TEMPORAL PARAMETERS DISCRIMINATE BETTER BETWEEN READ FROM NARRATED SPEECH IN BRAZILIAN PORTUGUESE

Plinio A. Barbosa

Speech Prosody Studies Group, Instituto de Estudos da Linguagem, Univ. of Campinas, Brazil
pabarbosa.unicampbr@gmail.com

ABSTRACT

This work shows that 5 out of 6 acoustic parameters that correctly classify read and narrated speech in Brazilian Portuguese are temporal parameters. Several statistical models showed that significant differences between the styles are revealed by: speech rate, a measure of articulation rate, duration-related salience rate, mean and standard-deviation of degree of duration-related salience and mean of F0 first derivative. A set of 161 excerpts of narrated and read speech from ten speakers was used for training an LDA model. Another set of 57 excerpts with different subjects was used for testing the same model. Its performance with the six aforementioned parameters has achieved in the case of read speech an accuracy rate of 90 % for the training subset and 94 % for the test subset and in the case of narrated speech 70 % for the training subset and 27 % for the test subset.

Keywords: speech rhythm, speaking style, prosody, Brazilian Portuguese.

1. INTRODUCTION

Even for native listeners, it is not straightforward to correctly identify a stretch of running speech as being from read or spontaneous speech. The work by [12], for instance, shows an overall performance of 76 % for identification by native listeners of short personal interviews (spontaneous) and reading of the transcripts of the interviews by the same Dutch subjects. The study considered two subjects. After analysing the acoustic differences between the two aforementioned speaking styles, the author concluded that "overall, read speech compared to spontaneous speech had a lower articulation rate, more F_0 variation, more F_0 declination, less shimmer, and less vowel reduction". The author points out the difficulty of separating the two styles due to the high inter-subject variability. The kind of material (interviews and reading of their transcripts) is likely to be responsible for that. In fact, a recent study reveals a great similarity between the values of acoustic correlates of lexical stress in read and spontaneous interviews [2].

In a study with four female subjects in two dialects of German (two subjects per dialect), [14] compared spontaneous narratives with read speech. In their case read speech was represented by 12 read isolated utterances. The duration of each material amounted to 1 min of speech per style and per subject. Using the Fujisaki model to gain in generalisation from the F_0 traces, the authors concluded that, in read utterances, subjects generally accent more frequently and assign less prominence. They observed in the data that read speech was slower than spontaneous speech.

In contrast, [10] found, in a comparison between read and spontaneous speech in 16 speakers of American English, a speaking rate from 4 to 57 % faster in the former style. The author says that, whether the slower speaking rate in spontaneous speech is due to a higher number of salient pauses or slower articulation rate, remains to be examined. She also shows that F_0 maxima and average for intermediate phrase are higher in read speech, though the analysis is restricted to one of the speakers.

The study by [16] aimed at automatically classifying and detecting speaking styles in European Portuguese. They worked with read and spontaneous speech extracted from 30 daily news from the Portuguese television. Read speech was obtained from voice-over and anchors in daily news, and spontaneous speech from interviews and commentaries. Accuracies of 93.7 % for read and 69.5 % for spontaneous speech were obtained by using automatically-extracted phonetic and prosodic parameters in a total of 322 features from 35 phones. The temporal and prosodic parameters used were mean, median, standard variation, maximum and minimum of phone durations and likelihoods from the hidden Markov model used to segment the phones. Prosodic parameters were the first and second order statistics of the F_0/HNR envelope in every voiced portion of the segment as well as the parameters of a polynomial fit of order 1 and 2 of that envelope. The rate of voiced portions was also used.

The present paper adds to the studies in Dutch and

German by working with 10 Brazilian Portuguese speakers (both male and female) in read and narrated speech. By avoiding the use of transcripts from interviews, we allowed eliciting a more prototypical reading style. The use of more speakers was useful to disantangle within-subject from between-subject variation. All subjects taken into account, 5 temporal and one melodic parameter discriminate between read and narrated speech with a performance similar to that of [16]'s study.

2. METHODOLOGY

Reading and narrating styles were chosen for representing two manners of speaking. In an informal assessment by hearing, narrating speech distinguishes from read speech by: (1) the larger use of pausing, (2) the larger use of rising and high pitch due to nonterminal boundaries, and (3) the higher number of hesitations. The choice of these two styles is motivated by the fact that narration presents elements which can be found in spontaneous conversation, such as hesitations due to macro- and microplanning of the discourse. Though hesitations can occur in read speech, they are much less frequent than in the case of narration. This feature is important to be considered in developing an approach to describe speech rhythm in natural conditions and to investigate the possible differences between less and more controlled situations of utterance production.

2.1. Corpora

The corpora consist of two kinds of style: read and spontaneous speech. Each of them have two subsets: training and test, yielding training read speech, test read speech, narrating training speech and spontaneous (narration and interview) test speech. The training subset consists of speech material obtained from ten speakers (five female and five male) of Brazilian Portuguese (henceforth BP). First of all the speakers read a 1,600-word text on the origin of the Belém pastries (reading style, RE). After the reading, the same subjects told what the text was about (narrating style, ST). The speakers were Linguistics students aged between 30 and 45 years at the time of recording. Excerpts from 10 to 20 seconds were extracted from several parts of this material in order to make up the training subset with 161 excerpts. The test subset was formed by two kinds of material. One of the same nature of the training set and another with speech from spontaneous speech (short interviews). The length of the excerpts was the same as in the training material. Read excerpts were obtained from five male speakers and one female speaker. Narration excerpts were obtained from two female speakers (one being a professional actress narrating for a commercial CD). The rest of the test subset is additionally formed by excerpts of short interviews from four male speakers. A total of 57 excerpts were obtained, from which only 13 excerpts are constituted by read speech. The main goal of the excerpts of interviews was to test their automatic classification, since it contains several instances of narration.

2.2. Measuring Techniques and Parameters Extracted

According to a traditional approach in speech research [6, 13, 8], syllables were phonetically segmented by tracking two consecutive vowel onsets (VO). The segmentation was performed semi-automatically in Praat [5] in two stages: automatic VO detection by using the BeatExtractor Praat script available in [1], followed by manual correction, where applicable. Two consecutive VOs define a VV unit, which contains only a single vowel, starting at the first VO. The BeatExtractor script detects points in the speech signal where changes in the previously filtered energy envelope are relatively fast and positive (from low to high energy). This constitutes a single tier in the corresponding TextGrid file.

Duration-related stress groups were then delimitated by automatically detecting duration-related phrase stress positions throughout the utterances. The sequence of phrase stress positions was automatically tracked by serially applying two techniques for normalising the VV durations: a $z-score$ (z) transform (equation 1):

$$(1) \quad z = \frac{dur - \sum_i \mu_i}{\sqrt{\sum_i var_i}},$$

where dur is the VV duration in ms, the pair (μ_i, var_i), the reference mean and variance in ms of the phones within the corresponding VV unit. These references are found in [1, p. 489] for BP, followed by a 5-point moving average filtering (equation 2):

$$(2) \quad z^i_{filt} = \frac{5.z^i + 3.z^{i-1} + 3.z^{i+1} + 1.z^{i-2} + 1.z^{i+2}}{13}$$

The normalisation technique and the detection of duration-related phrase stress positions (detection of z_{filt} maxima) were performed by Praat ProsodyDescriptor script which extracts 14 parameters for each excerpt by using a pair of Sound and TextGrid files in Praat. The script was implemented by the author. This two-step normalisation technique aims

at minimising the effects of phoneme-size intrinsic duration in the VV unit. This normalised duration maxima signal both prominence degree and prosodic boundary strength, indistinctly. This is not a drawback of this approach, since the salience of these two prosodic functions on a particular word, equally signals perceived rhythm. Listeners of Romance languages often attribute both functions to a prominent or a pre-boundary word when evaluating these functions in their own languages [4].

After these steps, six temporal parameters were extracted. The first one is speech rate in VV units per second, extracted from the corresponding TextGrid file. Silent pauses are included in the VV intervals. The second to fourth parameters are the mean, standard-deviation and skewness of the z_{filt} maxima (mean-z, sd-z, sk-z), which reveal the structure of duration-related pooled salience degree and boundary strength in the excerpt. The use of salience/boundary distribution is crucial to produce an accurate description of speech rhythm, as recently claimed by [11, 7]. The fifth parameter is the rate of the z_{filt} maxima in peaks per second (max-zrate) which is meant to stand for the salience or prosodic boundary rate, for the reasons mentioned before. The sixth temporal parameter is a measure of articulation rate, non-salient VV rate (non-salient-rate). It was computed by taking the VV intervals not containing silent pauses or final-lengthened acoustic segments. Their rate was computed by dividing the number of such units in a particular utterance by the total duration delimited between the first and the last VO of the utterance.

To these temporal parameters, seven F_0-related parameters were extracted automatically: the rate of F_0 peaks in peaks per second (F_0 rate). This sequence of F_0 peaks is obtained from the audiofile in five steps: (1) extracting the F_0 trace using Praat (limits between 75 and 600 Hz), (2) smoothing the obtained contour with a 1.5-Hz filter, (3) interpolating the gaps due to unvoiced segments, (4) automatically counting the number of peaks in the contour, and (5) dividing the number of peaks by the total duration of the excerpt. The other six parameters are represented by three statistical descriptors of F_0 and its first derivative: median (F_0med and dF_0med), standard-deviation and skewness. The values of F_0 are computed in semitones re 1 Hz.

Spectral emphasis was computed according to Eriksson et al. [9]. Since it did not produce any significant results, no further details of its computation are given.

ProsodyDescriptor script delivers a text file with 19 columns where the first five inform language (in this case, BP), speaker identity, speaking style, gender and number of token from a same speaker/style added to the 14 columns corresponding to the respective 14 extracted prosodic parameters. Two files were generated at the output, one for the training subset and the other for the test subset.

The training subset output file was used to compute models of 2-Way ANOVA for each one of the 14 parameters with gender and style as factors. These analyses were carried out using the R package [15]. No violations of the ANOVA assumptions were found. η^2 effect sizes for the two factors and their interaction were also computed for each ANOVA model. The results point our mainly speaking style differences, but differences due to gender and speaker are also discussed.

The same subset was used to build LDA models with the relevant acoustic parameters for discriminating style as predictor variables and the speaking style as the predicted variable. Both the training and test subsets were used to predict style from the prosodic parameters.

3. RESULTS

We will concentrate here in results that point out to style differences (F_0med exhibited the highest effect size, 87 %, but associated with the GENDER factor). Five temporal parameters: speech rate ($R^2 = 0.29$), non-salient-rate ($R^2 = 0.28$), mean-z ($R^2 = 0.27$), sd-z ($R^2 = 0.21$), max-zrate ($R^2 = 0.20$) and one melodic parameter, dF_0med ($R^2 = 0.09$) were the only six parameters with an effect size higher than 5 % for the STYLE factor. The interaction between GENDER and STYLE was non significant for all models. Figure 1 gives a box plot showing results for speech rate. Read speech is faster in both genders and male speakers are faster in both styles. Non-salient-rate mirrors the results for speech rate with slightly higher figures (adding 1 VV/s in mean). Style is by far the main reason for the differences found (for speech rate it explains 28 % of the variance out of 29 % total variance explained). This tendency is confirmed in all subjects in different degrees, as can be seen in Fig. 2.

As for parameter mean-z, narrated speech has higher values, indicating the presence of longer silent pauses (VV intervals from which z-scores are computed may contain this kind of segment). Furthermore, female speakers have higher values indicating that their pauses and lengthened intervals are longer than in males. Since there is no interaction, this inter-gender behaviour is found in both styles.

Parameter max-zrate revealed higher values for

Figure 1: Speech rate according to style (RE and ST) and gender (MA=male and FE=female).

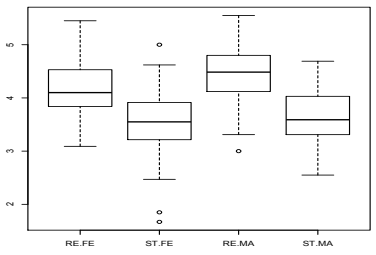
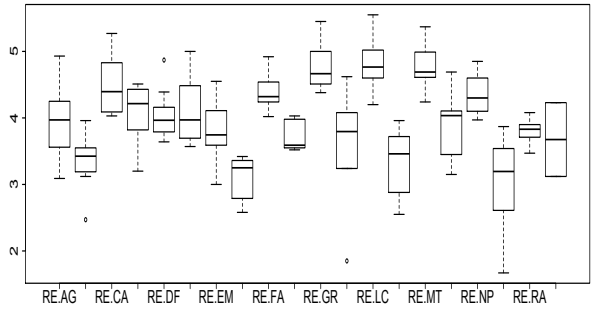


Figure 2: Speech rate across subjects according to style (RE and ST: tick to the right of the labelled one). Subjects AG, DF, GR, NP and RA are female.



males and read speech. This means that, in comparison with female and narrated speech, salient duration-related peaks have higher rates in male speakers and in read speech. The higher frequency of prominent units in read speech was pointed out by [14] for German. Parameter sd-z revealed higher values for females and narrated speech. This means that, in realising duration-related salience, female speakers vary more and there is more variation on that in narrated speech in comparison with read speech.

As for parameter dF_0med, factor GENDER is non-significant. Narrated speech has higher values for this parameter, which is related to a more frequent use of rising contours and high-pitched stretches in BP narration [3].

Because non-salient-rate is very close to speech rate in behaviour, the four remaining temporal parameters and the melodic parameters analysed here were used to build a LDA model for classifying style irrespective of gender. In the training set, the model classifies 90 % of read speech and 70 % of narrated speech correctly, a result very close to the results showed by [16] for European Portuguese.

The test subset, composed by 13 excerpts of read speech and 44 of narration and interviews, was predicted by the LDA model with 92 % and 27 % of correctness respectively for the read and non-read speech. Recall that there are no common subjects between the two subsets and that the majority of the excerpts in the test subset are from interviews. Furthermore, one of the narrators was a professional actress who did not produce any hesitation in her interpretation of a traditional story. When the two subsets are combined for training in a new LDA model the percentages of correct classification are 79 % and 73 % respectively for read and non-read speech. Which gives room for the possibility of automatic classification.

4. DISCUSSION AND CONCLUSION

This work shows that it is possible to classify read and narrated speech with a percentage of correctness of at least of 70 % for the training set, This is done by using a number of parameters far inferior to that based on HMM models (322 features in [16]).

The reason for the differences in performance between the two styles is certainly related to inter-subject variability in narration, irrespective of gender, as well as a higher variability of VV duration. A part of the values of the parameters for read speech overlap with those for narrated speech, which makes some excerpts of narrated speech look like read speech.

The presence of hesitation in non-professional narration, producing a behaviour more similar to the narrated speech in the training set, explains why professional narration was predicted as read speech in the test subset. Read speech has much lesser instances of hesitation than narrated speech.

We have found temporal variables to have a more determinant role in classifying read and narrated speech, suggesting that the use of F_0-related parameters are not so different across these two styles. The focus on the previous literature on melody instead of rhythm for disentangling the two styles could explain the relative lack of success in this initiative, as well as the use of a limited number of subjects.

The results of this work are relevant to building automatic classifiers of speaking styles and to understanding the factors affecting stylistic choices.

5. ACKNOWLEDGEMENT

The author thanks grant 301387/2011-7 from CNPq.

6. REFERENCES

[1] Barbosa, P. A. 2006. *Incursões em torno do Ritmo da Fala*. Campinas: Pontes/FAPESP.

[2] Barbosa, P. A., Eriksson, A., Åkesson, J. 2013. Cross-linguistic similarities and differences of lexical stress realisation in Swedish and Brazilian Portuguese. In: Asu, E., Lippus, P., (eds), *Nordic prosody. Proceedings from the XIth conference Tartu 2012*. Frankfurt am Main: Peter Lang 97–106.

[3] Barbosa, P. A., Mixdorff, H., Madureira, S. 2011. Applying the quantitative target approximation model (qTA) to German and Brazilian Portuguese. *Proc. of Interspeech 2011* Florence. 2065–2068.

[4] Beckman, M. E. 1992. Evidence for speech rhythms across languages. In: Tohkura, Y. e. a., (ed), *Speech Perception, Production and Linguistic Structure*. New York: IOS Press 457–463.

[5] Boersma, P., Weenink, D. 2014. Praat: Doing phonetics by computer. http://www.praat.org. Version 5.4.04.

[6] Classe, A. 1939. *The Rhythm of English Prose*. Oxford: Blackwell.

[7] Cumming, R. 2011. The language specific interdependence of tonal and durational cues in perceived rhythmicality. *Phonetica* 68, 1–25.

[8] Dogil, G., Braun, G. 1988. *The PIVOT Model of Speech Parsing*. Wien: Verlag.

[9] Eriksson, A., Thunberg, G. C., Traunmüller, H. 2001. Syllable prominence: A matter of vocal effort, phonetic distinctness and topdown processing. *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech '01)* Aalborg. 399–402.

[10] Hirschberg, J. 2000. A corpus-based approach to the study of speaking style. In: Horne, M., (ed), *Prosody: Theory and Experiment. Studies Presented to Gösta Bruce*. Dordrecht: Kluwer Academic Publishers 335–350.

[11] Kohler, K. 2009. Rhythm in speech and language: A new research paradigm. *Phonetica* 66, 29–45.

[12] Laan, G. P. 1997. The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style. *Speech Communication* 22(1), 43–65.

[13] Lehiste, I. 1970. *Suprasegmentals*. Cambridge, USA: MIT Press.

[14] Mixdorff, H., Pfitzinger, H. R., Grauwinkel, K. 2005. Towards objective measures for comparing speaking styles. *Proc SPECOM* Patras, Greece. 131–134.

[15] The R foundation for statistical computing. the R project for statistical computing. http://www.r-project.org/.

[16] Veiga, A., Celorico, D., Proença, J., Candeias, S., Perdigão, F. 2012. Prosodic and phonetic features for speaking styles classification and detection. *Advances in Speech and Language Technologies for Iberian Languages. Communications in Computer and Information Science* volume 328 89–98.