Neutral Speech Corpora – a test for neutrality

Ana Cristina Fricke Matte

Faculdade de Letras
University of Minas Gerais, Brasil
a9fm@yahoo.com

Abstract

What is neutral speech? At the horizon of this research there was the uncertainty that an ideal speech exists. This writing reports the results obtained of research in phonostylistics of Brazilian Portuguese with the objective of determining the necessary experimental conditions for recording so-called neutral speech. The experiment was designed to test these two hypotheses: 1) The phrase, or sentence, the minimal prosodic unit, is also the minimal unit of meaning in studies of expressing emotion in speech, even when our focus is on the production of complete texts that should be taken as a single unit of meaning. 2) The speaker's reported self-impressions can indicate certain sentences that have been affected by the reactions of the speaker, which conflict with the objective of recording neutral speech, and therefore should be rejected from a corpus of referential speech. The results obtained validated both of the hypotheses and enabled us to formulate a single unique test for neutral speech, recommended for the process of purging of referential corpora in experimental phonology, which is described in this work..

1.Introduction

This research is part of a larger study on emotion expression in speech, focusing on Brazilian Portuguese. Its methodology called for a comparison between data with emotions simulated and data with so-called neutral speech. Due to the lingering doubt whether obtaining and controlling a corpus of data for reference on neutral speech would be truly possible, a pilot study was conducted as described below.

The objective of the pilot study was to ascertain the best conditions for recording neutral speech data in order to specifically analyze duration. The initial hypothesis was that, even when reading aloud texts, expressing emotions and organizing phrases (isolated, in paragraphs, or as a whole original text would affect the duration of the speech segments to differing degrees. The epistemological foundation on which the research in question was developed is French linguistic semiotic theory. As regards the first hypothesis, the theory considers emotion to be a perceivable physical disturbance whose reference for comparison should be a socially accepted standard of behavior, which for purposes of this study is called neutral speech. Regarding the second hypothesis, we may cite the dictionary of Semiotics:

"Structural linguistics confers syntactic independence on the phrase. Thus, for Bloomfield, the phrase, even though built up from constituent elements, is not in itself, a constituent of any larger unit. L. Hjelmslev, in contrast, defines the phrase as the largest syntactic unit having an interactive character within an infinite text, considering it thus to be the only unit capable of being analyzed. Whether the process be bottom-up, beginning at the minimal elements (Bloomfield), or top-down and involve segmentation (Hjelmslev), the result will be, in both cases, comparable: the phrase emerges as a totality that covers a syntactic hierarchy." (Courtés & Greimas, undated, pp. 196-197)

Thus, the phrase – or sentence – semiotically as much as for prose, can be considered the minimal unit of meaning.

The test detailed here was designed to relate the conditions of text presentation with the emotions reported to changes in change the duration pattern of V-V units.

2.*Corpus*, segmentation and labeling

The corpus consists of the recording of neutral speech of a speaker (PA) experienced in acoustic phonetics and inexperienced in recounting stories. Texts were chosen of poetry and prose, both written by a different speaker (FM) and chosen for this research itself. After recording, the speaker (PA) was asked to take notes beside the sentences of his impressions, emotions and any commentaries he thought appropriate about the sentences that seemed important during the recording. Such notes were made exclusively according to the speaker's judgment and without restrictions, and later grouped into categories.

Of the prose sentences, 61% were left uncommented by the speaker. Impressions about the prose sentences were classified into the following categories: a) difficult to pronounce; b) tenderness; c) disappointment; d) certainty; e) humor; and f) pride.

For poetry, 32% of the sentences had notations beside them, all later classified in the category "humor." The poetic sentences that had been presented in paragraph form, and therefore, susceptible to prose influences in terms of duration were removed from the corpus.

The texts were presented in three ways:

a) in sentences separated by white spaces and in random sequences (henceforth phrase presentation), b) in

paragraphs in random order (henceforth paragraph presentation) and c) as a text organized in paragraphs with the sentences in their original order (henceforth textual presentation). Each format was read twice consecutively. Moreover, the sequences in random order were compared with those in the text's original order.

The research was conducted in two stages, the first composed of only 543 segments, with these being the most frequent vowel-consonant configurations in the corpus, in order to determine the appropriate statistical tests for the analysis. The second stage covered all the V-Vs, that is, 2432 segments.

The V-V unit was chosen because its segmentation corresponds to the segmentation perceived (Barbosa, 1996). The corpus was segmented by Praat version 4.1.17, with the script BeatExtractor.psc, implemented by Barbosa (2004), based originally on the Fred Cummins initial vowel extractor partially smoothed and incrementado by an additional filter and another technique that searches for "beats."

Each segment received a phonic label following the notation proposed by Albano and Moreira (1996). Despite being based on a phonetic transcription based on parameters of produced acoustics, the assessment of the listener (transcriber) was not disregarded. In order to do this, certain tests were conducted comparing different productions of the same phonological segment, especially those for which the acoustic parameters presents high wide variation the several times they were produced, giving preference to the "fullest" production of the segment, for being the one that most likely reflects the "acoustic image" of the segment for that speaker. In stretches for which a linear transcription would be impossible, we chose to segment in larger groups larger than V-V, so that, for example, "*se admirava*" (wondered) was segmented and labeled as follows
[s] [IadImiR] [av] [A]

For the reasons cited above, we kept the same label for all productions lacking acoustic signs of a certain segment, as in the example above, of the first vowel [I] ([sadImiRavA]) in higher rates of elocution, in order to evaluate the reduction/expansion of the segment in different productions.

Since V-V units consist of different numbers of phonic segments, we opted for relative measurements of duration obtained from the data of duration in milliseconds. These are the measurements relative to the z-score: measurement of lengthening or shortening of the duration of the V-V unit, in relation to the intrinsic duration of its phonic components; and the z-smoothed, that relativizes the value of the z-score, in reference to the surrounding z-scores by means of a smoothing of five points, resulting in the suppression of prominences irrelevant to the perception of phrasal accent (Barbosa, 2004). The z-score and the smoothed z-score were calculated using as a reference the table of durations for the speaker Zaldo (Barbosa, 2004), in which the same phonic notation is employed for the labels. Accented groups were detected with help of the program SGdetector, created by Barbosa (2004) for implementation in the MatLab.

3.Statistical Results

Despite the low correlations found for all the parameters analyzed of the z-score and smoothed z-score of neutral speech (as table 1 shows) indicating a lack of correlation between V-V units, the ANOVA results show that some of these factors do influence the average z-score and smoothed z-score.

Table 1: *Z-score and smoothed z-score for neutral speech*

Pearson Correlation	smoothed z-score
Presentation	r = -0.0623 p = 0.002
Position in the accent group	r = 0.1817 p = 0
Number of V-V units in the accent group	r = -0.0436 p = 0.032
Impression	r = 0.1744 p = 0
Order	r = -0.0569 p = 0.005
Reading	r = 0.0273 p = 0.178
Prose-poetry	r = 0.3725 p = 0

ANOVA one-way using the z-score over the entire corpus as a dependent variable and the impressions reported as a statistically significant factor (p = 0.0000), leading us to conclude that the speaker's impressions about the sentences read affect the results of normalized durations of the V-V units. Despite post hoc tests pointing to groups of different kinds of impressions, including those for which no impression was reported, the difference between z-scores for prose and poetry were highly significant, indicating the need to test separately for the distinct portion of the corpus consisting of prose separately from that of poetry.

The tests t comparing the averages of z-scores and z-smoothed of prose with those of poetry show a significant difference between the two groups (both with a p of 0.00). For prose, $_{M}$z-score = 0.35 e $_{M}$z-smoothed = 0.36, whereas for poetry the average z-score was 1.94 and for z-smoothed is 1.93. The Scheffé test demonstrated a significant difference between prose and poetry, regarding the z-score (p = 0,00) and the z-smoothed (p = 0,00). The z-score variation (standard deviation) with poetry was also greater than for that of prose.

The following graphics[1] (figures 1 and 2)., show a variation of the smoothed z-score according to its position in the accent group (-1 = V-V accented; accented position; -2 = V-V immediately before the

1 All graphics were obtained with the use of the program Statistica 6.0, of Statsoft, Inc.

accented V-V; -3 = V-V immediately before position -2 and so forth; -2 to -8 correspond to unaccented positions that precede the phrasal accent). Figure 1, representing prose, shows that the V-V units have a slightly shortened relative duration (z-smoothed) to position -5, at which point it increases again until it culminates in the accent (position -1)[2]. For poetry, Figure 2 shows the average of the z-smoothed sustained from position -8 to position -3, with the accent having influence only over the position immediately before it (position -2). This is due to a technique of enunciation common to poetry, which requires a regularity of syllable duration. We may conclude, therefore, that the data for poetry should not be mixed with the data for prose (ANOVA one-way with p = 0.0000 for duration of prose and not for poetry).

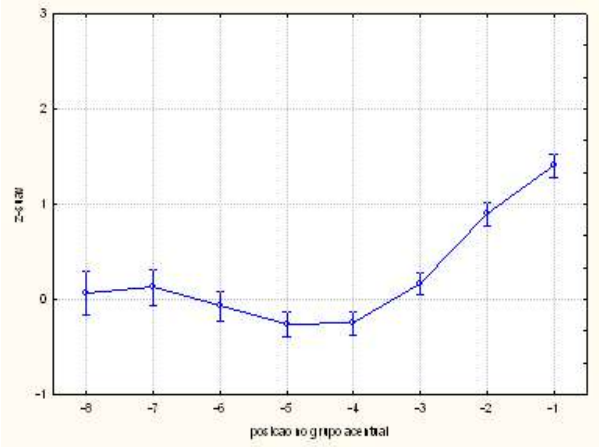

FIGURE 1 - PROSE Average value of z-smoothed according to position in accent group defined by the distance of the phrasally accented V-V unit.

Using ANOVA one-way for the factors of reading, (first or second time) presentation (phrase, paragraph, or complete text) and order (random or normal) for the poetry z-score, we can see that there is no statistical difference between the groups. The same result was obtained for prose. Although it was not statistically significant, it can be observed that those sentences for which no impression was reported are those with the smallest standard deviation both for prose and for poetry.

2 Positions with distances greater than 8 V-V units from the accent within the same accent group due to the reduced number of occurrences, equally for the prose data as for the poetry data.

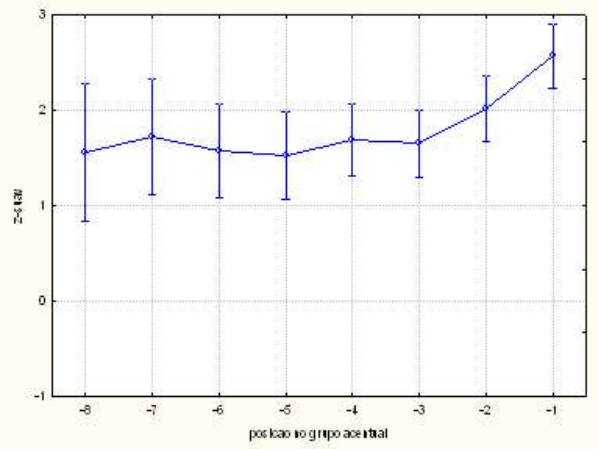

FIGURA 2 - POETRY Average value of z-smoothed according to position in accent group defined by the distance of the phrasally accented V-V unit.

The factorial ANOVA for z-smoothed with impression and accent group position as independent variables reveal that difficulty of pronunciation, humor, pride and disappointment result in durations similar to those with no report of impression (Figure 3), whereas tenderness and certainty affect the z-smoothed score, with the former shortening it and the second lengthening it, as compared to sentences without reportsuncommented sentences, especially in positions closest to the accent (Figure 4).

The results indicate that a test for neutral speech, should result in excising certain samples, which in the case of this pilot study, were all sentences reported by this speaker to have been spoken with “certainty” or “tenderness” in the prose corpus. On the other hand, for the case of poetic phrases/sentences, the only reported impression, humor, which, as mentioned above could be a possible effect of the reading technique common to the genre itself, no significant difference was found from uncommented sentences. Thus, for the purposes of this study, in terms of duration, all poetry samples may be considered examples of “neutral speech” by this speaker (Fig. 5). It may be noted that in prose, as well, humor appeared statistically similar to the uncommented sentences.

All ANOVA tests conducted to show the quality of z-smoothed according to accent group position, with these independent variables in addition to position: presentation, order and reading show that there is no significant difference between the groups either in prose or in poetry.

4.Final Considerations

For studies of prosody, our results point to the need for a test of reported impression by the speaker, that is a test for neutral speech, in order to determine after recording those sentences that were compromised by emotional effects, so that they may be excised from the corpus.

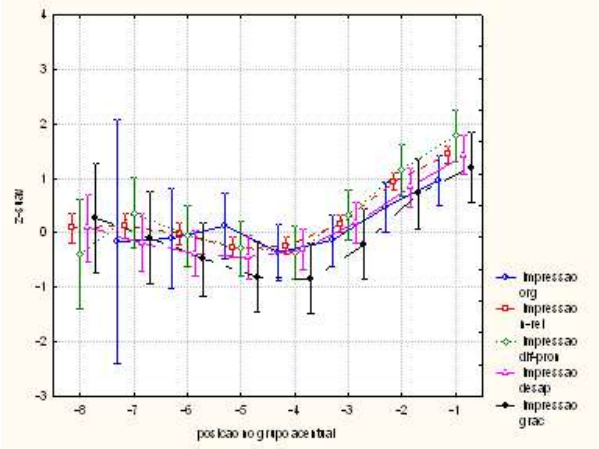

FIGURE 3 – Impressions about prose similar in quality to uncommented speech (for this speaker): disappointment, humor, pride, and difficulty of pronunciation. (pos inv AG = position in the accent group, using the V-V accented unit as reference, position -1).

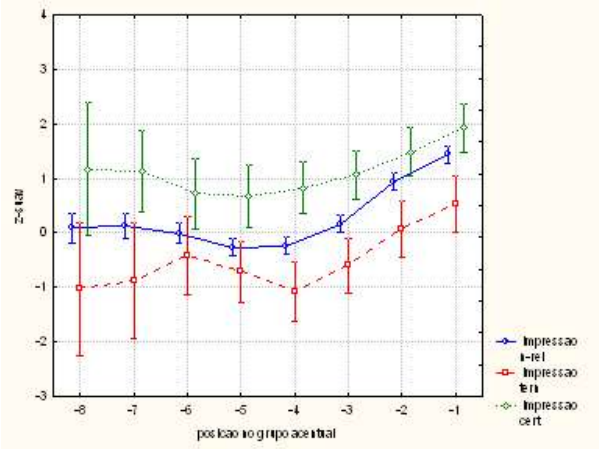

FIGURE 4 – Impressions about prose with statistically different durational quality than uncommented speech (for this speaker): tenderness and certainty. (pos inv AG = position in the accent group, using the V-V accented unit as reference, position -1).

The negative results of ANOVA for the format factor (in phrases, paragraphs or a complete text), indicate that there is no difference between the different formats of sentence presentations, corroborating the affirmation that the minimal prosodic unit corresponds to the minimal unit of meaning, that is to say that a text the size of a sentence is sufficient to insure the correct expression of its content. In this case, the preferential format for sentences to be presented is the original text, in paragraphs and in original order, since there was no significant influence between sentences, at least in terms of duration.

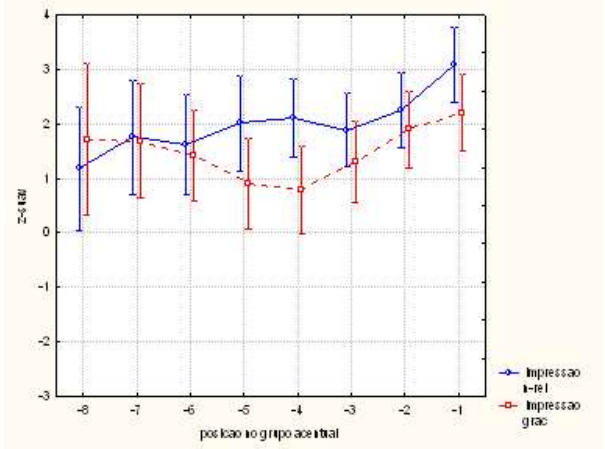

FIGURE 5 – Impression reported for poetry with durational quality statistically similar to uncommented speech (for this speaker): humor. (pos inv AG = position in the accent group, using the V-V accented unit as reference, position -1).

In conclusion, it may be said that the results are definitive for prose genre readings. For poetry, however, the question remains of whether control over a hypothetical production of neutral speech can be defined. Therefore, experimental investigation and an epistemological study are indicated to answer the question: what should be considered neutral speech for poetry?

5.References

1] ALBANO, E. & MOREIRA, A. A. "Archisegment-based letter-to-phone conversion for concatenative speech synthesis in Portuguese". *Proceedings ICSLP'96*, v.3, 1708-1711, 1996.

2] BARBOSA, P.A. "Elementos para uma tipologia do ritmo (lingüístico) da fala à luz de um modelo de osciladores acoplados", no prelo, 2004.

3] BARBOSA, P.A. "At least two macrorhythmic units are necessary for modeling Brazilian Portuguese duration: emphasis on segmental duration generation". *Cadernos de Estudos Lingüísticos.* Albano, E. C. (Ed.), 31, 33-53, 1996.

4] COURTÉS, J. & GREIMAS, A. J. *Dicionário de Semiótica*/trad. Alceu Dias Lima et all. São Paulo: Ed. Cultrix, (undated).