On the relationship between intra-oral pressure and speech sonority

Anne Cros[1], *Didier Demolin*[2], *Ana Georgina Flesia*[3], *Antonio Galves*[4]

(1,2,4) Universidade de São Paulo, São Paulo, Brasil
(3) Universidad de San Andres, Buenos Aires, Argentina
(1,3,4){cros,georgina,galves}@ime.usp.br, (2)ddemoli@ulb.ac.be

Abstract

We address the question of the relationship between two time series associated to the speech signal. The first one is the *sonority function* which was introduced in Galves *et al.* (2002) as an index of the local regularity of the speech signal. The second time series is the *intra-oral pressure* during the production of speech. We argue that the joint evolution of both time series can be well described by a simple probabilistic model. We show that our model is in good agreement with the results obtained by analyzing a linguistic corpus with recorded sentences in French and Kinyarwanda.

1. Introduction

In this paper we address the question of the relationship between two time series associated to the speech signal. The first one of these series is the *sonority function*, which was introduced in Galves *et al.* (2002) as an index of the local regularity of the speech signal. The second time series is the *intra-oral pressure* evolution during the production of speech. Both time series give information about high and low sonority regions of the speech signal. Indeed as a first approximation it can be observed that when sonority is high, the intra-oral pressure is low and conversely. Therefore the intra-oral pressure is a phonetic parameter related to the sonority function, and this makes it natural to ask about their relationship.

We address this question by identifying the regions in the time domain in which each one of the series performs a change indicating a jump between regions of different behavior.

To compare the time evolution of each time series we use two thresholds c_s and c_p in the domains of the sonority and of the intra-oral pressure respectively. It turns out that a typical joint behavior is that when one of the time series is above its threshold and the other one is below its own. The exceptions can be identified with the occurrence of phonetic segments in a set, typically that of voiced constrictive consonants.

The data we analyze is an original linguistic corpus with recorded sentences in French and Kinyarwanda.

2. The data

2.1. Sonority

The *sonority function* was introduced in Galves *et al.* (2002) as an index of local regularity of the speech signal. It was defined as 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 we compute 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.

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

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

Regular patterns characteristic of sonorant regions typically correspond to sequences of probability measures $\{p_t : t = 1, 2, ...\}$ close in the sense of relative entropy. We recall that the relative entropy for the column p_t with respect to the column p_{t-1} is defined by the formula

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

$S(t)$ is then defined as

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

so that S is close to 1 for spans displaying regular patterns, characteristic of sonorant portions of the signal, while S assumes values close to 0 for regions characterized by obstruency.

In our model we take time t belonging to the set $\{ku : k = 1, \ldots, T\}$, where u is the step unity of the spectrogram of the signal and T is the number of steps present in the spectrogram of the acoustic signal. In the present computation 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 60 and 800 Hz. We choose the tunning constant $\beta = 1.5$. Our computations were made with Praat (http://www.praat.org).

Cassandro *et al.* (2005) shows that the family of stochastic processes obtained by considering the sonority time evolution for different languages can be well described by a family of tied quantized chains. The chains are tied together by 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. We will use this model to codify the sonority time evolution obtained from our linguistic corpus with a binary symbolic chain.

2.2. Intra-oral pressure

The second time series we consider in this paper is the intra-oral pressure P_s. In our case study, P_s was recorded with a set of sentences from French and Kinyarwanda. P_s was measured via a small plastic tube (internal diameter 2 mm) inserted

through the nose up to the area behind the velum. The data were recorded on the workstation Physiologia (Teston, 1995) that allows synchronous recording of acoustic and aerodynamic parameters. The P_s signal was low-pass filtered at 70 Hz in order to obtain a smoothed curve. For details on the method, see Demolin *et al.* (2004).

3. Probabilistic model

We will denote by S the sonority function and P_s the pressure curve, preprocessed by low-pass filters based on wavelets and Fourier respectively, in order to consider them roughly in the same space of smoothness. We define the chain I_S that will codify S in zones of high and low sonority, and I_P that will codify P_s in regions of high and low pressure as follows

$$I_S(t) = \begin{cases} 1 & \text{if } S(t) > c_s \\ -1 & \text{if } S(t) \leq c_s \ ; \end{cases} \tag{4}$$

$$I_P(t) = \begin{cases} 1 & \text{if } P_s(t) \leq c_p \\ -1 & \text{if } P_s(t) > c_p \ ; \end{cases} \tag{5}$$

where c_s and c_p are two suitable cut points. We will also define the noise $\eta(t)$ which will account for processing errors

$$\eta(t) = \begin{cases} 1 & \text{with } \mathbb{P}(\eta(t) = 1) = 1 - \epsilon \\ -1 & \text{with } \mathbb{P}(\eta(t) = -1) = \epsilon \end{cases} \tag{6}$$

An initial model describing the relationship between these quantized versions of the sonority function and the intra-oral pressure is

$$\textbf{Model 1} \quad I_S(t) = I_P(t).\eta(t). \tag{7}$$

It means only that sonority is high ($I_S(t) = 1$) when pressure is low ($I_P(t) = 1$) and vice-versa. The errors in this explanation are accounted for by the action of the processing noise $\eta(t)$. Note that if the processing is done carefully, the noise should modify only a very small percentage of the data.

Figure 1 shows in the first plot a short sample of the speech signal, corresponding to the French words *le bateau*, and its segmentation. The second plot shows the sonority function and the third plot the pressure curve. We should notice that there are segments where the relationship "up" and "down" works well, as predicted by Model 1, and others where it seems to be just wrong. No noise will account for a whole segment where the stated relationship does not hold. As an example, we can observe in Fig. 1 that for the segments [b] and [l] both the sonority and the pressure are high. The reason is that these consonants are respectively voiced stop and lateral that close the vocal tract and make the pressure rise. We labeled this type of segments with $I_C = -1$ under them.

So the relationship between sonority and pressure is more complicated than predicted by Model 1. This leads to the introduction of a more complex model, that we will call Model 2.

$$\textbf{Model 2} \quad I_S(t) = I_P(t).I_C(t).\eta(t)\,, \tag{8}$$

where the auxiliary chain I_C is defined by

$$I_C(t) = \begin{cases} 1 & \text{if } \sigma(t) \in \mathcal{B} \\ -1 & \text{if } \sigma(t) \notin \mathcal{B} \ . \end{cases} \tag{9}$$

In the above definition $\sigma(t)$ denotes the speech signal, $\sigma(t) \in \mathcal{B}$ is a shorthand for "$\sigma(t)$ *is part of a segment belonging to the set* $\mathcal{B}$", where $\mathcal{B}$ is the set of "well behaved" phonetic segments.

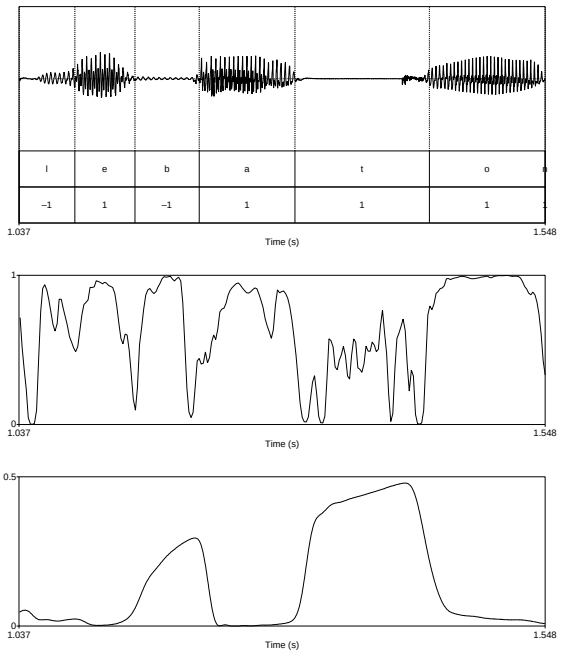

Figure 1: *First plot: audio waveform corresponding to the French words "le bateau", with its segmentation. Under each segment σ appears the value* 1 *if it belongs to the set $\mathcal{B}$, or* -1 *if it does not. Second plot: sonority function S. Third plot: intra-oral pressure P_s.*

Typically, the set of segments which are not in $\mathcal{B}$ and for which $I_C = -1$ are the voiced constrictive consonants.

The rationale of the formula (Eq. 8) is the following:

For instants t belonging to segments on $\mathcal{B}$, Model 1 holds, but for instants t that do not belong to $\mathcal{B}$, there is an inversion of the state of the sonority.

In Fig. 2 are represented three plots. The first one is the superposed plot of the smoothed sonority function and the pressure curve of the words *le bateau.* The second one shows the regions in which the data the two processes behave as predicted by Model 1. The third one shows the regions in which the two processes behave as predicted Model 2. We observe that in this plot $I_S.I_P = -1$ (thick line) for the instants where P_s and S are both high (or both low). So we expect that it corresponds to segments for which $I_C = -1$. In this plot like in the previous one, the points for which the model holds have got superposed coordinates.

Figure 2 shows that the percentage of points where there is concordance with Model 2 is higher than the percentage of points for which the Model 1 holds.

However we can observe on the second and third plots in Fig. 2, that Model 2 does not hold at the beginning and at the end of some segments. Figure 3 shows a zoom of the part [ba] of the French words *le bateau*, where the arrows indicate the points around the boundary between [b] and [a] for which $I_S(t) \neq I_P(t)I_C(t)$. Indeed in the segment [b], $I_C = -1$ so the arrow corresponds to $I_S(t) = I_P(t)$. On the contrary for [a], $I_C = 1$ and the arrow corresponds to the points for which $I_S(t) \neq I_P(t)$. We call these zones transition regions: they appear to be the price we have to pay in order to tie continuous signals with discrete binary ones.

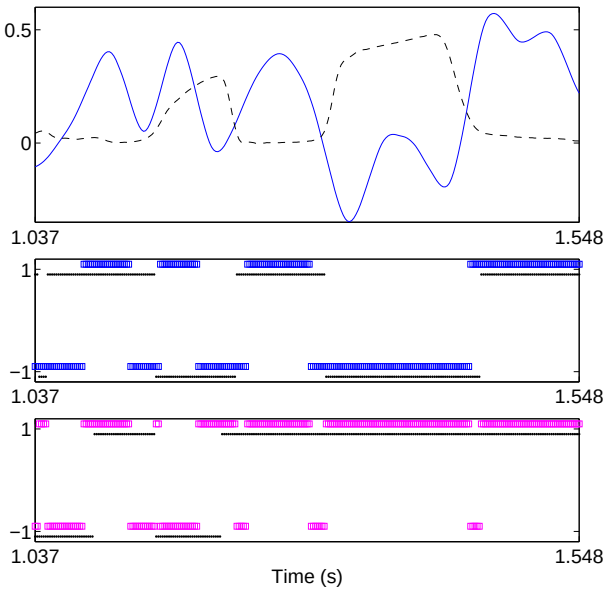

Figure 2: *First plot: superposition of the smoothed sonority function (solid line) and pressure curve (dashed line) of the words "le bateau". Second plot: map of agreement with Model 1. Thick line: symbol I_s. Thin line: I_p. Third plot: shows concordance with Model 1. Thick line: $I_s.I_p$. Thin line: I_c.*

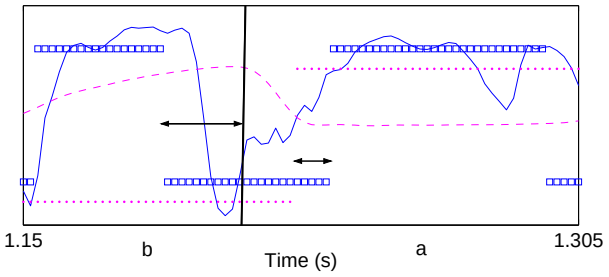

Figure 3: *Zoom of the part [ba] of the previous sample. Solid line: sonority function. Dashed line: pressure curve. Squares: symbol I_s. Points: I_p. Arrows: zones of transition for which $I_S(t) \neq I_P(t)I_C(t)$.*

Taking the transition zones into account, the percentage of points where there is no concordance between sonority and pressure should be reduced to the level of the processing noise.

Let us define $\mathcal{T}_T$ as the subset of points in the time domain set $\{1, \ldots T\}$ in which there is a transition step for the joint process $(S(t), Ps(t))$. An estimator of the proportion time spent in transition zones is given by

$$\widehat{\theta} = \frac{|\mathcal{T}_T|}{T} \,. \tag{10}$$

It is natural to define the mean error of Model 2 as

$$\widehat{e} = \frac{1}{T} \sum_{t=1}^{T} I_{I_S(t) \neq I_P(t) I_C(t)} \,. \tag{11}$$

Then the conjecture suggested by Model 2 together with the remark concerning the transition zone is that the difference between $\widehat{e}$ and $\widehat{\theta}$ coincides with the average noise since

$$|\widehat{e} - \widehat{\theta}| = \mathbb{P}(\eta(t) = -1) = \epsilon \,, \tag{12}$$

where t denounces a generic fixed point in the time domain.

In the next section, we will check if these predictions are in good agreement with the empirical results.

4. Empirical statistical analysis

As we stated in section 2, the intra-oral pressure was low pass filtered at 70 Hz when recorded, so in order to process the signals in (roughly) the same space of smoothness, we processed the sonority function with an orthogonal wavelet transform. $S(t)$ is analyzed on the 2^J middle points of the sentence, where J is the biggest integer so that $2^J \leq T$. Typically, $2^J = 512$ or 1024. The smoothing was performed spanning the signal on the Symmlet 8 basis and reconstructing it using only the 5 coarsest levels. An example of how the reconstructed function looks like is shown on Fig. 2.

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

To define the binary chain I_P associated to the intra-oral pressure we used the cut-off point $c_p = 0.05$, which seems to well discriminate zones of constant null pressure from zones in which the pressure is non null.

To check the validity of the model we will analyze two data sets. The first one is Kinyarwanda corpus with 27 sentences. The second one is a French corpus with 26 sentences. For every sentence we have the sonority function and the time evolution of the intra-oral pressure.

The results of the statistical analysis are summarized in Table 1. The second column of Table 1 shows the average percentages of points in the time domain for each language that behave according to Model 1. The third column of Table 1 shows the results for Model 2. In this case the percentage of agreement was estimated on a sub-corpus of sentences which have been previously hand-labeled with identification of the segments belonging to the set $\mathcal{B}$. The fourth column shows the percentage of time in which the processes are in transition zones. In this case also the transition zones are hand-labeled. The fifth column shows the values of estimated average noise ϵ, deduced from the Eq. 12.

	Model 1	Model 2	$\widehat{\theta}$	ϵ
Kinyarwanda	52.3%	$\simeq$ 89.8%	$\simeq$ 6%	$\simeq$ 4.2%
French	69.1%	$\simeq$ 78.5%	$\simeq$ 10%	$\simeq$ 11.5%

Table 1: *Statistical results.*

The results show that Model 1 holds for a greater proportion of points for the French sentences (69.1%) than for the Kinyarwanda (52.3%). This can be explained by the fact that the Kinyarwanda sentences in our data set contain more segments which do not belong to the set $\mathcal{B}$ (typically glottal stops, for which both pressure and sonority are low). Model 2 takes into account this feature, and the first estimated proportion of points for which Model 2 holds increases for this language until almost 90%. The results for the French sentences show that 78.5% of the points behave according to Model 2. This difference can be

partially explained by estimating $\widehat{\theta}$. The first estimations show that for the Kinyarwanda sentences, 6% of the points belong to transition zones, while for the French sentences $\widehat{\theta} = 10\%$. The calculation of the average noise leads to $\epsilon \simeq 4.2\%$ for the Kinyarwanda, and $\epsilon \simeq 11.5\%$ for the French sentences.

5. Discussion

The sonority function was introduced in Galves *et al.* (2002) as a tool to discriminate between rhythmic classes of languages. The goal was to reproduce in an entirely automatic way, with no need of previous hand labeling, the remarkable empirical results obtained by Ramus, Nespor and Mehler (1999). While remaining close to the spirit of Ramus *et al.* (1999), this new approach avoids the linguistic difficulties associated to the definition of the statistic parameters considered in Ramus *et al.* (1999) as well as in Duarte *et al.* (2001). For a discussion of this issue we refer the reader to Galves *et al.* (2002) and to Ramus (2002).

One can ask what could be a measurable phonetic correlate of this function. Leaving aside the amplitude of the speech signal (related itself to acoustic signal) and focusing on parameters linked with the production of speech, two parameters appear at first sight. The first is linked with biomechanical properties of the production of syllables and the second with aerodynamic parameters. Indeed since the introduction of the Frame/content theory by MacNeilage (1998) it became clear that the cyclic alternations of jaw lowering and closing account for patterns of speech observed both in language acquisition and in the patterns observed in the world's languages. However even if the jaw movements are a mechanical parameter it is not easy and straightforward to obtain quantified values of these movements. The jaw movements give rather the Frame in MacNeilage's theory while the segmentation of the speech signal would rather account for the content and this is what the sonority function accounts for. Therefore we thought that a good correlate for the sonority function would be the pressure measured in the vocal tract during the production of speech because each time that there is a constriction there is also a rise of pressure (except for nasal consonants) and each time that there is an opening there is a decrease of pressure. In addition, intra-oral pressure is an easily measurable and robust parameter (see e.g. Ohala (1974), Baken and Orlikoff (2000) for a more detailed discussion on this).

It is clear that there is a number of issues related to this work that have to be refined and discussed such as to refine the identification of some classes of sounds like sonorants and in particular nasals. However we think that the correlation between the sonority function and intra-oral pressure is well established.

We are also conscious of the existence of other works done in relation to rhythm in speech (e.g. Tajima and Port, 2003) but it seems premature at this moment to discuss those works in the perspective of the current study.

6. Conclusions

We addressed the issue of the relationship between the sonority function and the intra-oral pressure. We introduced a new model which derives a quantized binary chain associated to the sonority from a quantized binary chain associated to the intra-oral pressure, together with a third chain indicating the presence of segments of a certain type. Our model is in good agreement with the empirical data from French and Kinyarwanda we analyzed.

The results of this first work clearly shows that the relationship between the sonority function and intra-oral pressure can be well described by models whose construction is based on phonetic observations.

Our result suggests that it should be possible to discriminate rhythmic classes using samples of the intra-oral pressure, exactly as it was done using samples of the sonority function. This is a challenging issue to be investigated further.

7. Acknowledgments

This work was partially supported by CNPq grants 150244/2003-7 (AC), 301301/79 (AG), 303770/2003-1 (DD). It is part of PRONEX/FAPESP's Project *Stochastic behavior, critical phenomena and rhythmic pattern identification in natural languages* (grant number 03/09930-9) and also of CNPq's project *Stochastic modeling of speech* (grant number 475177/2004-5).

8. References

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

[2] Cassandro, M., Collet, P., Duarte, D.,Galves, A. and Garcia, J., "A stochastic model for the speech sonority: tied quantized chains and cross-linguistic estimation of the cut-points", Manuscript, 2005..

[3] Teston, B. and Galindo, B., "A diagnostic and rehabilitation aid workstation for speech and voice pathologies", Eurospeech 4, Eur. Speech Com. Ass. Madrid. 4, 1995.

[4] Demolin, D., Hassid, S. and Soquet, A., "Aerodynamics of French Consonants", Journal of the Acoustical Society of America, 2004.

[5] Ramus, F., Nespor, M. and Mehler, J., "Correlates of linguistic rhythm in the speech signal", Cognition, Vol. 73, 1999, p 265-292.

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

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

[8] MacNeilage, P. F., "The Frame/Content theory of evolution of speech production", Brain and Behavioral Sciences, Vol. 21, 1998, p 499-548.

[9] Ohala, J.J., "A mathematical model of speech aerodynamics", Proceedings of the Speech Communication Seminar, Stockholm, Vol. 2, Speech Production and Synthesis by Rule. G. Fant (ed.) Almqvist & Wiksell, 1974, p 65-72.

[10] Baken, R.J. and Orlikoff, R.E., "Clinical Measurement of Speech and Voice", San Diego, 2000, Singular.

[11] Tajima, K. and Port, R.F., "Speech rhythm in English and Japanese", J. Local, R. Ogden & R. Temple (eds.) Papers in laboratory Phonology VI, Phonetic Interpretation, Cambridge University Press, 2003, p 322-339.