INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Cross-linguistic Distinctions between Professional and Non-Professional Speaking Styles

Plínio A. Barbosa[1], *Sandra Madureira*[2], *Philippe Boula de Mareüil*[3],

[1]Instituto de Estudos da Linguagem, University of Campinas, Brazil
[2]LIACC, Catholic University of São Paulo, Brazil
[3]LIMSI, CNRS & Univ. Paris-Saclay, Orsay, France

pabarbosa.unicampbr@gmail.com, sandra.madureira.liaac@gmail.com,
Philippe.Boula.de.Mareuil@limsi.fr

Abstract

This work investigates acoustic and perceptual differences in four language varieties by using a corpus of professional and non-professional speaking styles. The professional stimuli are composed of excerpts of broadcast news and political discourses from six subjects in each case. The non-professional stimuli are made up of recordings of 10 subjects who read a long story and narrated it subsequently. All this material was obtained in four language varieties: Brazilian and European Portuguese, standard French and German. The corpus is balanced for gender. Eight melodic and intensity parameters were automatically obtained from excerpts of 10 to 20 seconds. We showed that 6 out of 8 parameters partially distinguish professional from non-professional style in the four language varieties. Classification and discrimination tests carried out with 12 Brazilian listeners using delexicalised speech showed that these subjects are able to distinguish professional style from non-professional style with about 2/3 of hits irrespective of language. In comparison, an automatic classification using an LDA model performed better in classifying non-professional (96 %) against professional styles, but not in classifying professional (42 %) against non-professional styles.

Index Terms: speaking style, professional speech, cross-linguistic prosody

1. Introduction

According to [1], a speaking style can be defined as a differentiation in the way of speaking. It is usually related to changes in voice quality, speech rate, intonation and rhythm associated with specific communicative acts such as professional speech. This includes acted speech, TV and radio news, sports broadcasting, professional reading and narration [2, 3]. Often, speech from professionals involved in these activities can be distinguished from non-professional speech [4], which is certainly related to differences both in segmental and prosodic aspects of speech. But this is not necessarily true for all professionals nor for all situations where speech is a tool for entertainment or calling the audience's attention.

By using delexicalised utterances from four professional speaking styles in standard French, namely, sport commentary, religious sermon, political discourse and broadcast news, Obin and colleagues [5] showed that it is possible to identify these styles above chance in a forced-choice test. The best identified style was sport commentary followed by broadcast news, whereas religious sermon and political discourse were often confused with each other. Both broadcast news and political discourse are partially confused with each other, as well. The authors pointed out that speech rate, intensity, pausing and prosodic prominence organisation are the main parameters guiding listeners for doing the classification.

In the same direction, but analysing delexicalised, non-professional speech, Barbosa and Silva [6] showed that listeners discriminate read from narrated speech in Brazilian Portuguese based on speech rate, stressed syllable rate and pausing. Thus, temporal parameters play a crucial role for discriminating at least these two styles, as confirmed by a follow-up study [7].

In the study presented here, a corpus of professional and non-professional speaking styles in four language varieties was used for three purposes: (1) investigating acoustic differences across styles and languages; (2) investigating if Brazilian listeners are able to classify and discriminate speaking styles in the absence of lexical information and whether their ability to do so depends on prosodic information of the language heard; (3) investigating if the classification resulting from human perception is somehow related to an automatic classification based on the acoustic parameters investigated here.

We decided to focus on melodic and intensity parameters in this study because no annotation for automatic parameter extraction is required. The reason for this choice is that, in Brazilian Portuguese, the algorithms tested for detecting vowels yield an error rate of approximately 20 % which is harmful for obtaining accurate speech rate and syllable-size duration measures. Furthermore, the automatic aligners tested so far, EasyAlign [8] and the aligner of the FalaBrasil project of the Federal University of Pará, failed in detecting post-stressed vowels and segments before or after silent pauses.

Our research issues and hypotheses are the following: (1) melodic and intensity parameters can discriminate the professional and non-professional speaking styles studied here; (2) professional styles can be identified without lexical information independently of the language of the original audio file; (3) the human classification is more accurate than an automatic classification based on acoustic parameters.

We think that it is worth investigating the role of non-temporal parameters for classifying and discriminating speaking styles focussing on the contrast between professional and non-professional speakers. In the case of classification, human and automatic classification are compared. These are the main goals of this work.

2. Methodology

2.1. Corpus

The CROSS-RHYTHM corpus comprises recordings of professional speech retrieved from YouTube and direct recordings

http://dx.doi.org/10.21437/Interspeech.2017-7

of non-professional speech in Brazilian Portuguese (BP), European Portuguese (EP), standard French (FR) and standard German (GE). Non-professional speech includes reading and narration material of ten subjects (5 males and 5 females) in each language variety. All subjects are university grads aged between 25 and 40. No subject uses his/her voice professionally. The text which was read was "The Awkard Monk", a circa 1,600-word text originally written in EP about the origin of the Belém pastries. The text was adapted to BP by Juva Batella and translated into French by the third author and into German by Hansjörg Mixdorff. Narration was produced by the same speakers just after the reading in each language. It yielded between circa 90 and 500 words depending on the speaker.

Professional speech includes political and broadcaster styles represented by recordings of six speakers per language (3 males and 3 females) downloaded from YouTube by collaborators in Brazil, Portugal, France and Germany. The audio files were extracted in Wave format from the videos and last up to 5 minutes. For the broadcaster style, we selected speech excerpts following the news headlines of TV channels. For the political style, we selected discourses addressed to public audiences (e.g. the Parliament or discourses on TV addressed to the Nation).

All audio files were resampled at 16 kHz with 16 bits of quantisation. For acoustic and perceptual analyses, at least three excerpts from 10 to 20 seconds were selected from each recording producing 517 excerpts for acoustic analysis (a subset of which was selected for perceptual evaluation). According to previous studies [6, 9] which assessed speaking style differences in both Portuguese and French using sound pairs in a discrimination test, the 10-to-20 second duration provides a good inter-rater reliability, meeting two edge conditions: to have enough material for evaluating the prosodic structure of an utterance, on the one hand, and not to exceed the limits of the working memory [10], on the other hand.

2.2. Acoustic parameters

A script (*ProsodyDescriptor2*), running in Praat [11] was designed for this work. It computes eight prosodic parameters from an audio file and writes their values down in a text file as a table. The parameters are: median (F0med), standard deviation (F0sd) and Pearson skewness (F0sk) of fundamental frequency (F0) in semitones with 1 Hz as the reference tone; median (dF0med), standard deviation (dF0sd) and Pearson skewness (dF0sk) of F0 first derivative (F0 rate) in semitones per second; rate of smoothed F0 peaks in peaks per second (F0r), and spectral emphasis (emph), a correlate of vocal effort, according to [12]. For computing F0 peak rates, a low-band filter of 1.5 Hz of cut frequency was used in order to highlight crucial F0 peaks related to pitch accents, following previous experiments. Spectral emphasis is defined as $L - L_0$, where L is the total energy of the sound file up to the Nyquist frequency and L_0 is the low-band-filtered sound energy. For our analysis, as suggested in [12], we set the cut frequency of the low band at 400 Hz. All 517 excerpts were used to test acoustic differences between the four speaking styles in the four language varieties by using the 2-Way Scheirer-Hare-Ray non-parametric test for comparing means. The fixed factors for statistical analyses were style and language variety. When necessary, gender differences were also evaluated and reported here.

2.3. Perceptual tests

A classification and two discrimination tests were carried out with 12 Brazilian listeners, all grad students of the Federal University of Alagoas, Catholic University of São Paulo and State University of Campinas aged from 25 to 40 years old. For the classification test, 94 stimuli were selected among the 517 audio excerpts used for the acoustic analysis. These stimuli cover the four speaking styles in the four language varieties, balanced for gender. The PURR algorithm [13] was used to delexicalise these 94 stimuli. The PURR method resynthesises the entire audio file by producing speech-like data with the same original prosody. This was done to avoid identification of known speakers by Brazilian listeners, in the case of professional speech. Without delexicalisation, it is likely that listeners would correctly recognise political and journalistic styles, guided by the understanding of the content and speaker identity rather than prosodic characteristics. A forced-choice test was set up using the PsyToolKit platform available at "http://www.psytoolkit.org/" , which is able: to generate a random sequence of stimuli, to play the audio files, and to save the subjects' responses. Listeners were instructed to listen to each stimulus and select one style between four possibilities: reading, narration, political discourse, and broadcast news. They were informed that lexical information was removed and that they would hear speech samples as through a wall.

Two discrimination tests were designed: one with read vs narrated speech pairs, and the other combining pairs of broadcast news with either political discourse or narration. The rationale for designing this second test as such was motivated by frequent confusions between the news announcer style and the latter two speaking styles in the classification test applied first. For the first test, 56 pairs of stimuli with original audio files were used. Since the speakers are unknown to the listeners we decided not to delexicalise these stimuli. For the second test, 66 pairs of stimuli with delexicalised utterances were used for comparing styles in BP only or BP compared with the other three languages. The two tests were balanced for gender (pairs of the same gender and different genders were prepared). The utterances in each pair were separated by a pure tone of 1,000 Hz. The instructions given to listeners in the first test were to listen to utterance pairs and to indicate which element corresponds to reading. The instructions given to listeners in the second test were to listen to each pair and to indicate which element corresponds to broadcast news.

3. Results

We will first present the results of the acoustic analyses, then the classification tests followed by the discrimination tests. As for statistical analysis of the acoustic data, the non-parametric Scheirer-Hare-Ray (SHR) test with factors LANGUAGE and STYLE was applied and showed significant main effects and interactions for all factors and almost all combinations (exception: df0med). The following sections give Wilcox post hoc tests results with the Bonferroni correction. In all cases, $\alpha = 0.05$

3.1. Results of the acoustic analyses

Table 1 compares professional with non-professional speech in terms of significant differences of means for two varieties of Portuguese. Acoustic parameter abbreviations are listed in section 2.2 . The styles are abbreviated as follows: broadcast news (BN), political discourse (PL), reading (RE) and narration (NR).

For BP, Table 1 reveals that the two professional styles group together with higher mean values than in the two non-professional styles for parameters F0sd, F0sk and dF0sd, which are related to the variabiility of F0 contours and their rate. Fur-

Table 1: *Significant differences between professional and non-professional speech for BP and EP. Inequalities indicate the higher/lesser parameter mean in a style or style grouping.*

parameter	BP sig./grouping	EP sig./grouping
F0med	ns	PL > (BN=RE=NR)
F0sd	(PL=BN) > (RE=NR)	BN > RE
F0sk	(PL=BN) >NR >RE	(PL=BN) < (RE=NR)
dF0med	ns	ns
dF0sd	(PL=BN) > (RE=NR)	PL > (RE=NR), BN=PL
dF0sk	ns	ns
F0r	PL > RE	(PL=BN) > (RE=NR)
emph	ns	BN>PL>(RE=NR)

thermore, political discourse features a higher F0 peak rate mean than reading does. As for EP, political discourse (for F0med), broadcast news (for F0sd and emph) or both professional styles (for F0sk, dF0sd, F0r) show higher mean values for these parameters in comparison with reading (for F0sd) or both non-professional styles. Summing up for the two varieties of Portuguese, 6 out of the 8 parameters have higher mean values in the professional styles.

Table 2 compares professional with non-professional speech for French and German.

Table 2: *Significant differences between professional and non-professional speech for FR and GE. Inequalities indicate the higher/lesser parameter mean in a style or style grouping.*

parameter	FR sig./grouping	GE sig./grouping
F0med	(PL=BN) > (RE=NR)	PL > (BN=RE=NR)
F0sd	(PL=BN) > (RE=NR)	BN < NR
F0sk	BN < NR	PL<BN<(RE=NR)
dF0med	ns	ns
dF0sd	(PL=BN) > (RE=NR)	ns
dF0sk	ns	ns
F0r	BN > (RE=PL=NR)	(PL=BN) > (RE=NR)
emph	PL>BN > (RE=NR)	PL > (LE=NR=BN)

For French, Table 2 reveals that the two professional styles group together with higher mean values than the two non-professional styles for F0med, F0sd, dF0sd and emph, as well as lesser mean value for F0sk in comparison with narration. In French, too, news announcers exhibit the highest F0 peak rate, which is related to a more lively style. In German, political discourse (for F0med and emph) or both professional styles (for F0r) show higher mean values in comparison with the two non-professional and lesser mean values for F0sd and F0sk. Summing up for the four languages, the two professional styles are significantly distinct from the two non-professional styles for the following parameters: F0sd, F0sk and F0r, irrespective of language. These parameters concern F0 variability and rate, which are related to liveliness in speech, as also found by [14] for the radio style. The authors found that standard deviation mean for radio was 7.2 ST against 5.4 ST for reading. Also for French, [15] did not find global parameters for distinguishing reading from journalistic discourse. However the greater use of initial prominence in comparison with reading can distinguish these two styles.

As far as the 8 acoustic parameters examined here are concerned, BP does not distinguish political discourse from broadcast news, whereas the other three languages distinguish them at least in terms of spectral emphasis, with higher mean values for political discourse: 4.2 dB against 2.0 dB in EP, 1.6 dB against 0.8 dB in French, and 1.9 dB against 0.4 dB in German. In the latter, this distinction is restricted to female speakers and likewise female speakers in EP exhibit a higher F0 median in the political discourse, whereas the mean value for this parameter is higher in both genders in German (95 ST for political against 88 ST for broadcast news). There is also a gender difference between the two professional styles in French, where F0 peak rate is higher in the female political discourse.

As for cross-linguistic differences in the news announcer style, EP and BP show much higher mean values for spectral emphasis (3.4 dB) than French and German (0.6 dB). The same is true for F0 and dF0 standard deviations: respectively 3.0 ST and 6.2 ST/s against 2.0 ST and 4.0 ST/s in French and German. This may make news announcers sound more lively in the two varieties of Portuguese than in French and German. In comparison with BP, French exhibits a higher F0 peak. As for German, it exhibits the smallest value of F0 median, which may makes it sound lower in tone in this broadcaster style.

3.2. Results of the perceptual tests and automatic classification

In the classification test, with delexicalised stimuli, the reading style was not identified above chance (25 %). The other styles were identified in the following proportions: PL = 47 %, and BN = NR = 40 % (proportion test, $X^2 = 5.9$, p = 0.015 and power test = 67 %). Thus, political discourse is more easily identified than the other three styles. The confusion matrix in Table 3 displays the responses (columns) for each style (lines).

Table 3: *Confusion matrix for the style classification test. Raw frequencies and relative ones (%) in parentheses. Higher response frequency in bold.*

	RE	NR	BN	PL
RE	84 (29)	**90** (**31**)	76 (26)	38 (14)
NR	78 (28)	**112** (**41**)	47 (17)	39 (14)
BN	42 (15)	61 (22)	**111** (**40**)	62 (23)
PL	47 (17)	54 (20)	45 (16)	**130** (**47**)

It is clear from Table 3 that reading is most confused with narration. This may be due to the fact that what is read is a story and that the reading of a story and narration share some acoustic characteristics such as long silent pauses and hesitations which can be inferred from delexicalised speech. Reading may also be confused with broadcast news, whereas there is less confusion between narration and professional style stimuli. Political discourse is better classified and is confused almost evenly with the other three styles. The news announcer style is often confused with political discourse and narration. These results are similar to the results reported by [5] for the news announcer style in French. The reason for a better identification of the political discourse here is probably due to the absence of religious sermon in our corpus. For BP, on the other hand, [16] found a classification rate higher than 90 % for a corpus composed of religious sermon, political discourse and TV broadcast news, also using delexicalised stimuli. As for listeners' performance according to language of the original samples, there is no statistical sig-

nificance for $\alpha = 5\%$. Since the stimuli are delexicalised, this result suggests that cross-linguistic prosodic differences are not that important for style classification in this corpus. This suggests that there are aspects in common for the style typology across the four languages studied here.

As for the two discrimination tests, they revealed that:

- in the first discrimination test with original speech of reading and narration, hits are higher than 90 % for reading choice, irrespective of the original language;
- in the second discrimination test with delexicalised speech of political and broadcast news, hits for the second choice are higher than 65 % (distinct from chance for $\alpha = 1$ %) within BP or with BP compared with the other languages;
- in the second discrimination test with delexicalised speech of narration and broadcast news, hits for the second choice are higher than 80 % (distinct from chance for $\alpha = 1$ %) within BP or with BP compared with the other languages.

Classification was also done with the acoustic parameters using the automatic technique of Linear Discriminant Analysis (LDA). By using the 517 excerpts and the eight acoustic parameters extracted from each stimulus to predict the four style categories, we obtained the following predictions, given in the confusion matrix in Table 4.

Table 4: *Confusion matrix for the LDA predictions. Raw frequencies and relative ones (%) in parentheses. Higher response frequency in bold.*

	RE	NR	BN	PL
RE	**165** (**76**)	44 (20)	7 (3.5)	1 (0.5)
NR	77 (44)	**91** (**52**)	5 (3)	2 (1)
BN	**29** (**48**)	8 (13)	14 (23)	9 (15)
PL	18 (28)	18 (28)	2 (3)	**27** (**41**)

By comparing the relative frequencies in Tables 3 and 4, reading and narration are better discriminated by the LDA model, that is, from acoustics only. Yet, narration is more often confused with reading than in the perception test shown in Table 3. As for broadcast news, listeners perform better with 40 % of hits against 23 % in the LDA model, where a higher confusion is observed both perceptually and automatically. An analysis broken down by language reveals that: (1) political discourse is better classified in EP perceptually (in BP and French it is often confused with narration), followed by German, where it is often confused with reading; (2) the news announcer style is better classified by the LDA model in EP (in BP it is often confused with political discourse and reading, whereas in German and French it is more often confused with reading); (3) reading is better classified by the LDA model in all languages, especially in BP, with less confusion with the other styles; (4) narration is better classified by the LDA model in EP, French and German, with more confusion with reading although less confused than in BP.

4. Discussion

The set of parameters analysed here, which deliberately excluded strictly temporal parameters, reveals that, with prosodic information only, listeners confuse reading with narration. This is not the case for the automatic classification based on the same acoustic-prosodic parameters, for which a hit rate higher than 75 % for reading stimuli and higher than 50 % for narration stimuli was obtained. The reason is probably due to the fact that the reading in our corpus is the reading of a long story where pausing and hesitations are frequent, similar to the ones in narration, which makes the forced-choice classification task more difficult. These acoustic cues can be perceived even in delexicalised speech. With both segmental and prosodic available in the first discrimination test, hit rate for reading is higher than 90 %, a performance certainly explained by the fact that listeners had access to the content of the spoken chain. Even in the case of delexicalised speech, as for the broadcaster style, perceptual discrimination is higher than the one obtained with the LDA model. Discrimination for the political discourse is very similar, irrespectively of method used.

It is important to highlight the closeness of the discrimination between reading and narration in the first discrimination test and the classification achieved by the LDA model. In a sense, this confirms that these two styles can be distinguished by the parameters examined here and not only by strictly temporal parameters. Language does not interfere in discrimination and classification hits, even if language may be identified based on prosody alone, according to [17]. Nevertheless, there is slight more confusion between reading and narration in the classification test in BP. Broadcast news and political discourse tend to be better classified in EP followed by German.

If we take only the two professional vs the two non-professional styles as pooled groups, the classification done by the listeners performs better than the one obtained by the LDA model in the case of the professional styles: 65 % of hits for non-professional-style stimuli and 63 % of hits for professional-style vs. 96 % of hits for non-professional-style and 42 % of hits for professional-style using the LDA model.

5. Conclusions

By investigating the acoustic differences across four styles and four language varieties, we showed that up to 6 out of 8 parameters distinguish professional from non-professional style in the four languages. This confirms the first hypothesis raised in the Introduction section. Brazilian listeners are partially able of identifying professional style against non-professional style with no lexical information at a proportion of about 2/3 of hits. They do that irrespectively of language, which confirms hypothesis 2. By investigating both perceived and automatic classification, we concluded that the LDA model is better in classifying non-professional styles, but not in classifying professional styles in comparison with the listeners. This partially confirms hypothesis 3. However, when the listeners have no available lexical information, the LDA model is far better in discriminating reading from narration.

6. Acknowledgements

This work was developed in the framework of the CNPq project "Cross-linguistic Analysis and Statistical Modelling of the Link between Speech Rhythm Production and Perception in Different Speaking Styles", grant # $476358/2013-2$. The first author also thanks the CNPq grant # $302657/2015 - 0$. We thank our colleagues H. Mixdorff and C. Oliveira for helping us in recording and selecting the material respectively for German and EP. R. Monteiro helped with the preparation of delexicalised speech in BP and EP. We also thank our speakers and listeners.

7. References

[1] J. T. Irvine, "Style as distinctiveness: the culture and ideology of linguistic differentiation," in *Style and sociolinguistic variation*, P. Eckert and J. R. Rickford, Eds. Cambridge: Cambridge University Press, 2001, pp. 21–43.

[2] M. Eskénazi, "Trends in speaking styles research," in *Proceedings of the Eurospeech 1993*, Berlin, Germany, 1993, pp. 501–509.

[3] P. Léon, "Variation situationnelle et voix professionnelles," in *Précis de phonostylistique*. Paris: Nathan, 1993, pp. 157–184.

[4] I. Fónagy and J. Fónagy, "Prosodie professionnelle et changements prosodiques," *Le Français Moderne*, vol. 44, pp. 193–228, 1976.

[5] N. Obin, P. Lanchantin, and X. R. A. Lacheret, "Discrete/continuous modelling of speaking style in HMM-based speech synthesis: Design and evaluation," in *Proceedings of the Interspeech 2011*, Florence, Italy, 2011, pp. 2785–2788.

[6] P. A. Barbosa and W. da Silva, "A new methodology for comparing speech rhythm structure between utterances: Beyond typological approaches," in *PROPOR 2012, LNAI 7243*, H. Caseli *et al.*, Eds. Heidelberg: Springer, 2012, pp. 329–337.

[7] P. A. Barbosa, "Temporal parameters discriminate better between read from narrated speech in brazilian portuguese," in *Proc. of the 18^{th} International Congress of Phonetic Sciences*. Glasgow, UK: The Scottish Consortium for ICPhS 2015, 2015.

[8] J.-P. Goldman, "Easyalign: an automatic phonetic alignment tool under praat," in *Proceedings of the 12^{th} Annual Conference of the International Speech Communication Association*, Florence, 2011, pp. 3233–3236.

[9] P. B. de Mareüil, A. Rilliard, and A. Allauzen, "A diachronic study of initial stress and other prosodic features in the french news announcer style: corpus-based measurements and perceptual experiments," *Language and Speech*, vol. 55, no. 2, pp. 263–293, 2011.

[10] N. Cowan, *Attention and Memory. An Integrated Framework*. New York: Oxford University Press, 1997.

[11] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," http://www.praat.org/, 2016.

[12] H. Traunmüller and A. Eriksson, "Acoustic effects of variation in vocal effort by men, women, and children," *J. Acoust. Soc. Am.*, vol. 107, pp. 3438–3451, 2000.

[13] G. P. Sonntag and T. Portele, "PURR - a method for prosody evaluation and investigation," *Journal of Computer Speech and Language*, vol. 12, no. 4, pp. 437–451, 1998.

[14] J.-P. Goldman, A. Auchlin, M. Avanzi, and A.-C. Simon, "Prosoreport: an automatic tool for prosodic description. application to a radio style," in *Proceedings of Speech Prosody 2008*, Campinas, Brazil, 2008, pp. 701–704.

[15] J.-P. Goldman, A. Auchlin, and A.-C. Simon, "Description prosodique semi-automatique et discrimination de styles de parole," in *Actes du Colloque Interfaces Discours-Prosodie 2009*, H.-Y. Yoo and E. Delais-Roussarie, Eds., Paris, 2011, pp. 207–221.

[16] L. Castro, "O comportamento dos parâmetros duração e frequência fundamental nos fonoestilos político, sermonário e telejornalístico," Ph.D. dissertation, Federal University of Rio de Janeiro, 2008.

[17] J. Ohala and J. B. Gilbert, "Listeners' ability to identify languages by their prosody," in *Problèmes de prosodie, Vol. II: Experimentations, modèles et fonctions*, P. Léon and M. Rossi, Eds. Ottawa: Didier, 1981, pp. 123–131.