Multimodal perception and production of attitudinal meaning in Brazilian Portuguese

João Antônio de Moraes[1], Albert Rilliard[2], Bruno Alberto de Oliveira Mota[1], Takaaki Shochi[3]

[1] Laboratório de Fonética Acústica, FL/UFRJ/CNPq, Brazil
[2] LIMSI-CNRS, Orsay, France
[3] Kumamoto University, Japan

jamoraes2@br.inter.net, albert.rilliard@limsi.fr, baomgama@gmail.com, shochi38@gmail.com

Abstract

This paper presents a perceptual and acoustic analysis of a set of 12 different prosodic attitudes of Brazilian Portuguese, separated between 6 social, 5 propositional plus a neutral expressions. Audio-visual performances of these attitudes by two native speakers are described as well as their recognition by Brazilian listeners. The results show better performances for the propositional attitudes, particularly in the audio modality, while visual information brings the main indices for social expressions.

Index Terms: Prosodic attitudes, audio-visual prosody, Brazilian Portuguese.

1. Introduction

The concept of prosodic attitude generally refers to the expression of social affects, voluntarily controlled by the speaker. As any social expression, their acoustic manifestations are linked to the culture and the language of the speaker, and differ on that point from basic emotional expressions, which may be seen as more spontaneous and universal expressions [7, 11, 8]. Inside this set of attitudinal expressions, two cognitively different categories of attitudes can be distinguished (cf. [4, 2]): on the one hand propositional attitudes, whose expressions participate in the propositional content of the sentence presented to the interlocutor (e.g. with irony, incredulity, obviousness…); while on the other hand social attitudes refer to the social interpersonal relationship established by a speaker addressing his interlocutor, owing to these attitudes (e.g. he speaks with politeness or arrogance).

Prosody has been proved to be produced and perceived through multimodal cues [10], as both auditory and facial indices simultaneously convey it. The importance of this multimodality is certainly more crucial in the case of the expressive function of prosody. For example, [9] have shown that visual information is important to disambiguate culturally specific expressions of Japanese politeness for occidental listeners. We hypothesize here a different implication of both audio and visual modalities in the two types of attitudes, propositional or social: in the former expressions, directly linked to the linguistic meaning of the utterance, audio cues may have a more important role than in the case of socially motivated attitudes, in which the visual cues played a more crucial role during the face-to-face interaction.

In this paper, we will describe a set of attitudes in Brazilian Portuguese. A perception test with native listeners is described; it allows rating the perceptual differences between attitudes. It will be used as an indicator of their cognitive and pragmatic differences. Some acoustic correlates that characterize these attitudes are shortly described and give relevant information on the relative importance of both visual and audio modalities.

2. Corpus

Starting from the work of [5], 12 different attitudes based on an assertive mode, and including 6 social attitudes, 5 propositional ones plus a neutral assertion, were performed by two native Brazilian Portuguese speakers (1 female and 1 male), namely:

- For social attitudes: *arrogance* (ARR), *authority* (AUT), *contempt* (CON), *irritation* (IRR), *politeness* (POL) and *seduction* (SED);
- For propositional attitudes: *doubt* (DOU), *irony* (IRO), *incredulity* (INC), *obviousness* (OBV) and *surprise* (SUR);
- Neutral assertion (NEU).

Each attitudinal label was completed by a longer description, in order to define a precise and semantically unambiguous concept. Each attitude was performed on the same semantically neutral 6 syllable long Portuguese sentence "*Roberta dançava*." ("*Roberta was dancing*."). Speakers were standing in a sound-proof room, in front of a video camera (JVC, model GY-DV300) and a high quality microphone. Audio was digitalized at 48 kHz (down-sampled at 22.5kHz for the stimuli), and video was encoded using the *cinepack* codec with 784*576 pixels resolution.

3. Perception test

3.1. Paradigm

A recognition test was set up in order to validate the speakers' performances. Attitudes, grouped in two categories (social or propositional attitudes plus the neutral sentence added to both groups), were presented to listeners in the three possible modalities (audio-only, visual-only and audio-visual). Subjects had to recognize the presented attitude during a forced-choice paradigm, amongst a list of 7 or 6 possible answers, which include all the attitude of the given category, plus the neutral expression. The presentation order of the attitudinal categories and of the modalities was balanced across subjects: half of them were presented with modalities in the order audio-only, visual-only, audio-visual, while the other half were presented with visual-only, audio-only, audio-visual. In each group, half the subjects were presented with social attitudes first, and the other half with propositional attitudes first. Each stimulus was played twice on each run, in order to ease the listeners' task. Subjects had to give their answers by selecting on a slider the relative intensity of the perceived attitude (one slider per possible attitude was provided). The scale ranged from "*barely marked attitude*" to "*very marked attitude*". The attitudes performed by both speakers were randomized in each group of attitudinal category and modality. Each subject has to rate 52 different stimuli.

3.2. Subjects

30 listeners (23 women, 7 men, with a mean age of 31 years), all native speakers of Brazilian Portuguese, participated in the experiment. None have reported any perception problem.

3.3. Results Analysis

3.3.1. Analysis of variance

Two analyses of variance were run on the perception test's results in order to evaluate the influence of the different factors on the subjects' perception, separately for propositional and social attitudes. The GLM repeated-measure procedure of SPSS was used. There was one between-subject factor: the order of presentation of the three modalities (OM, fixed), and three within-subject factors: the modality (M) of presentation (3 levels: audio, visual, audiovisual), the speaker (S, 2 levels: female, male) and the attitudes (At, 6 or 7 levels according to the attitudes' type). The intensity score given to attitudes was used as the dependent variable. Results are presented in Table 1.

Table 1. *Repeated measures ANOVAs on intensity scores. Insignificant interactions of within-subjects and between-subjects factors are omitted.*

Propositional attitudes					
	df	***df error***	***F***	***p***	***Partial η^2***
Between-Subjects Effects					
OM	1	28	0,93	0,344	0,03
Within-Subjects Effects					
Modality	2	56	53,65	**0,000**	0,66
Speaker	1	28	18,19	**0,000**	0,39
Attitude	5	140	12,70	**0,000**	0,31
M * S	2	56	4,33	0,018	0,13
M * At	10	280	4,32	**0,000**	0,13
S * At	5	140	16,97	**0,000**	0,38
M * S * At	10	280	8,01	**0,000**	0,22
Social attitudes					
Between-Subjects Effects					
OM	1	28	0,46	0,503	0,02
Within-Subjects Effects					
Modality	2	56	103,15	**0,000**	0,79
Speaker	1	28	0,43	0,516	0,02
Attitude	6	168	14,94	**0,000**	0,35
M * S	2	56	0,02	0,983	0,00
M * At	12	336	6,59	**0,000**	0,19
S * At	6	168	38,78	**0,000**	0,58
M * S * At	12	336	3,93	**0,000**	0,12

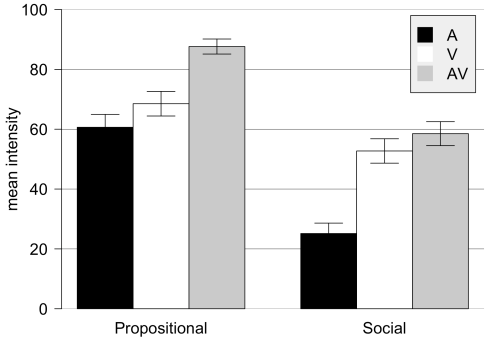

Figure 1: *Mean intensity rating in each modality, for both types of attitudes.*

These analyses show that the main sources of variance for both sets of attitudes are the modality of presentation: for social attitudes, audio (A) presentation received significantly lower scores than both visual (V) and audio-visual (AV) ones (post-hoc Tukey's HSD test with an α level of 1%); for propositional attitudes the intensity of the perceived expressions significantly increases from A to V and especially for AV (post-hoc Tukey's HSD test with an α level of 1%) – cf. fig. 1. However, even the lower score obtained for the audio only modality for the social attitude is significantly higher than chance level (here 14.28% – one-sample T-test: t_{419}=6.135, p<0.001), and is clearly above chance for propositional ones. Therefore, one can assume a comparable perceptual role of both audio (mean int.=60.8%) and video (mean int.=68.5%) modalities for the propositional attitudes, with respect to the level reached by audio-visual presentation (mean int.=87.6%).

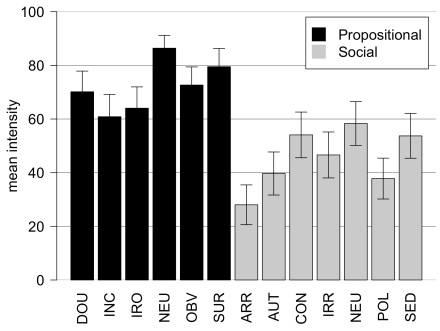

Figure 2: *Mean intensity rating for each attitude.*

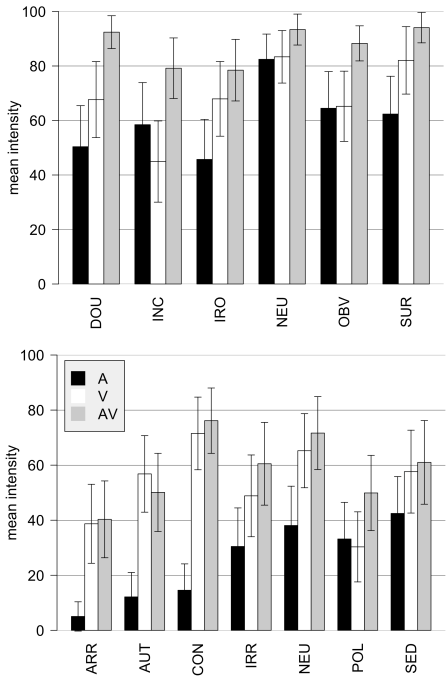

Figure 3: *Mean intensity for attitudes (top: propositional; bottom: social), for both speakers in each of the three modalities.*

The attitudes also have a main effect on scores (fig. 2). For propositional attitudes, the speaker's performance also has an important effect (the male speaker receives higher mean

intensity score), but not for the social attitudes. Significant interactions were also found between modalities & attitudes, between speakers & attitudes and between modalities, speakers & attitudes. The order of presentation of modalities has no effect on the results.

Fig. 3 presents the interaction between modality and attitudes for propositional and for social attitudes. It gives a good indication on the importance of multimodal presentation regarding speech expressivity: even if audio and visual modality may be preferred for some expressions, audio-visual presentation always received higher or equivalent scores than the best modality alone. For propositional attitudes, audio and visual modalities were found to be equivalent – except for irony. Audio-visual presentation received significantly higher scores for doubt and obviousness, was equivalent to visual for irony and surprise, to audio for incredulity and to both audio and visual for neutral (differences rated thanks to Tukey HSD post-hoc tests). Conversely, for social attitudes, audio played a less important role, receiving significantly lower scores than audio-visual and visual for all attitudes except seduction and politeness (and scores similar to visual for irritation), whereas no significant difference were observed between visual and audio-visual.

Finally, both types of attitudes received mean scores that differ significantly (2-tailed T-test for independent samples, performed on the intensity scores corrected by the chance level for each set: t_{2338}=14.525, p<0.001): propositional attitudes were better recognized than social attitudes and, as already mentioned, the recognition of social attitudes was particularly low when based on the auditory information only.

3.3.2. Confusion matrices

Confusion between attitudes is a main aspect of such perception tests. They are analyzed owing to a hierarchical clustering made on the confusion matrices obtained from the experiment (both speakers together): the number of answer received by each stimuli for each possible attitudes are used as input for the algorithm, using an Euclidean distance between vectors and the Ward grouping method (cf. figure 4). The boundary was set to half the maximum distance. Using this criterion, the following observations can be made.

For the propositional attitudes: in the audio-only condition, listeners show confusion between incredulity and irony, but distinguish all the others attitudes; in the visual-only condition, there is only confusion between doubt and incredulity, while with the audio-visual stimuli, listeners perfectly discriminate all attitudes – with weak confusion between incredulity and irony, showing the perceptive importance of auditory information.

Regarding social attitudes, that received weaker recognition scores, confusion is more important. For audio-only stimuli, three groups of stimuli emerge: seduction with politeness, authority with irritation, and contempt with arrogance & neutral. While the first two groups show important mutual confusion, the third one is mainly due to the low recognition scores of arrogance and contempt, recognized as neutral in this audio condition. Visual stimuli show a different pattern: arrogance and contempt are grouped, now with strong mutual confusion; authority, irritation and seduction are recognized, and politeness is mixed up with neutral.

The audio-visual presentation shows a clearer figure: subjects are here able to distinguish all attitudes, while still showing some confusion tendencies, but only between semantically similar expressions (arrogance & contempt, authority & irritation and seduction & politeness).

This confusion analysis allows us to consider that prosodic features are less important for distinguishing social attitudes – even if they are still important for some expressions such as seduction and politeness. A look at the prosodic parameters of these attitudes may help us to understand this behavior.

4. Acoustical analysis

Prosodic parameters of all acoustic stimuli were analyzed in order to extract values of fundamental frequency (F0), intensity and syllabic duration. Syllabic segmentation was hand done with the PRAAT [1] software, while extraction of F0 and intensity parameters was done thanks to the STRAIGHT [3] program, using MATLAB® scripts. F0 was expressed in semitones, intensity in dB and syllabic duration in seconds.

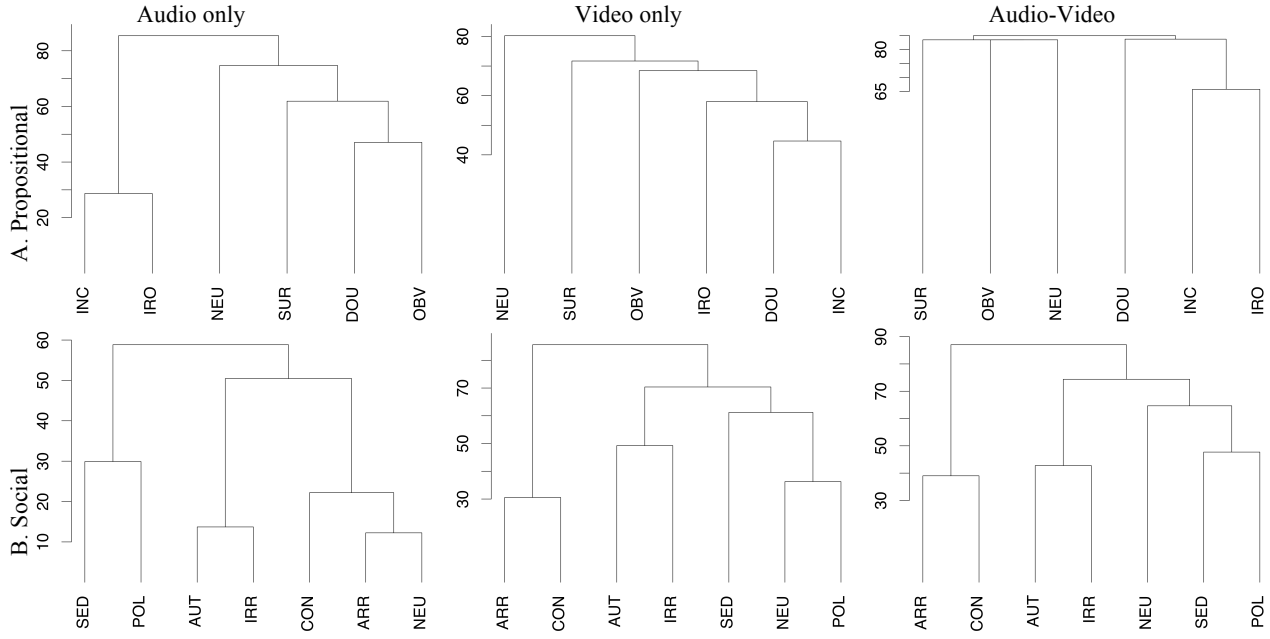

Figure 4: Hierarchical clustering for each group of attitudes, in the three modalities. Results for both speakers are averaged.

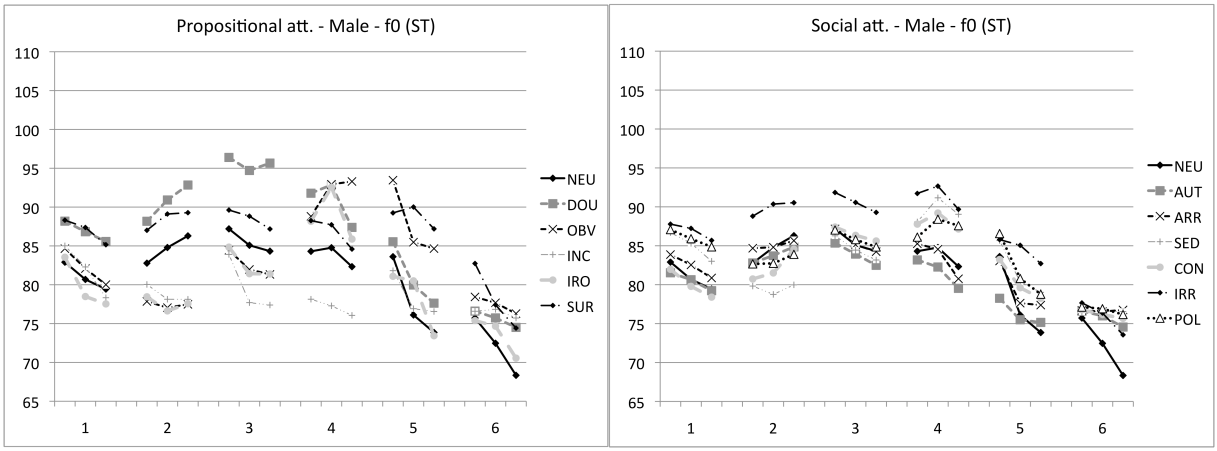

Figure 5: *Plot of the F0 (expressed in semitones and calculated on the vowels) for all modalities, for the male speaker.*

A comparison of the F0 contours of the utterances corresponding to social attitudes with those of propositional attitudes (see examples provided for the male speaker at figure 5), shows clear differences between the melodic configurations observed among the latter, which cannot be observed among the former. Such narrower variations of F0 for social attitudes may explain the less important role played by audio presentation for social attitudes, compared to propositional ones.

Amongst the propositional attitudes, some salient traits were found, already mentioned in preceding work. For example, an important lengthening of the penultimate syllable for irony and incredulity, observed on both speakers' performances, was already proposed by [5]. A complete analysis, including facial settings and voice quality parameters deserves to be done on this data, in order to extract the most pertinent features, but cannot fit in this paper.

5. Discussion & conclusion

The comparison of both speakers' performances underlines the importance of individual variation of expressivity. If speakers show overall comparable recognition scores, the different strategies they developed result in important perception differences. For example, the seduction performed by the female speaker received high scores, but not for the male speaker – whereas the inverse arises for irritation. The results for seduction may also be related to a gender difference. For doubt, they adopt different dominant modality: the male speaker uses mainly audio cues, while the female speaker relies primarily on visual ones – but both received comparable mean scores in the audio-visual presentation.

This last observation is a good example of the bimodality of attitudinal prosody's expressivity: both kinds of cues are used by speakers, and combined in different fashions, to achieve an efficient encoding of communication. These individual choices are also linked to and constrained by both linguistic and cultural factors [6]. Cross-cultural perception tests [9] should be performed on this data in order to measure the ability of foreign subjects to understand both kinds of cues.

Interestingly, and in accordance with our hypotheses, the propositional and social attitudes show different perceptual behaviours. As the former are directly linked to the linguistic content of the utterance, they were expected to be more strongly related to acoustic variations – and the results clearly support this expectation: subjects rely clearly on both audio and visual cues for propositional expressions, while they mainly use visual cues for social attitudes. However, audio cues have an important role in perception, if not a primary one: they are used to disambiguate some visual expressions, as well as to construct the detailed meaning of each expressivity.

An important difference has also to be mentioned here: the neutral sentence was presented in both group of attitudes, but its recognition scores differ significantly in both conditions (independent sample T-test performed on means corrected from chance level: $t_{289.426}$=7.050, p<0.001). This may be due to a more difficult task for listeners, who had to judge the more subtle differences in social expressions, than the ones present in propositional attitudes

This study only investigated assertive attitudes. The complete corpus also contains a similar set of expressions based on the interrogative mode. Perception tests are currently run in order to evaluate them, and these results will interestingly be compared to the one presented here.

6. References

[1] Boersma, P. & Weenink, D. "Praat: doing phonetics by computer (Version 5.0.23) [Computer program]". Retrieved May 10, 2008, from http://www.praat.org/

[2] Fónagy, I., Bérard, E. and Fónagy, J. Clichés mélodiques. Folia Linguistica, 17:153-185, 1984.

[3] Kawahara, H. "TANDEM-STRAIGHT, a research tool for L2 study enabling flexible manipulations of prosodic information". In Proceedings of Speech Prosody 2008, 619-628, 2008.

[4] Martins-Baltar, M. "De l'énoncé à l'énonciation: une approche des fonctions intonatives". Paris: Didier, 1977.

[5] de Moraes, J.A. "The Pitch Accents in Brazilian Portuguese: analysis by synthesis". In Proceedings of Speech Prosody 2008, eds. P.A. Barbosa, S. Madureira, & C. Reis, Campinas, Brazil: Editora RG/CNPq, 389-397, 2008.

[6] Pavlenko, A. "Emotions and multilingualism". Cambridge (U.K.): Cambridge University Press, 2005.

[7] Scherer, K. R. and Wallbott, H. G. Evidence for universality and cultural variation of differential emotion response patterning. Journal of Personality and Social Psychology, 66(2):310–328, 1994.

[8] Scherer, K. R. and Brosch, T. Culture-specific appraisal biases contribute to emotion dispositions. European Journal of Personality, 23:265-288, 2009.

[9] Shochi, T., Rilliard, A., Aubergé, V. and Erickson, D. "Intercultural Perception of English, French and Japanese Social Affective Prosody". In The role of prosody in Affective Speech, ed. S. Hancil, pp.31-59, Linguistic Insights 97, Peter Lang AG, Bern, 2009.

[10] Swerts, M. and Krahmer, E. "Audiovisual prosody and feeling of knowing. Journal of Memory and Language, 53(1):81–94, 2005.

[11] Zinck, A. and Newen A. "Classifying emotion: a developmental account". Synthese, 161:1–25, 2008.