 304	18 717
NARR	10	44	11 396	9 546
PARL	10	20	5 710	3 613
READ	16	36	9 932	6 648
RELG	7	54	8 726	6 141
RPRW	15	95	26 359	17 531
SPOR	5	35	7 601	5 305
WFOR	10	9	2 861	1 947
WISH	15	98	18 614	12 578
TOTAL	**105**	**491**	**117 503**	**82 026**

3. Methodology

3.1. Data processing

After manual orthographic transcription in Praat [14], we obtained a phonetic transcription as well as a segmentation of words, syllables, phones and pauses, automatically using EasyAlign [15]. Manual corrections were made to reach a high quality alignment between the segments and the speech signal. A unique annotator manually added a <*delivery*> tier in order to enhance downstream data processing. It contains four types of annotation: i) disfluencies, articulation and phonological phenomena: schwa; vowel lengthening (whether associated to hesitation or not); creaky voice; liaison and elision; ii) symbols to distinguish between complete silence, audible and less audible breaths, and mouth noises; iii) indices of paralinguistic phenomena (laugh, cough) and external sounds; iv) overlapping segments and syntactic plan interruptions. Part-of-speech tagging and multi-word unit detection was obtained automatically using DisMo [16]; this annotation is used to study the interface between speaking styles and grammar. A five-level degree of prominence for each syllable was calculated using ProsoProm [17]. Three additional tiers were automatically added: <*lex*> distinguishing between lexical and functional words; <*if*> localising initial vs. final lexical words' syllables; and <*ap*> which is an automatically generated segmentation into accentual phrases (in Mertens' sense [18], i.e. phonological words), including an annotation of the initial and final syllables of each accentual phrase. Pitch was corrected manually for the entire corpus, since the accuracy of several acoustic measures and prominence detection depend on it.

3.2. Acoustic measures

Acoustic and prosodic features were extracted for the entire corpus. Initially, ProsoGram's [18] two-step algorithm for pitch stylisation was applied: for each syllable, vocalic nuclei are detected based on intensity and voicing, and then the F0 curve on the nucleus is stylised into a static or dynamic tone, based on a perceptual glissando approach. ProsoReport [8] summarises this information, taking into account information contained in other tiers (such as <*delivery*> and <*lex*>) to produce a detailed collection of descriptive statistical measures for each corpus sample. These measures can be grouped into four main families: temporal measures (e.g. articulation rate); pitch measures (e.g. pitch register and movement); syllabic prominence measures (e.g. percentage of prominent syllables in various positions); correlational measures (e.g. percentage of accentual-group-initial prominent syllables). Additionally, DurationAnalyser [19] produced a set of statistics based on segmental information (e.g. variance coefficients for vowels or consonants, nPVI etc.). All this information was compiled for further analysis; in total, 129 prosodic descriptors for each corpus sample were retained.

4. Results

Based on the aforementioned prosodic descriptors, many results appear to be significant. We show only some of them based on the *genre* classification, then on different prosodic domains (duration, intonation and accentuation of initial and final syllables, across genres and situational features).

4.1. Results by phonogenre

The articulation ratio at the genre level sets apart WISH and RELG, reflecting that both situations are solemn. SPOR also stands out, but for another reason: the commentator has to pause while the ball moves from one player to the next. Conversely, WFOR, and to a lesser extent RPRW, show the highest articulation ratio, because of the time pressure imposed by broadcasting media.

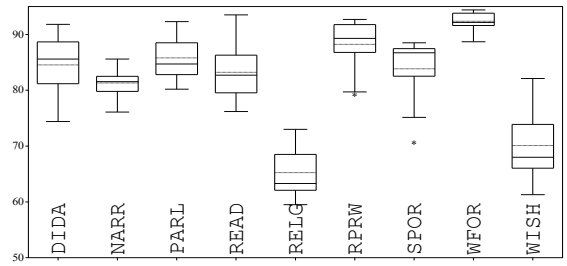

Figure 1: *Articulation ratio for the 9 phonogenres*

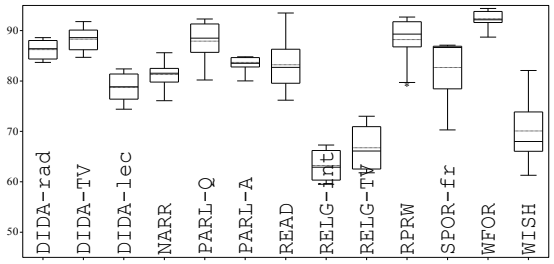

Figure 2: *Articulation ratio for the 13 sub-genres*

At the sub-genre level, articulation ratio also illustrates interesting contrasts. Parliamentary answers (PARL-A) clearly occupy more speech time than questions, because the listeners react during the answer and the speaker has to take into account this feedback. The three instructional sub-genres also show differences, mainly between the Radio and TV sub-genres. A possible explanation of this difference may be that radio samples are shorter than TV samples (DIDA-rad: approx. 3 min., DIDA-TV: 10-20 min.), as well as the fact that TV delivers images with speech, sharing time with visual flow [20]. The DIDA-lec subgenre (non-broadcast university lectures) is closer to DIDA-rad, suggesting that the media dimension is less important than the instructional one. RELG subgenres differ slightly on articulation ratio, though other prosodic measurements distinguish them much more clearly.

4.2. Segmental duration

Among the acoustic parameters based on segmental durations, the variance of vowel duration is the one exhibiting the best

discriminative power. Figure 3 is a box-plot of vowel duration variance for different sub-genres. The NARR genre detaches from others probably because of its spontaneous nature: at the syllable level, this results in frequent hesitation-related lengthening; at the discourse level, an irregular speech rate is entailed by the progressive construction of discourse. In contrast, the READ and WFOR genres have a lower variation of vowel duration, but for different reasons. READ readers are non-professionals and thus adopt a monotonous rhythm; whereas the WFOR genre has a very high speaking rate, causing a ceiling effect on vowel duration. Interesting differences occur again between in parliamentary sub-genres, showing a significantly lower variation for the answer [PARL-A] than for the question [PARL-Q], due to the increased interactivity (see 4.1).

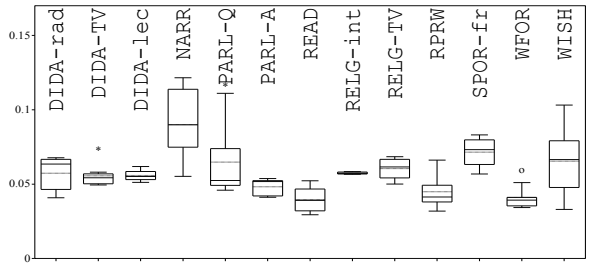

Figure 3: *Variation of vowel duration for the 13 sub-genres*

The *preparation* feature also shows a lower variation of vowel duration for prepared recordings (F(2,102)=50; p<0.001). This is explained by more lengthened hesitations in spontaneous speech. The *interactivity* feature shows similarly a greater variation (F(2,102)=31.4 p<0.001) for interactive recordings, usually the spontaneous ones (NARR, SPOR).

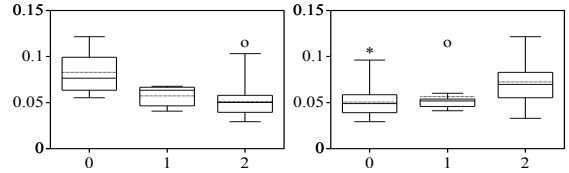

Figure 4: *Variation of vowel duration for 3 levels of preparation (left) and interactivity (right*)

4.3. Intonation

Intonational properties indicate a lower relative F0 variation for phonogenres with a larger audience (F(2,102)=10.5; p<0.001); this is surprising, as we hypothesised that public speaking would entail greater speaker involvement. However, this acoustic parameter varies according to our predictions across the media feature (F(2,102)=12.06; p<0.001).

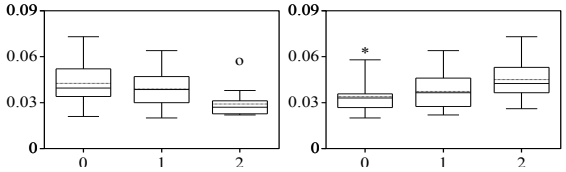

Figure 5: *Relative F0 variation for 3 degrees of situational features of audience (left) and media (right)*

4.4. Prominence in initial and final position

The study of initial and final positions of prominent syllables results in differentiating phonogenres based on their situational features.

4.4.1. Situational features

The percentage of prominent final syllables is decreasing as the phonogenre is getting more *interactive* (F(2,102)=8.88; p<0.001). This can be explained by a high score of hesitation in NARR and vowel lengthening, typical for sport commentaries SPOR (Figure 6, left).

The percentage of prominent initial syllables is getting higher if a phonogenre falls is *media*, where it is important to clearly distinguish discourse segments (Fig.6, right). The initial prominent syllables of AP shows similar results (F(2,102)=5.88; p<0.001).

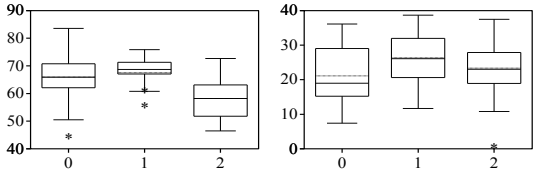

Figure 6: Percentage of prominent final syllables for interactivity (left) and percentage of prominent initial syllables for media (right) for each of the 3 degrees

The relative length of initial and final syllables of the AP varies in a significant manner across the *preparation* dimension (initial syllables F(2,102)=5.42 p<0.001; final F(2,102)=10.65 p<0.001). The variation is inverted: initial syllables of AP tend to be shorter in prepared discourse than in non-prepared, but final syllables become longer (Figure 7).

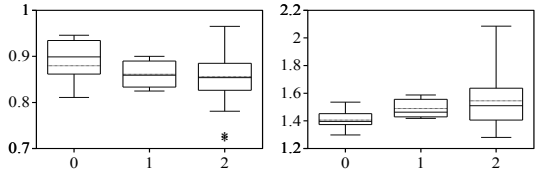

Figure 7: *Relative length of initial (left) and final (right) syllables of the AP for preparation*

The physical presence of *audience* implies lower percentage of initial prominent syllables per AP. This is the case for parliamentary speeches where one member of the house is talking directly to the government minister [PARL]; for university lectures, where teacher addresses the students [DIDA-lec], or sermons where the priest is addressing the faithful [RELG-TV] (F(2,102)=12.8; p<0.001).

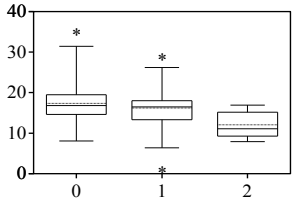

Figure 8: *Percentage of initial prominent syllables per AP for audience for each of the 3 degrees*

4.4.2. Sub-genres

Sub-genres level distribution of initial prominences (Fig. 9) partly reflects the values of situational features. Sub-genres in which an audience is present show the lowest level (PARL; DIDA-lec, RELG-TV, SPOR), whereas media ones (DIDA-rad, DIDA-TV, WFOR) show the highest level. RELG-int behaves like a media style, although it was graded as an intermediate style across the media dimension. Despite RPRW being a prototypical case of broadcast media style, it shows a low level of initial prominence. In fact, RPRW shows a high rate of prominent initial syllables of words, not of AP, which reflects the stylistic choice of marking initial syllables.

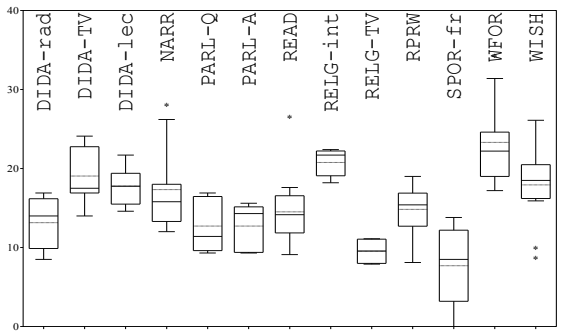

Figure 9: The distribution of the percentage of initial prominent syllables per AP for the 13 sub-genres

4.5. Principal Component Analysis

Since no unique prosodic/acoustic parameter is sufficient to clearly distinguish speaking styles, we explored a global statistical approach which finds the optimal linear combination of all parameters. A Principal Components Analysis (PCA) was thus performed, to model phonogenres and situational features with the parameters, knowing that some of them are high correlated. The first two principal components (PC) explain only 43% of the variance, while the first 8 explain 78.2%. A discriminating analysis with 8 PCs over 9 phonogenres showed that 93.3% of recordings were identified as belonging to the correct genre. Table 3 shows the percentage of correct classification decisions, over genre, sub-genre and the 4 situational features.

Table 3. *Percentage of correct classification with 8 principal components for Genre and Sub-Genre distinction as well the 4 situational features*

Genre	93.3 %
Sub-Genre	90.5 %
Audience	90 %
Media	84.8 %
Preparation	92.4 %
Interactivity	92.3 %

Figure 10 shows the distribution of the 105 speech recordings along the first two PCs. The first PC appears to be highly correlated with the articulation ratio, as the WFOR genre is clearly opposed to the genres WISH and RELG. The bottom of the figure shows the confidence ellipses for the 'media' situational feature according to the first 2 PCs with a clear distinction between semi-media (1) and media (2) condition, while the non-media (0) condition is over the first two conditions.

5. Conclusion

Our study led to two concrete outcomes: (a) a detailed methodology for future studies in *phonostylistics* and socio-phonetics, which is primarily based on automated tools and the study of prosody (an extension to other levels, e.g. the lexical level, is envisaged); and (b) a spoken corpus of high scientific value, due to its thorough multi-level annotation.

The results show that, while no single prosodic measure is sufficient to separate and classify speaking styles, a linear combination of several measures leads to a robust clustering of samples belonging to different genres. We next will explore feature selection techniques to determine which set of prosodic measures is the most relevant (instead of a PCA analysis that produces a global, but opaque, combination of measures).

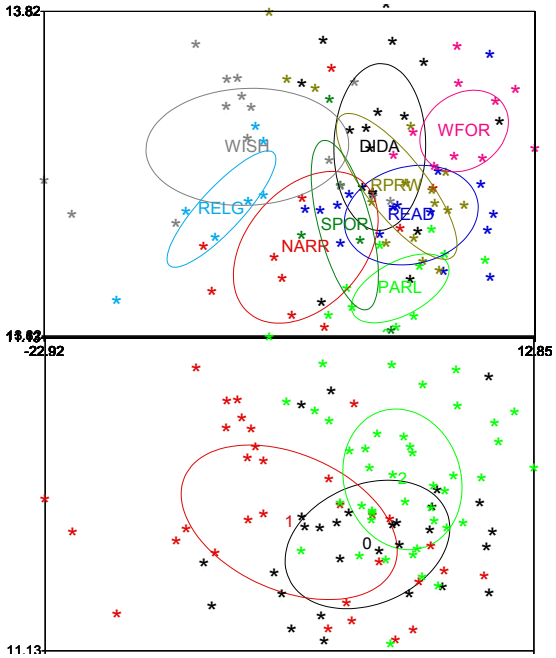

Figure 10: The 105 recordings according to the first two PC and confidence ellipses by phonogenres (top); the confidence ellipses for the media situational feature (bottom)

The study has also shown that following an iterative, adaptive procedure is necessary: while the initial, top-down approach was to select samples in order to create a balanced corpus (based on a predefined array of situational features), subsequent data analysis led to the observation that samples within a given genre could and should be further classified in *sub-genres*. Therefore, the interplay of prosodic measures and situational features gave rise to an *a posteriori* subdivision of genres (bottom-up approach) in order to ensure compact definitions and to reduce the excessive heterogeneity of some speaking situations. Results show (Fig. 1, 2, 3) that sub-genre groupings *transcend genre differences* (e.g. PARL-A and DIDA-rad, in Fig. 2), and that some of them are related to common, controlled situational features (cf. Fig. 9, DIDA-lec: non-media, public audience). They also present evidence for groupings due to unpredicted, or hidden, situational features, like "external time pressure" (cf. Fig. 1 PARL, WFOR, RPRW), "speech sequence duration", or "solemnity / ritual conventions" (RELG and WISH), that belong to the prototypical image of speaking style. The differences observed between questions and answers within the PARL genre suggest a situational feature [+ interactivity] at the sub-genre level. They reveal a prosodic reflection of a discursive (not situational) feature, namely the presence of other listeners that react to the exchange even though they do not participate in it directly. The results also indicate that several different speakers per genre must be included in the corpus for the genre-specific features to rise above individual variation.

Although some annotation steps remain manual, most of our methodology is automatic. This framework was built in a very generic way: we plan to provide additional prosodic measures and test corpora in other languages. Such research enables both verification of linguistic hypotheses and automatic genre identification. Finally, we should mention that the C-PhonoGenre corpus will be made available to the community for any other research purposes.

6. Acknowledgements

This research is funded by Swiss National Science Foundation – FNS Grant nr 100012_134818.

7. References

[1] Beacco, J.-C. “Trois perspectives linguistiques sur la notion de genre discursif”, Langages 38(153): 109–119, 2004.

[2] Solin, A. “Genre”, in J. Zienkowski, J.-O. Östman and J. Verschueren [Eds], Discursive Pragmatics, 119–134, John Benjamins, 2011.

[3] Bawarshi, A. and Reiff, M. J. Genre: An Introduction to History, Theory, Research, and Pedagogy, Indiana, Parlor Press, 2010.

[4] Fónagy, I. and Fónagy J. “Prosodie professionnelle et changements prosodiques”, Le Français Moderne 44: 193–228, 1976.

[5] Fónagy, I. La vive voix. Précis de psycho-phonétique, Paris, Payot, 1983.

[6] Léon, P. Précis de phonostylistique, Parole et expressivité, Nathan Université, Paris, 1993.

[7] 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.

[8] Goldman, J.-Ph., Auchlin, A. and Simon A.C., “Discrimination de styles de parole par analyse prosodique semi-automatique”, in H.-Y. Yoo and E. Delais-Roussarie [Eds] Actes d’IDP 2009, Septembre 2009, Paris, 2011.

[9] Lucci, V., “Étude phonétique du français contemporain à travers la variation situationnelle”, Université des langues et lettres, Grenoble, 1983.

[10] Koch, P. and Oesterreicher, W., “Langage parlé et langage écrit”, in G. Holtus, M. Metzeltin and Ch. Schmitt [Eds], Lexikon der Romanistischen Linguistik, I/2, 584-627, Niemeyer, Tübingen, 2001.

[11] Pršir, T., Goldman, J.-Ph. and Auchlin A., ”Variation prosodique situationnelle: étude sur corpus de huit phonogenres en français”, in P. Mertens and A.C. Simon [Eds], Proceedings of the Prosody-Discourse Interface Conference 2013, 107-111, Leuven, September 2013.

[12] Avanzi, M., Christodoulides, G., Schwab S., Bardiaux A., Goldman J.-Ph., “La variation prosodique régionale et stylistique en français – Analyse de neuf points d'enquête PFC”, Journées PFC, Paris, December 2013.

[13] Simon, A. C. [Ed], La variation prosodique régionale en français, DeBoeck, 2012.

[14] Boersma, P. and Weenink, D., “Praat: doing phonetics by computer”, online at http://www.praat.org

[15] Goldman, J.-Ph. EasyAlign: an automatic phonetic alignment tool under Praat, Proceedings of InterSpeech, Florence, Italy, 2011.

[16] Christodoulides, G., Avanzi, M. and Goldman, J.-Ph., “DisMo: A Morphosyntactic, Disfluency and Multi-Word Unit Annotator: An Evaluation on a Corpus of French Spontaneous and Read Speech”, Proceedings of the Language Resources and Evaluation Conference (LREC) conference, Reykjavik, Iceland, 26-31 May 2014.

[17] Goldman, J.-Ph., Avanzi, M., Simon, A.C. and Auchlin, A., “A continuous prominence score based on acoustic features”, Proceedings of InterSpeech 2012, 9-13 September, 2012.

[18] Mertens, P., “The Prosogram: Semi-Automatic Transcription of Prosody based on a Tonal Perception Model” in B. Bel and I. Marlien [Eds], Proceedings of Speech Prosody 2004, Nara, Japan, 23-26 March, 2004.

[19] Dellwo, V., Influences of speech rate on the acoustic correlates of speech rhythm: An experimental phonetic study based on acoustic and perceptual evidence, PhD Dissertation, Universität Bonn, 2010.

[20] Kern, F., “Speaking dramatically. The prosody of life radio commentary of football matches”, in Barth-Weingarten, D., Reber E., and Selting M. [Eds] Prosody in interaction, John Benjamins, 217-237, 2010.