# About Speech Overlaps: Prosodic Cues Contribution in Predicting a Change of Speaker

Roxane Bertrand, Robert Espesser

# About Speech Overlaps

## Prosodic cues contribution in predicting a change of speaker

Bertrand Roxane & Espesser Robert

Laboratoire Parole et Langage, UMR 6057, CNRS
29 Avenue Robert Schuman, 13621 Aix-en-Provence, France
e-mail : roxane.bertrand@lpl.univ-aix.fr;robert.espesser@lpl.univ-aix.fr

ABSTRACT
This paper concerns « non-smooth transitions » in turn-taking exchanges, i.e. *overlaps speech* phases. We wish to characterize this phenomenon in six dialogs. In this first study, we try to determine a few relevant parameters in predicting a speaker change after the overlap phenomenon.

## Introduction

Interaction is a co-construction by all the speakers. Meaning depends both on discourse content and the relationship between interactants. Showing how speakers organize turn-taking exchanges is one of the way to account for meaning and speaker's involvement. The main focus here is to characterize the means used by speakers in expressing their involvement in this construction.

## 1. Theoretical frame.

Traditionally, the majority of studies on discourse analysis have focused their interest on « smooth transitions », that is without overlaps or interruptions. We don't discuss here the difference between these notions except to say that the second has not a single and formal definition [1], [2]. Speech overlap -formal phenomenon- is then the purpose of this work.
Studies on turn-taking have disregarded and marginalized this phenomenon. But overlaps are very frequent and relevant in spontaneous dialogs. From an interactive standpoint, we consider them as crucial because they implie such important notions as negociation, dominance, involvement and cooperation.

## 2. Corpus

9 speakers in 6 dialogs were recorded. Duration of each dialog is 20 minutes. This type of experiment implies a specific device: several channels for voice. Subjects were equipped with laryngophons which allowed the recording of each voice on separate tracks. They were sitted face to face in the sound-proof room. Despite these relative « unnatural » speaking conditions, interactions between speakers sounded natural and fluent.We avoided to give them a discussion topic. They were free to speak or not. They were finally in a « normal interactive situation » in which they had to « occupy the scene » and eventually interact and speak with one another.

## 3. Formal analytic units and events automatic detection procedure

### 3.1 The phonatory group (GP)

Among the different units of analysis (such as intonational or grammatical units, turn) for dialog's study, we have choiced the *phonatory group* (GP) [3] which is based on formal criteria. It corresponds to the speech production of each speaker bounded by pauses longer than 200 ms (this duration can varie according to the study[1]). The GP corresponds to a specific class of physiologic events (such as breath group or final pause). Its choice has been determined by its easily detectable character.
The speech signals were automatically divided into GP and overlaps.

### 3.2 The POC sequence

3 phases are selected around overlaps phenomena. We called *POC* sequence these 'moments' in which we have P as *preliminary phase*, O as *overlap*, and C as *continuing phase*. Other types of sequences are appeared[2] but we focused here only on this kind of pattern which is obtained from an automatic pattern detection from label sequences (pattern detection is based on initial and final GP labels). So with a pattern like P or C we have 2 simultaneous information: speaker identity and the duration of this pattern. O concerns overlap (2 simultaneous voices : the initiator (IN) of the overlap and the non initiator (IN) who is P).
Therefore we selected here a category of overlaps related on one GP by each speaker. This category is the most frequent in this corpus (about 2/minute).
The 2 following figures illustrate the POC sequence : in the first, P and C are the same, in the second they differ.

```
          P          O          C
A  xxxxxxxxxxxxx|xxxxxxxxxxx|xxxxxxxxxxxxx   GP/A/
B               |xxxxxxxxxxx|               GP/B/
```

**figure 1: C=P; no change speaker**

```
          P          O
A  xxxxxxxxxxxxx|xxxxxxxxxxx|      C         GP/A/
B               |xxxxxxxxxxx|xxxxxxxxxxxxx  GP/B/
```

**figure 2: C ≠ P; change speaker**

Each phase must have a minimal duration of 130 ms (about mean duration of a typical unaccented french syllable). P or C <130 ms belong to « simultaneous events » which are not analysed in this study. 267 POC patterns are then retained here.

[1] For Koiso and al. [4], pause duration is 100 ms, which seems us too short to distinguish in french between a silent pause and an unvoiced segment.
[2] We hypothezise that if they are formally different, they function differently too.

## 4. Model analysis and parameters

The logistic model we test relates the probability of speaker change to the following parameters :

-durP, durO, durC : durations (in second) of each phase

-mf0P : normalized mean f0 of the last 130 ms of P

-mf0IN, mf0noIN : normalized mean f0 of the last 130 ms in the O phase for the both speakers (IN and noIN).

Normalized mean f0 is defined as : mean f0-token / mean f0-speaker,

with mean f0-token : mean f0 (Hz) of the 130 ms window (in Hz),

mean f0-speaker : mean f0 (Hz) of the speaker over the whole duration of the dialog.

The choice of f0 parameter is based on the study of Koiso and al. [4] on smooth transitions in turn-taking: in the same way like them, we assume that the information relevant to change speaker is localized at a point just before the overlap (end of P) and another point at the end of the overlap (mf0IN and mf0noIN).

See figure 3 and table 1 (p.6) for illustration.

## 5. Results

Logistic regression results

| *Parameter* | *Parameter estimate* | *Standard error* | *T statistic* |
|---|---|---|---|
| intercept | -1.295626 | 0.906585 | -1.429127 |
| durP | -0.932970 | 0.258482 | -3.609427 |
| durO | 1.754288 | 0.610717 | 2.872503 |
| durC | 0.596902 | 0.232852 | 2.563444 |
| mf0P | -1.737269 | 0.639688 | -2.715808 |
| mf0IN | 2.723869 | 0.596604 | 4.565623 |
| mf0noIN | -0.541377 | 0.486584 | -1.112608 |

*null deviance* : 364.99 on 266 df

*residual deviance* : 304.46 on 260 df

Analysis of deviance table

| | dev resid | Res dev | Pr (Chi) |
|---|---|---|---|
| durP | 12.268700 | 352.727 | 0.000461 |
| durO | 7.377250 | 345.350 | 0.006605 |
| durC | 7.005130 | 338.345 | 0.008128 |
| mf0P | 10.280000 | 328.065 | 0.001345 |
| mf0IN | 22.347400 | 305.718 | 0.000002 |
| mf0noIN | 1.250730 | 304.467 | 0.263413 |

$R^2$ Nagelkerke: 0.272220

Starting from the Null model, each parameter is sequentially added to the model. The analysis of deviance agrees with the partial t-test : except mf0noIN each parameter is a valuable predictor.

It is worth noting that the model gives much better results when the 3 worst items (in terms of residual) are eliminated from the regression [$R^2$Nagel : 0.315 ; Chi-2 (df = 258) =289.13, p= 0.08].

## 6. Discussion

Significative part of the variability of the model is explained by the retained parameters : duration of P, O and C and mean f0 on P, IN and noIN.
Speaker change probability increases when the main speaker is interrupted early (short P) and when O and C are longer : it's easier to take the turn if the other didn't have sufficient time to get involved in his own speech. The interrupter can consider it rightful to intervene in speech especially with the presence of a pause (before P) which can be the signal of a potential transitional place for two speakers. O could then be considered as resulting from a longer reaction time for the interruptor. The change speaker probability increases when meanf0 on the end of P is lower. This confirms the point that the interruptor can perceive here a potential transitional place (lower value is a cue of finality). Short P and lower values f0 can be explained as a minimal involvement of the main speaker which contribute to the interruption by the other. Concerning meanf0 on the end of O, speaker change probability increases when meanf0IN is higher : this may be interpreted not only as a cue of non finality of the turn (turn in progress) but also as a cue related to the intention of the initiator : his will to manifest he wants to take (and keep) the floor (dominance). Mean f0 values of noIN are not significant : this may be interpreted here as the fact that the main speaker is interrupted and have then to stop his discourse randomly.
On the opposite side, the probability of speaker change decreases when P increases and O and C decrease (in terms of duration). The overlap may be then considered here such as a back-channel signal (« minimal signal » as hum, OK). Back-channel doesn't interrupt the main speakerit is especially used to show to the speaker listener's involvement in the discourse. This type of back-channel signal is often realized in the lower values of f0 (mf0IN). The higher meanf0 on the end of P contribute to this point : it can be explained by the presence of a phatic signal (produced by the main speaker) which the main function is to capture attention of the listener (using high values of f0). Phatic and back-channel signals are usually produced in complementary distribution.
Extreme results of the continuum on which analyzed parameters here are valuable predictors illustrate two types of overlap : « intrusive » and « cooperative » overlap. These notions are borrowed from Murata [2] who classifies interruptions in intrusive and cooperative. For us, these notions are only based on formal criteria (change or not change speaker). In a future study, we will try to account for the pragmatic value of each overlap (in

considering the content of speech for example). One of our first goal is to analyze more specifically back-channel signals to confirm these first results and account for them statistically.

## Conclusion

This paper highlights how a few overlaps are used in the turn-taking exchanges with respect to prosodic parameters.
We showed that duration and fundamental frequency can be valuable predictors to account for an eventual change of speaker after the overlap.
Our results can help us to classify the kinds of overlaps used by conversational participants (in cooperative/intrusive overlap).

Acknoledgement : to E. Flachaire, who writes the ects [5] script for logit regression.

REFERENCES

[1] Makri-Tsilipakou, M. (1994) : Interruption revisited : affiliative vs. disaffiliative intervention, *Journal of Pragmatics*, 21, PP 401-426.
[2] Murata, K., (1994) : Intrusive or cooperative ? A cross cultural study of interruption, *Journal of Pragmatics*, 21, pp. 385-400.
[3] Autesserre, D., Y. Nishinuma, S. Delran-Bado (1992) : Respiration et phonation dans le dialogue oral spontané : valeurs de la fréquence fondamentale au contact des pauses, Prépublications du *Séminaire de Prosodie*, La Baume les Aix, pp. 63-72.
[4] Koiso, H., Y. Horiuchi, S. Tutiya, A. Ichikawa, Y. Den (1998) : An analysis of turn-taking and backchannels based on prosodic and syntactic features in japanese map task dialogs, *Language and Speech*, 41 [3-4], pp. 295-321.
[5] Davidson Russell, 1999 : http://russell.cnrs-mrs.fr/pub/ects3

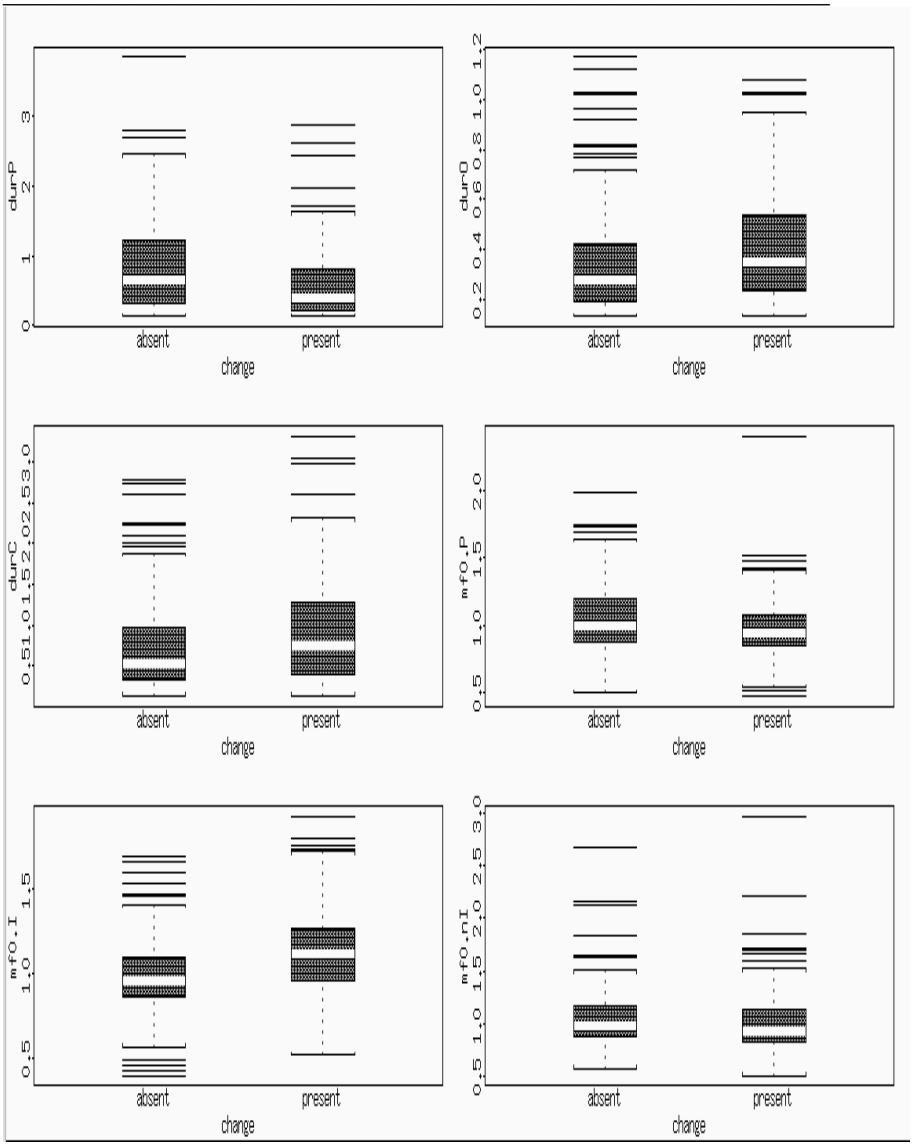


**Figure 3 : Boxplots of the predictors of change**

The horizontal line in the interior of the box is located at the median of the data. The height of the box is equal to the interquartile distance (IQD). The horizontal lines at the top of the graph represent outliers (>1,5. IQD).

| *Change* | *P* | *O* | *C* | *mf0P* | *mf0IN* | *mf0noIN* |
|---|---|---|---|---|---|---|
| mean | 610.2 | 420.7 | 913.3 | 0.97 | 1.12 | 1.01 |
| std error | 537.5 | 240.1 | 661.8 | 0.23 | 0.24 | 0.33 |
| *No change* | | | | | | |
| mean | 881.8 | 349.4 | 713.9 | 1.07 | 0.98 | 1.05 |
| std error | 700.2 | 218.1 | 559.3 | 0.33 | 0.22 | 0.27 |

**Mean, standard error for each predictor/ change or no change speaker**