INTONATION AS A CUE TO EMOTIONAL SPEECH PERCEPTION: AN EXPERIMENT WITH NORMAL AND DELEXICALISED SPEECH

Daniel Oliveira Peres

DLCV, University of Sao Paulo, Brazil
danielperes@usp.br

ABSTRACT

This study aims to investigate the role of intonation in the perception of emotion in spoken Brazilian Portuguese (BP). Recordings of 32 BP speakers were downloaded from the internet and presented to 36 listeners divided equally into two nationality groups (Brazilian vs. English). The participants performed two perception experiments, one based on normal speech and the other on delexicalised (low-pass-filtered) speech. Aspects of the speakers' productions were subjected to automatic analysis and compared to the listeners' judgments. The results show that: (i) Brazilians perform well in the normal speech condition, while the English participants' performance is moderately accurate; (ii) for both groups, performance was lower in the delexicalised speech condition; (iii) simple and multiple linear regressions show that median F0, F0 dispersion, the duration of intonation units, and related parameters can predict the perception of emotional speech by native and non-native speakers of BP.

Keywords: speech perception, emotional speech, intonation.

1. INTRODUCTION

The manifestation of emotions in humans is an object of study across a wide variety of disciplines. Studies in areas such as psychology, neurology, social interaction studies and linguistics have attempted to explain how emotions are expressed and perceived by speakers and listeners. The study of emotion can be divided into three major areas: evolutionary, social and cognitive (internal processes). In the present study, evolutionary (universal / biological) [3] and social (cultural / linguistic) [9] approaches are taken so as to investigate the role of linguistic knowledge on the perception of emotions. Despite general agreement about the role of biological and cultural features in the expression of emotion, the contribution of each set of factors is not yet clearly understood [11]. Perception experiments and acoustic analysis of relevant speech data have been used in order to shed light on this issue [2, 10, 7].

2. EMOTIONAL SPEECH

From a biological viewpoint, emotions act as an interface between the body and external stimuli, and serve as a mechanism that helps organisms to deal with problems that affect them. In the context of human language, the expression of emotion is more complex [6]. The analysis of emotional speech demands a high level of accuracy, because language can convey emotional meaning that is independent of simpler physiological reactions.

In spite of the important results of previous work on emotional speech, some problems in this area are still in need of attention: namely, the type of stimuli (acted vs. spontaneous speech), the design of the experiments in which language samples are administered, and reliable ways of measuring and comparing acoustic data and perception responses.

Previous studies of emotional speech have tended to use acted speech or spoken material gathered using other types of elicitation [2, 10]. The main problem with these stimuli can be the biased results they produce both in production and perception analysis, which can be the consequence of stereotypical patterns of intonation and voice quality performed by the actors [8].

In a language-as-context environment, experiments on emotional speech need to have an orthogonal quality such that the researcher can manipulate the features of the language samples, and try to establish how each feature relates to the participants' judgments.

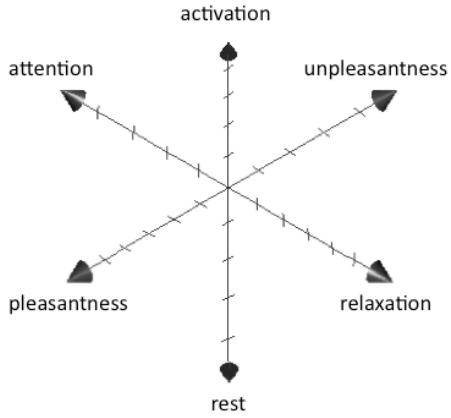
For this reason, the present study seeks to reduce the effects of these factors by using spontaneous speech and evaluating real situations in which emotional speech occurs. Moreover, the intonation of the test utterances will be isolated (delexicalised speech) in order to provide native and non-native speakers with equal conditions under which to evaluate the emotional speech samples.

2.1. Dimensional approach

A dimensional model has been proposed for the analysis of emotions (see one of the first proposals in [15]). This model assumes that emotional states can be explained by their position in a tridimensional space defined by *pleasantness* -

unpleasantness, rest - activation, and relaxation - attention.

Figure 1: Representation of the three-dimensional model presented in [15].



Other authors have suggested a two-dimensional model of emotion (*pleasantness - unpleasantness* and *attention - rejection*), as in [12].

Based on [14], the present study adopted a tridimensional approach using the two commonest dimensions found in emotional studies, namely *valence* (unpleasant - pleasant) and *activation* (non-agitated - agitated). A *dominance* dimension (submissive - non-submissive) has also been included, notwithstanding its relatively infrequent use.

3. DATA AND METHODOLOGY

3.1. Automatic analysis

Speech analysis was done automatically via *ExProsodia*, a VBA routine that analyses intonation and selects units of speech that contain substantial prosodic information, i.e. intonation units (IU; for more details see [4]). The current analysis is based on a proposal in [16], which claims that intonation can be divided into *mechanical-physiological* and *expressive* components. The first is related to individual physiological features, while the second pertains to intentional F0 variation produced by the speaker.

The *ExProsodia* analysis is based on F0 and intensity data extracted using the *Speech Filing System* [5] at intervals of 5 milliseconds. Each IU is selected taking into account the measurements of F0 (Hz) and intensity (RMS). In addition, the researcher can also set the duration (ms) values. Thus, the IU is the combination of these three parameters. The thresholds set for this analysis are:

- Lower threshold of F0: 50 Hz
- Upper threshold F0: 350 Hz
- Lower threshold of duration: 20ms
- Upper threshold of duration: 300ms
- Threshold of intensity: 2000RMS

The automatic analysis specifies the acoustic parameters to be used in the analysis of intonation.

3.1.1. Acoustic parameters of intonation

The acoustic parameters of intonation obtained by automatic analysis were:

F0 parameters:

- median F0 of the sentences (MT)
- standard deviation of median F0 (sdMT)
- skewness of median F0 (sMT)
- coefficient of variation of MT (cvMT)
- lower value (Hz) of intonation unit (lvIU)

Duration parameters:

- duration (ms) of intonation unit (IU)
- standard deviation of intonation unit (sdIU)
- interval duration of intonation unit (idIU)
- standard deviation of interval of intonation unit (sdiIU)

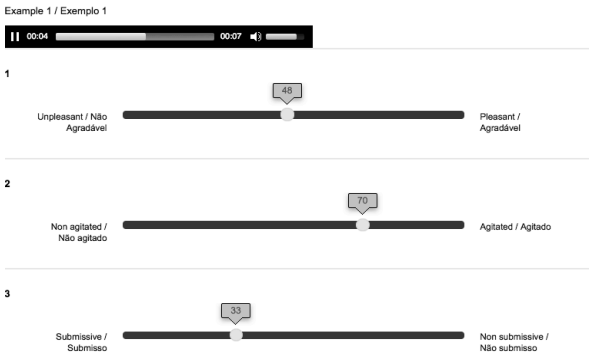
3.2. Experiments of perception

The data for this study were collected from the website *www.youtube.com*. Thirty-two excerpts of spontaneous emotional speech in BP were chosen and converted into .mp3 audio files (320 kbps). The recordings were delexicalised using the *Praat* script PURR (*Prosody Unveiling through Restricted Representation*) [1, 13].

Two experiments were presented to 18 Brazilians (from São Paulo) and 18 non-Lusophone English participants. The first employed samples of normal (non-manipulated) speech, and the other used delexicalised speech. For each stimulus, the participants gave a score on each of the dimensions *valence* (unpleasant – pleasant), *activation* (non-agitated – agitated), and *dominance* (submissive – non-submissive).

Using the computer mouse, the participants dragged virtual buttons along horizontal tracks displayed in an on-screen response form. While the participants slid the button, they could monitor the value of their scores from 0 to 100. This allowed them a wide range of variation in their judgments.

Figure 2: Screenshot of on-screen form for stimulus evaluation.



3.2. Data analysis

For the data analysis, two kinds of linear regression – simple and multiple – were performed. The first involved each dimension (*valence*, *activation* and *dominance*), and each acoustic parameter of intonation. The second took account of the three dimensions and pairs of acoustic parameters. In total, 204 linear regressions were carried out, of which 108 were simple and 96 multiple**.**

4. RESULTS AND DISCUSSION

4.1. Normal speech

4.1.1. Simple linear regression

For normal speech, significant results were obtained with respect to certain of the acoustic parameters and the participants' judgments. Among the Brazilian listeners (BR), the parameters cvMT and IU accounted for more than 80% of the variation in the judgments of *activation*. The judgments of *dominance* could be explained by MT, cvMT and IU. In the case of the English participants (ENG), *activation* and *dominance* yielded significant results in relation to cvMT and IU. *Valence* returned no significant results.

Table 1: Simple linear regressions – normal speech.

Dimension	Parameters	BR	ENG
Activation	MT	—	—
	cvMT	$R^2 = 0.82$	$R^2 = 0.73$
	IU	$R^2 = 0.84$	$R^2 = 0.81$
Dominance	MT	$R^2 = 0.61$	—
	cvMT	—	$R^2 = 0.57$
	IU	—	$R^2 = 0.62$

4.1.2. Multiple linear regression

Multiple linear regressions yielded significant results for the judgments of the dimensions given by the BR and ENG groups and some combinations of acoustic parameters. The dimensions were the same as those found for simple linear regression, and the relevant acoustic parameters were MT, cvMT, IU, sdIU, idIU, and sdiIU.

Table 2: Multiple linear regressions – normal speech.

Dimension	Parameters	BR	ENG
Activation	MT + IU	$R^2 = 0.86$	$R^2 = 0.85$
	MT + sdIU	—	—
	MT + idIU	—	$R^2 = 0.69$
	MT + sdiIU	—	—
	cvMT + IU	$R^2 = 0.92$	$R^2 = 0.86$
	cvMT + sdIU	$R^2 = 0.81$	$R^2 = 0.77$
	cvMT + idIU	$R^2 = 0.85$	$R^2 = 0.87$
	cvMT + sdiIU	$R^2 = 0.82$	$R^2 = 0.78$
Dominance	MT + IU	$R^2 = 0.66$	$R^2 = 0.71$
	MT + sdIU	$R^2 = 0.61$	—
	MT + idIU	$R^2 = 0.73$	$R^2 = 0.75$
	MT + sdiIU	$R^2 = 0.66$	—
	cvMT + IU	—	$R^2 = 0.68$
	cvMT + sdIU	—	$R^2 = 0.61$
	cvMT + idIU		$R^2 = 0.79$

4.2 Delexicalised speech

4.2.1. Simple linear regression

As per what was found for normal speech, the relevant dimensions in delexicalised speech were *activation* and *dominance*. *Valence* did not result in any significant effects. There were, however, significant results for the judgments given by the BR group and a number of acoustic parameters, viz. cvMT, IU, and idIU. There were no significant results related to the ENG group's judgments on the dimensions (R^2 values were below 0.50).

Table 3: Simple linear regressions – delexicalised speech.

Dimension	Parameters	BR	ENG
Activation	idIU	$R^2 = 0.66$	—
Dominance	cvMT	$R^2 = 0.54$	—
	IU	$R^2 = 0.53$	—

4.2.1. Multiple linear regression

The situation for multiple linear regressions with delexicalised speech was similar to that found for simple linear regression. Only BR participants returned significant results. The relevant acoustic parameters were MT, cvMT, IU, and idIU.

Dimension	Parameters	BR	ENG
Activation	MT + idIU	$R^2 = 0.66$	—
	cvMT + idIU	$R^2 = 0.71$	—
Dominance	MT + IU	$R^2 = 0.61$	—
	MT + idIU	$R^2 = 0.59$	—

Taking into account the results shown previously, in normal speech high levels of activation and dominance (BR and ENG) were related to low values of IU (shorter units in ms) and high values of

MT and cvMT (higher F0 values and variability). This combination is more frequent for the emotions *happiness* and *anger*. Low levels of *activation* and *dominance* (BR and ENG) are related to high values of IU (larger units) and low values of MT and cvMT (smaller F0 values and variability). This combination is more common in *fear* and *sadness*. These results consistently improved when the acoustic parameters of intonation were combined. As a result, clusters can be seen in multiple linear regressions that had high values of R^2 (Fig. 3)[i].

Figure 3: Scatterplot of degrees of activation, MT and IU (BR/normal speech); $R^2 = 0.86$.

Native BP speakers performed fairly well when their delexicalised speech results are compared to their normal speech results. The results for non-native speakers were random. The judgments of *valence* played no significant role in the perception of emotional speech. The causes of this outcome need to be investigated. The clustering of the dimensions was significantly affected by the absence of lexicon in both categories of listeners (native and non-native).

Figure 4: Scatterplot of the judgments of dimensions (Brazilian/normal speech).

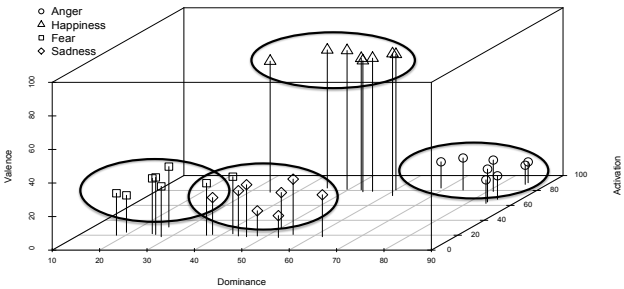


Figure 5: Scatterplot of the judgments of dimensions (BR/delexicalised speech).

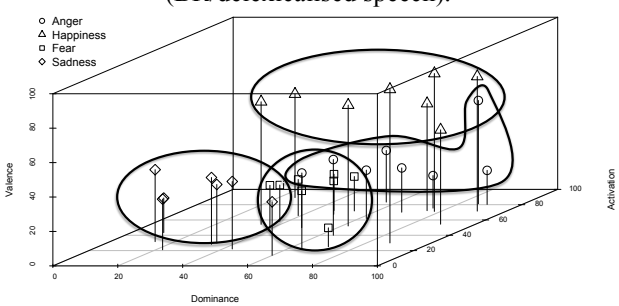


Figure 6: Scatterplot of the judgments of dimensions (ENG/normal speech).

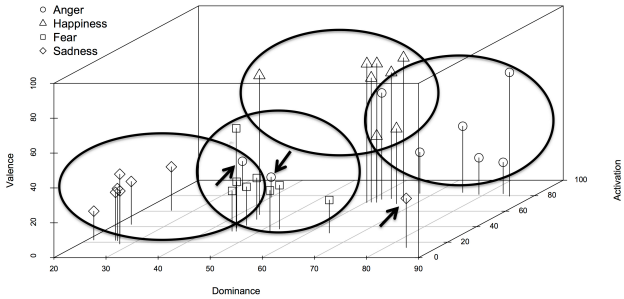
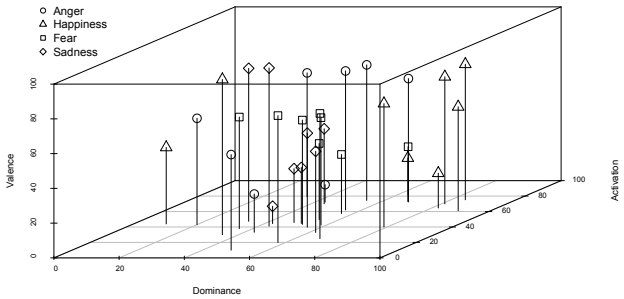


Figure 7: Scatterplot of the judgments of dimensions (ENG/delexicalised speech).



In the BR group's responses, some clusters were found, albeit with some overlapping data and outliers; in ENG, the data were almost completely overlapping.

5. FINAL REMARKS

Based on the present results, some features can be highlighted. In general, intonation serves as a perceptual cue for emotional speech. The performance of native and non-native BP speakers was above chance when they heard normal speech. The results of simple and multiple linear regressions, as well as the interaction between the dimensions supported the initial hypotheses in this regard.

However, some interesting issues remain unanswered: if the language presumably did not play any role in the perceptions of non-native speakers (non-Lusophones), why was their performance random when they judged delexicalised speech? Normal and delexicalised speech would not have been expected to have had such different results.

Despite of the limitations of such 'unfamiliar' stimuli, native BP speakers were still able to perform the experiment fairly well. It seems that the lexicon is not the only factor; other acoustic parameters probably also influenced the participants' performance. Further measurements of voice quality can improve our understanding of the perception of emotional speech. Moreover, other methods of delexicalisation that preserve voice quality must be considered.

6. ACKNOWLEDGEMENTS

This work was supported by CAPES (Brazilian Federal Agency for Support and Evaluation of Graduate Education). Process: 99999.007276/2014-01. I am grateful to Dominic Watt for his comments and suggestions. The remaining mistakes are my own.

7. REFERENCES

[1] Boersma, P., Weenink, D. 2013.Praat: doing phonetics by computer [Computer program]. Version 5.3.83, retrieved 11 May 2014 from http://www.praat.org.

[2] Costanzo, F. S., Merkel, N. N., Costanzo P. R. 1969. Voice Quality Profile and Perceived Emotion. *Journal of Counseling Psychology* 16/3, 267-270.

[3] Darwin, C. 1965 (1872). *The expression of the emotions in man and animals.* Chicago, IL: University of Chicago Press.

[4] Ferreira Netto, W. 2010. ExProsodia. *Revista da Propriedade Industrial – RPI* 2038, p. 167. Rio de Janeiro.

[5] Huckvale, M. A. (2008). Speech Filing System v.4.7/Windows SFSWin. Version 1.7.

[6] Keltner, D., Haidt, J., Shiota, M. N. 2006. Social functionalism and the evolution of emotions. In: Schaller, M., Simpson, J. A., Kenrick, D. T. (eds.) *Evolution and social psychology*. New York: Psychology Press, 115–142.

[7] Peres, D. O. 2014. Perception of emotional speech in Brazilian Portuguese: an intonational and multidimensional approach. *Nouveaux Cahiers de Linguistique Française* 31, 153-196.

[8] Roberts, L. 2011. Acoustics effects of authentic and acted distress on fundamental frequency and vowel quality. *Proceedings of The 17th International Congress of Phonetic Sciences (ICPhS XVII)*, 1694-1697).

[9] Russell, J. A. 1991. Culture and the Categorization of Emotions. *Psychological Bulletin* 110/3, 426-450.

[10] Scherer, K. R. 2000. A cross-cultural investigation of emotion inferences from voice and speech: Implications for speech technology. *International Conference on Spoken Language Processing. Proceedings of ICSLP 2000*, Beijing, China, 137-162.

[11] Scherer, K. R., Banse, R., Wallbott, H. G. 2001. Inferences from vocal Expression Correlate Across Languages and Cultures. *Journal of Cross-Cultural Psychology* 32, 76-92.

[12] Schlosberg, H. 1952. Three dimensions of emotions. *The psychological review* 61/2, 81-88.

[13] Sonntag, G. P., Portele, T. 1998. PURR – a method for prosody evaluation and investigation. *Computer Speech and Language, Special Issue on Evaluation*, v. 12, n. 3, 437-451.

[14] Trnka, R. 2011. How many dimensions does emotional experience have? The theory of multidimensional emotional experience. In: Trnka R., Balcar K., Kuska, M. (eds.) *Re-Constructing Emotional Spaces: From Experience to Regulation*. Prague College of Psychosocial Studies Press, 33-40.

[15] Wundt, W. M. 1897. *Outlines of Psychology*. Trans. by Charles Hubbard Judd. Leipzig, W. Engelmann; New York: G.E. Stechert.

[16] Xu, Y., Wang, Q. E. 1997. Component of intonation: what are linguistic, what are mechanical/physiological? *International Conference on Voice Physiology and Biomechanics*. Evanston Illinois.

---

[i] The data analysed in this study were also presented in a pre-test. Two Brazilians and two non-Lusophone English speakers classified the stimuli into four basic emotional categories. The labels (*happiness*, *sadness*, *fear* and *anger*) used in the figures are based on these results.