Report of Research Work

Identification of rhythmic classes of languages using adaptative representations of the speech signal

Visitor: Marcela Morvidone
Scientific referent: Antonio Galves
Laboratory: Instituto de Matemática e Estatística - USP

During my research stay at the IME I have been working on the implementation of an algorithm to decompose signals in different components to be used in the study of the rhythmic properties of languages.

Under the advising of Prof. Antonio Galves I have been working with Luis Fernandes Baumann, one of his students at the IME. I have also visited Jesús García at the Instituto de Matemática of the Universidade Estadual de Campinas, and Georgina Flesia at the Instituto Argentino de Matemática (IAM) at Buenos Aires, Argentina.

I have presented two seminars: *"Representación adaptativa de señales usando un diccionario de bases o marcos"* at the IAM, Buenos Aires on August 17th, and *"Adaptive signal representation using a dictionary of bases or frames"* at the IME, on August 24th.

A summary of the results obtained is included in the following attachment.

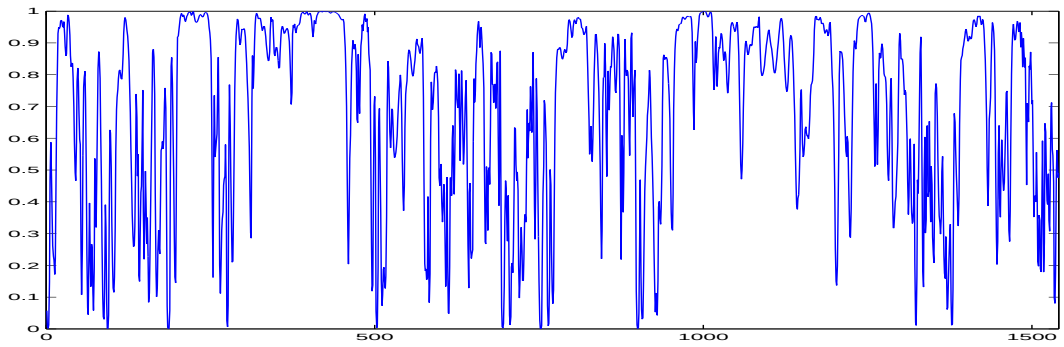

Figure 1: Sonority function

1 Studying the linearity of the speech sonority

In 2002 Galves *et al.* [1] introduced a rough measure of sonority as a tool to discriminate between rhythmic classes of languages. The goal was to reproduce in an entirely automatic way, with no need of previous hand labelling, the empirical results obtained by Ramus, Nespor and Mehler in 1999 [2]. The sonority was defined as a function which maps local windows of the acoustic signal on the interval $[0, 1]$. This function is close to 1 for regions displaying regular patterns characteristic of sonorant portions of the signal. In opposition, the function assigns values close to 0 for regions characterized by obstruency. Figure 1 shows an example of the sonority function corresponding to the Spanish phrase *Hace cinco minutos que el tren ha llegado a la ciudad*.

An empirical analysis of a multi-lingual corpus puts in evidence a linear relationship between the mean sonority and the mean increment of the sonority in absolute value across sentences of the sample. This corpus has 1973 sentences from 10 different languages. Figure 2 shows this striking relationship.

Prof. Galves and collaborators have suggested a stochastic model to give a simple explanation for this linear relationship. My research work consisted in implementing recent signal processing tools to automatically separate the intervals of sonority and obstruency in the sonority function in order to estimate the parameters that characterize the model and to use these estimations as a possible way to test the validity of the model.

In Section 2 we describe the stochastic model for the speech sonority, in Section 3 we present the algorithm used for the processing of the data, finally in Section 4 we show our preliminary results.

2 A model for the speech sonority

Two families of stochastic chains are considered $\{(S_t^l)_{t\in\mathbb{Z}} : l \in \mathscr{L}\}$ and $\{(X_t^l)_{t\in\mathbb{Z}} : l \in \mathscr{L}\}$ where $\mathscr{L}$ is a fixed but otherwise arbitrary set. The chains $(S_t^l)_{t\in\mathbb{Z}}$ in the first family take values in the interval $[0, 1]$. They represent the sonority contours of the different languages. The chains $(X_t^l)_{t\in\mathbb{Z}}$ take values in the finite alphabet $A = \{0, 1\}$. It is assumed that these processes are stationary and ergodic. They are tied together by the following assumption:
There exist probability distributions π_i and $\pi_{(i,j)}$ on $[0, 1]$ and $[0, 1]^2$ respectively indexed by symbols i and j in the alphabet A which are language independent and such that for any $l \in \mathscr{L}$

$$\mathbb{P}\{S_t^l \in B | X_t^l = j\} = \pi_j(B), \tag{1}$$

and

$$\mathbb{P}\{S_t^l \in B, S_{t+1}^l \in C | X_t^l = i, X_{t+1}^l = j\} = \pi_{i,j}(B, C), \tag{2}$$

where B and C are Borel subsets of $[0, 1]$.

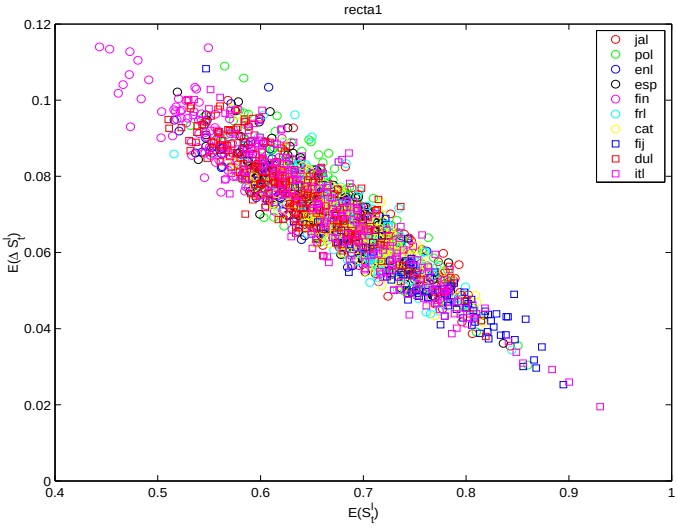

Figure 2:

Let's introduce the notation

$$p^l(i) = \mathbb{P}\{X_t^l = i\},\ p^l(i,j) = \mathbb{P}\{X_t^l = i,\ X_{t+1}^l = j\},$$

and

$$\theta(i) = \mathbb{E}\{S_t^l | X_t^l = i\} = \int s_1 d\pi_i(s_1) \text{ and } \theta(i,j) = \mathbb{E}(|S_t^l - S_{t+1}^l||X_t^l = i, X_{t+1}^l = j) = \int |s_1 - s_2| d\pi_{ij}(s_1, s_2). \tag{3}$$

Assumptions (1) and (2) imply that the expectations $\theta(i)$ and $\theta(i,j)$ are language independent. The following theorem, together with the ergodicity of the chains (S_t^l) may explain the linear relationship in Figure 2

Theorem *Under assumptions (1) and (2) it follows that*

$$\mathbb{E}\left(|S_t^l - S_{t+1}^l|\right) = a + b\,\mathbb{E}(S_t^l) + \epsilon^l,$$

where the constants a and b are language independent and defined as

$$a = \theta(0,0) - \theta(0)\frac{\theta(1,1) - \theta(0,0)}{\theta(1) - \theta(0)},\ \ b = \frac{\theta(1,1) - \theta(0,0)}{\theta(1) - \theta(0)}.$$

and the correction ϵ^l is language dependent and defined as

$$\epsilon^l = p_{0,1}^l\left(\theta(1,0) + \theta(0,1) - \theta(0,0) - \theta(1,1)\right).$$

Proof.

Using the notation stated in assumptions (1) and (2) we obtain

$$\begin{aligned}
\mathbb{E}\left(|S_t^l - S_{t+1}^l|\right) &= \sum_{i,j=0,1} \mathbb{E}\left(|S_t^l - S_{t+1}^l|/(|X_t^l = i, X_{t+1}^l = j)\right) \mathbb{P}\{X_t^l = i, X_{t+1}^l = j\} \\
&= \sum_{i,j=0,1} \theta(i,j) p^l(i,j) \\
&= \sum_{i=0,1} \theta(i,i) p^l(i,i) + \sum_{i \neq j} \theta(i,j) p^l(i,j) \\
&= \sum_{i=0,1} \theta(i,i) p^l(i,i) + \sum_{i=0,1} \theta(i,i) p^l(i,1-i) - \sum_{i=0,1} \theta(i,i) p^l(i,1-i) + \sum_{i \neq j} \theta(i,j) p^l(i,j) \\
&= \sum_{i=0,1} \theta(i,i)[p^l(i,i) + p^l(i,1-i)] + \epsilon^l
\end{aligned}$$

Since $p^l(i) = p^l(i,1) + p^l(i,0) = p^l(i,i) + p^l(i,1-i)$, it follows that

$$\mathbb{E}\left(|S_t^l - S_{t+1}^l|\right) = \sum_{i=0,1} \theta(i,i) p^l(i) + \epsilon^l$$

In the other hand

$$\mathbb{E}\left(S_t^l\right) = \sum_{i,j=0,1} \mathbb{E}\left(S_t^l | (|X_t^l = i, X_{t+1}^l = j)\right) \mathbb{P}\{X_t^l = i, X_{t+1}^l = j\}$$

and

$$\begin{aligned}
\mathbb{E}\left(S_t^l | (|X_t^l = i, X_{t+1}^l = j)\right) &= \int\int s_1 \pi_{i,j}(s_1, s_2) ds_1 ds_2 \\
&= \int s_1 [\int \pi_{i,j}(s_1, s_2) ds_2] ds_1
\end{aligned}$$

If we suppose that

$$\int \pi_{i,1}(s_1, s_2) ds_2 = \int \pi_{i,0}(s_1, s_2) ds_2 = \pi_i(s_1)$$

then

$$\begin{aligned}
\mathbb{E}\left(S_t^l\right) &= \sum_{i,j=0,1} \mathbb{E}\left(S_t^l | (|X_t^l = i, X_{t+1}^l = j)\right) \mathbb{P}\{X_t^l = i, X_{t+1}^l = j\} \\
&= \sum_{i,j=0,1} \theta(i) p^l(i,j) \\
&= \theta(0) p^l(0,0) + \theta(0) p^l(0,1) + \theta(1) p^l(1,0) + \theta(1) p^l(1,1) \\
&= \theta(0)[p^l(0,0) + p^l(0,1)] + \theta(1)[p^l(1,0) + p^l(1,1)] \\
&= \theta(0) p^l(0) + \theta(1) p^l(1) \\
&= \theta(0)(1 - p^l(1)) + \theta(1) p^l(1) \\
&= \theta(0) + p^l(1)[\theta(1) - \theta(0)]
\end{aligned}$$

In the same way we may derive the equation

$$\mathbb{E}\left(S_t^l\right) = \theta(1) + p^l(0)[\theta(0) - \theta(1)]$$

Then

$$\begin{aligned}
\mathbb{E}\left(|S_t^l - S_{t+1}^l|\right) &= \theta(0,0)p^l(0) + \theta(1,1)p^l(1) + \epsilon^l \\
&= \theta(0,0)\left(\frac{\mathbb{E}\left(S_t^l\right) - \theta(1)}{\theta(0) - \theta(1)}\right) + \left(\frac{\mathbb{E}\left(S_t^l\right) - \theta(0)}{\theta(1) - \theta(0)}\right)\theta(1,1) + \epsilon^l \\
&= -\frac{\mathbb{E}\left(S_t^l\right)\theta(0,0)}{\theta(1) - \theta(0)} + \frac{\theta(1)\theta(0,0)}{\theta(1) - \theta(0)} + \frac{\mathbb{E}\left(S_t^l\right)\theta(1,1))}{\theta(1) - \theta(0)} - \frac{\theta(0)\theta(1,1)}{\theta(1) - \theta(0)} + \epsilon^l \\
&= \mathbb{E}\left(S_t^l\right)\left[\frac{\theta(1,1) - \theta(0,0)}{\theta(1) - \theta(0)}\right] + \frac{\theta(1)\theta(0,0)}{\theta(1) - \theta(0)} - \frac{\theta(0)\theta(1,1)}{\theta(1) - \theta(0)} + \epsilon^l \\
&= \mathbb{E}\left(S_t^l\right)\left[\frac{\theta(1,1) - \theta(0,0)}{\theta(1) - \theta(0)}\right] + \frac{\theta(1)\theta(0,0) - \theta(0)\theta(1,1)}{\theta(1) - \theta(0)} + \epsilon^l \\
&= \mathbb{E}\left(S_t^l\right)b + a + \epsilon^l,
\end{aligned}$$

since

$$\begin{aligned}
a &= \theta(0,0) - \theta(0)\frac{\theta(1,1) - \theta(0,0)}{\theta(1) - \theta(0)} \\
&= \frac{\theta(0,0)[\,\theta(1) - \theta(0)]}{\theta(1) - \theta(0)} - \theta(0)\frac{\theta(1,1) - \theta(0,0)}{\theta(1) - \theta(0)} \\
&= \frac{\theta(0,0)\theta(1) - \theta(0,0)\theta(0) - \theta(0)\theta(1,1) + \theta(0)\theta(0,0)}{\theta(1) - \theta(0)} \\
&= \frac{\theta(1)\theta(0,0) - \theta(0)\theta(1,1)}{\theta(1) - \theta(0)}
\end{aligned}$$

Let note also that

$$\begin{aligned}
|\epsilon^l| &= \left|-\sum \theta(i,i)p^l(i,1-i) + \sum_{i\neq j}\theta(i,j)p^l(i,j)\right| \\
&= \left|\sum_{i\neq j}(\theta(i,j) - \theta(i,i))p^l(i,j)\right| \\
&\leq \sum_{i\neq j}|\theta(i,j) - \theta(i,i)|\,p^l(i,j) \\
&\leq p^l(0,1) + p^l(1,0)
\end{aligned}$$

3 The algorithm for adaptive signal decomposition

For estimating the parameters a and b we must be able to determine sonorant and obstruency regions for each realization of S_t^l. To do so we have used an algorithm for adaptive signal decomposition recently introduced by Gerd Teschke in [3]; we present here a brief description.

Let $\mathscr{H}$ be a Hilbert space and $f \in \mathscr{H}$, which is assumed to have a sparse representation in terms of a dictionary of bases or frames $\{\varphi_j^i\}_{j\in J_i}$ $(i = 1, 2, \ldots, n)$ in $\mathscr{H}$. For each frame we have the associated analysis operator $F_i : \mathscr{H} \to l_2,\ v \mapsto v^i = \left\{\langle v, \varphi_j^i\rangle\right\}_{j\in J_i}$, with upper frame bound B_i. All the frame operators are

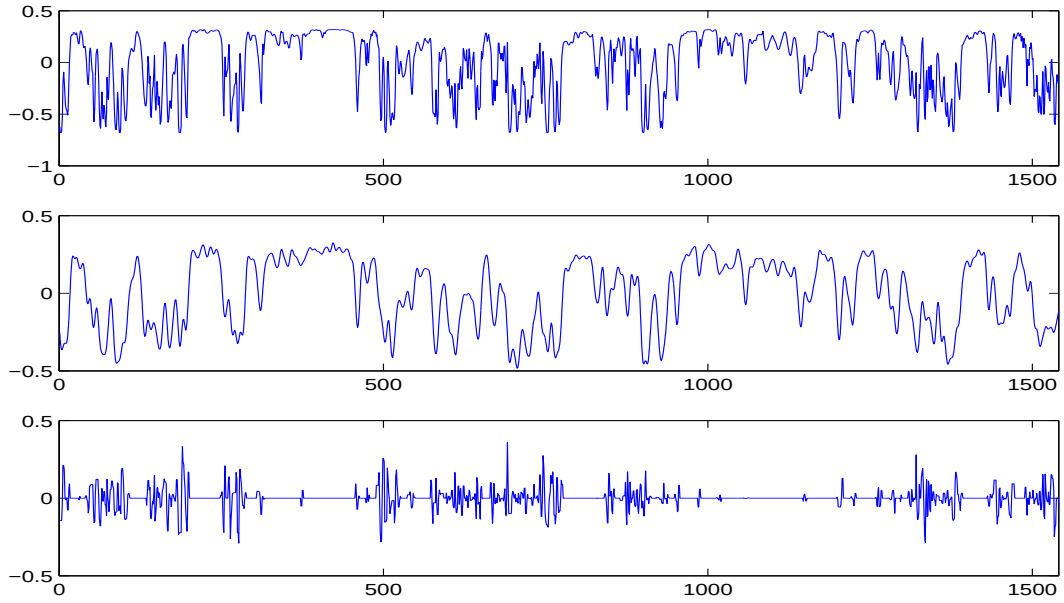

Figure 3: Decomposition of the sonority function into sonorant and obstruency components

related by considering the following reconstruction operator:

$$K : (l_2)^n = l_2 \times ... \times l_2 \to \mathscr{H}$$
$$(v^1, ..., v^n) \mapsto \sum_{i=1}^{n} F_i^* v^i$$

The adjoint operator of K is given by $K^* : \mathscr{H} \to (l_2)^n,\ f \mapsto (F_1 f, ..., F_n f)$

The problem of decomposing f into its different components may be formulated as the minimization of the functional

$$\Phi(c) = \|f - Kc\|^2_{\mathscr{H}} + \alpha_1 \|c^1\|_{l_1} + ... + \alpha_n \|c^n\|_{l_1}.$$

This is a convex functional so there exists a minimum.

Let's introduce some notation: the soft-shrinkage operator is defined by

$$S_t(x) = \begin{cases} x - t\,\mathrm{sign}(x) & \text{if } |x| \geq t \\ 0 & \text{if } |x| < t \end{cases}$$

For some $g \in l_2$ we also introduce the soft-shrinkage operation acting component-wise $S_t(g) = \{S_t(g_j)\}_{j \in J}$ Finally, for some vector of sequences $(g^1, ..., g^n) \in (l_2)^n$ and a multi- parameter $t = (t_1, ..., t_n)$:

$$\mathbf{S}_t(g) = (S_{t_1}(g^1), ..., S_{t_n}(g^n)).$$

The following result proved in [3] gives an algorithm to obtain the decomposition of a signal $f \in \mathscr{H}$ in terms of a family of bases or frames.

Theorem *Let $\{\varphi_j^i\}_{j \in J}$ $(i = 1, 2, ..., n)$ be a family of n frames in $\mathscr{H}$ where the respective analysis operators $F_i : \mathscr{H} \to l_2$ have upper frame bounds B_i and let $C = B_1 + ... + B_n$. Consider $f \in \mathscr{H}$. Then the sequence of iterates*

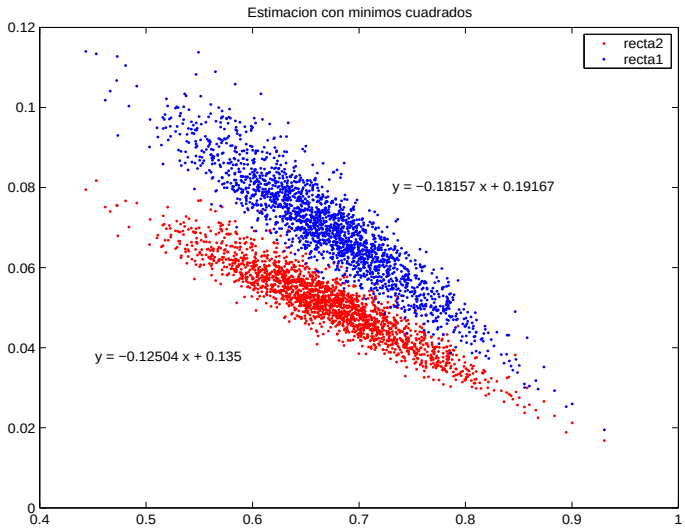

Figure 4:

$$c_{m+1} = \mathbf{S}_{\frac{\alpha}{2C^2}} \left(C^{-2} \left[K^* f + C^2 c_m - K^* K c_m \right] \right), \; m = 0, 1, ...$$

with $c_0 \in (l_2)^n$ *arbitrarily chosen, converges in norm to a minimizer of the functional*

$$\Phi(c) = \| f - Kc \|^2_{\mathscr{H}} + \alpha_1 \|c^1\|_{l_1} + ... + \alpha_n \|c^n\|_{l_1}.$$

In our applications we have used a family of two bases to obtain the decomposition of the sonority function: a Fourier bases, to represent the more regular part (the sonorant component) and a Haar wavelets bases to represent the irregular part (the obstruency component). Figure 3 shows the decomposition of the sonority function associated to the Spanish phrase *Hace cinco minutos que el tren ha llegado a la ciudad*. The upper image represents the original signal (we have worked with the sonority function minus its mean for practical reasons), the middle and the bottom images are the sonorant and obstruency components, respectively. We observe that the obstruency component may be used as a reference to discriminate between sonorant and non-sonorant regions considering the intervals of zeros and the intervals with significant values respectively.

4 The results

In collaboration with Dr. Georgina Flesia from the IAM, the estimation of the model parameters a and b has been performed using the multi-lingual corpus previously described. Figure 4 shows our results together with the empirical results of Figure 2. Specifically, the upper set of points represents the empirical estimation $\hat{\mathbb{E}}(S_t^l)$ against $\hat{\mathbb{E}}(|S_t^l - S_{t+1}^l|)$ while the other set plots the same $\hat{\mathbb{E}}(S_t^l)$ against $\hat{\mathbb{E}}(S_t^l) * \hat{b} + \hat{a}$, where $\hat{a}$ and $\hat{b}$ are our estimations. We can see that our estimations do not fit well the empirical results. Further analysis by Dr. Flesia is in progress to get a deeper insight of this behavior.

References

[1] A. Galves, J. Garcia, D. Duarte and C. Galves (2002). Sonority as a basis for rhythmic class discrimination. Paper presented at *Speech Prosody 2002*, Aix-en-Provence. Can be donwloaded from
`http://www.lpl.univ-aix.fr/sp2002/pdf/galves-etal.pdf`

[2] F. Ramus, M. Nespor and J. Mehler (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, **73**, 265-292.

[3] G. Teschke (2005). Multi-frame representations in linear inverse problems with mixed multi-constraints, *preprint* (submitted). Can be donwloaded from
`http://www.math.uni-bremen.de/~teschke/ps/audio_sparse.ps.gz`