C-PhonoGenre, a 7-hour corpus of 7 speaking styles in French: relations between situational features and prosodic properties

Jean-Philippe Goldman[1], Tea Pršir[1,2], Antoine Auchlin[1]
jean-philippe.goldman@unige.ch, tea.prsir@unige.ch, antoine.auchlin@unige.ch
1 Département de Linguistique, Université de Genève, Switzerland
2 Institut Langage & Communication, UCLouvain, Belgium

Abstract

Phonogenres, or speaking styles, are typified acoustic images associated to types of language activities, causing prosodic and phonostylistic variations. This communication presents a large speech corpus (7 hours) in French, extending a previous work by Goldman et al. (2011a), Simon et al. (2010), with a greater number and complementary repertoire of considered phonogenres. The corpus is available with segmentation at phonetic, syllabic and word levels, as well as manual annotation. Segmentations and annotations were achieved semi-automatically, through a set of Praat implemented tools, and manual steps.
The phonogenres are also described with a reduced set of situational dimensions as in Lucci (1983) and Koch & Oesterreicher's (2001). A preliminary acoustic study, joining rhythmical comparative measurements (Dellwo 2010) to Goldman et al.'s (2007a) ProsoReport, reports acoustic differences between phonogenres.

Keywords: phonostyle, prosody, speaking style

1. Introduction

Situations in which speakers utter surely have influence in the resulting speaking style. The goal of studying the variation in speech with a « situational » point of view is to establish correlations between situational features and prosodic properties. On one side, situations gather in groups according to an implicit typology, which still has to be determined; on the other, typical prosodic features tend to stabilize speaking styles (e.g. live sport report, church speech; etc.), and make them highly recognizable.

Large background for present research is the growing interest for genre in language sciences (Beacco 2004; Solin 2011), and more narrowly the study of oral genres, or phonogenres, and the so-called situational variation (Simon et al. 2010; Goldman et al. 2011; Boula de Mareüil 2012; Obin et al. 2008, among others). The research enlarges 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. It further joins rhythmical comparative measurements (Dellwo 2010) to Goldman et al.'s (2007) ProsoReport.

Following Goldman et al. (2011), we distinguish phonogenre, defined as a typified acoustic image associated to a situation and speech activity, from phonostyle, the properties of a given speech sample within a genre. Our speech samples are collected and grouped according to shared situational features, inspired from Lucci (1983) situational invariants, and Koch & Oesterreicher (2001) speech conception features (« traits conceptionnels »). Four dimensions of speech situation are considered, each one yields three values as in Table 2:

1. type of direct **audience** (public, face-to-face, microphone only/booth)
2. **media** (exclusively, semi-media, ou non-média)
3. degree of **preparation** (read, semi-prepared, spontaneous)
4. degree of **interactivity** (interactive, semi-interactive, or no interactivy).

2. Corpus collection and annotation

After C-Prom corpus (Avanzi et al. 2010), two conclusions were obvious: 1/ situational features should be more constrained to avoid dispersion in the speaking style. 2/ each speaking style should be represented by a greater number of speakers by phonogenre to avoid idiosyncrasy. Therefore, C-PhonoGenre corpus is composed of phonogenre, or speaking styles, or situational features that are more constrained, with at least 10 speakers per style.

2.1 Corpus collection

The corpus is composed of 7 phonogenres: parliamentary speech [ASS] (questions to government at French National Assembly), didactic speech [DID], liturgy [LIT], radio press review [RPR], live sport report [SPO], presidential New Year's wishes [VXP] and spontaneous narration [NAR]. The two phonogenres ASS and VXP could be considered as belonging to a unique "politic" macro-genre. The VXP part also has a diachronic dimension (from De Gaulle 1968 to Sarkozy 2007 for French presidents and from Dreifuss 1999 to Calmy-Rey 2011 for Swiss presidents).

PhonoGenre	Number of recordings	Duration (mn.)
ASS	10	20
DID	14	85
LIT	7	54
NAR	10	35
RPR	15	93
SPO	5	35
VXP	15	95
TOTAL	**76**	**417**

Table 1. Number and duration of recordings

2.2 Situational features

Seven phonogenres are described following their situational feature in Table 3. Three of them where split into two groups because of the specificity of their situational features. This splitting was done a posteriori since we realized that acoustic features of the same phonogenre are set apart in a significant manner if there is only one feature that changes. Table 2 describes three degrees of situational features.

degree	audience	media	preparation	Interactivity
0	Microphone	No	Spontaneous	Non (mono)
1	Face-to-face	Semi-	Semi-prepared	Semi-interactive
2	Public	Media	Read/Prepared	Interactive

Table 2. Three degrees of situational features

2.3 Segmentation and annotation

2.3.1 Segmentation

Beyond data collection, the main value of a speech corpus is its annotations. In our case, the whole corpus was segmented at the levels of phones, syllables and words with the EasyAlign tool (Goldman 2011b) on the basis of the orthographic transcription. This tool gives high quality segmentation, as it requires some manual adjustments between the successive automatic steps. Therefore, the human expert overviews the segmentation process and corrects the errors. The following table shows the vital stats of this corpus in terms of phone, syllable and word intervals. Over 96044 syllabic intervals, 86700 (90.3%) are plain articulated syllable and 9344 (9.7%) are pauses intervals.

	76 files	Articulated speech
Phones	205675	196331 (95.5%)
Syllables	96044	86700 (90.3% -2670 types)
Words	65423	56079 (85.3% - 8638 types)

Table 4. Counting of intervals at phonetic, syllabic and word levels

2.3.2 Delivery

Phonological and stylistic variations such as liaisons, elision and hesitation, as well as breath taking in pauses and mouth noises were manually annotated in an extra tier named *delivery*.

This tier is likewise the syllable tier, i.e. the number of intervals and the boundaries are exactly the same. The following table shows the various symbols used for the delivery tier.

Among the 86700 articulated syllables, 10887 have as delivery symbol (12.6 % of the articulated intervals), while, among the 9344 pauses, 5249 have as delivery symbol besides the main pause symbol "_" (i.e. 56 % of the pause intervals). Over all the syllabic intervals, 16136 have a delivery symbol (16.8%).

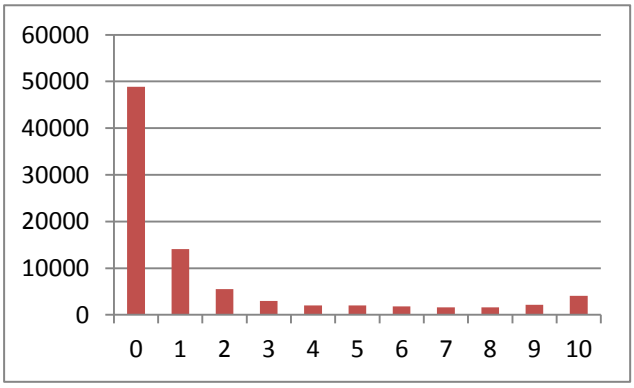
Articulated syllables related symbols		n	%
@	post-tonic syllabic schwa	2537	2.64
z	Hesitation	480	0.50
c	creaky voice	78	0.08
l	Liaison	2292	2.39
e	Elision	990	1.03
a	Non-hesitation lengthening (sport)	225	0.23
Silence related symbols			
_	Silence	9344	9.74
*	Breath	3106	3.23
o	less audible breath	1099	1.14
t	mouth noise	585	0.61
Others symbols			
#	Human noise (laugh, cough)	156	0.16
%	Other noise	980	1.02
+	Overlapping	77	0.08
!	Syntactic interruption	122	0.13

Table 5. Description and counting for delivery symbols

2.3.3 Prominence

An additional automatic process was applied to the whole corpus. Its goal is to calculate for each plain syllable, a gradual score of acoustic prominence. The ProsoProm tool (Goldman 2007b) was used and could yield a linear prominence score from 0 to 10 as in Figure 1.

Figure 1. Distribution of degree of prominence for the



86700 articulated syllables

2.3.4 Grammatical annotation

The word tier was also duplicated into a part-of-speech tier (POS) with grammatical annotation for further studies on the phonostylistic-grammatical interface. The tool Dismo (Christodoulides 2014) was used to automatically tag the 76 recordings. A simplified version of the tag set is showed in the following table with counts.

The total number of words (59285) is greater than the word counting during the segmentation with EasyAlign, as Dismo correctly makes a lexical distinction for agglutinated words like DET+NOUN pairs like "l'autonomie" as in Figure 2.

All in all, the TextGrid have one tier at phone level, three tiers at syllabic levels (syllable, delivery, prominence) and two tiers at word levels (words and POS), as shown in Figure 2.

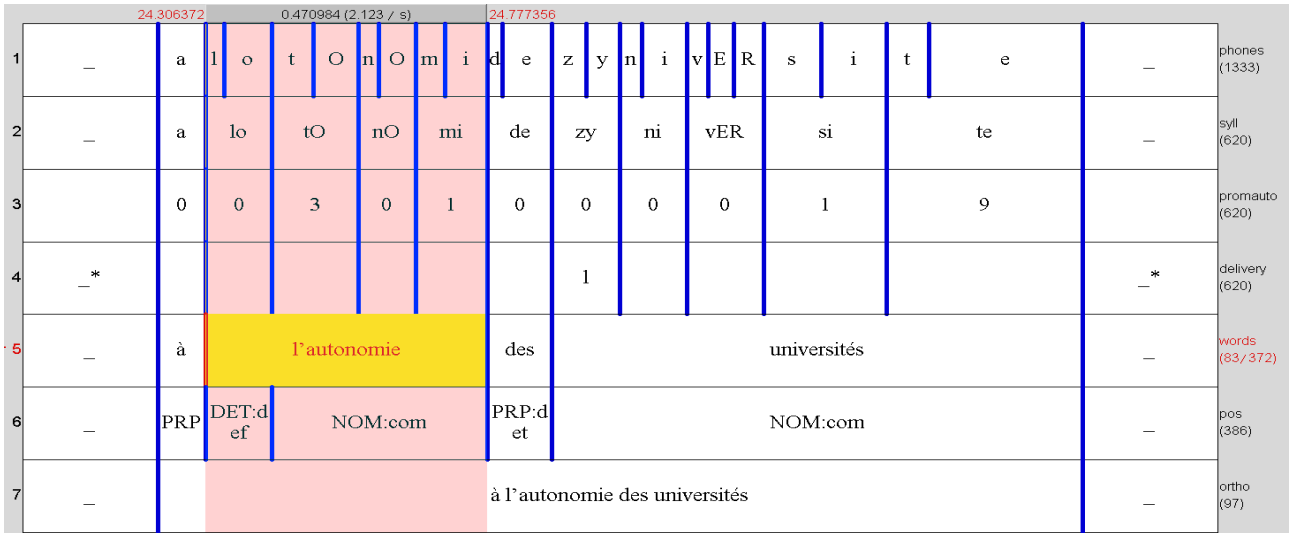


Figure 2. Multi-tier annotation for C-PHONOGENRE corpus at levels of phones, syllables (+prominence detection + delivery manual annotation) and words (+part-of-speech)

POS	N	%
ADJ (adjective)	3711	6.26
ADV (adverb)	3954	6.67
CON (conjunction)	3481	5.87
DET (determiner)	9273	15.64
FRG (foreign word)	13	0.02
INTJ (interjection)	532	0.90
NOM (noun)	15582	26.28
PFX (prefix)	26	0.04
PRO (pronoun)	6321	10.66
PRP (preposition)	7886	13.30
VER (verb)	8506	14.35
TOTAL	59285	100

Table 6. Description and counting for delivery symbols

2.3.5 Pitch

First, pitch was automatically detected by Praat. After the examination of distribution of pitch for each speaker, the floor and ceiling of pitch range were set at 50-500Hz for men and 60/70-550Hz for women. Then, for the entirety of data the pitch detection errors were corrected manually within the Praat. These data are also available with the corpus.All these automatic and manual preliminary steps are necessary for any acoustical analyses and further results.

3. Acoustic analysis and prosodic report

The goal of this preliminary study on the whole corpus is to describe the global prosodic characteristics for each phonogenre. Two tools were used to produce 128 acoustics measures. The first one is called ProsoReport (Goldman et al. 2007a) and provides a detailed prosodic report on two prosodic domains such as tones and rhythm. Global prosodic variables are computed for various sizes of speech units like phones, syllables, pauses as well as PSUs (pause-separated units) and finally the whole recordings (articulation rate, duration mean and deviation of phones, syllables, PSUs, pitch distribution, among others). Moreover, the automatic prominence detection tool gives even more complementary measurements (tonal and rhythmic spread of prominent and non-prominent syllables). Among the 64 acoustic measures, only 51 are relevant for our study as they represent rationalized measures (mean, rate, percentage) and not raw measures (total duration, number of syllables, etc.). Groups of recordings can be compared, while the size and number of recordings as well as the speaker individual properties can be ignored.

The second tool, described in Dellwo (2010), computes exclusively rhythmical variability and temporal measures, on the basis of vocalic, consonantal and syllabic intervals. For these, only 51 measures are taken in account out of 58 provided by the tool.

In this part of study we did not include NAR because it is still in annotation procedure. For that reason we associated a part of an external corpus (Avanzi et al 2012) composed of 20 recordings of neutral reading [LEC] and 20 recording of spontaneous speech during an informal conversation [CNV]. These 2 parts are described in duration and situational features as In Table 7: Thus, together with these two supplementary phonogenres, the data sum up in a table of 102 measures for 116 recordings, representing 8.5 hours of speech.

Genre	#rec	Dur	Audience	Media	Prep	Interac.
CNV	20	70	1	0	0	2
LEC	20	56	0	0	2	0

Table 7. Number and duration of two additional phonogenres with situational features

A Principal Components Analysis was used to model difference between phonogenres and situational features. The first two principal components (CP) explain 58% of

the variation, while the first eight explain 90.5%. A discriminating analysis for an automatic classification with 8 PCs over 8 phonogenres showed that 95% of recordings were identified correctly.

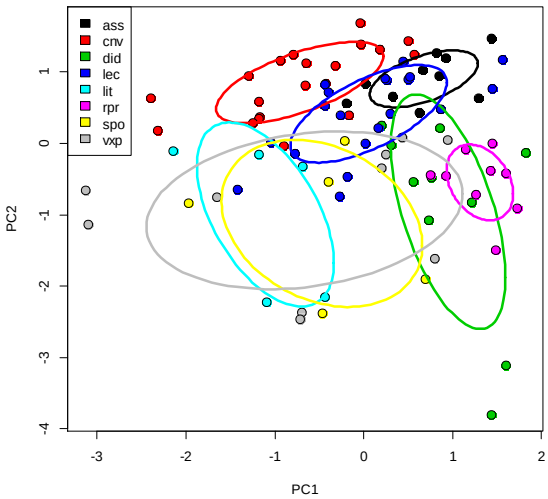


Figure 3 Distribution of 116 recordings in the first two Principal Components for 8 phonogenres

The graphical distribution of dots (Figure 3) shows the projection of the selection of 116 recordings onto first two Principal Components. This reduction of the initial data consists in eliminating the recordings of the same speaker since the aim was to analyze the phonogenre scatter and not the one of the idiostyle. It can be observed that ASS and RPR are the most compact phonogenres. The dispersion of CNV and LEC is slightly larger. DID and LIT are much less compact: this is probably because of the differences in speech situation explained above. Even bigger dispersion of data can be observed for SPO, not because of speech situation, but because of important differences in the nature of three sports: basketball, rugby and football. Each one has its own kinetic dynamics that is audible in the prosody of sport commentators (Audrit et al. 2012). Finally, VXP presents more than one particularity: 1/ the grouping of French presidents into the three chronological periods – 1970's, 1980-1990's and 2000's; 2/ the net separation of the discourse of Swiss and French presidents that implies the impact of geographical dimension.

4. Discussion

We presented here a large corpus grouping a variety of 7 phonogenres, or speaking styles. Each of these has at least 10 speakers, so that acoustic studies focus on the speaking style itself and get rid of individual characteristics. Further analyses are scheduled in order to model the differences between phonogenres as well as situational features. Some other aspects of phonostylistic variation are tackled such as spreading (i.e. punctual or non-continuous manifestation of phonogenre characteristics) that suggests to focus on dynamic acoustic measures, i.e. breath group-sized or syllable-sized rather than recording-sized.

5. Acknowledgements

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

6. References

Avanzi, M., A.C. Simon, J.-P. Goldman & A. Auchlin (2010). C-PROM. Un corpus de français parlé annoté pour l'étude des proéminences, *Actes des 23èmes journées d'étude sur la parole*, Belgique, 2010.

Avanzi, M., S. Schwab, P. Dubosson & J.-P. Goldman (2012). La prosodie de quelques variétés de français parlées en Suisse romande. Simon, A.C. (éd.). *La variation prosodique régionale en français*. Bruxelles, De Boeck/Duculot, pp. 89-119.

Audrit, S., T. Pršir, A. Auchlin, & J.-P. Goldman (2012), Sport in the media: a contrasted study of three sport live media reports with semi-automatic Tools, *Speech Prosody* 2012.

Beacco, J.-C. (2004). Trois perspectives linguistiques sur la notion de genre discursif. *Langages* 38/153, pp. 109-119.

Boula de Mareüil, Ph. (2012). 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.

Christodoulides, G., M. Avanzi , J.-P. Goldman (2014) DisMo: A Morphosyntactic, Disfluency and Multi-Word Unit Annotator : An Evaluation on a Corpus of French Spontaneous and Read Speech - LREC Proceedings

Dellwo, V. (2010). 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.

Goldman, J.-P., A.C. Simon, A. Auchlin & M. Avanzi (2007a). Phonostylographe, un outil de description des phonostyles prosodiques. *NCLF* 28, pp. 219-237.

Goldman, J.-P., M. Avanzi, A. Lacheret-Dujour, A. C. Simon & A. Auchlin (2007b). "A Methodology for the Automatic Detection of Perceived Prominent Syllables in Spoken French". Proceedings of Interspeech'2007 Antwerp, Belgium, pp. 91-120

Goldman, J.-P., A. Auchlin & A.C. Simon (2011a). Discrimination de styles de parole par analyse prosodique semi-automatique. Yoo, H.-Y. & E. Delais-Roussarie (eds) *Proc. of d'IDP-2009*. Paris.

Goldman, J.-P. (2011b) EasyAlign: an automatic phonetic alignment tool under Praat. *InterSpeech* September 2011, Florence, Italy.

Koch, P. & W. Oesterreicher (2001). Langage parlé et langage écrit. Holtus G., M. Metzeltin & Ch. Schmitt (eds), *Lexikon der Romanistischen Linguistik*, I/2. Niemeyer, Tübingen, pp. 584-627.

Lucci, V. (1983). Étude phonétique du français contemporain à travers la variation situationnelle. Université des langues et lettres, Grenoble.

Obin, N., A. Lacheret-Dujour, C. Veaux, X. Rodet & A.C. Simon (2008). A Method for Automatic and Dynamic Estimation of Discourse Genre Typology with Prosodic Features. *InterSpeech Proceedings*, pp. 1204-1207.

Solin, A. (2011). Genre. Zienkowski, J., J.-A. Ostman & J. Verschueren (eds), *Discursive Pragmatics*. John Benjamins, Amsterdam, pp. 119-134.

Simon, A.C., A. Auchlin, M. Avanzi & J.-P. Goldman (2010). Les phonostyles: une description prosodique des styles de parole en français. Abecassis, M. & G. Ledegen (eds), *Les voix des Français. En parlant, en écrivant*. Peter Lang, Berne, pp. 71-88.