Classifying speech sonority functional data using a projected Kolmogorov-Smirnov approach

Juan Antonio Cuesta-Albertos
Universidad de Cantabria

Ricardo Fraiman*
Universidad de San Andrés

Antonio Galves
Universidade de São Paulo

Jesús Garcia
Universidade Estadual de Campinas

Marcela Svarc
Universidad de San Andrés

December 26, 2005

Abstract

This paper addresses a linguistically motivated question of classification of functional data, namely the statistical classification of languages according to their rhythmic features. This is an important open problem in phonology. The analysis is based on the information provided by the sonority, which is an index of local regularity of the speech signal. Our main tool is the projected Kolmogorov-Smirnov test. This is a new goodness of fit test for functional data. The result obtained supports the linguistic conjecture of the existence of three rhythmic classes.

1 Introduction

It has been conjectured in the linguistic literature that languages are divided into three classes according to their rhythmic properties (Lloyd 1940, Pike 1945, Abercrombie 1967, among others). The intuition was that these classes were characterized by the special role played by the *stress*, or the *syllable*, or the *mora* in the emergence of rhythmic units in the language. This intuition justified the names of *stress-timed*, *syllable-timed* or *mora-timed* associated to the three conjectured classes.

During half a century, neither a precise definition of each class, nor any reliable phonetic evidence of the existence of the classes was presented in the linguistic literature. The situation started changing at the end of the century. First of all, Mehler *et al.* (1996) gave empirical evidence that newborn babies are able to discriminate rhythmic classes. Then Ramus, Nespor and Mehler (1999), from now on RNM, gave

This work was partially supported by Ministerio de Ciencia y Tecnología (Spain), grant BFM2002-04430-C02-02 (J.A.C.), CNPq grant 301301/79 (A.G.) and is part of PRONEX/FAPESP's Project *Stochastic behavior, critical phenomena and rhythmic pattern identification in natural languages* (grant number 03/09930-9) and of CNPq's project *Stochastic modeling of speech* (grant number 475177/2004-5).

AMS 2000 subject classification. Primary : 62M07, 62P99 ; *secondary*: 62M99.

Keywords and phrases: rhythmic classes of languages, sonority, projected Kolmogorov-Smirnov test

Running head: Rhythmic classes identification.

*On leave, Universidad de la República

for the first time evidence that simple statistics of the speech signal could discriminate between different rhythmic classes.

RNM's approach is based on two statistics of the speech signal: the proportion of time spent in vocalic intervals and the empirical standard deviation of the durations of the consonantal intervals, denoted $\%V$ and ΔC, respectively. The choice of these parameters is guided by the following linguistic facts. Languages conjectured to be stress-timed, as English, spend a smaller proportion of time in vocalic intervals, than languages conjectured to be syllable-timed, as Italian. Languages conjectured to be stress-timed display a much bigger variety of types of consonantal intervals than languages conjectured to be syllable-timed. Finally, languages conjectured to be mora-timed, like Japanese, behave as super syllable-timed languages.

Figure 1 shows the averages values of $\%V$ and ΔC on a sample of 20 sentences produced by 4 speakers of each of eight languages, English, Polish, Dutch, Catalan, Spanish, Italian, French and Japanese. It turns out that the empirical values of $\%V, \Delta C$ for the eight languages considered appear to cluster into three groups which correspond precisely to the intuitive notion of rhythmic classes. English, Polish and Dutch conjectured to be stress-timed languages appear together, French, Spanish, Catalan and Italian conjectured to be syllable-timed languages appear in a separate group, and finally, Japanese, conjectured to be moraic, appears isolated.

Insert Figure 1 around here

Successful as it was, RNM's approach has two major drawbacks. First of all, it depends on a previous hand-made identification of the boundaries of the vocalic intervals in the acoustic signal. The problem is that this boundary identification depends in many cases on decisions which are very difficult to reproduce in a homogeneous way.

The second drawback has a linguistic nature. In fact it has been shown by psycho-linguists that babies' ability to discriminate the phonotactic properties of their own language emerge between 6 and 9 months. Therefore the fine-grained discrimination between vowels and consonants necessary to perform the analysis proposed in RNM seems to be beyond their linguistic ability. However Mehler *et al.* (1996) shows that newborn babies are able to discriminate rhythmic classes with a signal filtered at 400Hz. In the signal so severely filtered, it is hard to distinguish nasals from vowels and glides from consonants. This strongly suggests that the discrimination of rhythmic classes by babies relies not on fine-grained distinctions between vowels and consonants, but on a coarse-grained perception of sonority in opposition to obstruency.

This was the motivation for the introduction in Galves *et al.* (2002) of a local index of regularity of the speech signal which was called *sonority*. This index is a function which maps local windows of the acoustic signal on the interval $[0, 1]$. This function assumes values close to 1 when the region displays regular patterns characteristic of sonorant portions of the signal. In contrast, the function will assign values close to 0 to regions characterized by obstruency.

In Galves *et al.* (2002) it was suggested that it was possible to recover the conjectured rhythmic classes directly from the analysis of the trajectories of the sonority in the sample considered in RNM. To give a sound statistical basis to this claim is the goal of the present paper. We refer the reader to Ramus (2002) for an illuminating discussion of the rhythmic class conjecture.

The main tool we will use in what follows is the projected Kolmogorov-Smirnov test which was proposed recently in Cuesta-Albertos, Fraiman and Ransford (2004). This paper is a step forward in the approach initiated in Bélisle *et al.* (1997). The projected Kolmogorov-Smirnov test makes possible to compare the laws of the stochastic processes producing the time evolutions of the sonority for the different sentences and languages.

This paper is organized as follows. In Section 2 we define the sonority, and introduce the data we will analyze. In Section 3 we present the projected Kolmogorov-Smirnov test, which is the main statistical tool that will be used in our analysis. In Section 4 we present the results of the projected Kolmogorov-Smirnov test applied to the linguistic corpus considered in RNM. A general discussion of the issues considered here and perspectives of future research are presented in Section 6. The data sets and computer codes used in this paper can be obtained at the site `www.ime.usp.br/∼tycho/prosody/sonority/KS`.

2 The data

In Galves *et al.* (2002) an index of local regularity of the speech signal was introduced under the name of *sonority*. This is a mapping of the spectrogram of the acoustic signal into a function of time taking values in the interval $[0,1]$. At each time step it is computed the relative entropy between neighboring normalized columns of the spectrogram. A local average of these relative entropies is then mapped through a fixed decreasing function to define the current value of the sonority.

Formally denote by $c_t(f)$ the power spectral density at time t and frequency f. Time is discretized in steps of 2 milliseconds. The values of the spectrogram are estimated using a 25 milliseconds Gaussian window. Only frequencies from 80 Hz to 800 Hz, by steps of 20 Hz, were considered. The normalized power spectral density is defined by

$$p_t(f) = \frac{c_t(f)}{\sum_{f'} c_t(f')} \, .$$

This defines a sequence of probability measures $\{p_t : t = 1, \ldots, T\}$.

The sonority is defined as

$$S(t) = e^{-\beta \sum\limits_{i=1}^{3} h(p_t \mid p_{t-i})} \, ,$$

where h denotes the relative entropy between two probability measures and β is a free parameter taking positive real values.

We recall that the relative entropy for the column p_t with respect to the column p_{t-i} is defined by the formula

$$h\left(p_t | p_{t-i}\right) = \sum_f p_t\left(i\right) \log\left(\frac{p_t\left(f\right)}{p_{t-i}\left(f\right)}\right) \, . \tag{1}$$

The relative entropy is always a positive number (by Jensen's inequality), and it is close to 0 when the probability measures are similar.

Figure 2 shows the synchronized time evolutions of the pressure (top), of the spectrogram (middle) and of the sonority (bottom) for a piece of a Japanese sentence.

Insert Figure 2 around here

The definition of the sonority is motivated by the fact that regular patterns characteristic of sonorant spans typically will correspond to sequences of probability measures which are close in the sense of relative entropy. Therefore if the window around time t covers a region of the acoustic signal which is regular, and therefore sonorant, then $S(t)$ will be close to 1. In contrast, regions in which the acoustic signal present a chaotic behavior, for instance regions corresponding to stop consonants, will correspond to intervals in which $S(t)$ will assume values close to 0, with important variations. It has been observed recently that there exists a strong relationship between the time evolution of the sonority and the time evolution of the intra-oral pressure during the production of speech (cf. Cros *et al.* 2005). From this last paper we borrow the choice of $\beta = 2.5$.

The spectrograms used in the present analysis were produced by the software Praat (`www.praat.org`) and the sonority was calculated using the software Piccolo (`www.ime.usp.br/~tycho/prosody/piccolo`).

The linguistic data we use in the present article is the one analyzed in the Ramus *et al* (1999). It is a set of 160 sentences from eight different languages: English, Polish, Dutch, Catalan, Spanish, Italian, French and Japanese. For each language a set of of 20 sentences was selected among 54 sentences controlled with respect to the number of syllables (from 15 to 21) produced by 4 female speakers. The selection of the sentences was justified by the need to eliminate outliers produced by different rates of speech. To achieve this goal Ramus *et al.* (1999) selected the sentences whose duration is closer to the mean duration of the sentences with the same number of syllables in the set. The sentences were read in a soundproof booth, were low-pass filtered and digitized at 16 kHz and recorded directly in the hard disk. This multi-lingual corpus belongs to the *Laboratoire de Sciences Cognitives et Psycholinguistique (EHESS/CNRS).*

3 The projected Kolmogorov-Smirnov test

Kolmogorov-Smirnov type goodness of fit test has been broadly studied for one dimensional data. However, once we leave the one dimensional setting, the problem becomes a much more difficult task, and there are no quite satisfactory results even for two dimensional data. Recently, Cuesta-Albertos, Fraiman and Ransford (2004), based on the approach in Bélisle *et al.* (1997), have proposed a way to tackle this problem in the infinite dimensional space. The ideas of the results are based on one dimensional projections. In particular, in that paper, a Kolmogorov-Smirnov type goodness of fit is derived, which roughly speaking is based on performing a one dimensional Kolmogorov-Smirnov test for the projections of the data on a randomly selected direction.

In this paper we will use these results to study the sonority sample paths of the sentences in RNM corpus. The procedure is based on the following theorem presented in Cuesta-Albertos *et al.* (2004). In the statement of the theorem $P_{\langle x \rangle}$ stands for the distribution of the projection of P on one-dimensional subspace spanned by x.

Theorem 3.1 (Cuesta-Albertos, Fraiman and Ransford, 2004). *Let H be a separable Hilbert space, and let λ be a non-degenerate Gaussian measure on H. Let P, Q be Borel probability measures on H. Assume that:*

- *the absolute moments of P, $m_n := \int \|x\|^n \, dP(x)$, $n \in I\!N$, are finite and satisfy Carleman's condition*

$$\sum_{n \geq 1} m_n^{-1/n} = \infty \, ;$$

- *the set $\{x \in H : P_{\langle x \rangle} = Q_{\langle x \rangle}\}$, is of positive λ-measure.*

Then $P = Q$.

To apply the above theorem in the classification of the sonority samples, we will consider each sonority path as the realization of a given probability measure defined on the Hilbert space of square integrable functions. This is natural since the sonority sample paths are positive bounded functions. To put all the sonority sample paths in the same space, we will only consider the sample paths in a fixed interval of time $[0, T]$. This means that in what follows we consider the Hilbert space $H = L^2([0, T])$.

Let $\mathcal{L}$ denote the set of 8 languages under consideration. For each $l \in \mathcal{L}$, denote by $\mathcal{S}_l$ the set of recorded sentences from language l in the corpus. I.e.

$$\mathcal{S}_l = \{S^{(l,i)} : i = 1, \ldots, n_l\},$$

where $S^{(l,i)} = (S^{(l,i)}(t))_{0 \le t \le T}$ is the sonority time evolution of sentence i^{th} of language l in the corpus and n_l is the total number of recorded sentences of language l. We assume that the sonority time evolutions corresponding to the different sentences $S^{(l,i)} \in \mathcal{S}_l$ are independent realizations of the same stochastic process

$$S^l = (S^l(t))_{0 \le t \le T}\,.$$

We will be concerned with two-sample goodness of fit problems. Take two different languages $l \neq l'$ and consider the samples $\mathcal{S}_l$ and $\mathcal{S}_{l'}$. We want to check whether the two samples come from the same population. This means to decide between the null hypothesis $P^l = P^{l'}$ against the alternative hypothesis $P^l \neq P^{l'}$, where P^l and $P^{l'}$ stand for the probability laws of the processes S^l and $S^{l'}$ respectively.

As sonority curves are bounded, the distribution which produces them satisfies Carleman's condition and we can apply Theorem 3.1 to this problem. Thus, following Cuesta *et al.* (2004), we will use the following procedure to perform the two-sample test.

- First choose at random a realization of a standard Brownian motion $W = (W(t))_{t \in [0,T]}$. We assume without loss of generality that this Brownian motion is defined in the same probability space as the family of sonority processes. The realization of the Brownian motion will play the role of random direction in which we will project the sonority.

- Then calculate the two samples projected Kolmogorov-Smirnov statistic

$$D_W(\mathcal{S}_l, \mathcal{S}_{l'}) = \sup_{x \in I\!R} \sqrt{\frac{n_l n_{l'}}{n_l + n_{l'}}} \left| \frac{1}{n_l} \sum_{i=1}^{n_l} \mathbf{1}\{\langle S^{(l,i)}, W \rangle \le x\} - \frac{1}{n_{l'}} \sum_{i=1}^{n_{l'}} \mathbf{1}\{\langle S^{(l',i)}, W \rangle) \le x\} \right| . \qquad (2)$$

  In equation (2), $\langle \cdot, \cdot \rangle$ denotes the usual inner product in the Hilbert space $L^2([0,T])$

$$\langle S^{(l,i)}, W \rangle = \int_0^T S^{(l,i)}(t) W(t) dt,$$

  and n_l and $n_{l'}$ are the sizes of the samples $\mathcal{S}_l$ and $\mathcal{S}_{l'}$ respectively.

- Reject the null hypothesis if $D_W(\mathcal{S}_l, \mathcal{S}_{l'})$ is large enough. Otherwise accept it.

The big advantage of this test is that, under the null hypothesis, if the common distribution has continuous projections, then the distribution of $D_W(\mathcal{S}_l, \mathcal{S}_{l'})$ does not depend on the realization W of the Brownian motion. Moreover, even without the continuity hypothesis, the asymptotic distribution of $D_W(\mathcal{S}_l, \mathcal{S}_{l'})$ does not depend on the realization W and it is known to be

$$\lim_{n_l \wedge n_{l'} \to \infty} \mathbb{P}\{D_W(\mathcal{S}_l, \mathcal{S}_{l'}) \le t\} = 1 - 2 \sum_{k=1}^{\infty} (-1)^{k+1} e^{-2k^2 t^2}.$$

Therefore, given a level α, we can find c_α, such that

$$\lim_{n_l \wedge n_{l'} \to \infty} \mathbb{P}\{D_W(\mathcal{S}_l, \mathcal{S}_{l'}) > c_\alpha\} = \alpha,$$

for an asymptotic α-level conditional test.

The test is consistent, *i.e.* under the alternative hypothesis $P^l \neq P^{l'}$ we have

$$\lim_{n_l \wedge n_{l'} \to \infty} \mathbb{P}\left\{D_W(\mathcal{S}_l, \mathcal{S}_{l'}) > c_\alpha\right\} = 1,$$

for almost all realizations W of the Brownian motion.

4 Statistical analysis of the sonority data

The application of the projected Kolmogorov-Smirnov test to the sample of sonority sample paths is made more difficult by the small size of the sample (only 20 sentences for each language). An additional difficulty comes from the short length of the sonority sample paths. In effect, the length of the original sentences is between 1 and 3 seconds. The sentences have been digitized with a sampling rate of 16 kHz. Finally their sonority was computed in steps of 2 milliseconds. Therefore the shortest sonority path has only 800 values. To have all the paths in the sample with the same length we will only consider the first 800 values of each sonority path. Therefore in order to numerically implement the test, all the calculations will be done with this finite grid, with discrete time $t = 1, \dots, 800$.

To make the test more stable, instead of taking only one random direction we will take many of them. For each pair of languages $l \neq l'$ we proceed as follows.

- Choose N independent realizations $W_i, i = 1, ..., N$ of the Brownian motion $W = (W(t))_{t\in[0,T]}$. These directions will remain fixed from now on.
- For each realization W_i, $i = 1, \dots, N$ we test the null hypothesis $P^l = P^{l'}$ at level η by projecting the samples $\mathcal{S}_l$ and $\mathcal{S}_{l'}$ on direction W_i, using the statistic defined in formula (2).
- For each realization W_i, $i = 1, \dots, N$ build up the auxiliary random variable $Z_i(l, l')$ which takes the value 1 if the projected test rejects the null hypothesis, and takes the value 0 otherwise.
- Define the average value

  $$\bar{Z}(l, l') = \frac{1}{N}\sum_{i=1}^{N} Z_i(l, l')$$

  and reject the null hypothesis if $\bar{Z}(l, l') \geq d_\alpha$.

We recall that in our data set the size of the sample of sentences $n_l = 20$ for any $l \in \mathcal{L}$.

The question now is which is the value of d_α which assures that we have a test of level α? The auxiliary random variables $Z_i(l, l'), i = 1, \dots, N$ are equally distributed but not independent (since for each of the N directions we use the same data from the languages). Therefore $\sum_{i=1}^{N} Z_i(l, l')$ is not a binomial random variable and d_α cannot be obtained from a binomial table.

To face this difficulty we will use a bootstrap procedure which mimics the standard bootstrap approach to the two samples problem (cf. Efron and Tibshirani [10]). The validity of this procedure will be discussed in the final section.

Our goal is to obtain an approximated quantile of the distribution of the sum of the Bernoulli variables for a given pair of languages l and l' under the null hypothesis. To do this, we

- First for each bootstrap replication, we build up a pair of independent bootstrap samples of size 20 from the pooled sample $\mathcal{S}_l \cup \mathcal{S}_{l'}$.

- Now we compute the statistic for the two bootstrap samples, using the same N directions that has been fixed, and obtain the bootstrap statistic $\bar{Z}^*(l, l')$.

Finally take d_α for the languages l and l', as the $(1-\alpha)$-quantile of the values $\bar{Z}^*(l, l')$ obtained in the bootstrap replicates.

We applied the test to each pair of languages l and l', with $N = 100, B = 1000, \eta = 0.05$. Tabla 1 reports the p-values of the test for each pair of languages. To simplify the lecture of the table, we present in boldface the entries in which the test at level $\alpha = 0.1$ rejected the null hypothesis of equality of the two populations.

language	pol	ital	fren	span	dut	eng	cat
jap	0.176	**0.009**	**0.094**	**0.087**	**0**	**0.001**	0.456
pol		0.202	1	1	**0.05**	**0.029**	0.234
ital			0.227	0.242	**0.056**	1	0.135
fren				1	**0.091**	**0.029**	0.122
span					**0.09**	**0.086**	0.226
dut						1	**0.02**
eng							**0.05**

Table 1: *Bootstrap p-values for* $N = 100$, $B = 1000$, *and* $\eta = 0.05$. *Significant differences appear in boldface.*

We observe that at level $\alpha = 0.1$, if we only consider five languages (Dutch, English, French, Japanese and Spanish) the test produces three clear clusters. The first cluster contains French and Spanish. The second cluster contains Dutch and English. Finally Japanese appears isolated as the test rejects the null hypothesis that the Japanese sample and any one of the other five samples have been produced with the same law. This clustering is compatible with the linguistic conjecture which classifies Dutch and English as stress-timed languages, French and Spanish as syllable-timed languages and Japanese as a mora-timed language.

Aside of the comparison with English, Italian belongs to the same cluster as Spanish and French, as conjectured in the linguistic literature.

The situation is less clear with respect to Catalan and Polish. The test accepts at level $\alpha = 0.1$ the null hypothesis of identity of Catalan with any of the other languages, with the exception of Dutch and English. This would go in the direction of considering that mora-timed languages are actually super-syllable timed languages, and this is linguistically appealing. However this is incoherent with the distinction between Italian and Japanese even at level .01.

A similar result is obtained with Polish. The test accepts the identity between the law of Polish with all the other languages, with the exception of Dutch and English.

Based on these remarks we will perform a new test by grouping the sonority sample paths in three groups. In the first group we put together the 60 sonority sample paths of the conjectured syllable-timed languages, French, Italian and Spanish. The second group contains the 40 paths of the conjectured stress-timed languages, Dutch and English. Finally the 20 sonority paths of Japanese, which is conjectured to be a mora-timed language, remain in a third group.

We perform the projected test in the same way as before, now having as null hypothesis that groups i and j are samples from the same population, for each pair $i \neq j$, with $i, j = 1, 2, 3$. Table 2 shows the results obtained with $N = 100$, $B = 1000$ and $\eta = 0.05$.

Table 2 shows that the test found significant all the differences between groups. This second test

category	mora–timed	stress–timed
syllable–timed	0.025	0.002
stress–timed	0.00	

Table 2: *Bootstrap p-values for the three groups, with $N = 100$, $B = 1000$ and $\eta = 0.05$*

reinforces the linguistic conjecture of existence of three different rhythmic classes. However, this second test was done with the sonority data of only six languages. The case of Catalan and Polish requires further analysis. We will return to this point in the discussion section.

5 A simulation study

In this section we present a simulation study using a simple family of Markov chains. These chains mimic a quantized version of the sonority. The goal of the section is to provide additional evidence of the ability of the projected Kolmogorov-Smirnov test to classify functional data like the sonority functions.

In our model, (X_t^l) will be Markov chains with two states 0 and 1, representing the low and high sonority zones, respectively. Formally, for each language l we define

$$X_t^l = \mathbf{1}\left\{S^l(t) \geq c\right\},$$

where c is a suitable cut-point.

The transition probabilities will be denoted by

$$\mathbb{P}(X_t^l = y | X_{t-1}^l = x) = p^l(y|x), \quad x, y \in \{0, 1\},$$

while the invariant probability measure $\mathbb{P}(X_t^l = x)$ will be denoted by $p^l(x)$, for $x = 0, 1$.

We will assume that

$$p^l(0|0) = p(0|0)$$

is the same for all languages, while the invariant probability measure p^l, will depend on language l. Obviously this choice determines uniquely the value of the other probability transitions of the chain. Furthermore we will only consider three languages (English, Spanish and Japanese) selected from the three conjectured rhythmic classes.

The choice of only two symbols and the assumption that $p^l(0|0)$ is constant aim to make the model as parsimonious as possible, with only a parameter $p^l(1)$ distinguishing the different chains in the family. Actually, the choice of a binary alphabet is reminiscent of the basic binary linguistic classification of phonemes in two main types: consonants (which have small sonority) and vowels (which have high sonority).

This simplified model is inspired by the binary quantization of the sonority data presented in Cros et al.(2005). From this paper we borrow the value of the cut-point $c = 0.68$.

We estimate the probability transitions of the three binary chains by the usual maximum-likelihood method for Markov chains (cf. Billingsley 1961). Therefore the common probability transition $p(0|0)$ is estimated by the proportion of transitions from 0 to 0 in all languages, *i.e*

$$\hat{p}(0|0) = \frac{\sum_{l\in\mathcal{R}} \sum_{i=1}^{20} \sum_{t=2}^{T^{(l,i)}} \mathbf{1}\left\{X_{t-1}^{(l,i)} = 0, X_t^{(l,i)} = 0\right\}}{\sum_{l\in\mathcal{R}} \sum_{i=1}^{20} \sum_{t=2}^{T^{(l,i)}} \mathbf{1}\left\{X_{t-1}^{(l,i)} = 0\right\}}. \tag{3}$$

The parameter $p^l(1)$ is estimated for each language as the proportion of time the binary sequence corresponding to language l visits state 1, *i.e*

$$\hat{p}^l(1) = \frac{\sum_{l\in\mathcal{R}}\sum_{i=1}^{20}\sum_{t=1}^{T^{(l,i)}} \mathbf{1}\left\{X_t^{(l,i)} = 1\right\}}{\sum_{l\in\mathcal{R}}\sum_{i=1}^{20} T^{(l,i)}} . \tag{4}$$

In formulas (3) and (4) we denoted $X_t^{(l,i)}$ the binary symbol codifying the sonority value of sentence (l, i) of language l at time t and $T^{(l,i)}$ is total length of sentence (l, i). The estimated values are reported in Table 3.

$\widehat{p}\,(0\,\vert 0\,)$
0.93

l	$\widehat{p}^l\,(1)$
English	0.940
Spanish	0.945
Japanese	0.950

Table 3: *Maximum likelihood estimates for the parameters of the Markov model.*

We want to check the ability of the projection test to discriminate the simulated samples of the three Markov chains presented above. For each l we generate 50 independent realizations of length 6000 of the Markov chain $(\tilde{X}_t^l)$. Here $(\tilde{X}_t^l)$ stands for the Markov chain defined by the parameter $p^l(1)$ together with the common probability transition $p(0|0)$. Then we perform the projected Kolmogorov-Smirnov test using a unique direction, generated by a discrete version of the Brownian motion. We repeat the procedure 500 times. The null hypothesis is that the for each pair $l \neq l'$ of the three considered languages, the samples produced by $(\tilde{X}_t^l)$ and $(\tilde{X}_t^{l'})$ have the same law. Table 4 reports the percentage of the 500 replications in which we do not reject the null hypothesis.

	English	Spanish	Japanese
English	0.96	0.36	0.13
Spanish		0.97	0.29
Japanese			0.96

Table 4: *Proportion of the 500 realizations of* $\tilde{X}_t^l$. in which the null hypothesis is not rejected.

6 Discussion

Galves *et al.* (2002) suggested to use the sonority as a tool to discriminate rhythmic classes of languages. The main purpose of the present paper is to give a sound statistical basis to this approach.

The results in Table 2 reinforces the linguistic conjecture of the existence of three different rhythmic classes. The differences between groups found by the test are strongly significant for each pair.

If we consider only six languages (Dutch, English, French, Italian, Japanese and Spanish) the test based on the sonority paths suggests the existence of three clusters. The first one contains French, Italian and Spanish. The second one contains Dutch and English, while Japanese appears isolated as a third cluster. This clustering is compatible with the linguistic conjecture which classifies Dutch and English as stress-timed languages, French, Italian and Spanish as syllable-timed languages and Japanese as a mora-timed language.

The case of Catalan and Polish reveals an important linguistic difficulty. This point has been addressed, for instance, in Ramus *et al.* (1999), from where we quote the following. *"Even though they have most often been described as syllable-timed and stress-timed, respectively, Catalan and Polish (...) are languages whose features match neither those of typical stress-timed languages, nor those of typical syllable-timed languages.(...) Indeed, Catalan has the same syllabic structure and complexity as Spanish, and thus should be syllable-timed, but it also presents the vowel reduction phenomenon, which is consistently associated with stress-timed languages. Polish presents the opposite configuration, namely a great variety of syllable types and high syllabic complexity, like stress-timed languages, but no vowel reduction at normal speech rates. As a matter of fact, phonologists did not reach firm agreement on their rhythmic status."*

To have a closer look at the sensibility of the method when the parameters are close, we increased the length of the sonority sample paths S_t^l from 800 to 6000, ($t = 1, ..., 6000$). It is clear from the Markov structure of the model, that to have paths of larger length is equivalent to increase the sample size in the simulation for paths of smaller length. With a sample path of length 800, the problem becomes harder, in particular when the values of the parameters are very close, like for Spanish and English, or Spanish and Japanese. The sample size was 50 realizations for each language and we performed 500 replications. The results were reported in Table 4. The fact that a simple Markovian model with just one free parameter succeeds that well to fit the linguistic empirical binary sequences is remarkable.

The idea of studying the law of the one-dimensional projection in a direction chosen at random is reminiscent of the classical projection pursuit method in multivariate analysis. The difference is that in our case the direction is chosen at random, according to a non-degenerate Gaussian probability distribution, as proposed in Cuesta *et al.* (2004) (for a comparison between the employed procedure and the projection pursuit method in the one-sample case, see Barrio *et al.*, 2005). This procedure provides a consistent projected Kolmogorov-Smirnov test which is used here to cluster the languages based on their sonority curves.

To make stable the projected Kolmogorov-Smirnov test we have used many directions chosen at random as suggested in Cuesta *et al.* (2004). Those authors considered the maximum of the projected one-dimensional Kolmogorov–Smirnov statistic over the N directions. Instead of the maximum, in this paper we take the average of the auxiliary variables $Z_i, i \in 1, \ldots N$ taking value 1 if the projected Kolmogorov-Smirnov test rejects the null hypothesis at level η on the random direction W_i, and 0 otherwise. This procedure has shown in our case to work better than the maximum. Notice that a drawback of both approaches is that we lose the distribution-free property, since the distribution of both statistics will depend at least on the covariance function of the common underlying distribution. This difficulty was overcomed with a bootstrap procedure.

With respect to the bootstrap procedure, the basic question is: can we ensure that the sample distribution of our statistic $\bar{Z}(l, l')$ is properly approximated by its bootstrap counterpart? In our framework the result is far from being straightforward. For functional data Gine and Zin (1990) have proved a bootstrap version of Donsker Theorem for the empirical process. Some extensions of this result are given in Sheely and Wellner (1992); see also Politis *et al.* (1994). van der Vaart and Wellner (1996, Section 3.9.3) prove several theorems of bootstrap validity, based on the delta method for Hadamard (or Fréchet) differentiable statistical functionals taking values in normed spaces. These results can be used to show the bootstrap validity in our framework. See also Cuevas, Febrero and Fraiman (2005). However this goes far beyond the scope of the present work.

7 Acknowledgments

We thank Marzio Cassandro, Pierre Collet, Denise Duarte, Emmanuel Dupoux, Jacques Mehler, Marina Nespor, Janet Pierrehumbert, Frank Ramus and Sharon Pepperkamp for many illuminating discussions. Special thanks to Frank Ramus for making the data used in this paper available to us.

References

[1] Abercrombie, D. (1967) *Elements of general phonetics*, Chicago: Aldine.

[2] Barrio, del E., Cuesta-Albertos, J.A.,Fraiman, R. and Matrán, C. (2005). The random projection method in goodness of fit for functional data. *Manuscript.*

[3] Bélisle, C.,Massé, J. and Ransford, T. (1997) When is a probability measure determined by infinitely many projections? *Annals of Probability*, **25**, 767-786.

[4] Billingsley, P. (1961) Statistical methods in Markov chains. *Annals of Mathematical Statistics* , **32**, 12 - 40.

[5] Cover, T. and Thomas, J. A. (1991) *Elements of information theory.* John Wiley & Sons, Inc., New York

[6] Cros A., Demolin D., Flesia G. and Galves A. (2005) On the relationship between intra-oral pressure and speech sonority, *Interspeech 2005- Eurospeech*, Lisbon.

[7] Cuesta-Albertos, J. A., Fraiman and R. and Ransford, T. (2004) A sharp-form of the Cramér-Wold theorem. *Manuscript.*

[8] Cuevas, A., Febrero, M. and Fraiman R. (2005) On the use of the bootstrap for estimating functions with functional data. *Computational Statistics and Data Analisis*, to appear.

[9] Dudley, R. M. (1990) Nonlinear functionals of empirical measures and bootstrap. In: Eberlein, E., Kuelbs, J., Marcus, M.B. (Editors) *Probability in Banach Spaces* **7** 63-82, Birkhauser, Boston.

[10] Efron, B. and Tibshirani, R. J. (1993) An introduction to the bootstrap. *Monographs on Statistics and Applied Probability*, **57**, Chapman and Hall, New York.

[11] Galves, A., Garcia, J., Duarte, D. and Galves, C. (2002) Sonority as a basis for rhythmic class discrimination. In *Speech Prosody 2002*, Aix-en-Provence.

[12] Giné, E.and Zinn, J., (1990) Bootstrappring general empirical measures, *Annals of Probability* **18** 851-869.

[13] Gill, R.D. (1989) Non- and semi-parametric maximun likelihood estimators and the Von Mises method. *Scandinavian Journal of Statistics* **16** 97-128.

[14] Lloyd, J. (1940) *Speech signal in telephony*, London.

[15] Mehler, J., Dupoux, E., Nazzi, T. and Dehaene-Lambertz, G. (1996) Coping with linguistic diversity: the infant's viewpoint. *Signal to syntax: bootstrapping from speech to grammar in early acquisition*, J.L. Morgan and K. Demuth, eds.

[16] Pike, K.L. (1945) *The intonation of American English*, Ann Arbor: University of Michigan Press.

[17] Politis, D.N. and Romano, J.P. (1994) Limit theorems for weakly dependent Hilbert space valued random variables with application to the stationary bootstrap. *Statistica Sinica* **4** 461-476.

[18] Praat program and manuals. Can be downloaded from `www.praat.org`.

[19] Ramus, F. (2002) Acoustic correlates of linguistic rhythm: perspectives. *Speech Prosody 2002*, Aix-en-Provence.

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

[21] Sheely, A. and Wellner, J.(1992) Uniform Donsker classes of functions. *Annals of Probability* **20** 1983-2030.

[22] van der Vaart, A. and Wellner, J. (1996) *Weak Convergence and Empirical Processes.* Springer, New York.

Juan Antonio Cuesta-Albertos
Departamento de Matemáticas, Estadística y Computación
Universidad de Cantabria
Avda. los Castros s.n.
39005 Santander, Spain
e-mail: cuestaj@unican.es

Ricardo Fraiman
Departamento de Matemáticas
Universidad de San Andrés
Vito Dumas, 284
1644 Victória, Argentina
e-mail: rfraiman@udesa.edu.ar

Antonio Galves
Instituto de Matemática e Estatística,
Universidade de São Paulo
Rua do Matão, 1010,
05508-090 São Paulo SP, Brasil
e-mail: galves@ime.usp.br

Jesús Garcia
Instituto de Matemática, Estatística e Cálculo Científico,
Unicamp
Cidade Universitária *Zeferino Vaz*,
6166 Campinas SP, Brasil
e-mail: jg@ime.unicamp.br

Marcela Svarc
Departamento de Matemática
Universidad de San Andrés
Vito Dumas, 284
1644 Victória, Argentina
e-mail: msvarc@udesa.edu.ar

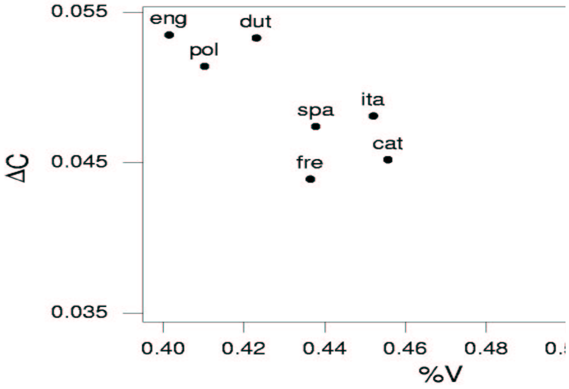

Figure 1: *Distribution of languages on the ($\%V, \Delta C$) plane (from Ramus et al. 1999).*

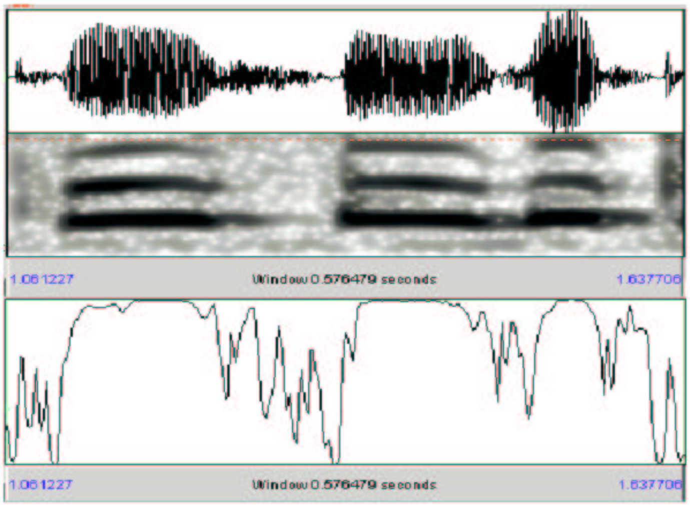

Figure 2: *Graphs of the acoustic signal (top), spectrogram (middle) and sonority (bottom) for a Japanese utterance. The horizontal axis represents time.*