Expectations for Discourse Genre Identification: a Prosodic Study

Nicolas Obin, Volker Dellwo, Anne Lacheret, Xavier Rodet

Expectations for Discourse Genre Identification: a Prosodic Study

Nicolas Obin [1,3]*, Volker Dellwo* [2]*, Anne Lacheret* [3]*, Xavier Rodet* [1]

[1] Analysis-Synthesis Team, IRCAM, Paris, France
[2] University College of London, London, UK.
[3] Modyco Lab., University of Paris Ouest - La Défense, Nanterre, France
nobin@ircam.fr, v.dellwo@ucl.ac.uk, anne.lacheret@u-paris10.fr, rodet@ircam.fr

Abstract

Speech can be divided into discourse genres based on the contextual environment it occurs in (e.g. political speech, sport commentary speech, etc.). The present study investigated whether listeners can distinguish between speech from different discourse genres on the basis of acoustic prosodic cues only [1]. In a perception experiment with delexicalized speech 70 listeners with varying experience in French (native speakers, non-native speakers, and non-speakers) were asked to identify four different types of discourse genres (church service, political, journal, and sport commentary). Results revealed a fair identification ability with a significant increase in performance with increasing experience in French. Identification confusion was used to cluster discourse genres according to their perceptual similarity.

Index Terms : discourse genre, speaking style, prosody, perception, speech synthesis.

1. Introduction

The concept of discourse genre (DG) has been studied widely in rhetoric and literature and more recently extended to the oral domain ([1, 2]).

It is generally hypothesized ([3, 4])that each situation and each given social context correspond with a specific mode of production which is associated to specific formal markers that bear the traces at all levels (semantic, syntactic, phonological) of the DG. Following these studies, research in textual typology aims to : 1) describe the diversity of discourses (literary , legal, political, religious, etc.) ; 2) understand their classification into genres ([5]) ; 3) estimate their formal markers, in particular the co-occurrence of specific cues that can be considered as being typical of a genre. The challenge is to provide a robust and shared DG's typology. However, it remains very difficult to go further than conventional generic types (private, professionnal, public speech) and subdivisions (face to face conversation, phone conversation, public debates, radio and TV broadcasts, unprepared vs. planned speech, etc. [2]). In the absence of a comprehensive representation of discourse genres and classes, each domain defines specific classification criteria which best account for its purpose (social, language activity, and formal spaces [6, 7]).

In strictly phonetic terms, studies focus on the description of phonostyles ([8, 9]). In particular, public discourse (such as political, religious, journalistic and sport), considered as cultural stereotypes, are related to expressive strategies that act as markers of a phonostyle ([10]).

Methods have been proposed recently to relate achievements in phonostylistic ([11]) to text segmentation and classification in semantic in order to provide a unified interpretation framework : thematic content and informational sequence, semantic-syntactic structures and prosodic patterns (for recent work on French, see [12, 13]).

This study aims at : 1) assessing whether the acoustic prosodic patterns of DGs are perceptually salient ; 2) estimating a classification of DG's on the basis of perceptual proximity ; 3) providing a reference for the evaluation of a speaking style speech synthesis system. Methodologically, a perceptual experiment was carried out in which listeners of different native language backgrounds had to identify different French DGs based on prosodic patterns.

The paper is organized as follows : section 2 presents the design of speaking style corpus from which the stimuli for the present experiment were derived ; section 3 presents the experimental setup ; finally, results are presented and discussed in sections 4 and 5 respectively.

2. Speaking Style Corpus Design

2.1. Corpus Design

For the purpose of speaking style speech synthesis, a 4 hours French multi-media corpus was designed from which the stimuli for the present perceptual experiment were selected. The corpus consists of four different DG's : catholic mass ceremony, political, journalistic, and sport commentary. In order to reduce the DG intra-variability, the different DGs were restricted to specific discourse contexts (see list below) and to male speakers only.

The following is a description of the four selected DG's :

- **mass** : christian church sermon (pilgrimage and sunday high-mass sermons) ; single speaker monologue, no interaction.
- **political** : New Year's speech ; single speaker monologue ; no interaction.
- **journal** : radio review (press review ; political, economical, technological chronicles) ; almost single speaker monologue with a few interactions with a lead journalist.
- **sport commentary** : soccer ; two speakers engaged in monologues with speech overlapping during intense soccer sequences and speech turn changes ; almost no interactions.

Speech samples were collected from *real speech* multi-media contents. Recordings date from the 2000's with the exception

[1]This study was supported by ANR Rhapsodie 07 Corp-030-01 ; reference prosody corpus of spoken French ; French National Agency of research ; 2008-2012.

of the political discourse which homogeneously ranges from 1975 to 2007. Speech samples are compressed audio (mp3 format at various and unknown bit rates) with strongly variable audio quality (background noise : crowd, audience, recording noise and reverberation). The sample selection was especially designed to provide well-balanced DG's corpora (total DG's duration ; mean duration per speaker)[2]. Corpus properties are summarized in table 1.

speaking style	media	# speaker	speaker gender	mean duration / sample	mean duration / speaker	total duration
mass	none	7	7M	12mn	11mn	**1h20**
political	TV	5	5M	12mn	14mn	**1h10**
journal	radio	5	5M	4mn	14mn	**1h10**
sport	radio	4	4M	20mn	9mn	**35mn**

TAB. 1 – *Description of the speaking style speech corpus.*

The corpus was enriched in particular with the following linguisitic annotations : orthographical transcription ; phonemic alignement and breath detection using *ircamAlign* [14], then manually corrected ; syllabification on inter-pausal groups.

3. Experimental Design

The experiment consists in a multiple choice DG's identification task based on prosodic cues. It was not possible to control the linguistic content of the speech utterances which is an evident cue for DG's identification (a single keyword may sometimes be sufficient to identify a DG). For this reason the stimuli were delexicalized using low pass-filtering. Stimuli were utterances that were extracted from the speaking style speech corpus (see section 2), filtered to remove semantic access, and then presented as a multiple choice identification experiment to multi-language speakers in a crowd-sourcing framework.

3.1. Subjects

70 subjects participated in this experiment : 37 native French speakers, 20 French speaker, 15 non-French speakers ; 46 expert subjects, 24 naïve subjects. *Expert* subjects had a variety of backgrounds : speech and audio technologies, linguistic, musicians. 7 subjects were removed because they produced mistakes in using the experimental interface. In the case of multiple participation of a subject, his first participation was used for analysis only. Subjects were aged from 20 to 65 years, with a strong proportion (65%) within the 20-35 year range.

3.2. Stimuli

40 speech utterances (10 per DG) were selected from the speaking style corpus. Such selection was conducted to provide various and representative prosodic patterns for each DG.
Firstly, segmentation into speech utterances was accomplished according to the *prosodic period*. Prosodic period is defined as being a sequence of inter-pausal groups which ends with a conclusive frontier (combination of three acoustic criteria : low pitch accent, long syllable duration, followed by a long pause). It is in particular supposed to be related to discursive speech units (a prosodic object as a marker of a communicative and discursive object). However analysis of speech samples reveals that the definition of the *prosodic period* fails to account for some specific speech sequences or more generally to some DG. In particular, it was regularly observed that journalistic utterances do not end with a major frontier (low pitch accent which is not followed by a pause, or even high-pitch accent which is followed or not by a pause, or inconsiderable long prosodic periods). This was especially observed for spontaneous speech such as the sport commentary where prosodic periods could not be directly associated to discursive units. Consequently, segmentation into speech utterances was finally decided by an expert linguist according to discursive-prosodic cues in such conflictual cases. In particular, speech segmentation was chosen as being prosodic objects which were not necessarily formally defined by prosodic constraints only but combining prosodic cues to weak discursive, syntactic and semantic dependencies to the discursive context.
Secondly, the selection criteria were derived from a classification of speech utterances into discursive sequences as well as prosodic structure and complexity. In particular, *archetypal speech utterances* were selected depending on the DG (mass : *"au nom du père et du fils, et du Saint-Esprit, ainsi soit-il"*, *in the name of the Father, the Son, and the Holy Spirit, Amen.* ; political : *"mes chers compatriotes, vive la République et vive la France"*, *my fellow countrymen, long live the republic ! Long live France !* ; sport commentary : *"oh le but de Babel ! le but de Babel ! le but de Babel !"*, *What a goal by Babel ! Goal by Babel ! Goal by Babel !* ; journal : no specific speech utterance was observed). Then, speech utterances were classified into *discursive sequences* depending on the DG. For instance, journalistic chronicles can be formally described as a sequence of topic sequences with punctual interaction with a lead speaker during topic changes. Speech utterances were thus classified into global introduction from a lead speaker (*"l'Eco du jour" : c'est l'actualité économique de ce lundi 26 octobre 2009 vue par Philippe Lefébure*, *"Eco du Jour" is the economic news for today, Monday, October 26, 2009 hosted by Philippe Lefébure*), and initial, medium, terminal and transitional sequences for each topic (initial : *"MacDo va quitter l'Islande : conséquence directe de la crise, raconté dans L 'Humanité"* , *MacDonald's is quiting Iceland : a direct result of the financial crisis, story in "l'Humanité".* ; conclusive : *"les français disent qu'on va dans la mauvaise direction, une seule réponse du côté de l'UMP : il faut y aller plus vite"*, *The French say we have taken the wrong path, the only reply from the UMP is : "We must go faster".*). Sport commentary sequences were classified according to actional, situational (current action, past action, off-line comments), and emotional (more or less intense) criteria. Other DG's speech utterances were classified in the same manner. Such classification interestingly relates to specific prosodic patterns. Then various *prosodic sequences* were chosen for each DG (in particular : low, medium, and high terminal pitch accent, intermediate lexical pitch accents as well as hesitations).
Thirdly, as it was observed that speech utterance's duration strongly depend on the DG, speech utterances were classified into short (4 ± 0.5s.) and long ($10 \pm 1s$.) utterances that were homogenously distributed for each DG.
Finally, 2 speech utterances were selected for each speaker to remove any identification based on speaker recognition.

Then, speech samples were processed as follows : a) background noise and reverberation were minimized using a noise cancelation algorithm ([15]) ; b) semantic access was removed using a band-pass filter. A pass-band was chosen that insured that the lowest frequency of the fundamental frequency and the highest frequency of its first harmonic was included ([15]). This was done to extract speech prosodic characteristics only

[2]This was reached with the exception of the sport commentary which has half duration than the other DG's

c) active speech mean level normalization at -20dBov [16] ; d) speech samples were compressed in mp3 format at 192Kb/s.

3.3. Procedure

The experiment consists of a multiple choice identification task from speech prosody perception. It was conducted according to source-crowding technique using web social networks[3]. Subjects were given a brief description of the different speaking styles without being exposed to actual speech examples. This approach was adopted in order to focus subjects on their own mental representation of the different speaking styles and their expected prosodic cues.
They were asked to associate a speaking style to each of the speech samples[4]. For this purpose, subjects were given three options :

- **total confidence** : select only one speaking style when certain of the choice ;
- **confusion** : select two different speaking styles when two speaking styles are possible ;
- **total indecision** : select "indecision" when completely unsure. Subjects were asked to use this possibility only as a very last resort.

Additional informations was gleaned from the participants : speech expertise (expert, naïve), language (native French speaking, French speaking, non-French speaking), age, and listening condition (headphones or not). Subjects were encouraged to use headphones.

4. Results

Identification performance was estimated using a newly developed measure based on Cohen's Kappa statistic. Cohen's Kappa provides agreement between two raters in the case of caterogical rating [17]. Our measure monitors the agreement between subjects' ratings and the correct answer. The resulting Kappa values are considered as a measure of identification performance. The measure varies from -1 to 1 : -1 is perfect disagreement ; 0 is chance ; 1 is perfect agreement.

Overall score reveals fair identification performance (K=0.45). Figure 1 shows that there are differences according to native language background of the listeners. French natives perform better than non-native French-speakers ; non-French speakers perform only slightly.
ANOVA analysis (*one-way analysis of variance* [18]) was conducted to assess whether identification performance depends on the language of the subjects.
Analysis reveals a significant effect of the language ($F(2,59) = 15; p < 0.001$). Post-hoc analysis reveals significant difference between native French speakers and the others ($F(1,52) = 13; p < 0.001$, $F(1,43) = 24; p < 0.001$) but no effect between non-native French speakers and non-French speakers ($F(1,23) = 3; p = 0.07$) (fig. 1).
Investigating the results in finer detail reveals that identification performance significantly depends on the DG. Substantial identification performance was observed for sport commentary (K=0.7) ; fair identification for the journalistic discourse (K=0.54) ; and only slight identification for mass discourse and political discourse (K = 0.38 and 0.34 respectively).

[3] *Ircam Analysis and Synthesis Perceptual Tests* on Facebook : `http://www.facebook.com/group.php?gid=150354679034&ref=ts`

[4] the experiment is available on : `http://recherche.ircam.fr/equipes/analyse-synthese/obin/pmwiki/pmwiki.php?n=Main.SSRecoProso`

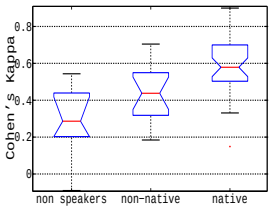

FIG. 1 – *Identification performance according to the subject's language background.*

Interestingly, identification performance reveals different configurations of the language effect depending on the DG (fig. 2).

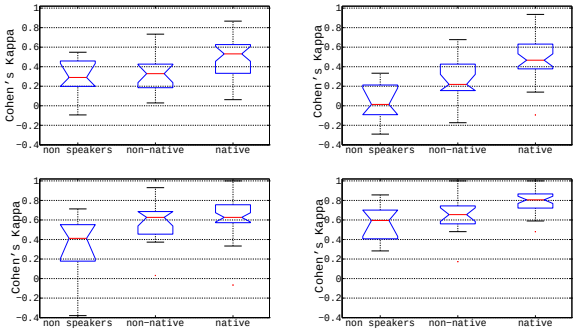

FIG. 2 – *Identification performance according to the subject's language background for each discourse genre. From top to down and left to right : mass, political, journalistic, sport.*

For the mass, a significant difference between native French speakers and the others was observed ($F(1,52) = 6.9; p = 0.01$, $F(1,43) = 5.8; p = 0.02$), but none among these ($F(1,23) = 0.19; p = 0.7$). For the political discourse, a significant difference between all language pairs was observed : native French vs. French speaking ($F(1,52) = 11; p = 0.001$), native French vs. non-French speaking ($F(1,43) = 30.5; p < 0.001$) as well as French speaking vs. non-French speaking ($F(1,23) = 7.8; p = 0.01$). For the journalistic discourse, no significant difference between native French and French speaking ($F(1,52) = 1; p = 0.3$) and a significant between these and non-French speaking ($F(1,43) = 13.5; p < 0.001$,$F(1,23) = 5.1; p = 0.03$) were observed. For the sport commentary, there was a significant difference between native French speakers and the other ($F(1,52) = 14; p < 0.001$,$F(1,43) = 17; p < 0.001$), but none between these ($F(1,23) = 0.8; p = 0.37$). Furthermore, it was noticed that non-French speaking participants provides slight and random identification performance in the case of journalistic and political discourse respectively.

Finally, *multi-dimensional scaling* (MDS, [19]) and *hierarchical clustering* ([20]) methods were used to represent and estimate DG's similarity according to the observed perceptual confusion. For each speech sample, the observed confusion matrix was used to define speech utterance coordinates. A similarity distance between speech utterances was then estimated according to the *city-block* metric. The set of pairwise speech utterances similarity was then used to represent speech utterances into a 2-dimensional space according to multi-dimensional scaling. Finally, DGs were clustered using the *complete linkage method*. In parallel, the 4 DG's have been discussed and rated by two expert linguists after the *Koch's conceptual scale* ([7]) with

3 degrees. Then DGs were clustered according to this conceptual description in the same manner. Figure 3 represents speech utterances into the resulting 2-dimensional space and the comparison of the resulting DG's typology.

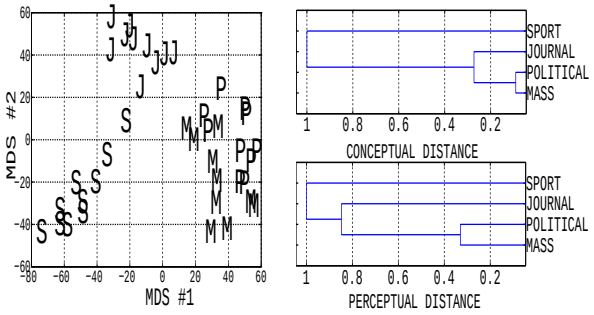

FIG. 3 – *(a) Representation of the speech utterances according to their similarity according to Multi Dimensional Scaling. (b) DG's clusters from : conceptual scale (top) and perceptual confusion (down)*

5. Discussion

The experiment reveals fair identification performance from prosodic perception. This confirms evidence for the hypothesis that DGs relate to prosodic patterns that do not depend on a specific speaker and that are shared by a group of listeners.

Not suprisingly, these prosodic objects dramatically depend on the language backgrounds. This shows that DG's abstract prosodic representation depends at least on the language. It could be hypothesized that such representation could more generally depend on language classes or even on culture [5].

More precisely, the experiment reveals that the language factor significantly depends on the DG. This suggests that DG relate to prosodic objects that could be more or less shared regardless of the language. In particular, prosodic cues related to sport commentary appeared clearly almost common across all languages while prosodic cues related to political discourse and mass discourse were language specific.

Interesting intermediate identification performance of non-native French speakers was observed : if identification performance is systematically situated between native French speakers and non-French speakers, different grouping were observed, depending on the DG (grouping with non-French speakers for mass discourse and sport commentary, grouping with native French speakers for the journalistic discourse, and no grouping for the political discourse). This suggests a native language effect as well as a cultural background dependency.

A comparison of DG clusters as estimated from the conceptual classification and from prosodic perception revealed a similar cluster structure (fig. 3). This confirms that discourse context (situational, spatio-temporal, ... context) consistently relate to prosodic strategies. Moreover, prosodic clusters precise the perceptual distance that in particular clearly distinguishes journalistic discourse from political discourse and mass discourse on the prosodic dimension. This result supports the hypothesis that prosodic strategies act as markers of a specific *speech act* ([21]) (for instance : neutrally describing an event with distanciation for the journalistic discourse vs. arguing and persuading for the political discourse and mass discourse). Sport commentary stands significantly apart from the other DGs. This confirms previous studies on the very specific nature of the sport commentary ([22]), in particular in its iconical dimension : sportscasters do not only describe but vocally mimic the action being observed. This is even more true in the case of radio sport commentary, where sportcaster must supply the absence of the image media.

[5] such hypothesis is supported by non-French speaking participants which comment that they could not represent themselves "how sounds" a Christian sermon (religious dependency) nor political new year's speech (cultural dependency)

6. Conclusion

A perceptual experiment on the identification of discourse genres on a speech prosodic perception basis was proposed. Identification performance confirms evidence for the hypothesis that DGs relate to prosodic patterns that do not depend on a specific speaker and that are shared among listeners. When overall factorial analysis reveals a significant language effect, it is clearly dependent on the DG. A comparison between DG clusters obtained from a conceptual description and perceptual confusion indicates that discourse context consistently relates to specific prosodic strategies. DG perceptual clusters even precise and suggest other discursive effects to explain observed differences of prosodic configuration. In a further study, these results will be used as a reference identification performance to evaluate a speaking style speech synthesis system.

7. References

[1] M. Halliday, *Spoken and written Language*. Oxford University Press, 1985.

[2] D. Biber, *Variation Across Speech and Writing*. Cambridge University Press, 1988.

[3] E. Benvéniste, *Problème de linguistique générale*. Gallimard, 1966.

[4] M. Bakhtine, *Esthétique de la création verbale*. Gallimard, 1984.

[5] F. Rastier, *Sens et textualité*. Hachette, 1989.

[6] S. Branca-Rosoff, "Types, modes et genres : entre langue et discours," *Educational and Psychological Measurement*, vol. 87, pp. 5–24, 1999.

[7] P. Koch and W. Oesterreicher, "Langage parlé et langage écrit," vol. 2, pp. 584–627, 2001.

[8] N. Obin, A. Lacheret, C. Veaux, , X. Rodet, and A.-C. Simon, "A method for automatic and dynamic estimation of discourse genre typology with prosodic features," in *Interspeech*, 2008.

[9] A.-C. Simon, A. Auchlin, M. Avanzi, , and J.-P. Goldman, *Les voix des Français*. Peter Lang, 2000, ch. Les phonostyles : une description prosodique des styles de parole en français.

[10] A. Lacheret-Dujour and F. Beaugendre, *La prosodie du français*. CNRS, 1999.

[11] Y. Fonagy, *La vive voix*. Payot, 1983.

[12] A. Lacheret, B. Victorri, and M. Avanzi, "Schématisation discursive et schématisation intonative : question de genre ?" *To be published*, 2009.

[13] L. Degand and S. A.-C., "Mapping prosody and syntax as discourse strategies : how basic discourse units vary across genres," *Where Prosody Meets Pragmatics*, pp. 81–107, 2009.

[14] P. Lanchantin, A. Morris, X. Rodet, and C. Veaux, "Automatic phoneme segmentation with relaxed textual constraints," in *LREC*, Marrakech, Morroco, 2008.

[15] N. Bogaards and A. Roebel, "An interface for analysis-driven sound processing," in *AES*, New York, USA, 2005.

[16] P. Kabal, "Measuring speech activity," MMSP Lab Technical Report, Dept. Electrical & Computer Engineering, McGill University, Tech. Rep., 1999.

[17] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[18] J. Hair, R. Anderson, M. Tatham, and W. Black, *Multivariate data analysis*. Prentice-Hall, 1995.

[19] I. Borg and P. Groenen, *Modern Multidimensional Scaling : theory and applications*. Springer-Verlag, 2005.

[20] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2009, ch. Hierarchical clustering, pp. 520–528.

[21] J. Searle, *Speech Acts*. Cambridge University Press, 1969.

[22] J. Deulofeu, "Les commentaires sportifs constituent-t-ils un genre ? au sens linguistique du terme ?" *Colloque Questions de méthode dans la linguistique sur corpus* , 1998.