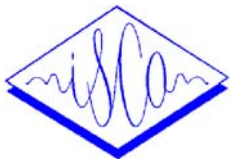

Speech Prosody 2004
Nara, Japan
March 23-26, 2004

Relating Emotional Content to Speech Rate in Brazilian Portuguese

Ana Cristina Fricke Matte

Phonetic and Psycholinguistics Lab. & Department of Linguistics
Universidade Estadual de Campinas, Brazil
acfm9000@aol.com

Abstract

Emotion is frequently conceived, in the speech science literature, in such a way as to organize the relationship between specific emotions taken as psychological concepts and phonetic parameters such as voice quality, speech rate, and prominence, in the phonetic domain. The main goal of this paper is to propose a language-based analysis of the emotional content of the text in order to get a more abstract and culturally independent approach of emotion in speech. This work presents an experiment that relates speech rate and the temporality as a constitutive element of emotion. We are able to quantify the temporal content of emotion in the text by a semiotics analysis.

10.21437/SpeechProsody.2004-153

1. Introduction: emotional gesture

In this study we introduce the notion of an emotional gesture in speech related to some aspects of meaning in the time domain. In order to do that, we adopt the conceptualization of meaning proposed by Saussure (1989) and Hjelmslev (1968), whose concept of meaning is composed by expression and content. As regards meaning, Saussure worked with the distinction between signifier and signified, defined as two inseparable sides of a paper sheet. Hjelmslev developed the static Saussurian model making it dynamic by defining a set of two independent plans with analogous structures: the content plan (signified) and the expression plan (signifier).

The relationship between these plans determines three kinds of systems: signic, symbolic and semi-symbolic. Verbal language is a signic system: it relates the two plans in an arbitrary way. For the symbolic system, the relationship is term-to-term dependent, each content corresponding to just one expression and vice-versa, as the color "red" for excitement. The notion of an emotional gesture in speech is related in this article to the third type of system, the semi-symbolism, for which the organizational principle is a categorical relationship between content and expression (Greimas & Courtés, 1986: 203-206). This suggests that emotional gesture is a nonparallel but a dependent movement between emotional content and prosodic expression.

Prosody is a component of the expression plan, which integrates the notion of rhythmic structure. In this respect, speech rate, the dependent variable focused here, is crucial to understand the timing of the units in the speech chain. The domain for speech rate computation is delimited from the second glottal period of the first phrasally-stressed vowel of the utterance to the second glottal period of the last phrasally-stressed vowel in the same utterance. Utterance size is expressed as the number of its phonetic syllables. The corpus is a narration with simulated emotion, based on the view of a vocal caricature.

The concept of vocal caricature has a semiotic basis: communication as the product of the interaction between two similar but non-equal language systems. For speech, the speaker works with a code that must be known by the hearer, but each one has their own sub-codes based in their own history of life and their own language acquisition processes. As idiolects, these different sub-codes produce a special kind of noise, the ideological noise in the communication process, present in different degrees. This notion of the communication scheme comes from a proposal by Silva (1972). The relevance of this notion relies on the concept of a minimal trace, structure or element that allows communication itself. These minimal elements are taken as a caricature, and must be present for a simulated emotion communicates something about the intended emotion.

We based our concept of emotion on Greimas & Fontanille (1993)'s theory for the analysis of passion in verbal texts. These authors present passion as a complex process. Vengeance, for instance, presupposes disappointment, which in turn presupposes the belief in an action that happens in the future. Emotion is the corporal (or gestural) perturbation that allows to perceive the passion behind someone's behavior. Passion can be named and is explicit in the discourse, but emotion can not be named and appears in the discourse as observations about time, space and about the subject. If emotion is just a perturbation, then the vocal caricature must be an abstract result of the text analysis, a result about different levels of tension in each part of the text. The tension flow modulation (*M*) quantified here operates with the temporal content of the text and transforms this kind of information into values of tension. We will return to that question later in this paper.

2. Characterizing speech rate patterns and tension flow by deviation

2.1. The *corpus*

In order to study the relationship between emotional content and the variation of speech rate in the expression plan, we chose the narrative of a story for children. *Cachinhos de Ouro* (*Goldilocks and the Three Bears*) was recorded in a disc in 1995 and told by a female adult from Rio de Janeiro (Brazil) who is a well-known Brazilian actress and a professional storyteller. The *corpus* contains 109 sentences from 1 to 55 syllables. As regards segmentation, V-V units (from one vowel onset to the next one) are used as syllable-sized units, instead of phonological syllables. Sentence boundary was determined by the presence of a silent pause. The following analyses are based on this *corpus*.

2.2. Analyzing speech rate

The type of relation between number of syllables in each sentence and speech rate can be observed in the Fig. 1 (The minimum and the maximum speech rates were obtained by adding one standard deviation to the mean- or median, in the

cases where the variation coefficient was greater then 30 % - and value). This kind of relation is explored in the following to assess speech rate variation sentence by sentence.

A cluster analysis technique (joining: tree clustering) allowed us to divide the corpus according to number of syllables in each sentence and the respective average of speech rate into three different groups. Group A contains sentences with 1 or 2 syllables; group B sentences with 3, 4, 6, 8, 10, 11, 13, 15 or more than 21 syllables; finally, group C sentences of 5, 7, 9, 12, 14 and 16 to 20 syllables. Due to a statistical similarity between groups B and C (Scheffé Test: $p < 0.05$), we grouped them together. The chosen groups are group 1, with small sentences (1 to 2 syllables), and group 2, with all the other sentences. Group 1 has a speech rate average of 3.4 syllables/s (SD = 1.13) and group 2, a speech rate average of 5.5 syllables/s (SD = 0.37) ($r = 0.83$, $p = 0.000$).

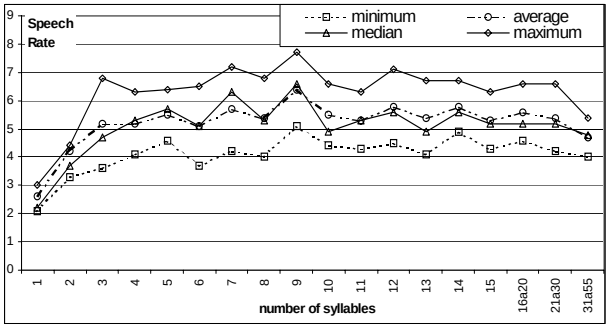

Figure 1: *Speech rate (syl./s) and number of syllables in each sentence.*

2.3. Semiotic analysis

In the semiotic field, emotion is a variation of tension, which is expressed by a variation in behavior. This behavior reveals that the subject is affected by a passion. Different levels of tension indicate different states for the subject, but it is necessary to consider the context in order to know which specific passion is being revealed by the corresponding corporal perturbation. Thus, the analysis of emotion in speech should consider that the perturbation found in the speech chain is only a clue of passion, not the entirely passion itself. When distinguishing simple from complex passion, emotion or corporal perturbation works more adequately as a clue for simple passions. For instance, it is easier to name sadness, a simple passion, than disappointment, one state of some complex passions, if one takes only corporal perturbation into account.

By focusing on temporal perturbations along the text, a function *M* (1), called deep temporal flow modulation, is proposed in order to compare speech rate deviation with the deviation of tension in the content of the text. This modulation intends to characterize the variation of tension in the content of the (written) text, deduced from the complex temporal dynamics suggested therein. In order to compute *M*, we work with the semiotics concepts of flow (f), tempo (α) and aspectualization (A). Each one of these concepts receives a value from a set of five, associated with this gradation: minimal, near to minimal, neutral or central, near to maximum and maximum ("near to" means tends to). The minimal and maximal ends are dichotomical: continuation vs break for flow (near to continuation is called "break of break" and near to break is called "break of continuation"), and deceleration ($\alpha = 0.9$) vs. acceleration ($\alpha = 1.1$) for tempo. Because aspectualization is a more superficial and concrete element of the discourse, this concept works with three dichotomies: prospectivity vs retrospectivity, gradation vs jump and duration vs punctuality. The values have been assigned considering the position in the scale taken from the semiotic analysis (table 1).

Table 1: *Values for* M *components*

	Min	Near to min	Near to max	Max
Flow	f = 1	f = 2	f = 4	f = 5
Tempo	$\alpha = 0.9$	$\alpha = 0.95$	$\alpha = 1.05$	$\alpha = 1.1$
Aspectual.	A = 1	A = 1.5	A = 2.5	A = 3

The semiotic analysis will provide each sentence with three values: *f*, *alfa* and *A*. Any combination is possible, but some are more rarely found, specially opposing the extreme values of tempo and aspectualization. Opositions with flow, however, are very common.

As way of illustration, the episode of the fall of Alice into the hole of the tree (*Alice in the Wonderland*) will be analyzed here. This story has different Brazilian versions, adapted for cinema, disc and books. Two versions for this episode are particularly relevant.

1. In the first version, the scene lasts a long time, with the description of objects seen by Alice and her thoughts about bats, the distance to the center of the Earth. Suppose that this episode has seven sentences.

2. In the second version, the scene is also described as lasting a long time, but no description is given: the duration is only informed by the narrator in one sentence, e. g. "Alice fall and fall in a long hole. After some time falling, she arrived to…"

In both versions the flow begin with a break of the continuation ($f = 4$), because the fall interrupts a previous making of the girl; any break (of break or of continuation) is a passage from one kind of action or state to another one. In the first version, the first sentence has $f = 4$, but in all other sentences for which the break continues, the flow is a continuation of the break ($f = 5$). Since the second version, has just one sentence, the assigned value for f is 4.

As regards tempo, the long description of the fall in the first version decelerates it. It does not matter if Alice falls one or one thousand kilometers, the point is that the description step by step allows the perception of the fall as a course, not a simple passage as in the second version. Then, the first version is characterized with a decelerated tempo ($\alpha = 0.9$). In the second version, on the other hand, even though the narrator says that the fall lasted a lot, there is no explanations or gradations, there is no time in the narration and the tempo is accelerated ($\alpha = 1.1$).

Aspectualization is a minor abstraction and depends on of the semantic meaning of the words. The main indicator for aspectualization is the adverb. Then, it is more convenient to analyze a specific text to find out the aspectualization. Anyway, based on previous analysis of Brazilian versions of Alice, it can be inferred that a gradation predominates in the decelerated versions (version 1) and duration in the accelerated versions (version 2), giving us a value *of* 1 in both versions.

The modulation of the deep temporal flow is obtained by using formula (1) that applies the categorical values illustrated with the two versions of Alice. The semiotic analysis focuses on the written text, in order to minimize the correlation between the prosodic-physical domain (the expression plan, e.g. speech rate) and the semio-liguistic domain (the content plan, e. g. the temporality in the written text).

$$M = \alpha f - 8\alpha^{1/n} + \frac{(A-1)}{A} + 8\alpha \qquad (1)$$

In (1), *M* is a real value that represents the variation of tension in the text sentence by sentence. It is important to note that it does not reveal a physical tension, but the tension in the temporal content of the written text.

The *flow* (*f*) stands for the breaks ($f = 5$) and continuations ($f = 1$) related to the action described in the text. This value is modulated by the *tempo* (α). The influence of α over *f* is stronger according to the duration of the specific kind of event concentration. Such duration is represented by *n*, the index of each sentence, which is incremented by 1 during the sequence of sentences with the same α. The *aspectualization* (*A*) plays a secondary role in *M* because it can reveal tension just for the cases of non-relaxation (gradation, prospectivity or duration = 1; jump, retrospectivity or punctuality = 3). The content values and the specific shape of the function are conceived in such a way as to obtain a scale compatible with the normalized durations and absolute values of speech rate.

Table 2 recuperates the values for the sentences in the first version of Alice's fall analyzed above. For the second version, we have $f = 4$, $\alpha = 1.1$, $n = 1$, and $A = 1$: $M = 4.4$.

Table 2: *Values of* M *for the seven sentences of the example "Alice's fall", first version.*

A	f	alfa	n = sentence	*M*
1	4	0,9	1	3,6
1	5	0,9	2	4,11
1	5	0,9	3	3,98
1	5	0,9	4	3,91
1	5	0,9	5	3,87
1	5	0,9	6	3,84
1	5	0,9	7	3,82

The temporal categories that constitute *M* are based on the temporal organization proposed by Zilberberg (1990), and the variation of the tension in the text is based on the graphic proposition of Fontanille and Zilberberg (2001). In this perspective, emotion is revealed as a temporal tension, namely the unexpected variation in the way communication occurs in each text.

The *M* function reveals the temporal content of the text, taken sentence by sentence. It is an empirically determined function which tries to express the logical relationship between the main temporal elements of tension (Matte, submitted): flow (breaks and continuations of the narrative), tempo (accelerations and decelerations of the discourse), and aspectualization (done by focus, orientation or segmentation in the discourse).

Up to now, the application of this formula requires the intervention of a semiotician. As seen above, the speaker cannot to infer *M* directly by reading the text, because the formula implicates a previous analysis of three levels of the temporality, two of them not immediately given. Even if the speaker were an expert in tensive analysis, s/he would need to calculate *M* before taking any decision about each sentence. If he tries it, the fluent reading of the text would be impossible even for neutral speech.

In the work presented here, *M* is applied to the sentences of Goldilocks and the three bears.

Figure 2 is a graphic of the moment when Goldilocks gets into the house. Two points of view are present in alternation: Goldilocks, to whom the invasion is just a curiosity, and the speaker, to whom the girl is intrusive and improper. *M* reveals this alternation as a variation of tension because the *M* values go up and down around 3.6, limit between tension and relaxation suggested in our previous data (Matte, 2002).

It is important to note that *M* does not inherit any phonetic information, because it considers only the conceptual part of the meaning, the content. Putting prosodic (speech rate) and emotional information (*M*) in the same graphic, an interesting picture emerges. The result is a graphic where the semi-symbolic system can be revealed, suggesting the correlated categories.

In semiotics we describe a semi-symbolic system by relating two or more categories of the two different plans, content and expression, and by determining the orientation of each category in the observed relationship. Both the application of the function *M* to the content plan and the quantitative results from the acoustic-phonetic analysis allows us to work quantitatively with the semi-symbolic system. It is thus possible to evaluate correlation between content and expression with statistical analysis.

3. Relating emotional content to duration

The next step of this experiment is to put together two deviations: speech rate deviation and tension flow deviation (expressed by *M*). Our experiment shows that there is no direct correlation between *M* and a deviation of speech rate, but there is an indirect relationship between them if observed different parts of the text are taken into account. The narrative analysis of the content allows the division of the text into four parts, statistically different from each other. The statistical analysis of *M* using the parts as levels shows a significant difference between the tense parts and the relaxed parts, with the following distribution (values for *M)*:
A - relaxed: mean = 1.14, SD = 0.84 (85% of $M <= 3,6$)
B - tense: mean = 4.96, SD = 1.55 (74% of $M > 3,6$)
C- tense: mean = 5.37, SD = 1.04 (89% of $M > 3,6$).
D - relaxed: mean = 2.70, SD = 1.20 (61,5% of $M <= 3,6$).
Parts A and C were labeled stable, whereas parts B and D unstable because of the percentage of M according to the limiar value of tension.

The deviation of speech rate not explained by the number of syllables in the sentence emerged mainly when the text was tense and unstable (figure 2).

In the relaxed parts of the text, A and D, the *instability* seemed to change the speech rate by deviating it from the expected, more neutral values. We obtained 29% of deviation for speech rate in the stable part A, and 38% of deviation in the unstable part D. Even in the tense parts, content=level *stability* or *instability* are factors that perturb speech rate. We obtained 38% of deviation in the stable part C and 53% of deviation for speech rate in the unstable part B (figure 2).

These results indicate that a high average of *M* induces deviation in the speech rate ($r < 0.005$), and this deviation is more notable with a high deviation pattern (*instability* of *M*). Moreover, there is a similar effect of *M* in the speech rate in a relaxed, but unstable part (D) and a tense, but stable part (C).

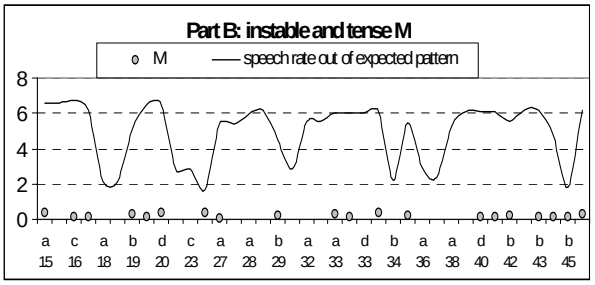

Figure 2- *Deviated speech rate (white) and M in the second part of the story, when the voices of the speaker and of the Goldilocks alternate tension and laxity, respectively. The deviation was considered if it was greater than or equal to 10% and corresponds to 53% of the sentences in this part of the text. It is relevant for* M *that the speech rate is perturbed, but the exact values of this perturbation are not. The points in the graphic indicate the sentences with perturbed speech rate*.

The relaxed sentences in stable sequence affect less the speech rate than any other. There would be a secondary role for *M*: it affects speech rate as a change in *M*'s average and standard deviation.

The post hoc tests for *M* indicate that there is no difference between the semiotic parts B and C (Scheffe test between C and D, $p < 0.8$; the others have $p < 0.05$). However, as regards speech rate, the LSD test indicates significant differences between the parts, except between D and {A, B}. Note that C is different from any other part.

4. Final observations

The results about the patterns of deviation of speech rate concerning the modulation of the deep temporal flow (laxity/tenseness) are not conclusive. It is important to note that our data had no neutral reference, then we could not effectively test the influence of the temporal flow perturbing speech rate ruling out the expected linguistic expected information. From the results we claim that emotional content can affect speech rate mainly in tense situations, but we are not able to define how strong is this influence yet. Work on synthesis of emotional prosody can help assess the strength and limits of this influence.

The analysis suggests the possible existence of one semi-symbolic system of emotion in speech correlating duration and tension flow. It also suggests taking into account the work with stress groups and duration of smaller groups in order to check the hypothesis of correlation between a higher domain in the content and a smaller domain in the expression, due to the importance of the linguistic rules in the communication process.

5. Acknowledgments

The author thanks Plinio Almeida Barbosa for his suggestions, Nick Campbell for his comments, and the FAPESP support.

6. References

[1] Barbosa, P. A., 1996, "At least two macrorhythmic units are necessary for modeling Brazilian Portuguese duration: emphasis on segmental duration generation", *Cadernos de Estudos Lingüísticos*, 31, 33-53.

[2] Barbosa, P. A., 2000, "Syllable-timing in Brazilian-Portuguese: uma crítica a Roy Major", *D.E.L.T.A.* Vol. 16, 2, 369-402.

[3] Fontanille, J.; Zilberberg, C., 2001, *Tensão e Significação* (Tension et signification)/trad. I. C. Lopes, L. Tatit e W. Beividas. Discurso Editorial/Humanitas, São Paulo.

[4] Goldstein, L., 2003, "Emergence of discrete gestures", *15th ICPhS*, Barcelona.

[5] Greimas, A.J.; Courtés, J., 1986, *Sémiotique – Dictionnaire Raisonné de la Théorie du Langage II*. Paris, Hachette.

[6] Greimas, A. J.; Fontanille, 1993, J. *Semiótica das Paixões - dos estados de coisas aos estados de alma*.(Sémiotique des passions)/trad. M. J. R. Coracini. Série Temas # 33. São Paulo, Ed. Ática.

[7] Hjelmslev, L., 1968, *Prolégomènes a une Théorie du Langage - et - La Structure Fondamentale du Langage* /trad. Anne-Marie Léonard, Arguments # 35, Les Editions De Minuit, Paris.

[8] Matte, A. C. F., 2002, *Vozes e canções infantis brasileiras: emoções no tempo*/ tese de doutorado defendida na FFLCH- USP, São Paulo, Brasil.

[9] Matte, A. C. F., 2003, "Tempo fonoestilístico e semi-simbólico (phonostylistic and semi-symbolic time)", *51.o Encontro do GEL*, Taubaté, SP.

[10] Murray, I. R.; Arnott, J. L., 1993, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion", *The Journal of the Acoustical Society of America*, Vol. 93, 2, 1097-1108.

[11] Saussure, F., 1989, *Cours de linguistique générale*/ Ed. critique par Rudolf Engler, Harrassowitz, Wiesbaden.

[12] Silva, I. A., 1972, *A dêixis pessoal*/ tese de doutorado defendida na FFLCH – USP, São Paulo, Brasil.

[13] Valença, F., 1995, *Suzana Vieira Conta Alice no País das Maravilhas*./ phonographic adaptation Fátima Valença, collection Olha Quem Está Contando, disc 01015/3.

[14] Zilberberg, C., 1990, "Relativité du rythme". *Protée – Théories et Pratiques Sémiotiques*. Département des Arts et Lettres de l'Université du Quebec à Chincontimi, Vol. 18, 1, 37-46.