Journal of Speech Sciences 4(1) 41-60.2014
Available at: www.journalofspeechsciences.org

Prosodic features of situational variation across nine speaking styles in French

PRŠIR, Tea*°
GOLDMAN, Jean-Philippe*
AUCHLIN, Antoine*

*Department of Linguistics, University of Geneva
°Institut Langage & Communication, UCLouvain

Abstract

This paper presents results from an on-going study of prosodic and phonostylistic variation across speaking styles, i.e., acoustic images associated to types of language production, also called phonogenres. It extends previous work in (1, 2) by enlarging the corpus (C-PhonoGenre, 8 hours) and by exploring a more comprehensive collection of genres. The situational parameters in (3, 4) are reduced to four situational features, each admitting three values, the combination of which differentiates sub-phonogenres. The main goal of this study is to establish correlations between the situational and prosodic features of discourse. Corpus processing, annotation and measure calculation are performed semi-automatically, through a set of tools implemented under Praat and manual steps. Rhythmical measurements by DurationAnalyser (5) combined with the output of ProsoReport (6) produce an acoustic analysis of the differences between phonogenres. A large number of micro- and macro-prosodic measures provide a fine-grained 'prosometric' description. This article presents the methodology for collecting the corpus, and results for the phonogenres.

Keywords: speaking style; prosody; situational features; corpus annotation; acoustic measurement

*Contact of the author: tea.prsir@gmail.com

1 Introduction

A speaking style does not depend only on the speaker's identity, such as his level of education or his place of origin. It also depends on the speaking situation that calls for a more or less determinate encoding of specific speech features, regarding the lexis, grammar and discourse structure, as well as phonetic and phonological features. The goal of the present study is to establish correlations between situational settings and prosodic features.

The general context for the present research is the growing interest in genre in language sciences (7-9), and more specifically for spoken language genres and situational variation (1, 2, 10-17). We use the term *phonogenre* for a spoken language genre, and we distinguish it from a *phonostyle*. The *phonogenre* is defined as a typified acoustic image associated to a situation and speech activity, whereas *phonostyle* refers to the features of a given speech sample within a phonogenre. The term *speaking style*, commonly used in this research field, is to be understood as embracing both phonogenre and phonostyle. By introducing the distinction between genre and style we are underlining that speech is necessarily produced within one phonogenre, but each speaker has his individual phonostyle, within each phonogenre.

Situational variation is approached from two points of view. On one hand, situations are grouped according to an implicit typology, which still needs to be determined; on the other hand, typical prosodic features tend to characterise phonogenres (e.g., sport commentary, religious sermons), and to make them highly recognisable.

Speech samples are collected and grouped according to shared situational features, inspired from "situational invariants" (3), and "speech conceptional features" (4) that situate each sample within a continuum that ranges from "language of immediacy" to "language of distance". Four features of speech situations are considered, each one admitting three degrees (Table 1):

Table 1: Three degrees of four situational features

Degree	Audience	Media	Preparation	Interactivity
0	Microphone only	Non-media	Spontaneous	Monologue
1	Face-to-face	Semi-media	Semi-prepared	Semi-interactive
2	Public	Media	Prepared	Interactive

The studied phonogenres are described according to their situational features in Table 3.

The degree *semi-* is introduced to reflect the complexity of certain speaking situations, where the discrete opposition presence *vs.* absence of one feature is not sufficient because the discourse can either share the two conditions or stand in-between. For example, the religious sermon would usually be considered as *non-media*, but as it is also broadcasted in the media (sermons on the Internet and televised church mass) we labelled it as *semi-media*. As for interactivity feature, we labelled the parliamentary speech as *semi-interactive* as it lies between a monologue and interactive dialogue: one member of the parliament addresses a *prepared* question (monologue), and then the government minister answers the question: this is once more a monologue, but it is *semi-prepared* since the minister has to improvise from a prepared draft. Therefore, the label *semi-interactive* was assigned for two reasons: (a) this question and answer exchange represent a sort of dialogue exchange and (b) the government minister has to respond to the reactions of the members of the opposition in the Parliament.

This research expands the set of previously studied phonogenres, as well as the corpus duration, both globally and per studied genre; it relies on the same improved semi-automatic

speech annotation methodology as (1, 2, 6). It further joins rhythmical measurements (5) to ProsoReport (6).

2. Corpus collection and annotation

Previous work on a smaller multi-speaking style speech corpus (C-Prom, 17) has oriented this research in two ways: in pointing out (a) the need for more homogeneous phonogenres by constraining their situational features, and (b) the need to avoid idiosyncrasy by studying a larger number of speakers for each phonogenre. Therefore, our C-PhonoGenre corpus is composed of speaking styles whose situational features are more constrained, with ideally 10 speakers per style.

2.1 Corpus collection

The corpus C-PhonoGenre is composed of eight phonogenres that we describe here in detail, providing their metadata (source, year of recording, variety of French). The total duration of each phonogenre, as well as number of recordings is summarised in Table 2.

- [ASS] parliamentary speech. Ten recordings dating from 2012-2013 have been retrieved in pairs: five questions by members of the parliament addressed to the government, at the French National Assembly, and the respective five answers by ministers, during “Question Time”. MPs and Ministers have at their disposal a maximum of 2 minutes speaking time.
- [DID] educational speech. Six recordings of two educational television programs (*Ce corps est le vôtre* 1980 and *Le dessous des cartes* 2010). Eight recordings of an educational radio program (*Les p'tits bateaux* 2012) where each invited speaker (teacher, researcher) has 3 minutes to answer a question asked by a child. Three audio recordings of scientific lectures by French, Swiss and Belgian university professors (2007).
- [LIT] religious sermons. Three recordings have been retrieved from Internet where one priest is preaching a sermon (2012). Four recordings were extracted from a church mass broadcast on catholic television KTOTV (2013).
- [MET] weather forecasts. Ten recordings from French, Swiss and Belgian radios (2013) with average duration less than a minute.
- [NAR] spontaneous narration. Ten native French speakers have been recorded (2013) while telling a personal story of their choice. Recording duration is between 3 and 10 minutes. All the speakers speak standard French.
- [RPR] radio press reviews. Fifteen recordings from four French radio stations (*France Musique*, *France Culture*, *France Inter*, *RFI*) and one Swiss (*Journal du Matin*) dating from 2004 to 2013 with duration between 4 and 11 minutes.
- [SPO] sport commentary. Five recordings: one for basketball (Belgian commentator, 2010), one for rugby (2007) and three for football (1998) with a couple of French commentators.
- [VXP] presidential New Year’s wishes. Fifteen recordings dating from 1968 to 2007 by six French presidents and from 1999 to 2011 by four Swiss Confederation presidents.

Both female and male speakers are represented equally in the corpus whenever achievable. In fact, LIT (religious sermons) and SPO (sports commentaries) consist exclusively of recordings of male speakers. Among the presidential New Year’s wishes genre, there are two Swiss female speakers.

Data cover three different French-speaking areas: Metropolitan France, Belgium and Switzerland. Regional variation is not explored in this study; nevertheless, the information is

present in the corpus and can be used for further study. In the discussion part, regional variation is taken into account as a partial explanation of the dispersion observed at the intra-genre and inter-speaker levels.

The average duration of the recordings is 4min 40sec (minimum: 37 sec; max: 13 min). 75% of the recordings have duration between 3 and 5 minutes. The corpus has been gathered from various media sources: television (DID, LIT, MET, SPO, VXP), radio (DID, RPR) or Internet (LIT, ASS), whereas NAR and LEC have been home-recorded.

For this comparative study of different phonogenres, the existing C-PhonoGenre corpus was increased by 16 recordings of "reading" [LEC] from the C-Prom-PFC corpus (18).

Table 2: Number and duration (in minutes) of recordings by phonogenre

Phonogenre	Num. of recordings	Duration (min.)
ASS	10	20
DID	17	98
LEC	16	36
LIT	7	54
MET	10	9
NAR	10	35
RPR	15	93
SPO	5	35
VXP	15	95
Total	105	490

2.2 Situational features

At a first level, genre identification is based on the type of speech activity. This allows for grouping under DID (educational speech) both media and non-media speech, as well as for considering that parliamentary speech is distinguished from presidential New Year's wishes – though they could have been united in a "political discourse" genre label. A description of the assumed phonogenres based on their four situational features (Table 3) produces internal sub-phonogenre levels for four of them: parliamentary speech [ASS], educational speech [DID], religious sermons [LIT], and sport commentaries [SPO]. Yet, identical values for situational features do not imply identity of genre: consider [DID-TV], weather forecasts [MET], radio press review [RPR]. Situational variation in [DID] is clearly due to distinctive contextual conditions: radio, television and a conference room; similarly for [LIT]: Internet vs. a church. On the other hand, the two sub-phonogenres of parliamentary speech [ASS] are determined by the position of the speaker in the exchange: he is either the one who asks or the one who answers. Therefore, these distinctions are to be considered at the interactional and psycho-sociological level. The division of [SPO] into a basketball sub-phonogenre on one hand, and a football and rugby on the other, was made because the two latter are interactive (several commentators), whereas in the case of basketball, there is only one speaker.

Table 3: Degrees of situational features for phonogenres and sub-phonogenres

Phonogenre	Sub-phonogenre	Audience	Media	Preparation	Interactivity
ASS-Q	Question	2	1	2	1
ASS-R	Answer	2	1	1	1
DID-Rad	Radio	1	2	2	2
DID-TV	TV	0	2	2	0
DID-cnf	University Conference	2	0	1	0
LEC	Reading	0	0	2	0
LIT-Int	Sermon on Internet	0	1	2	0
LIT-MTV	Mass on TV	2	1	2	0
MET	Weather Forecast	0	2	2	0
NAR	Narration	1	0	0	2
RPR	Radio Press Review	0	2	2	0
SPO-b	Basketball	0	2	0	0
SPO-foru	Rugby/Football	1	2	0	2
VXP	Pres. New Year's Wishes	0	1	2	0

2.3 Segmentation and annotation

2.3.1 Segmentation

A manual orthographic transcription in Praat (19) and a semi-automatic processing (EasyAlign, 20) result in a lexical, syllabic and phonemic segmentation of the corpus. Moreover, EasyAlign detects automatically pauses and provides a PSU (Pause-Separated Units) tier. This information is relevant for the study of discourse structure and for the prosodic boundaries. In order to retrieve reliable results from instrumental and acoustic analysis, the segments' boundaries of the whole corpus and of each level of segmentation have been manually corrected.

Of 129 566 intervals in tier *syllables*, 117 502 (90.7%) are plain articulated syllables and 12 064 (9.3%) are pause intervals. More details about phone and word levels are presented in Table 4.

Table 4: Counting of articulated and paused intervals at phonemic, syllabic and lexical levels

	Phones	Syllables	Words
Total	277 538	129 566	90 036
Articulated	265532 (95.7%)	117 502 (90.7%)	77 985 (86.6%)
Pause	12 064 (4.3%)	12 064 (9.3 %)	12 064 (13.4 %)

2.3.2 Delivery

Additionally to the above-mentioned tiers, an extra tier named *delivery* has been created by duplicating the syllable tier and annotating it manually with stylistic and phonological variations such as liaisons, elision and hesitation, breath and mouth noises in pauses, and post-tonic schwas. In French, a post-tonic schwa is optional, in the sense that it may be omitted. This

specificity makes it interesting for various levels of speech studies. For example, at a phonological level, if schwa is pronounced it means that there is an extra syllable added to the word (cf. Figure 1, schwa [@] is pronounced in word “groupes” [gRu.p@]). The various symbols used for the delivery tier are grouped in Table 5.

Table 5: Description and counting for delivery symbols

Articulated syllables related symbols		n	%
@	Post-tonic syllabic schwa	2825	2.18
z	Hesitation	1289	0.99
c	Creaky voice	267	0.21
l	Liaison	2594	2.00
e	Elision	1491	1.15
a	Non-hesitation lengthening (sport)	265	0.20
Silence related symbols			
_	Silence	11754	9.07
*	Breath	3715	2.87
o	Less audible breath	1142	0.88
t	Mouth noise	721	0.56
Others symbols			
#	Human noise (laugh, cough)	313	0.24
%	Other noise	1065	0.82
+	Overlapping	99	0.08
!	Discourse interruption and repair	168	0.13

Among the 117 502 articulated syllables, 8 490 (7.2%) are tagged with one or more delivery symbols, representing 8 731 delivery symbols (as one syllable may have several tags). The 11 754 silences are all tagged with the main silence symbol _ and 5 028 (42.7%) are tagged with one or more silence-related delivery symbols (* o t).

The information contained in the delivery tier is used for describing a phonogenre as well as for characterising the personal speaking style of each speaker. Furthermore, it is an indicator of speech fluency or disfluency. It is also used in the detection of prosodic prominence for distinguishing between effective vowel lengthening and hesitation when taking into consideration the parameter of duration (cf. 2.3.5). A detailed annotation of silence characteristics can be used for detecting the prosodic boundaries. Finally, the annotation grouped under *other symbols* in Table 5 eliminates invalid syllables in the process of acoustic analyses (cf. 2.4).

2.3.3 Grammatical annotation

The lexical segmentation tier is doubled with a part-of-speech (POS) tier (named *pos-min* in Figure 1). Each word has been labelled automatically with its grammatical category by the tool DisMo (21). The following simplified version of tag-set is used to set apart lexical/content words from functional/grammatical words:

- *lexical words*: {ADJ (adjective, except for pre-nominal position) + ADV (adverb) + NOM (noun) + FRG (foreign word) + ITJ (interjection) + VER (verb)};
- *functional words*: {CON (conjunction) + DET (determiner) + PFX (prefix) + PRO (pronoun) + PRP (preposition) + VER:*:aux (auxiliary verb)}.

The total number of words obtained by DisMo (82 025) is greater than the word count during segmentation with EasyAlign (77 985), as DisMo correctly makes a lexical separation

(in tier *tok-min*) for contracted forms, such as preposition-noun pairs (e.g. “d’opposition” illustrated in Figure 1).

Grammatical information is used subsequently for automatic creation of three additional tiers: i) in the *lex* tier, the symbol * indicates a lexical word. This information is used in the following processing steps to detect ii) potentially stressed groups in the *SG* tier. The potentially stressed group (“*groupe accentuable*”, or “stress group” as defined in 22) is considered to be the minimal potentially stressed unit in French and can contain more than one word (e.g., functional word(s) followed by lexical word). The term “potentially stressed” refers to the fact that the automatic detection is based on a predictive model (22) that postulates that each final full syllable of a lexical word may, but does not have to, be stressed. iii) The *if* tier indicates the different kinds of stress that can possibly take place within a SG: final stress *f* is the most frequent in French; initial stress is tagged *i* if it is at the beginning of a lexical word, or *1* if it is at the beginning of a SG; if the penultimate syllable is stressed, it’s tagged *p* in the *if* tier. The final syllables containing schwa are considered non-stressable. For that reason the syllable [p@] of [gRu.p@] is tagged @ in the *if* tier and the final stress *f* is assigned to [gRu] (cf. Figure 1).

In summary, the TextGrids have one tier at phone level, four tiers at syllabic level (syllable, if, prominence, delivery), four tiers at word level (words, tok-min, pos-min and lex), one SG tier and three tiers at pause-separated units level (phono, ortho and PSU) as illustrated in Figure 1.

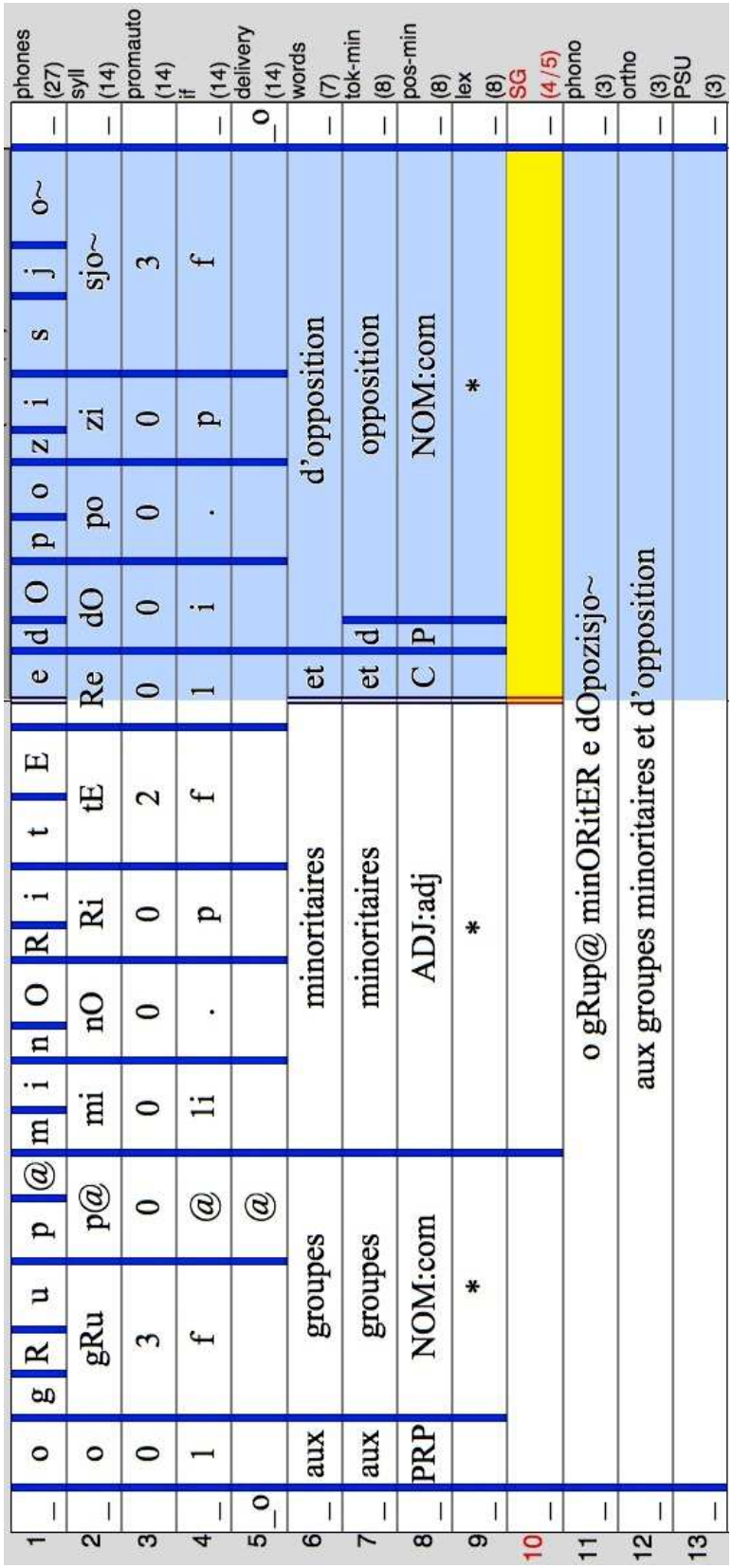

Figure 1: Multi-tier annotation for C-PhonoGenre corpus at levels of phones, syllables (+ if + prominence + delivery), words (+ tok-min + pos-min + lex), SG, phono, ortho and PS.

2.3.4 Pitch

Praat's automatic pitch detection method resulted in frequent errors for data with particularly noisy environments, such as sport commentary or church mass, and for data containing many hesitations, such as spontaneous narration. This proved to be an obstacle for some acoustic measurements, and therefore the pitch tier was recalculated and then manually corrected within Praat. For the sake of homogeneity of the corpus, pitch was corrected for all data and not only for those where the recording was problematic.

2.3.5 Prominence

After segmentation and alignment of the speech signal, the corpus received prosodic annotation of prominence. Each syllable was assigned a score of acoustic prominence from 0 to 4 using ProsoProm (23). This tool is built upon findings on acoustic correlates of perceived prominence in French, previously according to a protocol defining three degrees of prominence and applied to a manual annotation of the 70 minutes corpus C-Prom (17), and today according to another protocol defining five such degrees and applied to a manual annotation of an 18 minutes corpus. As mentioned under 2.3.2, the information from the delivery tier helps to improve the automatic detection of prominence by taking into account hesitation and the post-tonic syllabic schwa. Both phenomena were indicated as being problematic for prominence detection in (24). ProsoProm takes into account pitch and duration of syllables relatively to surrounding syllables, as well as pauses and pitch rises. For the entire corpus, 63.2% of the syllables were labelled [0] non-prominent, 11.9% scored [1], 5.8% scored [2], 5.5% scored [3] and 13.6% scored [4].

The automatic, semi-automatic and manual preliminary steps described under 2.3 are required for the acoustical analyses and further results.

2.4 Acoustic analysis and prosodic report

Three tools implemented as Praat scripts – Prosogram, ProsoReport and DurationAnalyser – are used for the acoustic and statistic treatment of the corpus.

Prosogram (25) is applied for pitch stylisation of the data. Its two-step algorithm first detects vocalic nuclei for each syllable based on voicing and intensity; and then the nucleus pitch curve is stylised into a static or dynamic tone based on a perceptual glissando approach (26).

Taking the pitch stylisation from Prosogram as a starting point, ProsoReport (6) incorporates information contained in the other tiers described under 2.3.1-2.3.3. ProsoReport has been under constant development for a couple of years by increasing the number of extracted acoustic and prosodic features. At present, it proposes 64 temporal and pitch measures in order to address different questions concerning phonetics, prosody and linguistics.

A detailed prosodic report provides measures at local (phones, syllables, pauses) and global (Pause-Separated Units, PSU) level, as well as measures and statistics for the entire recording (e.g., articulation rate/ratio, duration/pitch mean and deviation, pitch distribution). In some cases only relative measures (e.g., mean, rate, percentage) are considered, and in other cases, only absolute measures (e.g., number of phones, total duration). This could be useful if groups of recordings are compared, while the size and number of recordings as well as the speakers' individual properties need to be ignored. Thanks to the previous automatic prominence detection (cf. 2.3.5), ProsoReport also computes the tonal and rhythmic distribution

of prominent and non-prominent syllables (e.g., percentage of prominent syllables in various positions).

DurationAnalyser (5) computes exclusively temporal measures and rhythmical variability measures based on vocalic, consonantal and syllabic intervals.

3 Results

Data measures are grouped either by phonogenre, by sub-phonogenre, or by each situational feature. They are divided into three parts:
(3.1) measures related to the annotation and segmentation of the corpus;
(3.2) acoustic measures at macro- and micro-prosodic level, as well as segmental duration;
(3.3) global measures by using Principal Component Analysis.

3.1 Annotations

Measuring the number of syllables per pause-separated units (PSU) shows that religious sermon [LIT] has the shortest units, followed by presidential wishes [VXP] and sport commentary [SPO] (Figure 2 – $F(8,96)=8.55$ $p<0.001$). [LIT] and [VXP] have both a solemn character since they are addressed to the potential followers of a religion or citizens of a country. Sport commentary speech is depending on the game action and the PSU reflect these dynamics. The highest score of number of syllables per PSU, obtained for weather forecast [MET] and radio press review [RPR], reflects the time pressure characterising speech production in broadcast media. For these two phonogenres, one can also observe an important variability in their box plot span. This indicates the big diversity of journalist strategies when it comes to segmenting the speech flow and/or taking a breath. On the contrary, the box plot spans for [LIT] and [SPO] are quite compact suggesting that the different speakers would produce PSU containing a similar number of syllables.

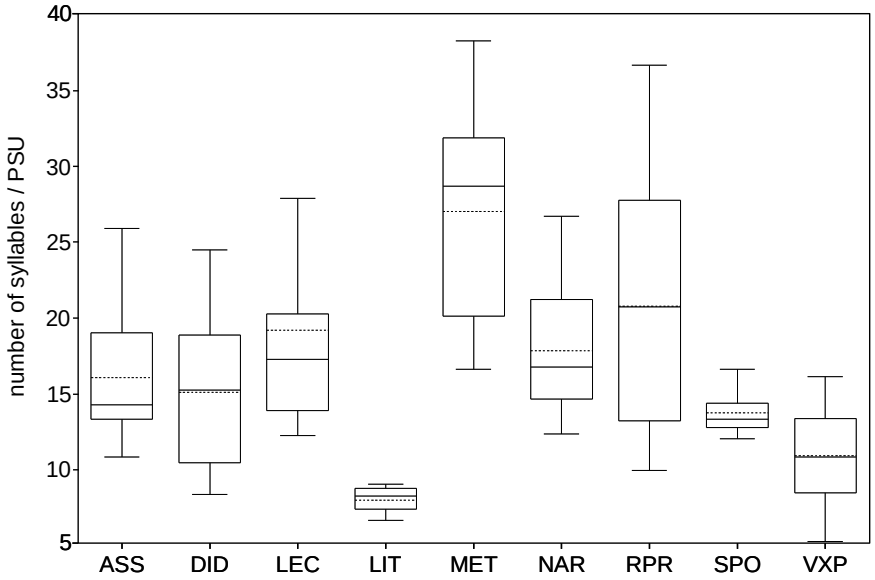

Figure 2: Number of syllables per pause-separated units (PSU) for nine phonogenres

3.2 Acoustics
3.2.1 Macroprosodic measures

The macroprosodic component refers to the speaker's choice of rhythm and intonation patterns.

Measures of the articulation ratio (proportion of articulated speech vs. silences) at phonogenre level show a global effect ($F(12,91)=30.96$ $p<0.001$) and oppose religious sermons [LIT], presidential wishes [VXP] and sport commentaries [SPO] to the others and thus corroborate their character announced under 3.1.

Articulation ratio at sub-phonogenre level brings some new insights illustrated in Figure 3. Ministers (or their delegates) who answer [ASS-R] during Question Time in parliamentary speech tend to occupy more speech time than deputies who formulate their question [ASS-Q]. It is probably because the answer provokes more or less loud reactions among deputies in the parliament. The speaker reacts in turn to this situation by reducing the number of pauses in order to maximise his use of the two minutes allocated to him. Educational speech [DID] as well presents differences at the sub-phonogenre level, mainly between Radio and TV. This might be for a couple of reasons: TV recordings are longer (10 minutes) than radio ones (3 minutes); speech time at television must be shared with the visual flow, and prosodic features are likely to be different in the situation where speech refers to image (27). A lecture at a scientific conference [DID-cnf] pronounced by university professors is naturally closer to [DID-Rad], that gathers answers from teachers and researchers, than to [DID-TV] with clearly media context, though instructional. The media framework can have an impact on the speech flow. For example, artificial silences are often introduced and the speech flow is cut and reorganised in order to create a documentary or reportage. The difference between two situations of religious sermons [LIT] speech is hardly distinguishable as for articulation ratio, nevertheless the two sub-phonogenres differ more clearly in other prosodic measures as discussed below.

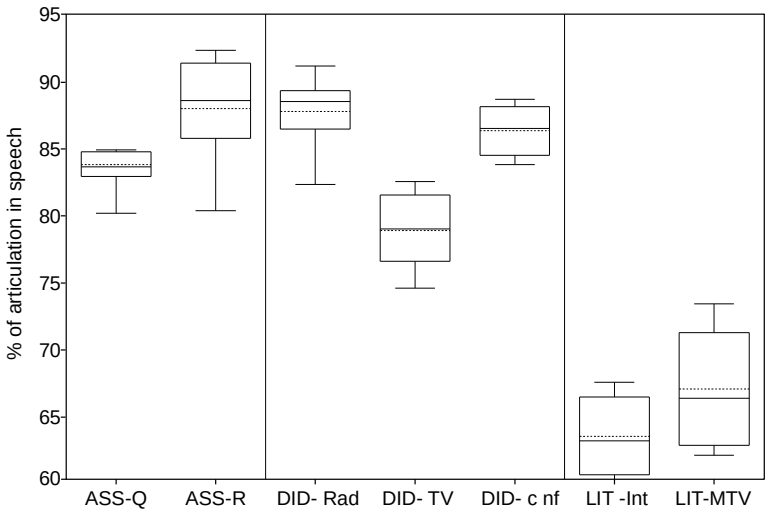

Figure 3: Articulation ratio for seven sub-phonogenres

Considering the above-mentioned seven sub-phonogenres, the articulation rate (number of syllables per second silence excluded) analysis gives an interesting difference (Figure 4). The parliamentary speech question [ASS-Q] manifests a faster articulation rate than parliamentary speech answer [ASS-R]. The same distinction holds for the two [LIT] sub-phonogenres: a sermon on Internet [LIT-Int] is read faster than a sermon served in a church [LIT-MTV]. Differences between the three educational [DID] sub-phonogenres are not relevant when it

comes to speech rate (number of syllables per second silence included). Post-hoc tests for the articulation rate ($F(12,91) = 9.33$, $p<0.001$) oppose significantly [LIT] and presidential wishes [VXP] to the other phonogenres (Figure 4). Weather forecast [MET] is detached from the others as phonogenre with the highest speech rate. Reading [LEC] is the most spread one: this can be explained by the fact that the recordings are equally shared between older and younger generations, as well as between speakers of Paris and Lyon (East of France). Studies in regional prosody of French (28) report that the speech rate is faster in the former than in the latter. The dispersion of narration [NAR] data is less important; nevertheless, it reflects the heterogeneity of story topics and of the speaking styles of storytellers. A similar dispersion is observed for presidential New Year's wishes [VXP] and reflects geographic and diachronic differences described under 2.1.

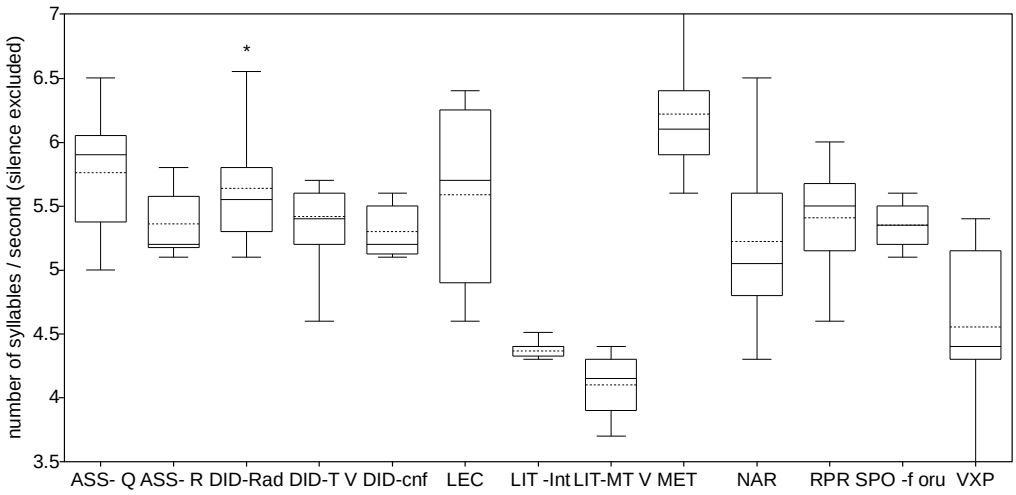

Figure 4: Articulation rate for thirteen sub-phonogenres

Pitch variation

Macroprosodic measures of pitch bring out similarities between educational speech [DID], radio press review [RPR] and presidential wishes [VXP]: the standard deviation of pitch is higher, which means that the melodic agitation is more important. Conversely, parliamentary speech [ASS], religious sermons [LIT] and reading [LEC] are less subject to melodic variation. For the same group – [DID], [RPR] and [VXP] – the pitch range is larger, which is a cue for a more significant prosodic expressivity than for [ASS], [LIT] and [LEC].

Intonational properties indicate a lower relative pitch variation (standard deviation σ of pitch / average $\bar{x}$ of pitch; measured in semitones) 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 a greater speaker's involvement. However, this acoustic parameter varies according to our predictions across the media feature ($F(2,102)=12.06$; $p<0.001$).

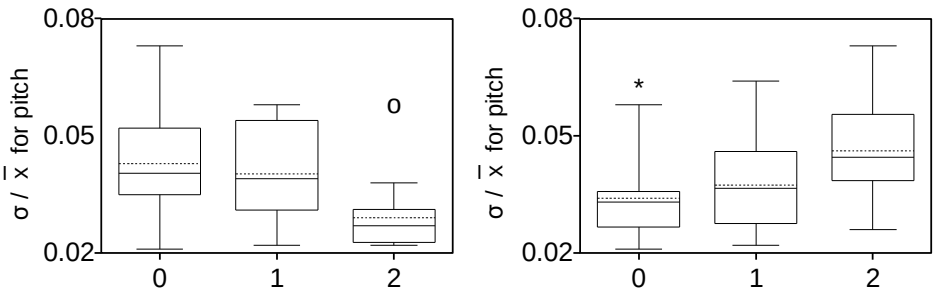

Figure 5: The relative pitch variation for 3 degrees (0 absent, 1 semi-, 2 present)of *audience* (left) and *media* (right) situational features

3.2.2 Microprosodic measures

The study of initial and final positions of prominent syllables permits to differentiate phonogenres according to their situational features.

The percentage of prominent final syllables is decreasing as the phonogenre is getting more interactive ($F(2,102)=10.43$; $p<0.001$). This can be explained by a high score of hesitation in narration [NAR] and of vowel lengthening, typical for sport commentaries [SPO] (Figure 6, left).

The percentage of prominent initial syllables is getting higher as a phonogenre falls within broadcast media speaking style, where it is important to clearly distinguish discourse segments (Figure 6, right). The initial prominent syllables of a potentially stressed group (SG) show similar results ($F(2,102)=5.88$; $p<0.001$).

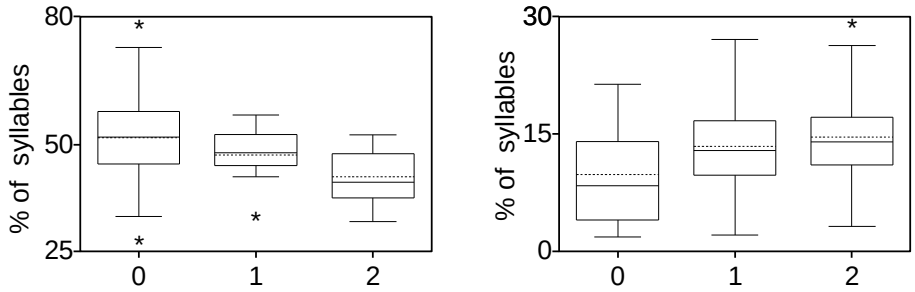

Figure 6: Percentage of prominent final syllables for *interactivity* (left) and percentage of prominent initial syllables for *media* (right) for each of the 3 degrees

The relative length of initial and final syllables of the potentially stressed group (SG) varies in a significant manner across the preparation dimension (initial syllables $F(2,102)=4.05$; $p<0.001$; final $F(2,102)=5.42$; $p<0.001$). Initial syllables of SGs tend to be shorter in prepared discourse than in non-prepared, but final syllables become longer (Figure 7).

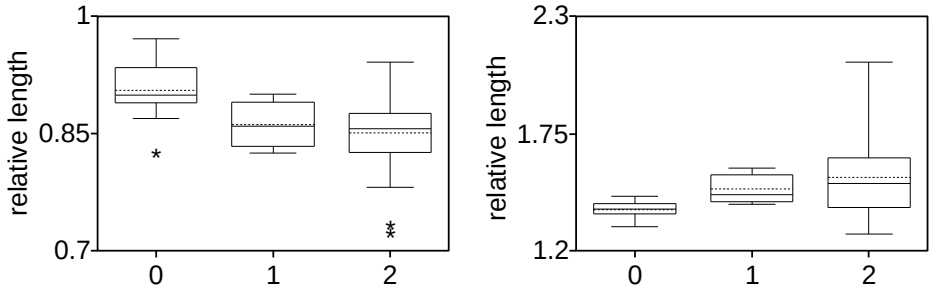

Figure 7: Relative length of initial (left) and final (right) syllables
of the potentially stressed group (SG) for *preparation* for each of the 3 degrees

The physical presence of audience implies lower percentage of initial prominent syllables per SG ($F(2,102)=12.8$; $p<0.001$). This is the case for parliamentary speech [ASS] where the deputy is talking directly to the minister; or for university lectures, where the teacher is addressing the students [DID-cnf]; or during sermons where the priest is addressing the believers [LIT-MTV]

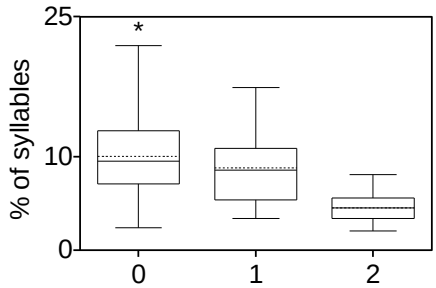

Figure 8: Percentage of initial prominent syllables per SG for *audience* for each of the 3 degrees

Sub-phonogenre level distribution of initial prominences (Figure 9) partly reflects the values of situational features ($F(12,91)=7.13$; $p<0.001$). Sub-phonogenres in which an audience is present, have the lowest percentage of initial prominent syllables – parliamentary speech [ASS], lecture [DID-cnf], church mass [LIT-MTV] –, whereas the media ones – educational speech on the radio [DID-Rad] and on the television [DID-TV], weather forecasts [MET] – show the highest percentage. Sermon on the Internet [LIT-Int] behaves like a media style, although it was graded as an intermediate style in the media dimension. Although the radio press review [RPR] is a prototypical case of broadcast media style, it shows a low level of initial prominence. Actually, [RPR] displays a high rate of prominent initial syllables of words, not of potentially stressed groups (SG); this reflects a stylistic choice in marking initial syllables.

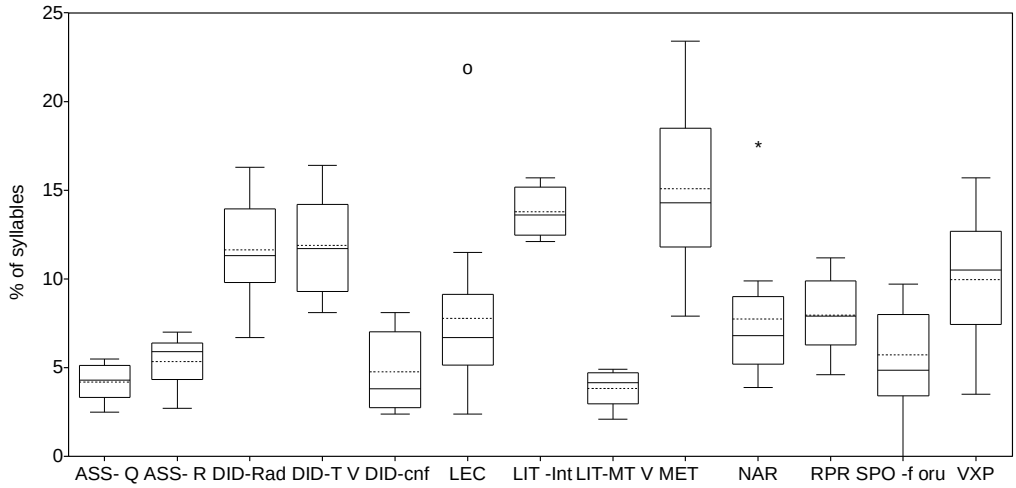

Figure 9: The distribution of the percentage of initial prominent syllables per SG for the thirteen sub-phonogenres

Educational speech [DID-Rad] and [–TV] (but not lectures), reading [LEC], weather forecasts [MET], and radio press review [RPR] show a greater proportion of rising syllables (Figure 10) than parliamentary speech [ASS], [DID-cnf], church mass [LIT-MTV], spontaneous narration [NAR], [SPO], and (to a lesser extent) presidential wishes [VXP] ($F(12,91)= 4.14$; $p<0.001$). Low rate for [VXP], [LIT] and [NAR] may be due to their 'empathic' dimension.
As for falling syllables ($F(12,91)= 2.74$; $p=0.003$), [LIT-MTV] and [VXP] are clearly set apart from the others, marking both their authoritative status as well as the fact that they are speaking of the future.

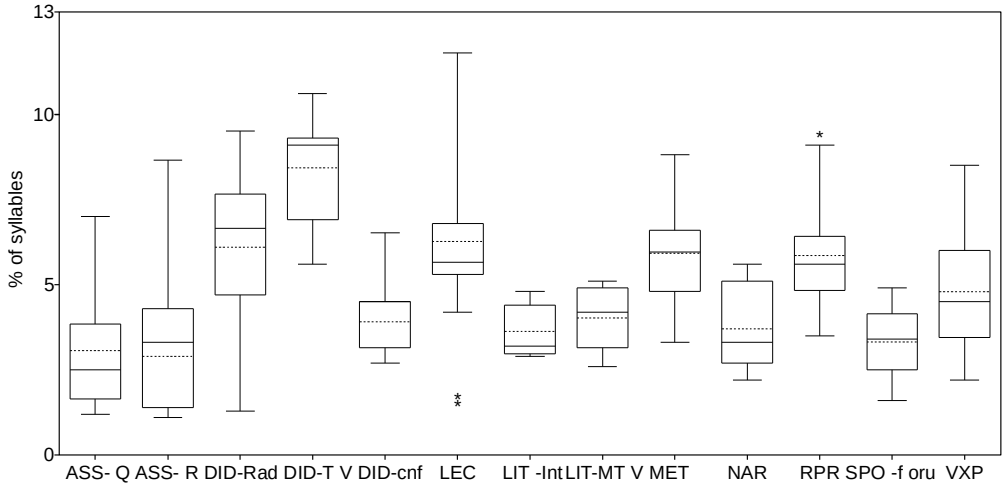

Figure 10: Percentage of rising syllables for thirteen sub-phonogenres

3.2.3 Segmental duration

The spontaneous narration [NAR] and religious sermon [LIT] phonogenres are those with higher vowel duration (Figure 11), but for different reasons: in the first case, this is only a side effect of hesitations, while it is a deliberate feature of liturgical speech, as the annotations in the delivery tier confirm. The weather forecast [MET] has the shortest vowels, followed by radio press review [RPR] and reading [LEC], for which the duration has a high variance. On the level

of sub-phonogenre, there are no strong differences. It is most marked for parliamentary speech [ASS], in which answers have vowel duration longer than questions. For [LIT], sermons have shorter vowels than church masses.

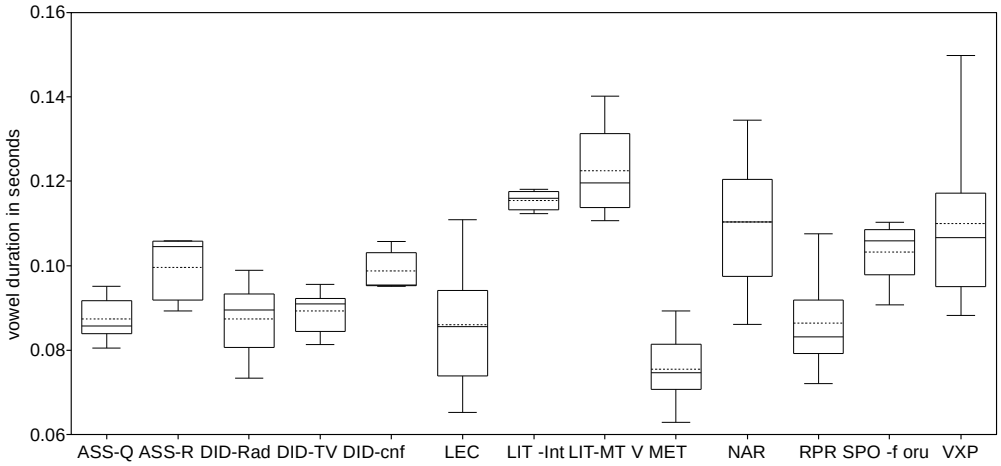

Figure 11: Vowel duration for thirteen sub-phonogenres

3.3 Principal Components Analysis

A Principal Components Analysis (PCA) was applied to investigate globally the differences between phonogenres and between each situational feature. This statistical technique makes an optimal linear combination of all the acoustics parameters. The resulting "principal components" (PCs) are dimensions of a normalised vector space but do not correspond to the original features. Instead, each principal component (a linear combination of various acoustic features) "explains" or improves the prediction of increasing parts of the population's variation. In our case, the parameters of the two tools ProsoReport and DurationAnalyser were grouped to model phonogenre distinction in the PCA. The first two principal components explain 58% of the variation, while the first eight explain 90.5%. A discriminating analysis for an automatic classification with those first eight PCs over nine phonogenres showed that 93% of recordings were identified correctly.

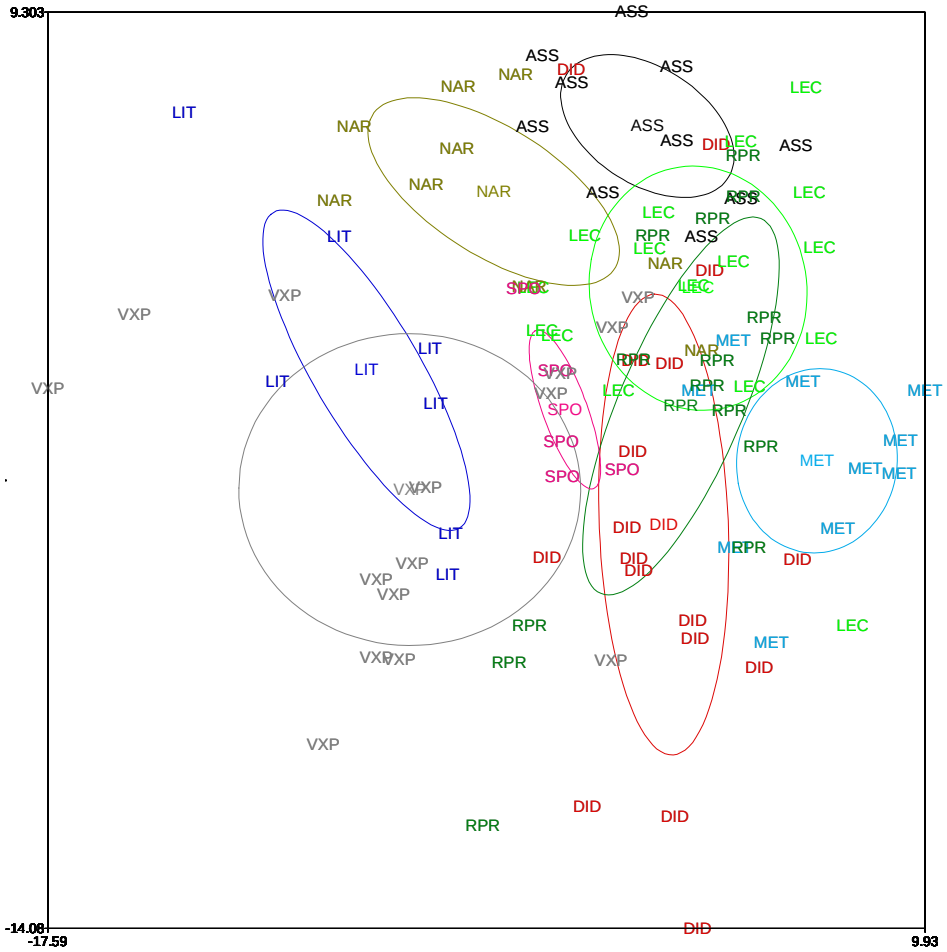

Figure 12: Distribution of 105 recordings in the first two Principal Components for nine phonogenres (the confidence interval of ellipses is 0.6)

The graphical distribution of phonogenres (represented by abbreviations in Figure 12) shows the projection of the selection of 105 recordings onto the first two Principal Components. It can be observed that parliamentary speech [ASS] and weather forecasts [MET] are the most compact phonogenres, probably because of the strict constraints of the situation. The dispersion of reading [LEC] and narration [NAR] is slightly larger and reflects the geographical and age differences among speakers. Educational speech [DID] and religious sermon [LIT] are even less compact: this is because of the differences in speech situation explained above. The same for radio press review [RPR] where the dispersion is probably due to one speaker with a particular speaking style represented 3 times in the corpus. Finally, presidential wishes [VXP] present more than one particularity: (a) the grouping of French presidents into the three chronological periods – 1970s, 1980-1990s and 2000s; (b) the clear separation of the discourse of Swiss and French presidents that shows the impact of geographical dimension.

4. Discussion

We have presented a large corpus consisting of a variety of nine phonogenres. Except for sport commentary [SPO] and religious sermon [LIT] phonogenres, each of these is represented by at least ten speakers, with the idea that we study the phonogenre itself and get rid of individual characteristics.

We have shown how acoustic measures give rise to groupings of phonogenres among themselves and according to situational features, and how they characterise phonogenres. This justifies a posteriori our choice of ten speakers per phonogenre, and is a progress with respect to the previous studies quoted in this paper. During this research it appeared that four situational features with three degrees each (cf. Table 3) are significant for the study of phonogenre. In this sense reducing a much larger set of features and degrees established in earlier work is justified.

Our results show that phonogenres present evidence for groupings according to unforeseen, or hidden, situational features that are part of their prototypical image. For example, external time pressure, reflected in the duration of speech runs, is inherent in parliamentary speech [ASS], weather forecasts [MET] and radio press review [RPR]; ritual pressure on solemnity is characteristic for religious sermons [LIT] and presidential wishes [VXP] and is reflected in the bigger proportion of falling syllables.

The sub-phonogenre level was introduced to ensure solid definitions and to reduce the excessive heterogeneity of some phonogenres. The differences observed between questions and answers within the parliamentary speech [ASS] phonogenre suggest the relevance of the interactivity situational feature at the sub-phonogenre distinction level. They reveal a prosodic reflection of a discursive (not situational) category, namely to be an initial or reactive member of an exchange (though not in direct interaction). This essentially shows that, when collecting a sub-phonogenre corpus, a general speaking style label should be avoided and that special attention should be given to the exact situation of each recording, i.e., accurately defining its situational features.

This study provides a methodology for broad prosodic investigation of a large and varied corpus by using a semi-automatic set of procedures. Although some annotation steps remain manual, most of the procedure is automatic. As this framework was built in a very generic way, future work should propose a more targeted selection of prosodic measures, and test corpora of other languages. Two kinds of applications can be considered: verification of linguistic hypotheses and automatic phonogenre identification.

Finally, we should mention that the corpus will be made available to the community for research purposes.

Acknowledgements

This research is funded by Swiss National Science Foundation – FNS Grant n° 100012_134818. The authors thank the anonymous reviewer for many useful and constructive suggestions. The authors also thank George Christodoulides for his contribution in data extension and treatment.

References

1. Goldman J-P, Auchlin A, Simon AC. Discrimination de styles de parole par analyse prosodique semi-automatique. In Yoo HY, Delais-Roussarie E, editors. Actes d'IDP 2009; Septembre 2009; Paris; 2011. p. 207–221. Available from: http://makino.linguist.jussieu.fr/idp09/docs/IDP_actes/Articles/Goldman.pdf.

2. Simon AC, Auchlin A, Avanzi M, Goldman J-P. Les phonostyles: une description prosodique des styles de parole en français. In: Abecassis M, Ledegen G, editors. Les voix des Français. En parlant, en écrivant. vol. 2. Berne: Peter Lang; 2010. p. 71–88.

3. Lucci V. Étude phonétique du français contemporain à travers la variation situationnelle. Grenoble: Université des langues et lettres de Grenoble; 1983.

4. Koch P, Oesterreicher W. Langage parlé et langage écrit. In: Holtus G, Metzeltin M, Schmitt CH, editors. Lexikon der Romanistischen Linguistik. vol. I/2. Tübingen: Niemeyer; 2001. p. 584–627.

5. 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. Bonn; 2010.

6. Goldman JP, Simon AC, Auchlin A, Avanzi M. Phonostylographe, un outil de description des phonostyles prosodiques. Nouveaux Cahiers de Linguistique Française. 2007;28:219–237. Available from: http://clf.unige.ch/display.php?numero=28&idFichier=110.

7. Beacco JC. Trois perspectives linguistiques sur la notion de genre discursif. Langages. 2004; 38/153:109–119.

8. Solin A. Genre. In: Zienkowski J, Ostman JA, Verschueren J, editors. Discursive Pragmatics. Amsterdam: John Benjamins; 2011. p. 119–134.

9. Bawarshi AS, Reiff MJ. Genre: An Introduction to History, Theory, Research, and Pedagogy. West Lafayette, Indiana: Parlor Press; 2010.

10. Léon P. Précis de phonostylistique. Parole et expressivité. Paris: Nathan Université; 1993.

11. Fónagy I, Fónagy J. Prosodie professionnelle et changements prosodiques. Le Français Moderne 44; 1976. p. 193–228.

12. Llisterri J. Speaking styles in speech research. In: ELSNET/ESCA/SALT Workshop on Integrating Speech and Natural Language. Dublin, Ireland; 1992. Available from: http://liceu.uab.cat/~joaquim/publicacions/SpeakingStyles_92.pdf.

13. Eskénazi M. Trends in Speaking Styles Research. In: Proceedings of Eurospeech '93: 3rd European conference on speech communication and technology. Berlin, Germany: ESCA European Speech Communication Association; 1993. Available from: http://www.isca-speech.org/archive/eurospeech_1993/e93_0501.html.

14. Boula de Mareüil P. Accents et styles. Une étude à base de perception et d'analyses acoustiques à travers le traitement automatique de la parole [HDR]. Université Paris 3. Paris; 2012.

15. Obin N, Lacheret-Dujour A, Veaux C, Rodet X, Simon AC. A Method for Automatic and Dynamic Estimation of Discourse Genre Typology with Prosodic Features. In: Proceedings of InterSpeech 2008. Brisbane, Australia: ISCA; 2008. p. 1204–1207. Available from: http://www.isca-speech.org/archive/archive_papers/interspeech_2008/i08_1204.pdf.

16. Hirschberg J. A corpus-based approach to the study of speaking style. In: Prosody: Theory and Experiment-Studies Presented to Gösta Bruce. Dordrecht: Kluwer; 2000.

17. Avanzi M, Simon AC, Goldman J-P, Auchlin A. C-PROM. Un corpus de français parlé annoté pour l'étude des proéminences. XXVIIIèmes Journées d'Etude sur la Parole JEP 2010 [Internet]. Mons, Belgique: Université de Mons; 2010. p. 73–76. Available from: http://www.afcp-parole.org/doc/Archives_JEP/2010_XXVIIIe_JEP_Mons/2010_XXVIIIe_JEP_Mons.pdf.

18. Avanzi M, Schwab S, Dubosson P, Goldman J-P. La prosodie de quelques variétés de français parlées en Suisse romande. In: Simon AC, editor. La variation prosodique régionale en français. Bruxelles: De Boeck/Duculot; 2012. p. 89–119.

19. Boersma P, Weenink D. Praat: doing phonetics by computer. Available from: http://www.praat.org

20. Goldman JP. EasyAlign: An Automatic Phonetic Alignment Tool Under Praat. In: Interspeech'11, 12th Annual Conference of the International Speech Communication Association. Firenze, Italy; 2011. p. 3233–3236. Available from: http://archive-ouverte.unige.ch/unige:18188.

21. Christodoulides G, Avanzi M, Goldman J-P. DisMo: A Morphosyntactic, Disfluency and Multi-Word Unit Annotator: An Evaluation on a Corpus of French Spontaneous and Read Speech. In: Proceedings of 9th Language Resources and Evaluation Conference (LREC) 2014. Reykjavik, Iceland; 26-31 May.

22. Mertens P. A Predictive Approach to the Analysis of Intonation in Discourse in French. In: Kawaguchi Y, Fónagy I, Moriguchi T, editors. Prosody and Syntax. Series "Usage-Based Linguistic Informatics" 3. Amsterdam: John Benjamins; 2006. p. 64-101.

23. Goldman J-P, Avanzi M, Simon AC, Auchlin A. A Continuous Prominence Score Based on Acoustic Features. In: Proceedings of Interspeech 2012. Portland, USA. p. 2454-2457.

24. Morel M, Lacheret-Dujour A, Lyche Ch, Poiré F. Vous avez dit proéminence ? Actes des XXVIes Journées d'études sur la parole (JEP) 2006. Dinard, France; 12-16 June.

25. Mertens P. The Prosogram: Semi-Automatic Transcription of Prosody based on a Tonal Perception Model. In: Bel B, Marlien I, editors. Proceedings of Speech Prosody (SP) 2004. Nara, Japan; 23-26 March.

26. d'Alessandro C, Mertens P. Automatic pitch contour stylization using a model of tonal perception. Computer Speech and Language 9(3); 1995. p. 257-288.

27. Kern F. Speaking dramatically: The prosody of live radio commentary of football matches. In: Barth-Weingarten D, Reber E, Selting M, editors. Prosody in interaction. Amsterdam: John Benjamins; 2010. p. 217–237.

28. Simon AC, editor. La variation prosodique régionale en français. Bruxelles: De Boeck/Duculot; 2012.

29. Audrit S, Pršir T, Auchlin A, Goldman J-P. Sport in the media: a contrasted study of three sport live media reports with semi-automatic Tools. Proceedings of the 6th International Conference on Speech Prosody 2012 [Internet]. Shanghai, China: Tongji University Press; 2012. Available from:

http://www.speechprosody2012.org/uploadfiles/file/sp2012_submission_170.pdf