Sonority as a Basis for Rhythmic Class Discrimination

Antonio Galves[1], *Jesus Garcia*[1], *Denise Duarte*[1] *& Charlotte Galves*[2]

Instituto de Matemática e Estatística, University of São Paulo [1]
Instituto de Estudos da Linguagem, University of Campinas [2]
{galves; jesusg; dduarte}@ime.usp.br, galvesc@obelix.unicamp.br

Abstract

Recently, several papers, starting with Ramus, Nespor and Mehler (1999), gave evidence that simple statistics of the speech signal could discriminate between different rhythmic classes.In the present paper, we propose a new approach to the problem of finding acoustic correlates of the rhythmic classes. Its main ingredient is a rough measure of sonority defined directly from the spectrogram of the signal. This approach has the major advantage that it can be implemented in an entirely automatic way, with no need of previous hand-labelling of the acoustic signal. Applied to the same linguistic samples considered in RNM, it produces the same clusters corresponding to the three conjectured rhythmic classes.

1. Introduction

It has been conjectured in the linguistic literature that languages are divided into different classes according to their rhythmic properties (Lloyd 1940, Pike 1945, Abercrombie 1967, among others). During half a century, no reliable phonetic evidence was presented to support this claim. Recently, several papers, starting with Ramus, Nespor and Mehler (1999), from now on RNM, gave evidence that simple statistics of the speech signal could discriminate between different rhythmic classes (cf. Grabe and Low 2001, Gibbon and Gut 2001, Frota and Vigário 2001).

The great interest of this line of research is well exemplified by the three clusters of languages suggested by the statistics presented in RNM, which correspond to the three rhythmic classes which have been conjectured in the linguistic literature. However the empirical basis of these approaches is far from being satisfactory since the statistical treatment is only of a descriptive nature and it is based on very small samples. Another problem for these analyses is that they are based on hand-made labelling of the acoustic signal which depends in many cases on decisions which are very difficult to reproduce in a homogeneous way.

In the present paper, we propose a new approach to the problem of finding acoustic correlates of the rhythmic classes. Its main ingredient is a rough measure of sonority defined directly from the spectrogram of the signal. Applied to the same linguistics samples considered in RNM, it produces the three conjectured clusters. This approach has the major advantage that it can be implemented in an entirely automatic way, with no need of previous hand-labelling of the acoustic signal.

2. RNM revisited

2.1. RNM's approach

RNM analyzed the acoustic signal of 20 sentences selected among 54 sentences produced by 4 speakers of each of the following languages: English, Polish, Dutch, Catalan, Spanish, Italian, French and Japanese. The selection of the sentences was justified by the need to eliminate outliers produced by different rates of speech. The chosen sentences were segmented into vocalic and consonantal intervals. For each language, the sample standard deviation of the durations of the consonantal intervals (ΔC) and of the vocalic intervals (ΔV), and the proportion of time spent in vocalic intervals ($\%V$) were computed.

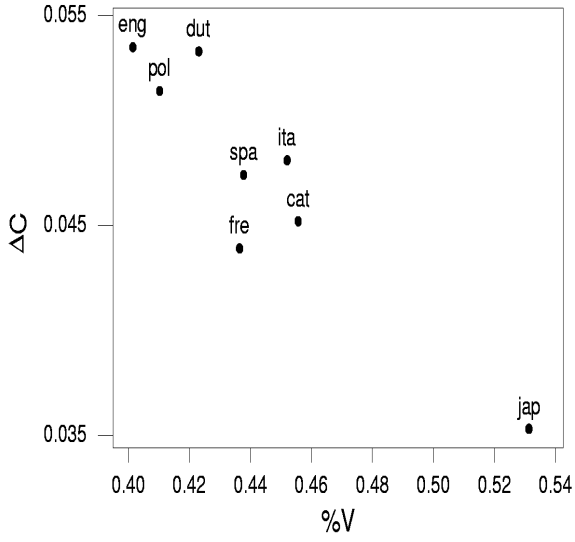

Figure 1: Distribution of languages on the ($\%V, \Delta C$) plane, based on Ramus et al.(1999.)

They found that the plot of the languages in the ($\%V$, ΔC) plan can be remarkably well fitted by a straight line with negative slope with a large linear correlation (-0.93, see figure 1).

Furthermore, 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.

RNM's approach is based on statistics of the distribution of a single interval, either consonantal or vocalic. This has been challenged by Grabe and Low (2001), and Gibbon and Gut (2001), who claim that correlations between two successive vocalic or consonantal intervals should be taken into account. However both correlation analysis and order identifica-

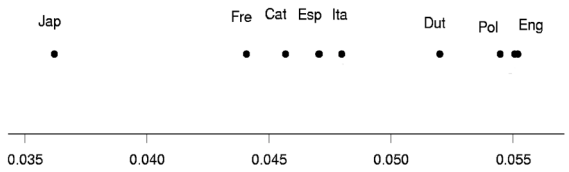

Figure 2: Estimated standard deviation of the Gamma distribution for consonantal intervals

tion based on the Bayesian Information Criterion seem to contradict this claim (cf. `http://www.ime.usp.br/~tycho/prosody/sonority/appendix.pdf`).

The statistics performed in RNM are of descriptive nature. This point was improved in Duarte et al. (2001) who analyzed the data presented in RNM using a parametric probabilistic model. This is the content of the next session.

2.2. A parametric probabilistic model for RNM

Duarte et al. (2001) propose a parametric family of probability distributions that closely fits the data in RNM. This has two advantages: it provides a deeper insight of the phenomena and makes it possible to perform statistical inference, i. e., to extend results from the sample (the data set) to the population (the set of all potential sentences).

The model is the following.

1. The durations of the successive consonantal intervals are independent and identically distributed random variables;
2. The duration of each consonantal intervals is distributed according to a Gamma distribution;
3. Different languages have Gamma distributions with different standard deviations;
4. The standard deviation is constant for all languages belonging to the same rhythmic class;
5. Standard deviations of different classes are different.

This model enables testing 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 are compatible with the data presented in RNM.

Figure 2 presents 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.

It is worth noting that the Gamma also fits well the sample distribution of the vocalic intervals. However, the parameters estimated from the samples do not support the expected clustering. For more details about this parametric model we refer the reader to Duarte et al. (2001).

2.3. The implementation of RNM on larger samples

The remarkable results presented by RNM are based on small samples of few languages. Its implementation on larger samples and its extension to other languages is a huge task, as it depends on a previous hand labelling of the speech signal. This is a time-consuming task. Moreover this hand labelling is often based on decisions which are difficult to reproduce in a homogeneous way, so that the data produced by different researchers may be not always comparable. A typical problem of labelling is raised by the cases in which it is hazardous to decide if a vocalic segment is present or not. This has strong consequences as it may dramatically change the value of ΔC.

This suffices to justify the need of an alternative approach which can be implemented on an entirely automatic way. This is precisely the goal of the next section.

3. A new approach to the problem

3.1. A rough measure of sonority

Newborn babies are able to discriminate rhythmic classes with a signal filtered at 400Hz (Mehler et al. 1996). At this level, 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.

Therefore a natural conjecture is that the identification of rhythmic classes must be possible using a rough measure of sonority. We will define a function which maps local windows of the acoustic signal on the interval $[0, 1]$. This function is close to 1 for spans displaying regular patterns, characteristic of sonorant portions of the signal. In opposition, the function will assign values close to 0 for regions characterized by obstruency.

The function is applied to the spectrogram of the signal. Let us call $s(t)$ this function, where t denotes time and belongs to the set $\{ku : k = 1, \ldots, T\}$, where u is the step unity of the spectrogram of the signal and T is the number of steps present in the spectrogram of the acoustic signal. In the present computation we took $u = 2$, where the units are counted in milliseconds. The values of the spectrogram are estimated with a 25ms Gaussian window. We only consider frequencies between 0 and 800 Hz. Our computations were made with Praat (http://www.praat.org).

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

$$p_t(i) = \frac{c_t(i)^2}{\sum_f c_t(f)^2} \tag{1}$$

This defines a sequence of probability measures $\{p_t : t = 1, \ldots, T\}$. Regular patterns characteristic of sonorant spans typically will correspond to sequences of probability measures which are close in the sense of relative entropy. We recall that the relative entropy for the column p_t with respect to the column p_{t-1} is defined by the formula

$$h\left(p_t | p_{t-1}\right) = \sum_i p_t\left(i\right) \log\left(\frac{p_t\left(i\right)}{p_{t-1}\left(i\right)}\right). \tag{2}$$

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

We now define the *sonority function* $s(t)$ as follows

$$s(t) = 1 - \min\left(1, \frac{1}{27} \sum_{u=t-4}^{t+4} \sum_{i=1}^{3} h(p_u | p_{u-i})\right). \tag{3}$$

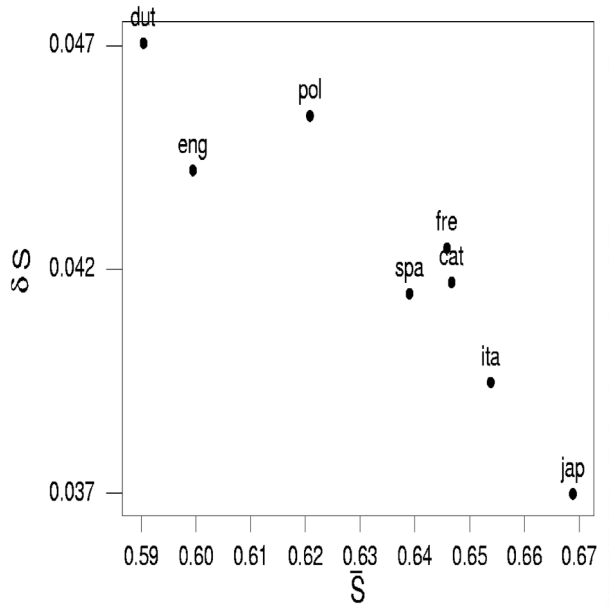

Figure 3: Distribution of the eight considered languages on the $(\bar{S}, \delta S)$ plane.

The sample mean the of $s(t)$ is defined as

$$\bar{S} = \frac{1}{T} \sum_{t=1}^{T} s(t)\,. \tag{4}$$

This will play the role of $\%V$.

The second statistics we consider is δS, defined as follows

$$\delta S = \frac{1}{T} \sum_{t=1}^{T} |s(t) - s(t-1)|\,. \tag{5}$$

The intuition behind this definition is the following. Typically the values of $p(t)$, and consequently $s(t)$, present large variations when t belongs to intervals with high obstruency, and are nearly constant in regions with high sonority. Therefore δS measures how important are the high obstruency regions in the sample. This estimator will play the role of ΔC.

In the next section we will compare the values of $\%V$ and $\bar{S}$ and δS and ΔC on the data considered in RNM.

3.2. Results

We computed $\bar{S}$ and δS for exactly the same sentences chosen by RNM. The result is presented in figure 3. To each one of the eight languages is associated a point in the plane. The first coordinate represents $\bar{S}$ and the second coordinate represents δS.

The graph in Figure 3 presents the same three clusters as Figure 1, even if the relative positions of the languages inside each group are not exactly the same. To understand the reasons of the similarities as well as of the differences between Figures 1 and 3 we compare our statistics with those in RNM. Figures 4 and 5 respectively show the plot of the 8 languages in the planes $(\%V, \bar{S})$ and $(\Delta C, \delta S)$.

Both graphs show that the statistics are correlated. Figure 4 shows that with the exception of Dutch, for the seven remaining languages, $\%V$ is an increasing function of $\bar{S}$. The inflection of the curve could be explained by the devoicing of final

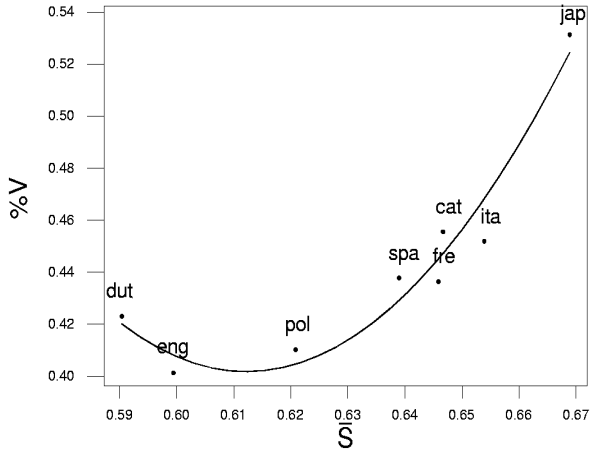

Figure 4: Distribution of the eight considered languages on the $(\bar{S}, \%V)$ plane.

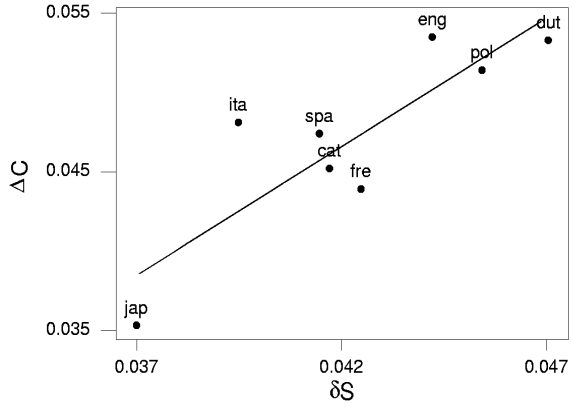

Figure 5: Distribution of the eight considered languages on the $(\delta S, \Delta C)$ plane.

voiced consonants in Dutch, as pointed out to us by Sharon Pepperkamp. Figure 5 shows that the average value per cluster of ΔC is an increasing function of the average value per cluster of δS.

A new feature emerges from the analysis of the histograms of the values assumed by s across the eight languages (cf. Table 1). First of all, the distance between the first and the third quartiles increases when we go from Japanese to Dutch. In other terms the distribution of sonority has a higher dispersion in the latter than in the former, with the syllable-timed languages occupying an intermediate position. This is summarized in the second column of Table 1. Extra information is provided by the region of small sonority of the distribution. The empirical probability of having sonority smaller than 0.3 also increases when we move from Japanese go Dutch, with again, the syllable-timed languages in an intermediate position. This is summarized in the second column of Table 1.

This last feature reinforces the idea already present in Duarte et al (2001) that the relevant information about rhythmic classes is contained in the less sonorant part of the signal, corresponding to the tail of the histogram.

Table 1: Probability of sonority smaller than 0.3 *and* $Q_3 - Q_1$ *for the eight languages*

lang	$P(s \leq 0.3)$	$Q_3 - Q_1$
Japanese	0.170	0.47
Catalan	0.177	0.47
Italian	0.182	0.48
Spanish	0.205	0.54
French	0.210	0.53
Polish	0.218	0.58
English	0.234	0.60
Dutch	0.254	0.64

4. Discussion

The results obtained here shed a new light on two claims which have been made in the psycholinguistic literature. The first one is that, as Mehler et al. (1996) put it, "vowels are the cornerstone of prosodic representation in very young infants" (op. cit., p. 112). Based on experiments about the perception of syllables by infants, they conclude that " infants may represent speech input as as a sequence of vowels including some information about their duration and energy" (idem, p.111). However it has also 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. Our approach opens a way of conciliating both claims, by replacing the subtle distinction between vowels and consonants by the rough distinction between sonority and obstruency.

The second claim is that the mechanism newborns use in order to discriminate rhythmic classes "cannot be based on statistical computations over large sample of speech", but "must rest on a rather simple procedure based on robust acoustic cues in the signal" (ibidem, p. 112). In RNM, since $\%V$ is strongly dependent on speech rate, ΔC appears as a more robust variable to discriminate rhythmic classes. However, ΔC depends on a complex computation, in contrast with our parameters which depend only on local information and can be extracted by simple comparison between successive values of the function $s(t)$.

Finally, while our analysis produces the same clusters as those in RNM, the value of $\bar{S}$ puts Polish an intermediate position between stress-timed and syllable-timed languages. This intermediate position could be related to the fact that in Polish there is no vowel reduction at normal speech rate (cf. RNM, section 2.3). This position is also coherent with the fact that adults discriminate Polish and English as well as they discriminate Spanish and English, as shown by Ramus et al. (submitted)

A striking feature appearing in the results obtained both in RNM and here is that the pairs of estimates $(\%V, \Delta C)$ and $(\bar{S}, \delta S)$, respectively, are linearly correlated. The linguistic meaning of this correlation remains to be clarified, but it suggests that the rhythmic class of a language can be specified by just one of the proposed sample statistics. In the framework of RNM, this has been challenged by Frota and Vigário (2001) who claim that this is not the case for Portuguese. This issue is outside the scope of the present discussion and will be treated in a forthcoming paper.

The main purpose of the present paper was to follow the way opened by RNM, showing that the relevant evidence about rhythmic classes can be automatically retrieved from the acoustic signal. In addition, our statistics are based on a coarse-grained treatment of the speech signal which is likely to be closer to the linguistic reality of early acquisition.

4.1. Acknowledgements

We thank Maria Bernadete Abaurre, Marzio Cassandro, Pierre Collet, Emmanuel Dupoux, Sónia Frota, Ulrike Gut, Ricardo Maronna, Jacques Mehler, Jean-Pierre Nadal, Marina Nespor, Nancy Lopes Garcia, Sharon Pepperkamp, Janet Pierrehumbert, and Frank Ramus for many illuminating discussions.

This work was partially supported by FAPESP (Projeto Temático *Rhythmic patterns, parameter setting and language change*, grant 98/3382-0), CNPq (Project *Probabilistics Tools for Pattern Identification Applied to Linguistics*, grant 465928/2000-5) and CAPES/PICDT and is part of the activities of the Núcleo de Excelência *Critical phenomena in probability and stochastic processes* (grant 66.2177/1996-6).

5. References

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

[2] Duarte, D; Galves, A.; Lopes, N.; 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

[3] Frota, S.; Vigário, M., 2001. On the correlates of rhythm distinctions: the European/ Brazilian Portuguese case. *Probus*, 13, 247-275.

[4] Gibbon, D.; Gut, U., (2001) Communication presented at the *Workshop on Rhythmic patterns, parameter setting and language change*, ZiF, University of Bielefeld.

[5] Grabe, E.; Low, E. L., 2000. Acoustic correlates in rhythmic class. Paper presented at the *7th conference on laboratory phonology*, Nijmegen.

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

[9] Ramus, F. (submitted). Perception of linguistic rhythm by newborn infants.

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

[11] Ramus, F.; Dupoux, E.; Zangl, R.; Mehler, J., (submitted). An empirical study of the perception of language rhythm.