Identifying rhythmic classes of languages using their sonority: a 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

May 15, 2005

Abstract

In this paper ...

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 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.

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 : 62M99 ; *secondary*: 62P99.

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

Running head: Rhythmic classes identification.

*On leave, Universidad de la República

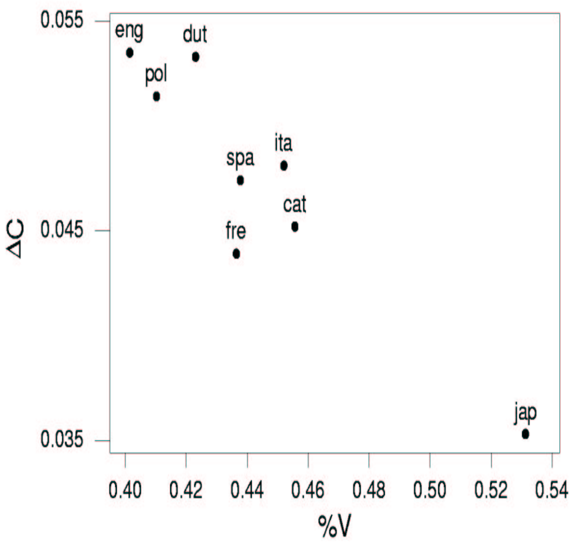

Figure 1: *Distribution of languages on the $(\%V, \Delta C)$ plane, based on Ramus et al.(1999)*

.

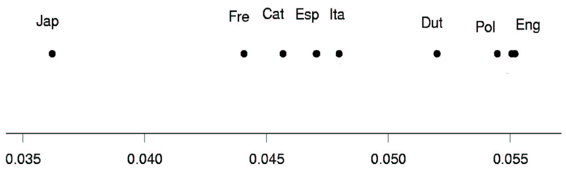

Figure 2: *Estimated standard deviation of the Gamma distribution for consonantal intervals*

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.

The statistics performed in RNM are of descriptive nature. This point was improved in Duarte et al. (2001) who proposed a parametric model that closely fits the data in RNM which made possible to perform an inferential analysis of the data. In this model the durations of the successive consonantal intervals are independent and identically Gamma distributed random variables. The model assumes that languages differ with respect to the values of the parameters of their Gamma distributions. In this framework the rhythmic classes conjecture can be rephrased as follows: the standard deviation of the Gamma is constant for all languages belonging to the same rhythmic class, but the standard deviations of different classes are different. With this model is possible to test the hypothesis that the 8 languages considered above are clustered as suggested in RNM descriptive statistics. The data support the model. The hypothesis that the standard deviations of Gamma distributions are constant within classes and differ among classes is compatible with the data presented in RNM.

Figure 2 shows the estimated standard deviations of the duration of consonantal intervals, for the 8 languages, using the Gamma distribution. The values of the standard deviations presented were obtained by maximum likelihood estimation. The figure displays the same three clusters already present in RNM's descriptive statistics.

Succesfull 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 main goal of the present paper.

The main tool we will use in what follows is the projected Kolmogorov-Smirnov test which follows from a recent result presented in Cuesta-Albertos, Fraiman and Ransforf (2004). This makes it possible to compare the laws of the stochastic processes producing the time evolutions of the sonority for the different sentences and languages.

Up to this point we are working with a non parametric point of view. Recently a parametric model for the family of stochastic processes producing the sonority time evolutions of the different languages was proposed by Cassandro *et al.*(2005) . This model is a family of tied quantized chains. The chains are tied together by the the assumption that there is a universal partition of the sonority domain, such that the distribution of the sonority, conditioned on each interval of the partition is language independent. In Cassandro *et al.*(2005) a consistent cross-linguistic estimator for the cut-points separating these intervals was also introduced.

It follows from this model that all the relevant linguistic information concerning the sonority should be retrieved from a symbolic stochastic chain taking values on a finite alphabet. This symbolic chain can be derived directly from the sonority process. In particular, the most important linguistic question of the existence of rhythmic classes should be decided using only the properties of the symbolic chains. The investigation of this issue is the second goal of this paper.

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. Section **??** recalls the notion of family of tied quantized chains and apply the projected Kolmogorov-Smirnov test to the symbolic chains obtained using the universal quantization suggested by this model. Section 6 uses a simple family of tied Markovian chains to check the adequacy of our classification procedure. A general discussion of the issues considered here and perspectives of future research are presented in Section 7. The data sets and computer codes used in this paper can be obtained at the site `www.ime.usp.br/~tycho/prosody/sonority/K-S classification`.

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

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

where h denotes the relative entropy between two probability measures and η is a free parameter taking positive real values. Following Cassandro *et al.* we take $\eta = 2.5$. This choice was guided by empirical considerations.

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 3 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.

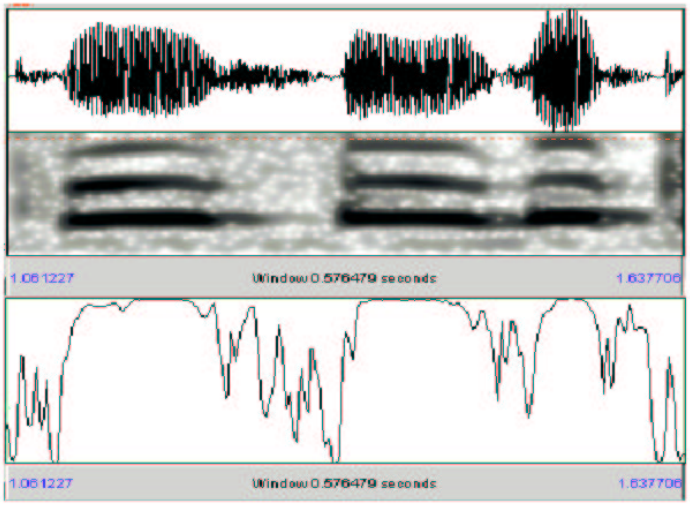

Figure 3: Graphs of the acoustic signal (top), spectrogram (middle) and sonority (bottom) for a Japanese utterance. The horizontal axis represents time.

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.

The spectrograms used in the present analysis were produced by the software Praat (`www.praat.org`). The computations of the sonority from the spectrogram and some basic statistics were carried out using the Free software Piccolo developed by Jesus Garcia.

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. 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) 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 from the linguistic data. 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 L 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 $\mathcal{H} = L^2([0, T])$.

Let $\mathcal{L}$ denote the set of 8 languages under consideration. For each $l \in \mathcal{L}$, denote by $\mathcal{U}_l$ the set of recorded sentences from language l in the corpus. It will be convenient to use the representation

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

where (l,i) denotes the i^{th} recorded sentences of language l in the corpus and n_l is the total number of recorded sentences of language l. Denote by

$$S^{(l,i)} = (S^{(l,i)}(t))_{0 \leq t \leq T}$$

the sonority time evolution of sentence (l,i). We assume that the sonority time evolutions corresponding to the different sentences $(l,i) \in \mathcal{U}_l$ are independent realizations of the same stochastic process

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

We will assume that these processes are stationary and ergodic. We will also assume that these processes are defined on the same, rich enough probability space.

We will be concerned with two-sample goodness of fit problems. To be more precise, for any language $l \in \mathcal{L}$ let

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

be the sample with all the sonority sample paths corresponding to the sentences in $\mathcal{U}_l$. Now 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.

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 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

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

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

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

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

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

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 $t = 1, \ldots, 800$.

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 $W = (W(t))_{t \in [0,T]}$.
- For each realization W_i, $i = 1, \ldots, N$ 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, $i = 1, \ldots, N$ 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 $\bar{Z}(l, l') \geq c_\alpha$.

. 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 c_α which assures that we have a test of level α? The auxiliary random variables $Z_i(l, l'), i = 1, \ldots, N$ 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 c_α cannot be obtained from a binomial table.

To face this difficulty we will use a bootstrap procedure. We want a quantile of the distribution of the sum of the Bernoulli variables for a given language l under the null hypothesis. To do this, we first fix the N directions we have chosen to perform the projected test. Next we bootstrap the statistic $\bar{Z}(l, l)$ under the null hypothesis. We proceed as follows. For each bootstrap replication, build up a pair of independent bootstrap samples of $\mathcal{S}_l$. Now we compute the statistic on the two bootstrap samples and obtain the bootstrap statistic $\bar{Z}^*(l, l)$.

More precisely we proceed as follows.

- Fix the N independent realizations $W_i : i = 1, ..., N$ of the Brownian motion $W = (W(t))_{t \in [0,T]}$.
- For each $b = 1, \ldots, B$, choose independent and uniformly distributed random indices taking values in the set $\{1, \ldots, 20\}$ I_i^b and J_i^b, $i = 1, \ldots, 20$. With these indices construct the two bootstrap samples

  $$\mathcal{S}_l^{\star, b} = \left\{ S^{(l, I_i^b)} : i = 1, \ldots, 20 \right\}$$

  and

  $$\mathcal{T}_l^{\star, b} = \left\{ S^{(l, J_i^b)} : i = 1, \ldots, 20 \right\} .$$

- For each realization W_i compute $D_{W_i}(\mathcal{S}_l^{\star, b}, \mathcal{T}_l^{\star, b})$ and define the corresponding $Z_i^{\star, b}(l)$ as 1 if the test rejects the null hypothesis at level η and as 0 otherwise.

- Define the average of the bootstrap auxiliary random variables

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

and define $c^\star_\alpha(l)$ as the $1-\alpha$ quantile of the vector

$$\bar{Z}^{\star,1}(l,l),\ldots,\bar{Z}^{\star,B}(l,l)\,.$$

In this way we obtain a bootstrap critical value $c^\star_\alpha(l)$ for each language l. To perform a two sample test for two given languages $l \neq l'$ we will use the maximum between $c^\star_\alpha(l)$ and $c^\star_\alpha(l')$ as critical value.

We applied the test to each pair of languages, with $N = 100, B = 1000, \eta = 0.05$ and for $\alpha = 0.1$ and $\alpha = 0.05$. Table 1 reports the result of the test for each pair of languages. In the table entry (l,l') presents the value of $\bar{Z}(l,l')$. The critical values calculated via bootstrap are reported on the last two columns and the last two rows. The critical values $c^\star_\alpha(l)$ appears both at entries $(l, c^\star_\alpha(l))$ and $(c^\star_\alpha(l), l)$. The test rejects the null hypothesis $P^l = P^{l'}$ at level α when the value at entry (l,l') is larger than $\max\{c^\star_\alpha(l), c^\star_\alpha(l')\}$. 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	$c^\star_{0.05}(l)$	$c^\star_{0.1}(l)$
jap	0.04	**0.43**	**0.09**	**0.08**	**0.77**	**0.74**	0.01	0.13	0.06
pol		0.03	0.0	0.0	**0.60**	**0.21**	0.03	0.13	0.05
ital			0.03	0.02	**0.14**	0.0	0.05	0.13	0.05
fren				0.0	**0.50**	**0.19**	0.06	0.12	0.06
span					**0.41**	**0.08**	0.03	0.11	0.05
dut						0.0	**0.74**	0.14	0.05
eng							**0.58**	0.12	0.04
$c^\star_{0.05}(l)$	0.13	0.13	0.12	0.11	0.14	0.12	0.12		
$c^\star_{0.1}(l)$	0.05	0.05	0.06	0.05	0.05	0.04	0.05		

Table 1: *Values of $\bar{Z}(l,l')$ and bootstrap critical 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 six languages (Dutch, English, French, Italian, Japanese and Spanish) the test produces three clusters. The first cluster contains French, Italian 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, Italian and Spanish as syllable-timed languages and Japanese as a mora-timed language.

The situation is less clear with respect to Catalan and Polish. The test accepts at both levels $\alpha = 0.05$ and $\alpha = 0.1$ 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 at both levels .05 and .1.

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 $\eta = 0.05$. We only report the result at level $\alpha = 0.05$ which is highly significant.

category	mora–timed	stress–timed	$c^{\star}_{0.05}(i)$
syllable–timed	0.32	0.70	0.24
stress–timed	0.82		0.09
$c^{\star}_{0.05}(j)$	0.06	0.04	

Table 2: *Values of* $\bar{Z}(i,j)$ *and bootstrap critical values for the three groups, 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 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 final section.

5 Statistical analysis of the quantized sonority

The notion of family of tied quantized chains was introduced in Cassandro *et. al.* (2005) as a model for the sonority time evolutions of a family of languages. The basic assumption of the model is the following.

Assumption There exist a positive integer N and an increasing sequence of cut-points $c_0 = 0 < c_1 < \ldots < c_N < c_{N+1} = 1$ and $N+1$ probability measures π_j, $j = 0, \ldots, N$, such that the support of π_j is contained in the interval $I_j = [c_j, c_{j+1}[$ and that at any time step t and for any $l \in \mathcal{L}$ we have

$$\mathbb{P}\left\{S^l_t \in B | S^l_t \in I_j\right\} = \pi_j(B) \ , \tag{3}$$

where B is any Borel subset of $[0, 1]$.

Here we are using the same notation as in section 3, namely $S^l = (S^l(t))_{0 \leq t \leq T}$ is the stochastic process which correspond to the sonority time evolution of language l and $\mathcal{L}$ is the set of eight languages represented in our sample.

By assumption, the cut-points c_j and the probabilities π_j, $j = 0, \ldots, N$ are independent of l. The intervals I_j will represent regions of different sonority levels.

The universal cut-points suggest the following natural quantization of the sonority process S^l. Let $\left(X^l_t\right)_{t \in \mathbf{Z}}$ be the chain taking values on the finite alphabet $\mathcal{A} = \{0, \ldots, N\}$, defined as follows

$$X^l_t = j \quad \text{if} \quad S^l_t \in I_j \ .$$

This chain tells which regions are successively visited by the sonority time evolution. The assumptions on $\left(S^l_t\right)$ imply that the chains $\left(X^l_t\right)$ are stationary and ergodic. Call

$$p^l(j) = \mathbb{P}\left\{X^l_t = j\right\} \ ,$$

the stationary marginal probability distribution of the chain.

The linguistic intuition behind this model is that all the linguistic relavant information concerning language l should be retrieved from the symbolic chain $\left(X^l_t\right)$. In this model the universality of the sonority regions and the corresponding probability distributions mimics the fact that the physiological features of the speech production apparatus are common to all human beings and therefore are language independent.

Cassandro *et. al.* (2005) introduced the following consistent procedure to estimate the universal cut-points c_j. For any pair of languages $l \neq l'$, any $T \geq 1$ and any $r \in [0,1]$ define

$$\widehat{W}_T^{l,l'}(r) = |\widehat{F}_T^l(r) - \widehat{F}_T^{l'}(r)| \ ,$$

where

$$\widehat{F}_T^l(r) = \frac{1}{T}\sum_{t=1}^{T}\{S_t^l \leq r\}$$

is the empirical marginal distribution of the sonority process $(S_t^l)_{t\geq 0}$. Define also

$$\widehat{c}_T^{\,l,l'} = \inf\left\{ v \in [0,1] \,\middle|\, \widehat{W}_T^{l,l'}(v) = \sup_r \widehat{W}_T^{l,l'}(r) \right\} \ .$$

Theorem (Cassandro *et al.* 2005). Assume that each probability π_j has no atom and that its support is the full interval I_j. If $p^l(j) \neq p^{l'}(j)$, for any $j \in \mathcal{A}$, then $\widehat{c}_T^{\,l,l'}$ converges almost surely to one of the universal cut-points, as T diverges.

To use the theorem when $N \geq 2$, Cassandro *et al.*(2005) consider the maxima of the empirical conditional functions $\widehat{W}_T^{l,l'}(r \mid [a_i, b_i])$ where the open intervals (a_i, b_i) form a covering of $(0,1)$. With this procedure, they identified four cut-points, which are estimated as $c_1 = 0.19$, $c_2 = 0.46$, $c_3 = 0.67$ and $c_4 = 0.93$.

Let us now apply the projected Kolmogorov-Smirnov test to the symbolic chains $(X_t^l)_{t\in\mathbf{Z}}$ taking values in the alphabet $A = \{0,1,2,3,4\}$ obtained from this quantization.

To have the data in the same functional space as before, we identify each symbolic chain produced by the quantization with a step function. Then we proceed as before, using the same $N = 100$ random directions, and doing exactly the same bootstrap procedure as in section 4 with $B = 1000$ and $\eta = 0.05$. Table 3 shows the results of the tests. As before, table entry (l, l') presents the value of $\bar{Z}(l, l')$. The critical values calculated via bootstrap are reported on the last two columns and the last two rows. The test rejects the null hypothesis $P^l = P^{l'}$ at level α when the value at entry (l, l') is larger than $\max\{c_\alpha^\star(l), c_\alpha^\star(l')\}$. 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	$c^\star_{0.05}(l)$	$c^\star_{0.1}(l)$
jap	0.04	**0.34**	**0.07**	**0.08**	**0.79**	**0.73**	0.02	0.15	0.06
pol		0.07	0	0	**0.73**	**0.26**	0.04	0.13	0.05
ital			0.03	0.02	**0.12**	0	0.02	0.18	0.08
fren				0	**0.48**	**0.09**	0.03	0.14	0.07
span					**0.37**	0.05	0.02	0.13	0.06
dut						0.01	**0.67**	0.11	0.06
eng							**0.5**	0.11	0.04
$c^\star_{o.05}(l)$	0.13	0.18	0.14	0.13	0.11	0.11	0.14		
$c^\star_{0.1}(l)$	0.05	0.08	0.07	0.06	0.06	0.04	0.06		

Table 3: *Values of* $\bar{Z}(l, l')$ *and bootstrap critical values for* $N = 100$, $B = 1000$ *and* $\eta = 0.05$. *Significant differences appear in boldface.*

With just the exception of entry (Spanish, English) all the other tests for the quantized chains give exactly the same results as the test with the sonority paths. Moreover the values at each entry are very close in both tables 1 and 3. This supports the idea that the quantized chains summarizes the relevant information conveyed by the sonority paths. The discussion of this issue was the second goal of this paper.

6 A simulation study

In this section we will present a very simple family of Markov chains which mimics in a very realistic way the behavior of the family of quantized sonority chains derived from the data. The goal of the section is to provide additional evidence of the validity of our approach. As a by-product we obtain an interesting simple model for the complex linguistic data.

In our model the quantized chains (X_t^l) will be Markov chains with two states 0 and 1, representing the low and high sonority zones, respectively. The transition probabilities will be denoted by

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

while the invariant probability measure $P(X_t^l = x)$ will be denoted by $p^l(x)$, for $i = 0, 1$. We will assume that

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

is the same for all languages, while the invariant probability measure $p^l(\dot{})$, 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 regions instead of the five regions suggested by the analysis developed in Cassandro *et al*(2005) aims to make the model as parsimonious as possible, with only a parameter $p^l(1)$ distinguishing the different chains in the family. Actually, this choice 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).

To obtain a binary quantization out from the data, we re-codify the empirical quantized chains, by replacing the former symbols 0,1 and 2 by symbol 0 and the former symbols 3 and 4 by symbol 1. Now we use the re-codified sequences to estimate the parameters by the usual likelihood method. More precisely, let $\mathcal{R}$ be the subset of $\mathcal{L}$ containing only the the three chosen languages, English, Japanese and Spanish. Assuming that the parameter $p(0|0)$ is the same for all languages, then its estimated value $\hat{p}(0|0)$ is just the percentage of all transitions from 0 to 0 in all languages, *i.e*

$$\hat{p}(0|0) = \frac{\sum_{l\in\mathcal{R}} \sum_{i=1}^{20} \sum_{t=2}^{T^{(l,i)}} \mathbf{1}\left\{X_{t-1}^{(l,i)} = 0, X_t^{(l,i)} = 0\right\}}{\sum_{l\in\mathcal{R}} \sum_{i=1}^{20} \sum_{t=2}^{T^{(l,i)}} \mathbf{1}\left\{X_{t-1}^{(l,i)} = 0\right\}}. \qquad (4)$$

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

$$\hat{p}^l(1) = \frac{\sum_{l\in\mathcal{R}} \sum_{i=1}^{20} \sum_{t=1}^{T^{(l,i)}} \mathbf{1}\left\{X_t^{(l,i)} = 1\right\}}{\sum_{l\in\mathcal{R}} \sum_{i=1}^{20} T^{(l,i)}}. \qquad (5)$$

In formulas (4) and (5) 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 4.

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

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

Table 4: *Maximun likelihood estimates for the parameters of the Markov model.*

We want to check the ability of the projection test to discriminate the samples of the three Markov chains presented above. For each $l \in \mathcal{R}$ 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 languages in $\mathcal{R}$, the samples produced by $(\tilde{X}_t^l)$ and $(\tilde{X}_t^{l'})$ have the same law. Table 5 reports the 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 5: Percentage of times in which we do not reject the null hypothesis among the 500 realization of $\tilde{X}_t^l$.

Finally, for each $l \in \mathcal{R}$, we will compare the sample with the empirical binary sequences $(X_t^{l,i}, t = 1, \ldots, 800)$, for $i = 1, \ldots, 20$ with a sample of 20 independent realizations of length 800 of the Markov chain $(\tilde{X}_t^l)$ generated with parameters presented in Table 5. Notice that in order to have realizations of the empirical and simulated chains with the same length we only considered the first 800 steps of the empirical chains. The choice of 800 follows from the fact that the minimal length of the empirical chains in the sample is precisely 800.

For each simulated sample we perform the projected Kolmogorov-Smirnov test using just one random direction generated by a discrete version of the Brownian motion. The null hypothesis is that the simulated and the real binary chains for each language have the same law. We consider 500 independent realizations of the samples of the Markov model, and for each one of the realizations we test the null hypothesis that the simulated sample and the empirical sample have the same law. Table 6 summarizes the results of the tests. The first column presents the average p-values obtained with the 500 replications. The second column shows the percentage of the replications in which we reject the null hypothesis.

	Average p-value	Rejection %
English	0.51	0.04
Spanish	0.65	0.01
Japanese	0.45	0.11

Table 6: *Goodness of fit of the Markov model*

The goal of this comparison is to investigate the adequacy of the order one Markov approximation to the law of the symbolic linguistic chains. The general conclusion is that this very simple Markovian is quite adequate for the languages we consider. However it is also clear that the Markovian model fits better the empirical English sequences than the empirical Japanese sequences. We will further develop this analysis in the final discussion.

7 Discussion

The main purpose of the present paper was to give a sound statistical basis to the claim rhythmic class discrimination using the sonority suggested in Galves *et al.* (2001).

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 directions are chosen at random, according to a non-degenerate Gaussian probability distribution. In this way, we are able to use the results in Cuesta *et al* (2004), that provide a consistent projected Kolmogorov-Smirnov test.

In our study, in order to make stable the projected Kolmogorov-Smirnov test we have used many directions chosen at random instead of just one. This has already been suggested in Cuesta *et al* (2004). A possibility considered there is to take

$$D(\mathcal{S}_l,\mathcal{S}_{l'})^{[N]} :=_{\max 1\leq i\leq N} D_{W_i}(\mathcal{S}_l,\mathcal{S}_{l'}),$$

the maximum of the projected one-dimensional Kolmogorov–Smirnov statistics over the N directions. Instead of the maximum of the projected one-dimensional Kolmogorov-Smirnov statistics over the N directions, 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 second way to make the test more stable, using many random directions, has shown in our case to work better. Notice that a drawback of both approaches is that we lose the distribution-free property, since the distribution of $D(\mathcal{S}_l,\mathcal{S}_{l'})^{[N]}$ will depend on the covariance function of the common underlying distribution. However, a bootstrap procedure is implemented to overcome this problem.

If we consider only six languages (Dutch, English, French, Italian, Japanese and Spanish)the test based on the sonority paths suggest 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 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.

The case of Catalan and Polish requires further analysis. Both languages can only be distinguish from Dutch and English by our test. On one hand, this would go in the direction of considering that mora-timed languages are actually super-syllable timed languages, and this is linguistic appealing. On the other hand, this is incoherent with the distinction between Italian and Japanese at both levels .05 and .1. A possible explanation of this phenomena is that the phrases chosen in Ramus *et al (1999)*, have still a great variability for each language, particularly within this two languages. A more detailed study based on a different criteria to select a subset of phrases of each language will be considered on a forthcoming paper.

The two models considered seems to be reasonably adequate.

With the data from the tied quantized chains we mainly replicate the results obtained with the sonority paths, except for the comparison between Spanish and English. Observe that the results in Table 3 are almost the same as in Table 1.

In the simple Markovian model used in the simulation presented in section 6 we used only two regions instead of the five regions suggested by the analysis developed in Cassandro *et al* (2005). This simplification aimed to make the model as parsimonious as possible, with only a parameter $p^l(1)$ distinguishing the different chains in the family. Actually, this choice 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). The fact this simplified model behaves in a realistic way raises the question of the necessity of the model with 5 symbols. This issue is discussed in a forthcoming paper.

To have a closer look at the sensibility of the method when the parameters are close, instead of increasing the sample size we have increase the length of the sonority sample paths S^l_t from 800 to 6000, $(t = 1, ..., 6000)$. The sample sizes was 50 realizations for each language and we perform 500 replications. The results were reported in Table 5. 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 fact that a simple Markovian model with just one free parameter succeeds that well to fit the linguistic empirical binary sequences is remarkable.

It is clear that since the parameters associated to the three different languages by this simple order one Markov chain model are very close we will need large samples of the chains to achieve this task.

8 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] Cassandro, M., Collet, P., Duarte, D., Galves, A., and Garcia, J. (2005). A stochastic model fot the speech sonority: tied quantized chains and cross-linguistic estimation of the cut-points. *Manuscript.*

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

[4] Duarte, D, Galves, A., Lopes, N. and Maronna, R.(2001). The statistical analysis of acoustic correlates of speech rhythm. Paper presented at the *Workshop on Rhythmic patterns, parameter setting and language change*, ZiF, University of Bielefeld. Can be downloaded from http://www.physik.uni-bielefeld.de/complexity/duarte.pdf

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

[6] Lloyd, J. 1940. *Speech signal in telephony*, London.

[7] Mehler, J.; Dupoux, E.; Nazzi, T.; 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.

[8] Nazzi, T., Bertoncini, J. and Mehler, J. (1998). Language discrimination by newborns: towards an understanding of the role of rhythm. *J. Experimental Psychology: human perception and performance*, **24**, 756-786.

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

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

[11] Ramus, F. (2002) Acoustic correlates of linguistic rhythm: perspectives. *Speech Prosody 2002*, Aix-en-Provence. Can be download from `www.lpl.univ-aix.fr/sp2002/pdf/ramus.pdf`.

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

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
Departamiento de Matemática
Universidad de San Andrés
Vito Dumas, 284
1644 Victória, Argentina
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
galves@ime.usp.br

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

Marcela Svarc
Departamiento de Matemática
Universidad de San Andrés
Vito Dumas, 284
1644 Victória, Argentina
msvarc@udesa.edu.ar