On the relationship between intra-oral pressure and speech sonority

Anne Cros, *Didier Demolin*, *Ana Georgina Flesia*,

Antonio Galves

Interspeech 2005

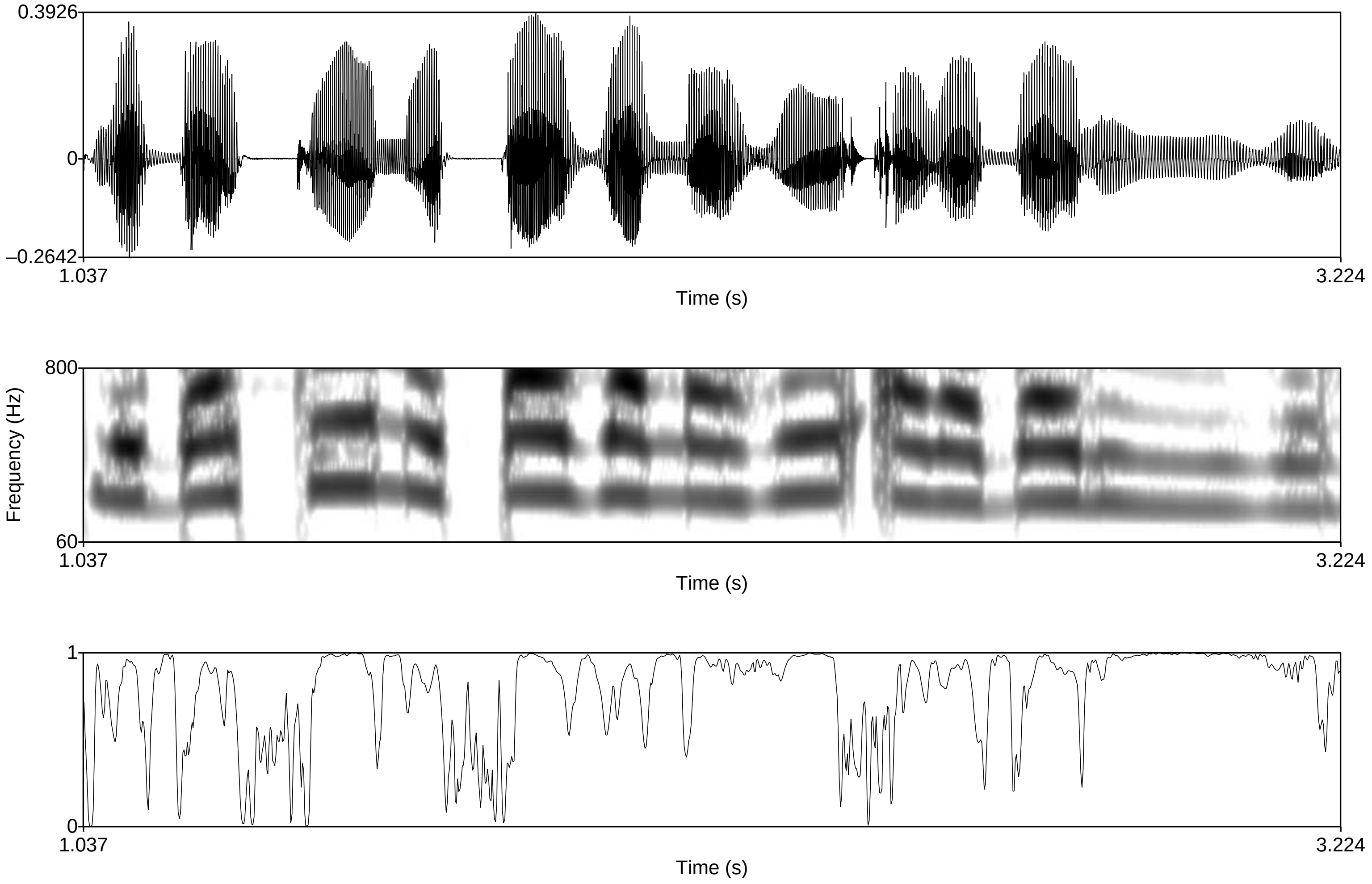
We address the question of the relationship between two time series associated to the speech signal.

- the sonority function
- the intra-oral pressure evolution during the production of speech.

The speech sonority

- The sonority function is an index of local regularity of the speech signal (Galves *et al.* 2002)
- It is defined as a mapping of the spectrogram of the acoustic signal into a function of time taking values in the interval $[0, 1]$.
- This function is close to 1 for spans displaying regular patterns, characteristic of sonorant portions of the signal.
- In contrast, in regions in which the acoustic signal present a chaotic behavior the sonority function will assume values closer to 0, with important variations.

le bateau n'est pas amarré à la balise



A formal definition

- Let $c_t(i)$ be the Fourier coefficient for the frequency i around time t in the spectrogram.
- We define the renormalized power spectrum by

$$p_t(i) = \frac{c_t(i)^2}{\sum_f c_t(f)^2} \, .$$

Regular patterns characteristic of sonorant regions typically correspond to sequences of probability measures $\{p_t : t = 1, 2, ...\}$ close in the sense of relative entropy.

The speech sonority

$$S(t) = \exp\{-\beta \sum_{i=1}^{3} h(p_t|p_{t-i})\} \, ,$$

where

$$h\left(p_t|p_{t-i}\right) = \sum_f p_t\left(f\right) \log\left(\frac{p_t\left(f\right)}{p_{t-i}\left(f\right)}\right)$$

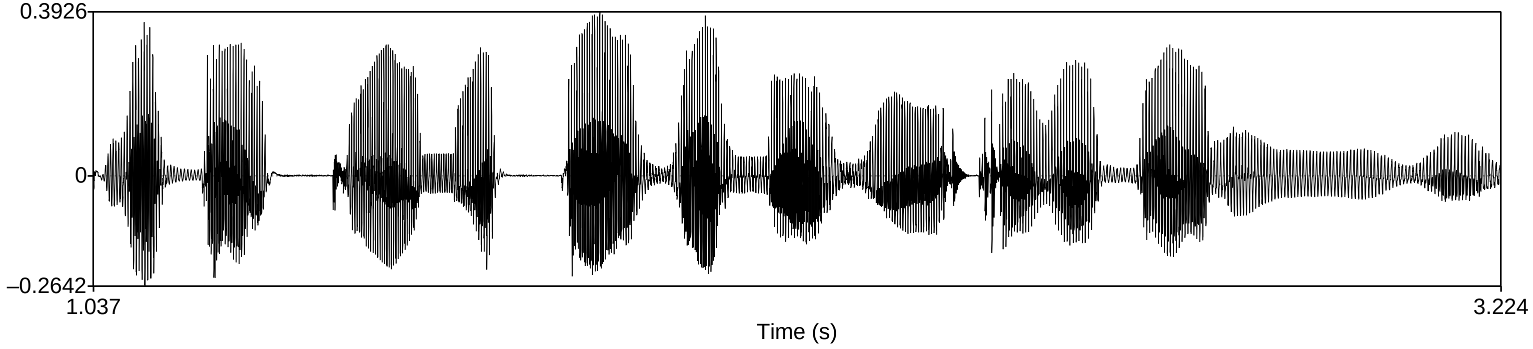
is the relative entropy of p_t with respect to p_{t-i}.

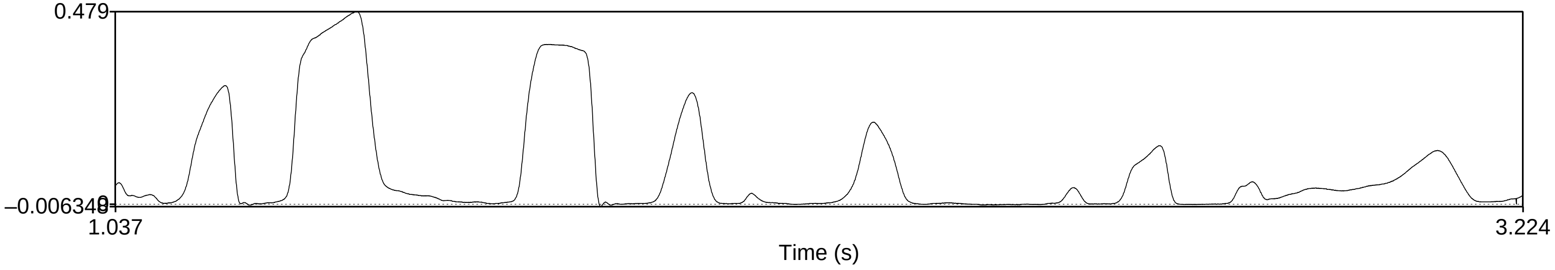
The intra-oral pressure $P_s(t)$.

- The time series $P_s(t)$ was measured via a small plastic tube (internal diameter 2 mm) inserted through the nose up to the area behind the velum.
- The data were recorded on the workstation Physiologia that allows synchronous recording of acoustic and aerodynamic parameters.

For details on the method, see Demolin *et al.* (2004).

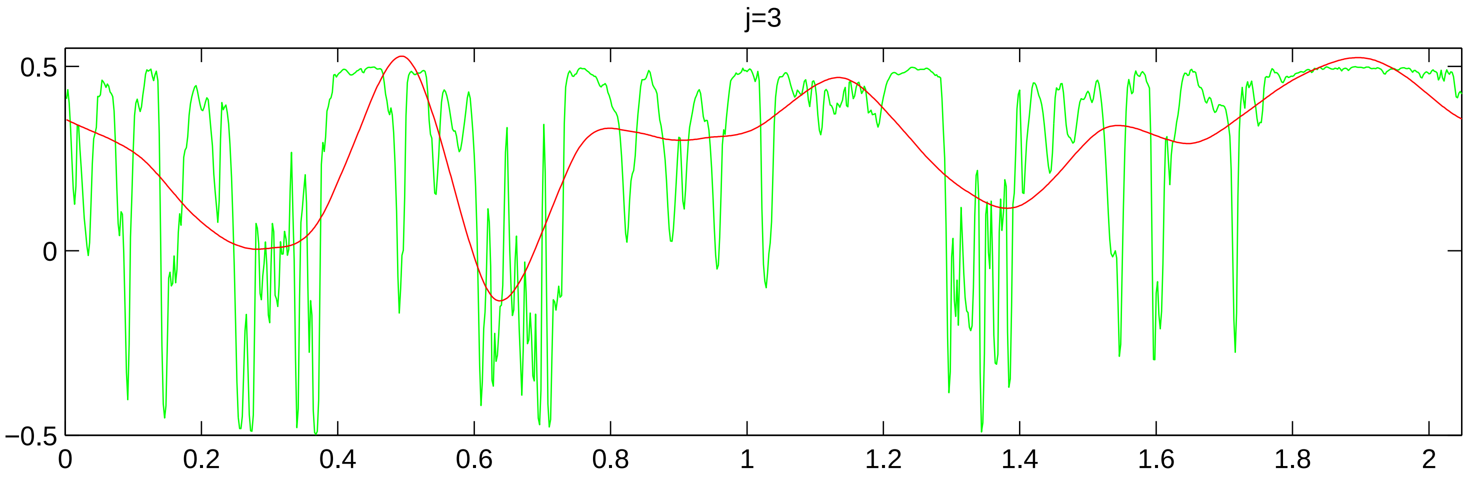
le bateau n'est pas amarré à la balise

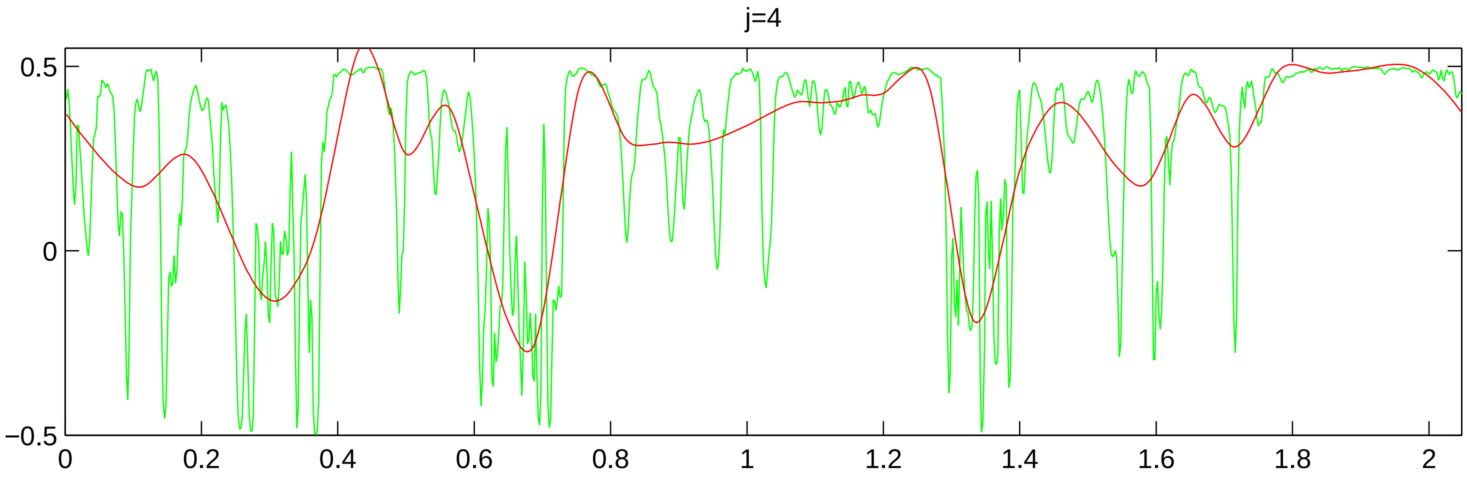


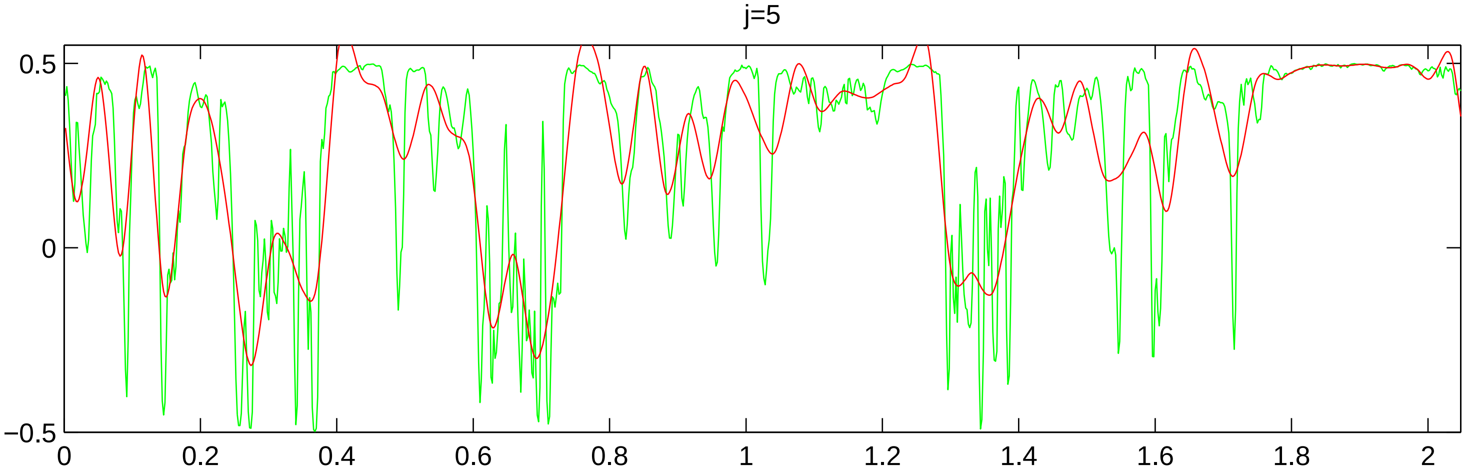


A signal processing ***intermezzo***

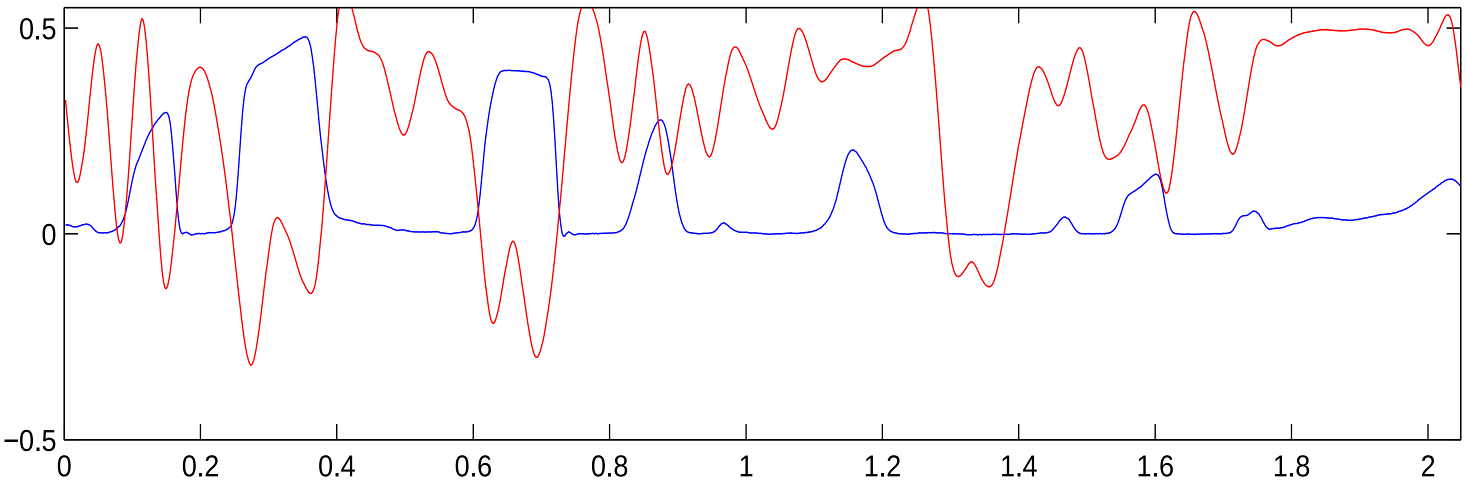
- In our data set, the intra-oral pressure was low-pass filtered at 70 Hz.
- Therefore to have the signals in the same space of smoothness, we processed the sonority function with an orthogonal wavelet transform.
- The smoothing was performed spanning the signal on the Symmlet 8 basis and reconstructing it using only the 5 coarsest levels.


j=3
0.5
0
−0.5
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2

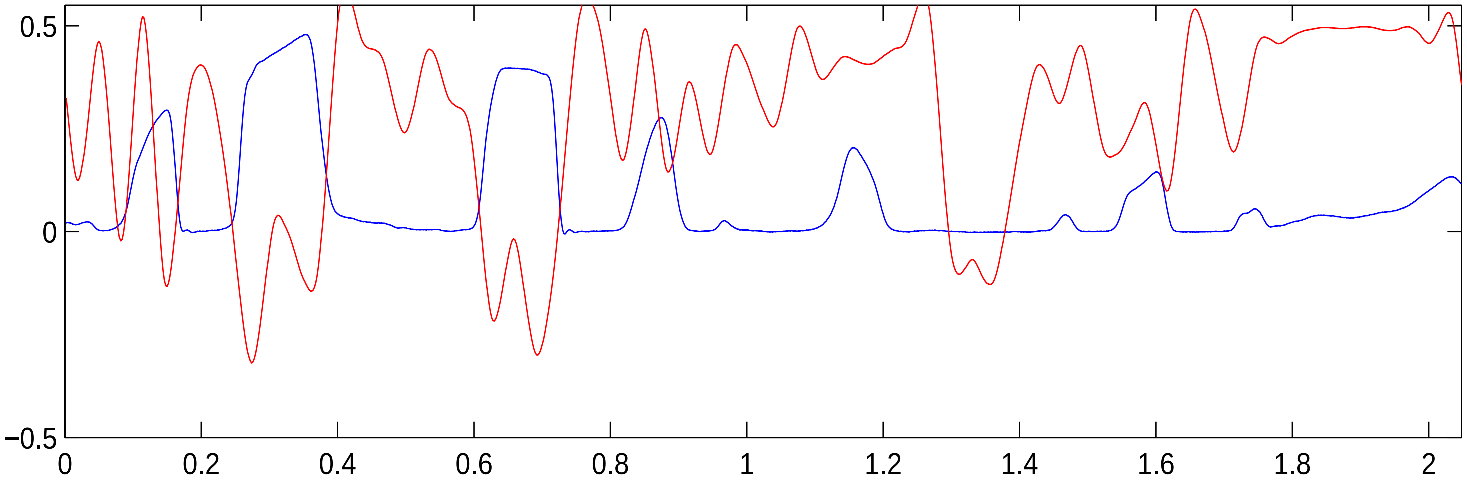

j=4
0.5
0
−0.5
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2


j=5
0.5
0
−0.5
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2

le bateau n'est pas amarré à la balise



le bateau n'est pas amarré à la balise



First guess: when the **sonority** *is* **high**, *the* **intra-oral pressure** *is* **low** *and conversely.*

Quantization of the sonority and the oral pressure

Define

- the chain $I_S(t)$ that codifies $S(t)$ in zones of high and low sonority:

$$I_S(t) = \begin{cases} 1 & \text{if } S(t) > c_s \\ -1 & \text{if } S(t) \leq c_s \ ; \end{cases}$$

- the chain $I_P(t)$ that codifies $P_s(t)$ in regions of high and low pressure:

$$I_P(t) = \begin{cases} 1 & \text{if } P_s(t) \leq c_p \\ -1 & \text{if } P_s(t) > c_p \ ; \end{cases}$$

where c_s and c_p are two suitable cut-points.

Model 1

$$I_S(t) = I_P(t).\eta(t)\,,$$

where $\eta(t)$ is independent of I_S and I_P with

$$\eta(t) = \begin{cases} 1 & \text{with} \quad \mathbb{P}(\eta(t) = 1) = 1 - \epsilon \\ -1 & \text{with} \quad \mathbb{P}(\eta(t) = -1) = \epsilon \end{cases}$$

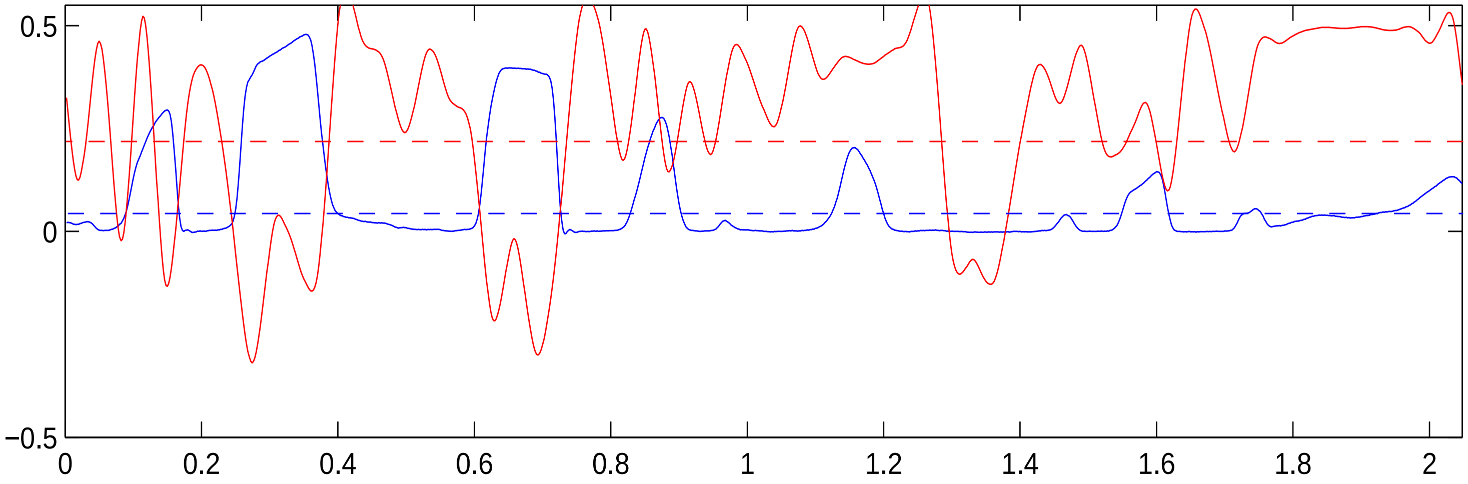
$\eta(t)$ accounts for processing errors.

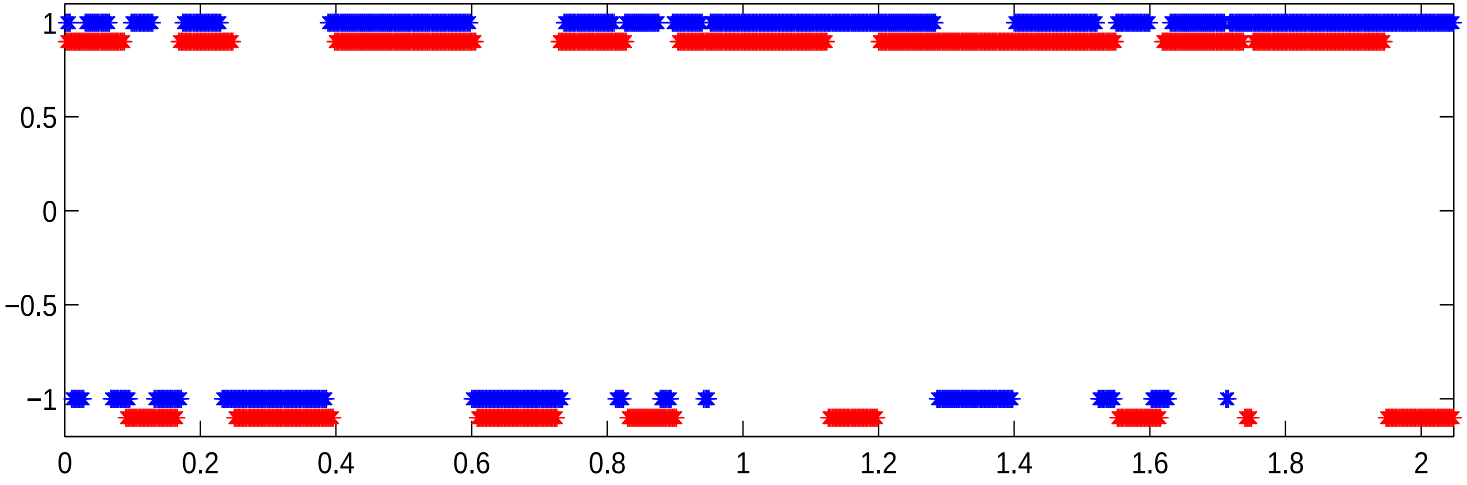
The data

To check the validity of the model we will analyze two data sets.

- The first one is a Kinyarwanda corpus with 27 sentences.
- The second one is a French corpus with 26 sentences.

le bateau n'est pas amarré à la balise





Proportion of time in which data behave as predicted by Model 1

French	69.1%
Kinyarwanda	52.3%

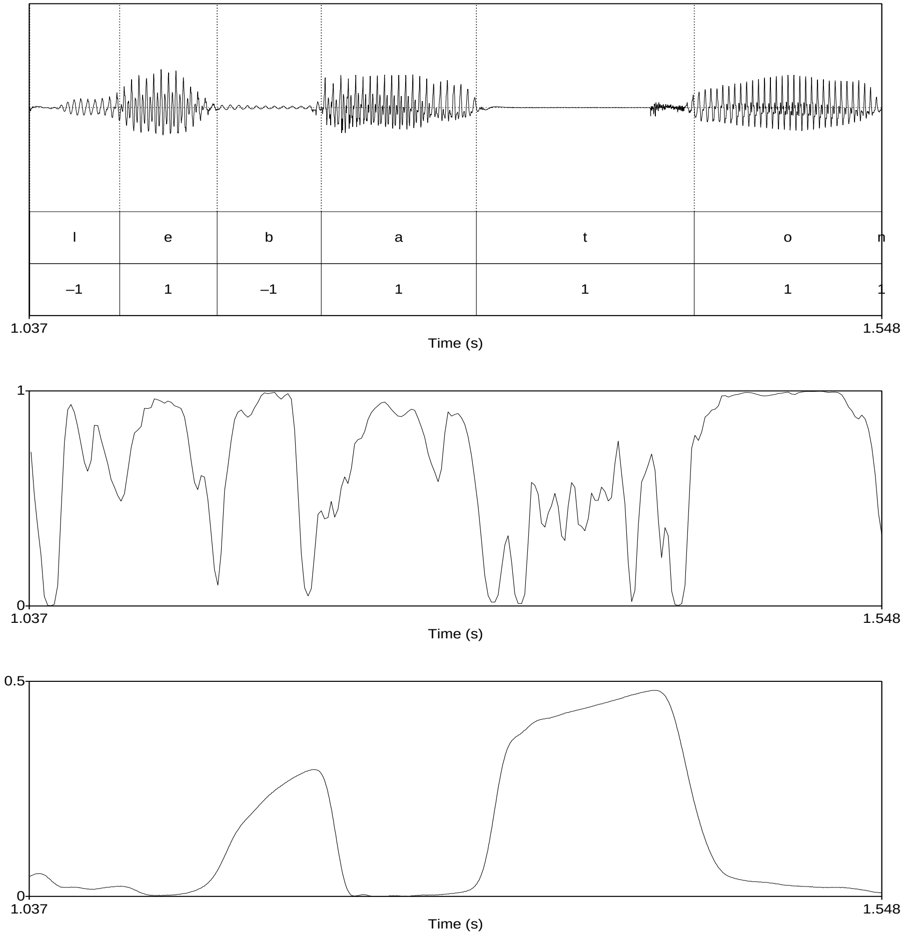
This is far from being satisfactory !!!

What is wrong with Model 1 ?

Model 1 does not describe the behaviour of certain phonetic segments, for which the pressure and the sonority are both high or low (for instance voiced constrictive consonants).

See what happens with [b] in *le bateau* (next figure).

le bateau...

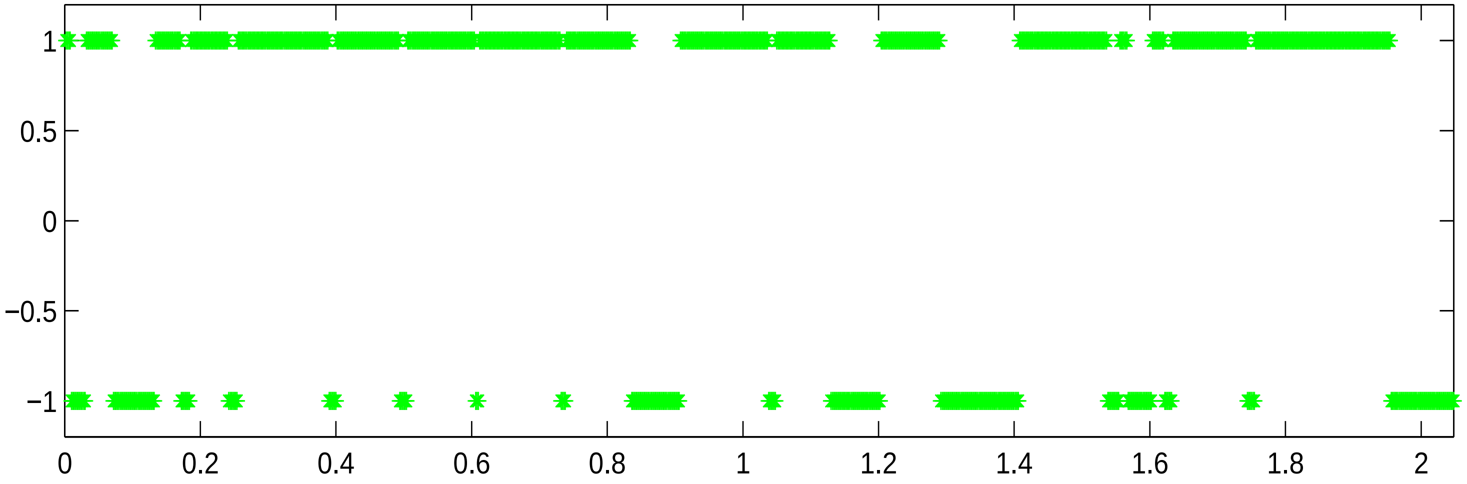

l
e
b
a
t
o
n
–1
1
–1
1
1
1
1
1.037
1.548
Time (s)
1
0
1.037
1.548
Time (s)
0.5
0
1.037
1.548
Time (s)

This helps understanding why Model 1 describes better French than Kinyarwanda.

French sentences	69.1%
Kinyarwanda	52.3%

Our data set of Kinyarwanda includes a lot of glottal stops for which both pressure and sonority are low.

Plot of $I_S(t).I_P(t)$



Well behaved segments for Model 1 are those for which in the absence of noise

$$I_S(t).I_P(t) = 1$$

.

Improving Model 1

- Call $\mathcal{B}$, the set of *well behaved* phonetic segments.
- Define the auxiliary chain $I_C(t)$ by

$$I_C(t) = \begin{cases} 1 & \text{if } \sigma(t) \in \mathcal{B} \\ -1 & \text{if } \sigma(t) \notin \mathcal{B} \end{cases} ;$$

where $\sigma(t)$ stands for the speech signal.

Model 2

$$I_S(t) = I_C(t).I_P(t).\eta(t)$$

With this model, if $\sigma(t)$ is part of a voiced constrictive consonant, then

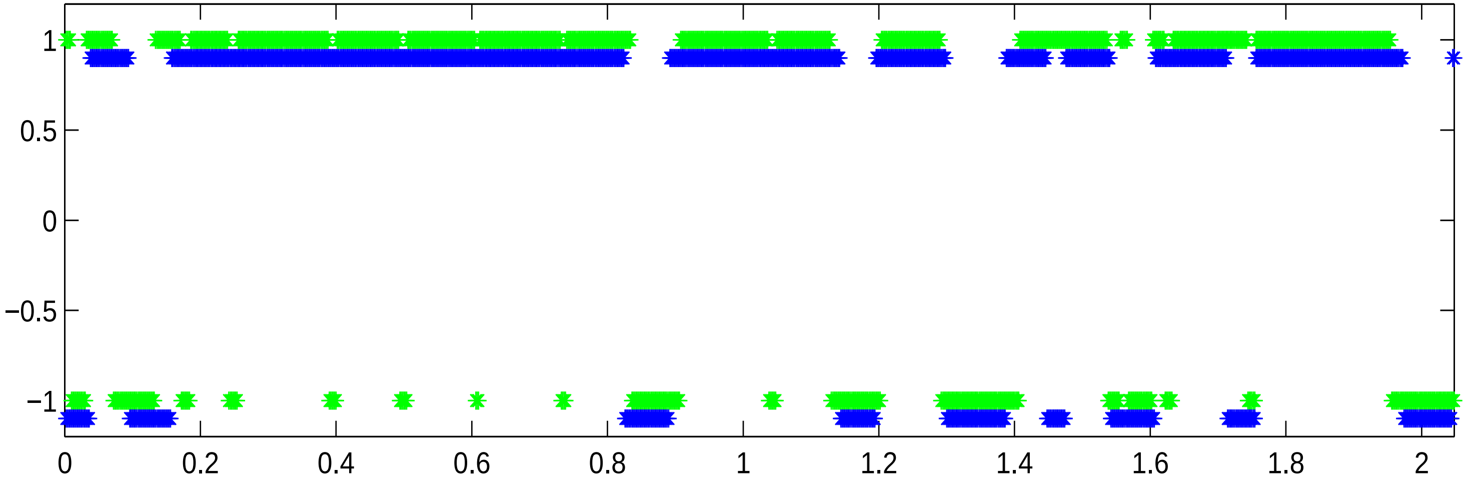
$$I_C(t) = -1$$

and we obtain

$$I_S(t) = -I_P(t).\eta(t)$$

which is the correct prediction for this segment.

Superposed plots of $I_S(t).I_P(t)$ and $I_C(t)$



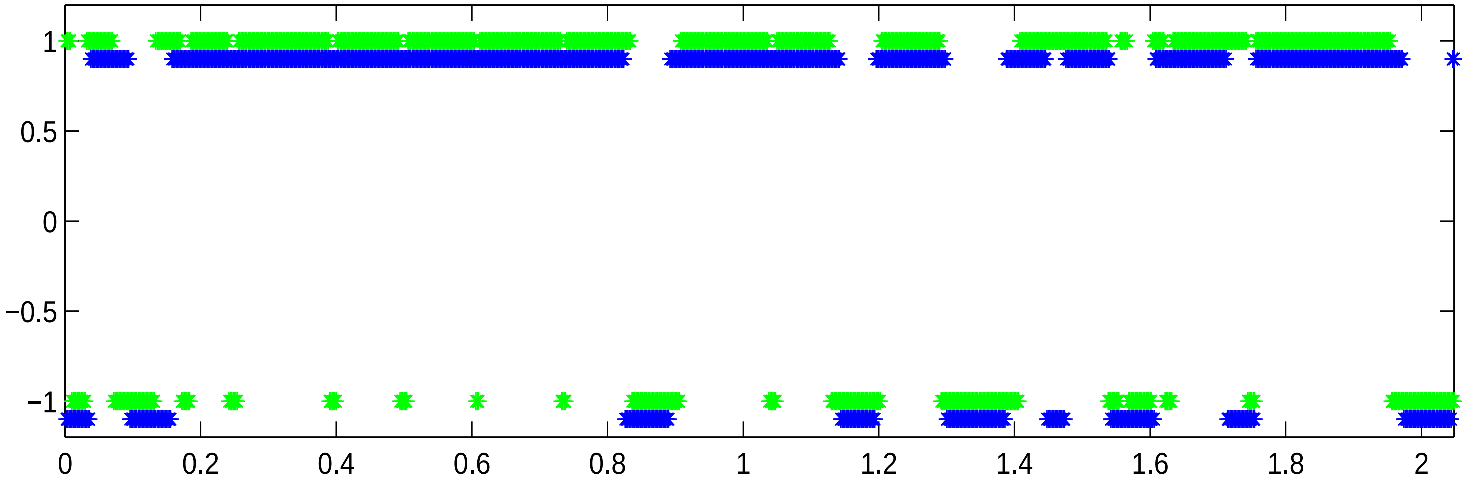
Model 2 describes well the data when $I_S(t).I_P(t) = I_C(t)$.

Model 1 only describes well the data when $I_S(t).I_P(t) = 1$.

Proportion of time in which data behave as predicted by Models 1 and 2

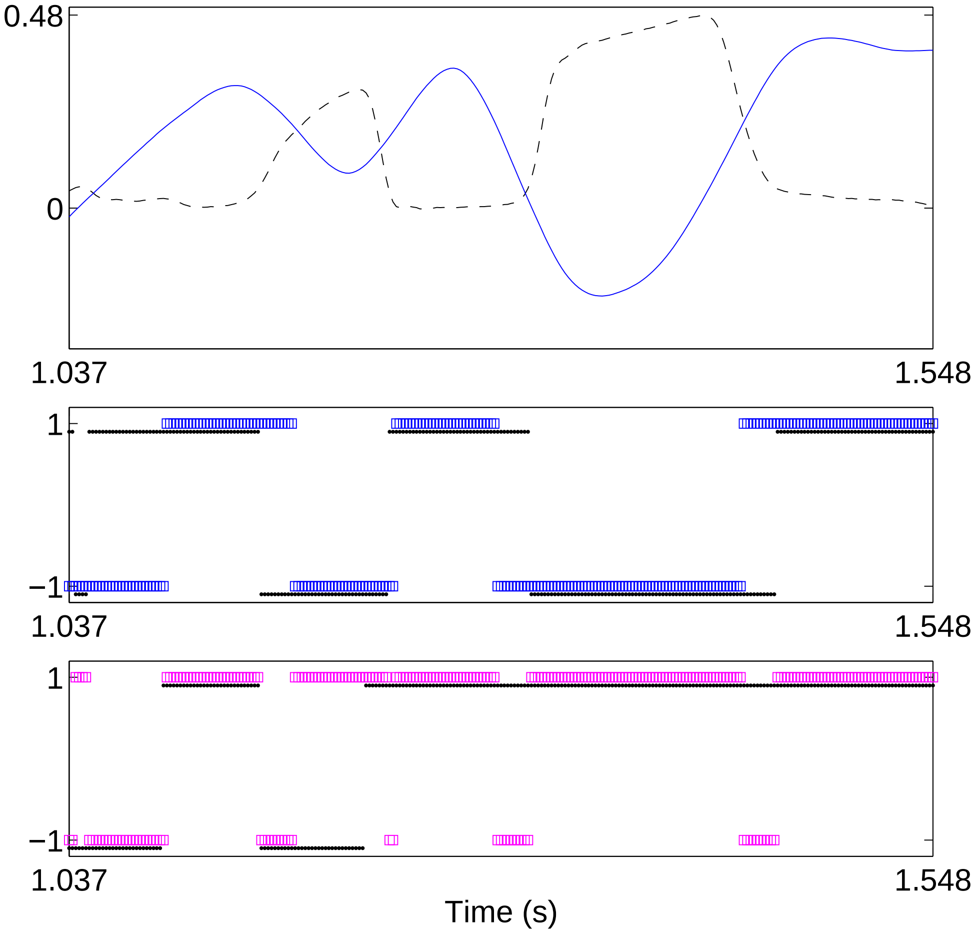
	Model 1	Model 2
French	69.1%	79.2%
Kinyarwanda	52.3%	81.6%

However...



...we can observe that Model 2 does not hold at the beginning and at the end of some segments.

Model 2 does not account for transition zones



A problem and a way to solve it

Transition regions appear to be the price we have to pay in order to tie continuous signals with discrete binary chains.

Let's take these regions into account...

A natural conjecture

Model 2 together with the remark concerning the transition zones suggests that if

$$I_S(t) \neq I_P(t) I_C(t)\,,$$

then

- either the processing noise is present at time t (*i.e.* $\eta(t) = -1$)
- or t belongs to a transition zone.

More formally...

Let e be the proportion of time in which

$$I_S(t) \neq I_P(t) I_C(t)\,,$$

and let θ be the proportion of time spent in transition zones.

The conjecture is that

$$e \simeq \epsilon + \theta\,,$$

where

$$\epsilon = \mathbb{P}(\eta(t) = -1)\,.$$

In the above formula we are neglecting second order corrections ...

A better estimation of ϵ

	Model 1	Model 2	θ	ϵ
French	69.1%	79.2%	10%	10.8%
Kinyarwanda	52.3%	81.6%	6%	12.4%

The proportion of points belonging to transition zones equals 6% for Kinyarwanda, and equals 10% for French.

The calculation of the average noise leads to $\epsilon \simeq 12.4\%$ for the Kinyarwanda, and $\epsilon \simeq 10.8\%$ for the French sentences.

This reinforces the conjecture that Model 2 describes quite well the situation outside transition zones.

To conclude

- It is clear that there is a number of issues related to Model 2 that have to be refined and discussed such as the identification of the class of well behaved phonetic segments.
- It also indicates that it should be possible to improve Model 2 (*i.e.* to define a new Model 3) by a suitable description of the behavior of both processes in the transition zones.
- However we think that the correlation between the sonority function and the intra-oral pressure is well established.

- PRONEX/FAPESP's Project *Stochastic behavior, critical phenomena and rhythmic pattern identification in natural languages* (grant number 03/09930-9)
- CNPq's project *Stochastic modeling of speech* (grant number 475177/2004-5)

Complements

- We take time t belonging to the set $\{ku : k = 1, \ldots, T\}$, where u is the step unity of the spectrogram of the signal and T is the number of steps present in the spectrogram of the acoustic signal.
- In the present computation $u = 2$, where the units are counted in milliseconds.
- The values of the spectrogram are estimated with a 25ms Gaussian window. We only consider frequencies between 60 and 800 Hz. We choose the tuning constant $\beta = 1.5$.
- Our computations were made with Praat (http://www.praat.org).

The cut-off point for the sonority

- To define the binary chain $I_S(t)$, we use the cut-off $c_s = 0.7$.
- This is one of the four universal cut-points identified and estimated in Cassandro *et al.* (2005).
- c_s seems to be the most relevant cut-point separating high and low sonority zones.

The cut-off point for the intra-oral pressure

- To define the binary chain I_P we use the cut-off point $c_p = 0.05$.
- This cut-off point seems to discriminate zones of constant null pressure from zones in which the pressure is different from zero.