

# A Method for Automatic and Dynamic Estimation of Discourse Genre Typology with Prosodic Features

Nicolas Obin<sup>1</sup>, Anne Lacheret-Dujour<sup>2</sup>, Christophe Veaux<sup>1</sup>, Xavier Rodet<sup>1</sup>, Anne-Catherine Simon<sup>3</sup>

<sup>1</sup>IRCAM, Analysis-Synthesis Team, Paris, France

<sup>2</sup>Modyco Lab., University of Nanterres, & Institut universitaire de France, Paris, France

<sup>3</sup>VALIBEL, Université catholique de Louvain, Belgique

nobin@ircam.fr, anne@lacheret.com, rodet@ircam.fr,

veaux@ircam.fr, anne-catherine.simon@uclouvain.be

## Abstract

This paper presents a work-in-progress on the automatic analysis of discourse genre in non-elicited speech. The study is focused on the development of *bottom-up* methods for automatic validation of discourse typologies found in linguistic descriptions (prosodic, syntactic, pragmatic and/or contextual and situational cues). The linguistic classification examined here opposes five discourse genres +/- controlled. To test this *a priori* classification under prosodic criteria, we propose a method that provides an automatic and dynamic estimation of discourse genre typology i.e. of prosodic similarities between discourse genres. This is achieved in a two-step procedure : a set of *discriminant prosodic patterns* is estimated and then used to raise a *typology* of discourse genres based on prosodic similarity criterion. The discriminant analysis reveals that a small number of prosodic patterns is sufficient to discriminate the 5 discourse genres. The typological analysis reveals some multi-level categorical oppositions on a continuous prosodic scale that can be interpreted in terms of +/- controlled speech.

**Index Terms:** discourse genre, prosody, typology, discriminant analysis, agglomerative clustering

## 1. Introduction

The concept of discourse genre (DG) is both trivial and complex. Being a basic notion in the linguistic field (each discourse is characterized by its own linguistic marks and then its own genre [1] [2]), it is also extremely complex to automatically estimate the prototypical linguistic features that distinguish among DG's and therefore validate the classifications proposed by linguists. The classification criteria are plural: sociological (sex, age, geographic area, profession, etc.) and linguistic (phonetical, syntactical, semantic): the present study focuses on the prosodic dimension. The main question addressed in this paper is the following: is it possible to distinguish among stated DG's (section 2) according to some precise prosodic cues? The proposed approach is based on statistical analysis of the observed prosodic features (section 3) among a given set of DG's. In a preliminary step (section 4) the discriminant ability of prosodic features is investigated. Then we propose a two-step statistical framework that estimate DG's differences on the prosodic dimension: a multiple analysis of variance is used to estimate a set of discriminant prosodic patterns among DG's (section 5); then an agglomerative clustering method (section 6) raises a DG's typology (i.e a hierarchical DG's structure) according to their prosodic similarity.

These questions are linked to a dual *top-down* and *bottom-up* process. Our automatic DG's discrimination and typology process lead to the creation of a tool that can be used to test and validate linguistic hypotheses on DGs. This approach can be seen as a control process for linguists to clarify their DG hypotheses and to create a methodology for a corpus design task in spontaneous speech ([3],[4]). Secondly, the use of linguistic knowledge is a great help when developing models of generation of variable prosodic discourse styles for speech synthesis and is useful for understanding results from automatic recognition of different discourse types.

## 2. Linguistic material

The corpus for this exploratory study - 52 minutes of speech, 3 investigation points (France (F), Belgium (B), and Switzerland (S)), 18 speakers (10 male, 8 female) - is divided in 5 DGs +/- controlled (see description in table 1) ([5],[6],[7],[8]). Each corpora has been segmented and transcribed in graphemic words and phonemic syllables using *PRAAT* and *EasyAlign*. Then, two expert annotators ([5],[7]) carried out a manual coding of prominence as well as dysfluencies which are characteristic of spontaneous speech: hesitations, filled pauses and false starts, final lengthening, and post-tonic schwas. Finally, the manual annotation was automatically validated ([9]).

	+/- controlled					Total
	Political discourse (PD)	Radio news (RN)	Radio interview (RI)	Map task (MT)	Life story (LS)	
Total dur. (min.)	10.30	11.30	10	9.30	10.30	52
Mean dur. (min.)	3	3	3.30	1	3	13.30
Syllables	2174	3164	2594	2251	2668	12851
Geo. area	F/B/S	F/B/S	F/B	F	F/B/S	F/B/S
Speakers	3	3	2	7	3	18
Sex (M/F)	2/1	2/1	1/1	5/2	0/3	10/8

Table 1: description of the studied corpus

## 3. Prosodic features

Prosodic analysis still remains a complex task. It raises various preliminary issues that has to be answered.

- The choice of the *observation unit* (syllable, prosodic group, minimal discursive unit, whole discourse),

- The size of the *analysis window* (one unit, several observation units or the whole discourse),
- The choice of the *prosodic features* considered as being relevant for the study (acoustic features, characteristic values on these features and temporal horizon for relativization of these values).

In the present study, we have chosen the prosodic group (PG, defined as the speech segment between two consecutive pauses and that end with a prominence) as the processing unit for two main reasons. First, for practical reasons: the corpus size (18 productions) does not allow robust statistical analyses if we take each production as a processing unit. Second, for theoretical reasons: we support the hypothesis that discourse types can be distinguished on the base of shorter units, such as PG. All analyses were carried out using a 10 PG analysis window centered on the targeted PG, in order to remove noise due to local inter-PG variations.

The prosodic description of the corpus centers around two levels of representation: i) the *phonological* information i.e. speech prosodic structure and ii) its *acoustic* correlates and their evolution in time. Additionally, the proposed description endeavors to report the evolution of the prosodic parameters over two distinct discourse temporal levels. On one hand an *intra-PG* description, i.e. characteristics and variations of the prosodic parameters within the PG; (see infra P1.1, P1.2, P2.1, A1.1, A2.1, A.3); on the other hand, an *inter-PG* description, i.e. variations of the prosodic parameters over the discourse (see infra A1.3, A2.2). The first level aims at measuring the *regularity* of the prosodic parameters and the *accentual contrasts* within the PG; the second aims at measuring the *regularity* of the prosodic parameters and the *discursive contrasts* within the discourse. Based on these remarks, the prosodic features used for the present study are organized as follows:

### 3.1. Phonological features

- PG length: Syllables number within the PG (P1)
- Prominence density: number of prominences within the PG, regarding total number of syllables of the PG (P2),
- Dysfluencies density - hesitations (P3.1), filled pauses and false starts (P3.2), final lengthening (P3.3), post-tonic schwas (P3.4),

### 3.2. Acoustic features

The set of acoustical features was chosen in order to describe variations in fundamental frequencies and local speech rate. Energy based features were rejected since this parameter shows large variations in recording conditions (level, noise-to-signal ratio, speaker distance).

- Local speech rate: mean local speech rate within the PG (A1.1); local speech rate variations (standard deviation) within the PG (A1.2); mean local speech rate within the PG to overall speech rate ratio (A1.3),
- Pitch: pitch variations (standard deviation) within the PG (A2.1); mean pitch within PG to overall mean pitch ratio (A2.2)
- Pauses and dysfluencies: Post PG pause duration (A3.1); Post PG pause duration to PG duration ratio (A3.2); Disfluencies duration within the PG (A3.3); Dysfluencies duration within the PG to PG duration ratio (A3.4). (see [10] for discussion on the choice of these coupled features)

Pitch range within the PG was also rejected, as it is more related to speaker identity than of discourse type. In addition, the corpus is disproportionate in terms of speaker gender (male vs. female) (see table 1); thus such a feature could introduce artificial differences between discourse types.

## 4. Effect of Discourse Type on Prosody

### 4.1. ANOVA framework

In order to test the effect of the discourse type on the considered prosodic parameters, we have conducted a *multi-class multiple one-way analysis of variance* (ANOVA) for each parameter individually. The analysis of variance aims to statistically compare the effect of independent categorical variables (discourse categories) on the mean of an ordinal variable (prosodic parameters) and to determine if there is at least one class, such as the mean value of the variable in this class differs significantly from the overall mean. This is achieved by means of the *null-hypothesis test* (i.e. difference between the respective means is null) which lie in the estimation of the probability of the observed difference between the classes conditionnaly to such an hypothesis (*p-value*). The separation criteria (*F-ratio*) is defined as the mean between-to-within class scatter ratio.

Let  $k$  be the class,  $N_k$  the number of observations for this class, and  $K$  the total number of classes. Let  $\mu_k$  be the feature mean for the class  $k$  and  $\mu$  the feature overall mean. The F-ratio is defined in the mono-dimensional case as follows:

$$F = \frac{\frac{SSB}{dfB}}{\frac{SSW}{dfW}} = \frac{\frac{1}{K-1} \sum_{k=1}^K (\mu_k - \mu)^2}{\frac{1}{N-K} \sum_{k=1}^K \sum_{n_k=1}^{N_k} (x_{n_k} - \mu_k)^2} \quad (1)$$

Then, the probability of observation of this separation under the *null-hypothesis* corresponds to the *p-value* such as the upper critical value of the *F-distribution* with (K-1, N-K) liberty degrees is equal to the observed *F-ratio*.

### 4.2. Results and discussion

The results of the analysis of variance are presented in terms of *F-ratio* and *p-value* in table 2 for each of the targeted prosodic parameters that were normalized according to their respective variance in a preprocessing step. All analyses were performed through *IrcamCorpusTool* [11] analysis framework.

P-Features	F-ratio	p-value	A-Features	F-ratio	p-value
P1	162.2	<10-6	A1.1	68.1	<10-6
P2	135.7	<10-6	A1.2	20.6	<10-6
P3.1	488.6	<10-6	A1.3	190.9	<10-6
P3.2	95.6	<10-6	A2.1	162.0	<10-6
P3.3	125.0	<10-6	A2.2	1.46	0.23
P3.4	231.9	<10-6	A3.1	238.7	<10-6
			A3.2	266.2	<10-6
			A3.3	48.8	<10-6
			A3.4	129.3	<10-6

Table 2: Results of anova analysis for each of the considered prosodic parameters.

The analysis of variance first shows that the discourse type has significant effect on the quasi-totality of each of the considered prosodic features, with the exception of the mean pitch variations within the PG, which is not relevant for discourse type discrimination. Such analysis reveals that discourse type can be distinguished both on the phonological level and on the acoustic space. A more detailed acoustic analysis sheds light on

the following points: i) Contrasts realized by the prosodic parameters (pitch and local speech rate) discriminate between discourse types on two different contrast levels: local speech rate shows better discrimination ability in terms of inter-PG variations than intra-PG variations; whereas for pitch, intra-PG variations are more discriminant than inter-PG variations. ii) While analyzing pause and dysfluencies durations: duration normalization previous PG duration is more discriminant in both cases (A3.2 vs. A3.1 and A3.4 vs. A3.3). This last point could be explained when recalling for example that long pause durations and PG characterize political discourses, while the radio news report is characterized, on the contrary, by short pause durations and long PG. Then coupled features can emphasize such phenomena.

## 5. Discriminant pattern matching of discourse type

As we have seen, discourse type has significant effects on prosodic features on both levels. However, when such an analysis allows the validation of the individual relevance of each of the considered prosodic features, it remains insufficient for two main reasons. First, it does not take into account the correlations between prosodic parameters, then implies redundant information. Second, it fulfills only partially the linguistic aim to characterize the prosodic properties which lead to the discrimination of each discourse. In other words, we would like to determine the set of prosodic patterns which best discriminate discourse types, and thus propose an easily interpretable representation of these differences by reducing the dimensionality of the initial prosodic space. These two problems can be tackled within the framework of a *multiple analysis of variance* (multi-class one-way MANOVA), i.e. by testing the hypothesis of difference on prosodic structures instead of on isolated cues.

### 5.1. MANOVA framework

The multiple analyses of variance are used for assessing group differences across multiple metric dependent variables simultaneously, based on a set of independent categorical variables [12]. MANOVA analysis is based primarily on the same principles as ANOVA but formulates the difference between classes in his matrixial form. The solution of this problem is achieved on the basis of the *canonical variate analysis*, i.e. the determination of a set of  $K - 1$  orthogonal linear combinations (*canonical variables*) of the  $D$  original dependent variables, which enables the optimal class separation by maximizing the *F-ratio*. This determination comes down mathematically to estimate eigenvectors and associated eigenvalues on the  $F$  matrix. Eigenvectors represent the discriminant direction and the eigenvalues are the  $F$ -ratios associated with these directions. The optimal dimension of the canonical space is then determined by reducing this dimensionality to the sub-space whose dimensions allow to reject the *null-hypothesis* to a desired *p-value* (here, 0.05). This has the effect of maintaining the dimensions that enable significant discrimination between classes, and to reject the noisy directions. Furthermore, the eigenvectors form an orthogonal basis and the resulting dimensions are uncorrelated, thus assuring non-redundancy.

Placing these properties back in a qualitative interpretation context:

i) Combining the prosodic parameters enables the understanding of the discourse type separation by means of discriminant patterns.

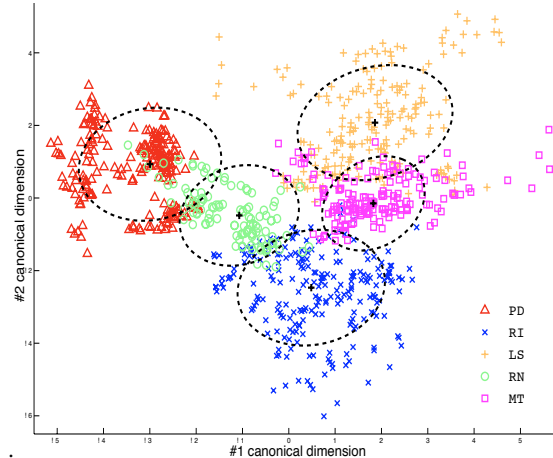


Figure 1: Scatter plot of discourse genre in the resulting canonical space (first two dimensions)

ii) Basis orthogonality of the canonical space enables the interpretation of the results by means of independent prosodic patterns.

iii) Dimensionality reduction facilitates a direct interpretation.

iv) The dimension of the resulting space gives an idea of the differential configuration of discourse types: a unique dimension indicates that a single pattern enables to discriminate all the classes, whereas the increase of the dimension suggests the possibility of a more complex configuration (i.e. hierarchical)

## 5.2. Results and discussion

The MANOVA analysis has been achieved with the same methodology than presented in section 4.2. The optimal dimension has been determined to 4 with the properties associated in the *F-ratios* decreasing order of  $F(4;496) = (995.6;664.9;451.8;104.67)$  with *p-value*  $< 10^{-6}$ . Such a dimensionality suggests a complex structure of prosodic discriminant patterns in discourse types. The estimated prosodic patterns outperform discriminant abilities compared to each feature considered separately. Figure 1 shows the data scatter according to the discourse type in the first two dimensions of the canonical space.

An overview of the observations scatter in the first two dimensions of the canonical space shows that discourse type is divided in well-separated zones within a *prosodic continuum* (partial overlapping with successive contiguity). A more detailed analysis leads to the following observations: the main discriminant direction (first canonical space dimension) divides discourse type distinction in two groups (*political discourse; radio news vs. radio interview; map task; life story*), whereas the second direction points out individual separations within each of these groups (*political discourse vs. radio news and radio interview vs. map task vs. life story*). This observation suggests that discourse types are hierarchically structured. Furthermore, map task and life story present the highest scatter in the canonical space, indicating a higher prosodic variability associated with free speech, as opposed to a lower scatter associated with controlled speech. Finally, political discourse reveals three relatively well separated distributions, suggesting that inter-individual variations (three speakers) remain significant and/or can be associated with distinct prosodic strategies.

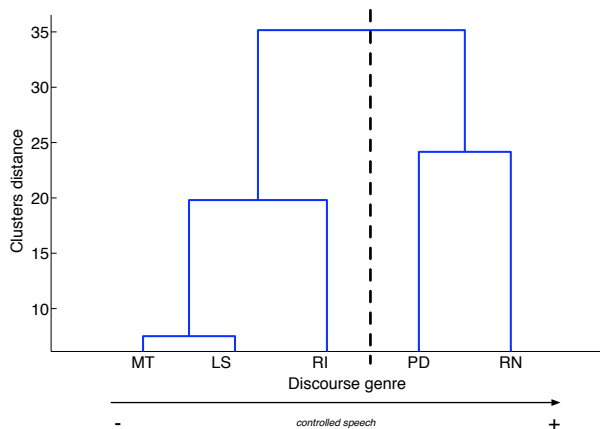


Figure 2: Resulting hierarchical typology of the studied DG's.

## 6. Prosodic Similarities of Discourse Type with Agglomerative Clustering

In order to specify the discriminant properties of discourse type and set up a discourse type typology according to prosody, we achieved an agglomerative clustering of discourse type within the canonical space.

### 6.1. Agglomerative Clustering Principles

The agglomerative clustering consists in producing a hierarchical clustering of the observed data by iteratively grouping them according to their similarities. This method requires (1) the definition of a norm which makes it possible to estimate the distance between each pair of observations within the considered space, and (2) an agglomerative method which enables to group these observations. Based on these two criteria, the method agglomerates iteratively the observation pairs that minimize the measured distance following the chosen criterion, then merges these observations into a single parent observation (node) and associates the distance with that node. The iteration is carried out until a unique observation is obtained.

### 6.2. Results and discussion

In the present study, the agglomerative clustering has been achieved on the basis of mahalanobis distance between discourse type centroid. The dendrogram of the figure 2 shows the resulting hierarchical structure. The analysis of the figure reveals a dominant binary discourse type structure (first node with maximal separation between two discourse types) with, on one hand, the controlled speech (political discourse, and radio news), and, on the other, spontaneous speech (radio interview, map task, and life story). In more details, this structure appears to be consistent as it indicates a three-group continuum from controlled to spontaneous speech: G1 (political discourse, radio news), G2 (radio interview), and G3 (map task and life story). This fact reveals the double discursive status of radio interview, both controlled in its socio-linguistic representation and free in its formal linguistic markers.

## 7. Conclusion

We have presented the first results of a *bottom-up* analysis of discourse genre. The statistical framework chosen is based on a discriminant prosodic patterns analysis in order to set up a methodology that enables to validate a given discourse genre ty-

pology and, on this base, to develop dynamic classifications. In this study, 5 types of speech, identified by linguists as 5 different discourse genres with the hypothesis that they have formal linguistic correlates, have been studied regarding to the prosodic level. In practice, we had to know how each genre is linked to specific prosodic strategies in terms of phonological structure on the one hand, of acoustic features on the other. Further investigations are necessary, based on larger corpora (more discourse genres, more speakers) and on automatic tools for accent and dysfluencies tagging. Then, the proposed method will be available both in an automatic detection task and in a modeling task. Other point: manova analysis has two major drawbacks: 1) it does not lead to a clear quantitative distinction of prosodic classes (we do not know which class is separated from which in the resulting dimension), 2) we still do not know how to interpret the discriminant prosodic patterns. In other words, we know that differences exist between our DGs, but we cannot produce a precise expertise of the function of each prosodic feature. A *post-hoc* analysis will help us to resolve these two major questions. After this first step, such a methodology will be applied to syntactical marks and then it will be possible to cross the two levels of representation cues in order to propose a linguistic characterization of different linguistic genres in French spontaneous speech.

## 8. Acknowledgements

This study was supported by:

- ANR Rhapsodie 07 Corp-030-01; reference prosody corpus of spoken French; French National Agency of research (ANR); 2008-2012.,
- Programmes Exploratoires Pluridisciplinaires (PEPS), CNRS/ST2I, 2008-2010.

Thank you to A.-C. Simon, M. Avanzi and J.-P. Goldman for the development of the corpus and the preprocessing of the data (transcription and manual annotation of prominences and dysfluencies).

## 9. References

- [1] M. Bakhtine, *Esthétique de la création verbale*. Paris: Gallimard, 1984.
- [2] D. Biber, *Variation Across Speech and Writing*. Cambridge University Press, 1988.
- [3] K. Kohler, "Prosodic phrasing in german spontaneous speech - categories, symbolization, empirical validation," Tech. Rep., 1997-2006.
- [4] L. Boves and N. Oostdijk, "Spontaneous speech in the spoken dutch corpus," in *Proceedings ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR)*, 2003.
- [5] A.-C. Simon, *La structuration prosodique du discours en français. Une approche multidimensionnelle et expérimentielle*. Berne: Peter Lang, 2004.
- [6] M. Avanzi, "L'indication d'itinéraire en français parlé. schématisation cognitive et organisation macro-syntaxique," Master Thesis, Grenoble 3, 2004.
- [7] M. Avanzi, J.-P. Goldman, A. Lacheret-Dujour, A.-C. Simon, and A. Auchlin, "Méthodologie et algorithmes pour la détection automatique des syllabes proéminentes dans les corpus de français parlé," *Cahiers of French Language Studies*, vol. 13, no. 2, pp. 2-30, 2007.
- [8] A.-C. Simon, A. M., and J.-P. Goldman, "La détection des proéminences syllabiques. un aller-retour entre l'annotation manuelle et le traitement automatique," in *Congrès Mondial de Linguistique Française*, Paris, France, 2008.
- [9] N. Obin, J.-P. Goldman, M. Avanzi, and A. Lacheret-Dujour, "Comparaison de trois outils de détection automatique de proéminence en français parlé," in *JEP'08*, Avignon, France, 2008.
- [10] J.-P. Goldman, A.-C. Simon, M. Avanzi, and A. Auchlin, "Phonostylographe, un outil de description des phonostyles prosodiques. chroniques radiophoniques et style lu," *Cahiers de linguistique française*, vol. 28, pp. 219-237, 2007.
- [11] C. Veaux, G. Beller, and X. Rodet, "Ircamcorpustools : an extensible platform for speech corpora exploitation," in *ELREC'08*, Marrakech, Maroc, 2008.
- [12] J. Hair, R. Anderson, M. Tatham, and W. Black, *Multivariate data analysis*. Prentice-Hall, 1995.