Speech Prosody 2016
31 May - 3 Jun 2106, Boston, USA

Elicitation techniques for cross-linguistic research on professional and non-professional speaking styles

Plínio A. Barbosa[1], *Sandra Madureira*[2]

[1]Speech Prosody Studies Group, Dep. of Linguistics, Univ. of Campinas, Brazil
[2]LIACC, Catholic Univ. of São Paulo, Brazil
pabarbosa.unicampbr@gmail.com, sandra.madureira.liaac@gmail.com

Abstract

This paper presents a detailed report on previous experiences with the elicitation of speech material for cross-linguistic research in both professional and non-professional speaking styles. Two main methodological issues arise in cross-linguistic research: to ensure a parallel corpus for appropriate comparisons across languages, and to ensure that perception tests based on the elicited material induce behaviours or assessments of listeners in similar conditions. Professional speaking styles usually require a control condition for comparison, and this is not easy to obtain. This problem is still more crucial when it needs to be addressed cross-linguistically. Both the collected material and the process for obtaining it taught us some important lessons over and above the inclusion of tasks such as asking for the help of phonetically trained native speakers of the involved languages: extended amounts of spontaneous speech can be obtained from interviews with close friends; these interviews can be read by the same speakers to ensure a comparison between read and spontaneous speech and such a comparison can be done in several languages; regarding perception, the stimuli used for testing can come from different languages if a delexicalisation procedure is used; both for reading and narration, text choice has consequences on listeners' behaviour.
Index Terms: elicitation technique, speaking style, cross-linguistic research

1. Introduction

In a thoroughly survey, Niebuhr and Michaud [1] have recently pointed out the challenges we should face and the care we should take to elicit appropriate speech material for research. Care should not be restricted to recording conditions and equipment quality alone, but should crucially consider the choice of speakers, the kind of task and the interaction between tasks and speakers. As the authors remind us, speakers are not machines, but very creative persons sensitive to environmental conditions. The consideration of subjects' impressions after having carried out tasks often unveil the presence of non-controlled nuisance variables or aspects to be taken into account that help advance our understanding of speech communication.

This paper considers a great deal of our previous experience in elicitation techniques for prosody research. Several methodological aspects concerning elicitation of speech material will be discussed in detail. These elicitation techniques helped us to tackle theoretical issues on prosody production and perception. The studies considered here deal with the cross-linguistic comparison of speech rhythm and intonation involving languages such as Brazilian Portuguese (henceforth BP), European Portuguese (EP), Swedish, Standard German and Standard French.

Section 2 presents five studies conducted on different aspects of rhythm and intonation research whose setup for collecting the speech material allowed us to learn important lessons as regards elicitation. These lessons are further presented in detail in section 3. We summarised the lessons to take home in section 4. The conclusion section closes the paper.

2. Theoretical issues and elicited material

In [2], we compared read and narrated speech in both Standard German and BP. Our main concern was to describe and model the way subjects in both languages use intonation for these two different tasks. A 1,600-word BP text on the origin of Belém pastries was translated into Standard German by the second author in [2]. The text was originally written in EP and adapted to BP mainly due to differences in lexical choice. Translation was made sentence by sentence in order to allow direct comparison across the two languages both for the sake of description and modelling. The Fujisaki model [3] was used to generate the melodic contours in the two languages and the two speaking styles. The size of the text, unknown to all speakers in both languages, allowed a narrated material ranging from circa 100 words to 600 words depending on the subject. Six subjects (two females and four males) in each language were recorded. This paired corpus, "The Belém Corpus", has been translated from the original EP text into several languages and its recordings can provide a multilingual corpus for future works. The speakers were students between 30 and 45 years old and had Linguistics (BP) or Computer Science (German) backgrounds. The analyses were based on excerpts from 150 to 200 words in each language and style. The experience acquired during the elicitation of both read and narrated speech concerns mostly the consequences of text choice and extension. The latter has consequences for the quantity of material obtained in narration tasks.

In a study [4, 5] part of a Swedish project coordinated by Anders Eriksson entitled "A typology for word stress and speech rhythm based on acoustic and perceptual considerations" and aiming at uncovering the production mechanisms and the perceptual cues signalling lexical stress as a function of speaking style in five languages (Swedish, English, French, Estonian and BP) a parallel corpus across languages was built and consisted of three speaking styles: word list reading, sentence reading and informal interview, the latter two in comparable phonic contexts. Such a corpus enabled the examination of the consequences of the level of familiarity between interviewee and interviewer for eliciting spontaneous speech.

In [6], it was necessary to compare a professional speaking style with a control condition. Intonation patterns in the speech

 doi: 10.21437/SpeechProsody.2016-103

produced by a radio speaker simulating broadcasting outside his place of work was compared with intonation patterns previously used by him during an informal interview. Twenty-nine headlines were carefully chosen from this interview to be announced by the same speaker in a broadcasting style. Lessons learned from this study concerned the finding that factors such as the syntactic structure of the headlines (use of words with maximal information load value and abbreviations) besides the order (interview should come before announcing) and the place (a cozy place for the interview and a studio for the broadcasting style) of the recording sessions influence speech production and therefore should be taken into account in recording these styles.

In [7], a study was carried out to compare the production and perception of rhythm in two speaking styles in BP by using the Belém Corpus. Yet another study [8] carried out within a larger project concerning professional and non-professional speaking styles in BP, EP, Standard French and Standard German aimed at describing prosodic differences used the Belém Corpus for reading and narration besides dialogues, TV broadcasts and political discourse as corpora. The choice of equivalent material for professional speakers was not simple and some key issues related to this will be discussed in the next section.

3. Eliciting corpora for prosody research

Several lessons were taken from our previous experience in eliciting speech material for prosody research which can be useful for newcomers. Most of the knowledge obtained came from the consequences of aspects not thought of before recording and experimental setup took place. We present here four main issues for obtaining appropriate data for production and perception studies.

3.1. Reading and immediate narration: text choice

As reported in the previous section, "The Belém Corpus" allowed us to obtain equivalent read and narrated material for EP, BP, French and German. No results regarding the data obtained for EP are discussed here nor does the comparison between narration and reading in EP or between EP narration and the other languages concerned since Portuguese speakers used episodic memory in their narratives as the text about the Belém pastries is part of their cultural heritage.

In fact, though almost all Portuguese speakers did not know the details about the origin of the pastis de Belém, (pastries from Belém), they, anyway, had something to tell about them because they are part of their culture. The consequence was that all Portuguese speakers had something to add when narrating the text, making use of their episodic memory. Narration was thus very different from narration by the speakers of other cultures, including Brazilian. Narration in EP was more colloquial and less hesitant than narration in the other languages.

Another crucial issue was text adaptation. Even if the original EP text is readable and understood by Brazilian subjects, text adaptation was necessary mainly due to lexical choice which causes more hesitations when BP speakers find unfamiliar words.

The size of the text allowed a more extensive narration, although this extension varied across subjects. There was also a trend for male speakers to narrate in less time than female speakers, because they gave less details about the story. The amount of material is not only important to allow generalisation in describing prosodic differences across styles and languages, but, crucially, to allow generalisation for modelling approaches as it was the case of analysis-by-synthesis with the Fujisaki model in [2].

Narration was done immediately after reading, using the same equipment for all speakers, in an attempt to avoid the fading of events over time. This is not a problem per se, but our interest was to ensure more narrated material for analysis.

All subjects were fluent in reading, which is an important issue when working with read material. This ensures less pausing and hesitation, if these phenomena are not the main ones to be considered for study. Of course, fluency is an issue in particular studies, such as research with children before they get fluent, dyslexic people, stutters. For these cases, a measure of the degree of fluency should be a variable to be considered for statistical analysis.

3.2. Contrasting spontaneous and read speech

One of the most challenging issues when contrasting prosodic aspects between read and some types of spontaneous speech is to respect the *ceteris paribus* condition. That is, it must be ensured that the prosodic features encountered are only due to stylistic differences and not to context variation in the syntactic, the phonological and the semantic levels.

Since it is not possible to obtain good results by "simulating" a spontaneous speech by reading texts, even if the reader of the text is also its writer, another strategy to elicit comparable data is to be found. In the framework of the aforementioned Swedish project, we decided to start by eliciting spontaneous interviews. Spontaneous conversation was chosen due to its pervasiveness in communication. To allow obtaining longer stretches of monologues, we decided to set up interviews. For the BP data, interviews were done between close friends, allowing to obtain fairly long periods in which only the interviewee speaks (a similar, successful experience using close friends was also described by [9]). Material from 15 to 45 minutes was obtained that way. In [4, 5] cited above, five male interviewees and five female interviewees in both BP and Swedish were recorded for analysis. All subjects were young people from the universities of Campinas and Gothenburg. All interviews took place in a lab environment, but this setup was not taken to be a potential cause of unexpected behaviour since not only the speakers were Linguistics students used to recording situations, but mainly because they were talking to their close friends. This is confirmed in the work by [10], based on two studies with field and laboratory recordings, which suggests that the recording equipment was found to matter less than the interlocutor. Interacting with a close friend or a stranger had a stronger influence on the manner of speaking.

All interviews were transcribed orthographically to be read up to two weeks later by the same speaker who participated in the interview. Fifteen selected chunks of the interviews not containing hesitations or long pauses were chosen for the reading task. The selection also considered that syntactic structures were compatible with reading. Reading was made easier because each speaker faced his/her own lexical choice. Fifteen isolated words were chosen to be read in isolated form just after the reading of the chunks of the interviews. These words were chosen from the read material, one from each selected chunk. These three kinds of speech material were used for studying the acoustic correlates of lexical stress in three conditions: isolated read word, read words in context and spontaneously produced words. In the case of isolated word reading, word list intonation was to be avoided by saying to the subjects to read each word as if it was the only one in the list.

Results revealed that the hierarchy and effect size of the acoustic correlates of lexical stress were the same for words in interviews and connected read speech (see a related issue in [11]). Differences in word list reading were related to final lengthening and low melodic levels marking boundaries corresponding to declarative intonation. Despite instructions, some speakers still used word list intonation at some degree. Unfortunately, this behaviour was not easy to verify during the task.

3.3. Contrasting professional and non-professional speech

Ensuring comparable conditions when comparing professional and non-professional speech is also a challenge, since, ideally, an equivalent text content should be obtained for professional and non-professional speech data.

In [6], the author decided to work with an AM-radio broadcaster who hosts a variety show in a radio station in Campinas. The author compared melodic contours of headlines read professionally with a very similar material spoken non-professionally. In order to elicit data, she recorded the speaker outside his place of work, in our lab. The strategy adopted was to make him speak about his professional life to her in an informal interview. The speaker reported several aspects of his experience during 45 minutes. The majority of the features of his professional style disappeared due to this setting. After transcribing the interview orthographically, 50 sentences intuitively chosen for being similar to the syntactic construction of headlines were given to a journalist to choose the ones acceptable for this kind of professional broadcasting. From this set, 29 headlines were chosen which were spoken by the radio speaker in the lab while instructing him to read them as if he were in the radio station during his program. These utterances sounded like professionally spoken material. They were read in normal and self-chosen fast speech. The main differences between the melodic contours of the headline-like utterances in the original interview and the headlines read professionally are related to the rate of pitch accents and speech rate.

In a recent, preliminary work for the "Cross-linguistic Analysis and Statistical Modelling of the Link between Speech Rhythm Production and Perception in Different Speaking Styles" (CROSSTYLE) project being conducted for comparing professional and non-professional styles in BP, EP, Standard German and Standard French, we decided to gather professional speech from podcasts available in the Internet. In the case of professional speech, political discourses and news commentaries were chosen for analysis and gender-balanced (three males and three females for each speaking style). In this case, there was no possible matching of content. But other challenges arose.

As for political discourses, we decided to choose discourses given to a general audience (that is, discourses addressed to the party general assemblies and private interviews for journalists were included in the corpus) for all languages. This is an issue related to content type or genre. The excerpts lasted between 1 to 3 minutes to match the non-interrupted news commentaries of TV professional speakers, which rarely lasted more than 3 minutes.

As for non-professional speakers, five male and five female subjects recorded material for the Belém corpus as well as participated in a Video Silent Task (VST) [12]. The difference between the VST from the usual Map Task [13] is that, for the former, the amount of speech material is balanced between the participants because they watch a short, silent movie and talk with each other about what they have seen to try to reconstruct the story. Dialogues were obtained that way for BP and French, so far. Although the material is not matched as for context, general melodic and rhythmic parameters can be extracted to examine what kinds of global prosodic descriptors reveal differences within and across languages for the same style contributing to the understanding of this kind of stylistic variation.

For ensuring descriptors be computed in the same way, a set of 14 melodic and rhythmic acoustic parameters were automatically obtained by a script running in Praat. Having as input files an audio file paired with its corresponding annotation in syllable-sized units, these descriptors allow a parallel comparison across styles and languages. In a preliminary result with 10 speakers where read and narrated speech in BP were compared [8], we used an LDA for discriminating the two styles, which shows that it is easier to predict read speech due to the great variability inherent to narrated speech.

3.4. Obtaining appropriate perceptual data for analysing the effectiveness of speaking style

A similar result for these two styles in BP was obtained in earlier work [7], which used a discrimination task to evaluate if the listeners were able to assess the difference between speech rhythm in read and narrated speech. There are three challenges here: (1) the kind of instruction to give to a lay person, (2) the size of the paired audio files to be compared, and (3) how to do comparisons across languages.

Lay persons have difficulties in understanding what speech rhythm means. Thus, we cannot ask people to judge differences in speech rhythm in two audio chunks. Even if it is not at all clear what people understand with the expression "manner of speaking" (*modo de falar* in Portuguese), we decided to ask them to judge the degree of difference between the manners of speaking in the two files in a five-point Likert scale ranging from identical to completely different. The comments made by each participant after the task and their consistence in giving the same response when the order of the paired stimuli was changed and presented in a random order revealed that the instructions were followed, that is, matched the expert expectations concerning rhythm. This seems to be the case because lay judgments are only correctly predicted in multiple correlations when classical acoustic parameters related to rhythm are used as predictor variables.

The reason for choosing relatively long stretches of speech to be evaluated (from 10 to 20 s) was guided by the high standard-deviations of the listeners responses obtained in a previous study for excerpts of 1 to 2 seconds [14]. The long-duration excerpts allow the listeners to more accurately evaluate the manner of speaking than short-duration excerpts (see a similar extension for voice similarity judgement in [15]).

Cross-linguistic comparison of manners of speaking is also scheduled for the CROSSTYLE project mentioned above. The main problem of assessing manners of speaking in comparing languages is that recognition of a given language could misguide listeners in their responses. That is why it is important to preserve prosodic information and make language recognition opaque for the listeners. This can be done by using a technique called delexicalisation. Delexicalisation is the method of suppressing the segmental information from the speech signal to render it unintelligible while preserving their prosodic characteristics. Vainio and colleagues method [16] combines inverse filtered glottal flow and all-pole modelling of the vocal tract with the advantage of preserving voice quality. For this reason we used it for delexicalising all audio files for the study in

[7]. A few weeks later, we repeated the same discrimination test with the original files and the results were statistically the same. It must be reiterated, though, that only BP was examined in this work.

Assessing behaviour of listeners in a perception task can be done by carefully examining the histogram of their responses to evaluate if one or more of them is behaving very differently from the others. This technique presupposes some reliability among the listeners (raters), which can be evaluated with inter-rater reliability tests such as kappa-based indices (Cohen's for two raters, Fleiss' for m raters) for graded scales such as the Likert scale. Raters taken as outliers can be found by computing z-scores and discarding raters whose absolute mean response z-scored values are higher than 3. The same technique can be used for discarding stimuli that triggered random or anomalous behaviour of the listeners. Preserving coherence across raters and stimuli allows a robust assessment of the perceptual test.

4. Discussion: lessons to take home

From the reports presented in the previous sections, the following aspects of appropriate eliciting of speech material for prosody research can be summarised.

1. The quality of read material depends crucially on the behaviour of the subject for the task of reading a text. This behaviour changes according to the chosen text and its extension, as well as the subject's fluency in reading;
2. To ensure the same context is obtained for comparing spontaneous with read speech, eliciting interviews with close friends is suggested. In our experience, this strategy produced longer monologues. After this session, the reading of the orthographic transcriptions of these interviews produced a good-quality read material comparable to the original interviews in terms of phonological and phonetic context. Care should be taken not to choose excerpts from the interviews not suitable for reading due to syntactic complexity, interruptions, reformulations and related issues;
3. The same procedure of doing an interview at first and then reading the transcription of this interview can be used to study professional styles. When reading, the professional speaker can be asked to read professionally, as if in a studio. Different kinds and rates of reading can be tested, depending on the speaker's professional experience;
4. To ensure comparability across different professional speech materials across languages and subjects, it is possible to match them as for content type, extension, gender, addressee, among other aspects;
5. Delexicalisation is a useful method to make language recognition opaque, which is important for doing perception test cross-linguistically, when language recognition could misguide the listeners' responses;
6. Robustness of statistical analyses from perception tests' responses can be ensured by trimming the responses as for outliers.

5. Conclusion: a general outlook

From what was examined here, it seems that comparability and ecological validity of recorded speech in cross-linguistic and cross-style research are crucial issues for the future. As for ecological validity, we mean the closeness of the recorded material to speech used in everyday life, especially in spontaneous conversations. This may be require, more often then it has been done so far, the use of discreet, good-quality recording equipment during spontaneous conversations. An example of a rich material obtained from these kinds of recording is the C-ORAL-Brasil corpus developed by Raso and Mello [17] for studying informal speech from Minas Gerais State Brazil. Comparability requires combining more controlled material with spontaneous material, which can be obtained by the kind of technique shown here and in [11].

6. Acknowledgements

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

7. References

[1] O. Niebuhr and A. Michaud, "Speech data acquisition - the underestimated challenge," *Kieler Arbeiten in Linguistik und Phonetik (KALIPHO)*, vol. 3, pp. 1–42, 2015.

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

[3] H. Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of japanese," *Journal of the Acoustical Society of Japan*, vol. 5, no. 4, pp. 233–241, 1984.

[4] P. A. Barbosa, A. Eriksson, and J. Åkesson, "On the robustness of some acoustic parameters for signalling word stress across styles in brazilian portuguese," in *Proceedings of the 14th Annual Conference of the International Speech Communication Association*. London: Causal Productions, 2013, pp. 282–285.

[5] A. Eriksson, P. A. Barbosa, and J. Åkesson, "The acoustics of word stress in swedish: A function of stress level, speaking style and word accent." in *Proceedings of the 14th Annual Conference of the International Speech Communication Association*. London: Causal Productions, 2013, pp. 778–781.

[6] L. C. P. Campos, "Radialista: análise acústica da variação entoacional na fala profissional e na fala coloquial," Master's thesis, University of Campinas, 2012.

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

[8] P. A. Barbosa, "Temporal parameters discriminate better between read from narrated speech in Brazilian Portuguese," in *Proceedings of the 18th International Congress of Phonetic Sciences*, T. S. C. for ICPhS 2015, Ed. Glasgow, UK: The University of Glasgow, 2015, pp. 1053: 1–5.

[9] O. Niebuhr, B. Peters, R. Landgraf, and G. Schmidt, "The Kiel Corpora of "Speech & Emotion" - a summary," in *Proceedings of the 41st Conference of the German Acoustical Society*, Nuremberg, Germany, 2015, pp. 1–4.

[10] J. M. Scobbie and J. Stuart-Smith, "Socially stratified sampling in laboratory-based phonological experimentation," in *The Oxford Handbook of Laboratory Phonology*. Oxford: Oxford University Press, 2015, pp. 607–621.

[11] R. Godement-Berline, "Using a replication task to study prosodic highlighting," Boston, 2016, these proceedings.

[12] K. J. Kohler, B. Peters, and M. Scheffers, "The Kiel corpus of spontaneous speech IV, German: Video task scenario (Kiel-DVD1)," Kiel: IPDS, Christian-Albrechts-University, 2006.

[13] A. Anderson, M. Bader, E. Bard, E. Boyle, G. M. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. S. Thompson, and R. Weinert, "The HCRC map task corpus," *Language and Speech*, vol. 34, pp. 351–366, 1991.

[14] W. da Silva and P. A. Barbosa, "Caracterização semiautomática da tipologia rítmica do português brasileiro," Anais do Colóquio Brasileiro de Prosódia da Fala. ID [2432011], 2011.

[15] L. Öhman, A. Eriksson, and P. A. Granhag, "Mobile phone quality vs direct phone quality: How the presentation format affects earwitness identification accuracy," *The European Journal of Psychology Applied to Legal Context*, vol. 2, no. 2, pp. 161–182, 2010.

[16] M. Vainio *et al.*, "New method for delexicalization and its application to prosodic tagging for text-to-speech synthesis," in *Proc. of Interspeech 2009 - Speech and Intelligence*, Brighton, UK, 2009, pp. 1703–1706.

[17] T. Raso and H. Mello, "The C-ORAL-BRASIL I: Reference corpus for informal Brazilian Portuguese," in *Lecture Notes in Computer Science*. Heidelberg: Springer, 2012, pp. 362–367.