ILLOCUTIONS AND ATTITUDES
Prosodic cross-linguistic perception of social affects in Mandarin Chinese by native, French and Vietnamese listeners

Yan LU¹, Véronique AUBERGÉ¹, Albert RILLIARD², Wentao GU³
¹GIPSA-Lab; Grenoble France; ²LIMSI; Orsay France; ³Institute of Linguistic Science and Technology, Nanjing Normal University, Nanjing, China
yan.lu@gipsa-lab.grenoble-inp.fr, veronique.auberge@gipsa-lab.grenoble-inp.fr, albert.rilliard@limsi.fr, wtgw@njnu.edu.cn

Abstract
Social affects play an important role in the face-to-face interaction and are implied in the realization of speech acts. The prosody is a main vector of social affects and its cross-language variability is a challenge for language description as well as for foreign language teaching. The present work aims at examining the perception of Chinese social affects in an intra-cultural perceptual experiment and the influences of tones on the perception of these social affects in another inter-cultural perceptual experiment. A speech corpus was designed with the variation of length, tone location and syntactic structures of utterances, and has been incorporated with 19 social affects. For each experiment, a specific sub-corpus was selected. The tests results show that the social affects were globally recognized over chance level by native and non-native listeners; “declaration” is the attitude which attracted the most confusions; all subject groups separated the 19 Chinese social affects in two subsets: a subset of “assertive” attitudes (represented by “declaration”) and a subset of “interrogative” attitudes (represented by “question” and “doubt”); more similarities were found between French and Vietnamese listeners in inter-cultural perception experiment.

Keywords: social affects; prosodic perception; tones; Mandarin Chinese; French; Vietnamese.

1. Introduction
The affects expressed in interactive speech imply two different levels of the speaker’s cognitive processing (Aubergé, 2002): the involuntarily controlled expressions of affects (so-called “emotions”), and the intentionally controlled expressions expressed through audio-visual prosody (so-called “social affects” or “attitudes”). Prosodic attitudes, -functions of the speaker’s opinion, beliefs or knowledge (Wichmann, 2000), are an integral part of the language interaction building and are performed through the audio-visual prosody. They need to be learned in infancy and would benefit to be explicitly taught in foreign language teaching. In the present work, some values of social affects, which potentially reveal the speaker’s opinion or some social and situational cues, e.g. the speaker-hearer relationship, were selected for two perceptual experiments in order to investigate the prosodic perception of social affects in Mandarin Chinese by native and non-native listeners.

Different hypotheses have been set up about the typologies of attitudinal expressions (Martins-Baltar,1977; Wichmann, 2000; de Moraes et al., 2010; Gu et al., 2011), and we propose to classify social affects into three categories: first, the attitude, intention or opinion of the speaker about what he says (even if he does not express any attitude by performing a simple declaration or question, it is then considered as the attitude to give no information on his own attitude, - Aubergé, 2002); second, some expressions characterising the social relation implied in the interaction, e.g. politeness, authority; third, the expressions depending on the socio-cultural context of interaction, typically for intimacy, infant-directed speech and seduction.

Mandarin Chinese (also referred to as Putonghua or Standard Chinese) has four tones which were defined customarily according to the characteristics of their fundamental frequency curve as: high level (tone 1), rising (tone 2), dipping (tone 3) and falling (tone 4). Belonging to different families of languages, Mandarin Chinese, Vietnamese and French have their own specific linguistic structures. Both Mandarin and Vietnamese are tonal languages; French is not tonal (and not stressed). Compared to French, and from the prosodic and cultural point of view, Vietnamese could be considered as closer to Chinese. Therefore, it is supposed that Chinese lexical tones could influence to some extent the prosodic perception of Chinese social affects by subjects of different language backgrounds.

Hence, an intra-cultural perceptual experiment was designed to examine how prosodic social affects in Chinese can be perceived by native Chinese, and another inter-cultural perceptual experiment was required to investigate how these social affects can be perceived by French and Vietnamese listeners and if the effect of tones can be shown on the perception of social affects outside of any morphosyntactic and semantic influences.

2. Corpus
2.1 Speech corpus design
In order to compare the parameters implied in the variability of prosody, a dedicated and controlled corpus was built to convey different social affects.

The corpus was designed with consideration of utterances’ length (in syllables), of tones location and of syntactic structure, which were systematically varied in order to analyze further the variation of one parameter in the same context for the others. As the social affects could not be produced without reference to context, a dedicated context of interaction was described for each social affect, in order to help the speaker to express them...
as naturally as possible. All utterances were constructed to bear a literally neutral meaning (i.e. not conveying any meaning which implies a specific social affect nor emotion) but in the same time could be expressed with all the social affects studied. The complete corpus contains 152 utterances performed with 19 attitudes, i.e. 2888 stimuli.

2.2 Selected social affects

Some social affects (attitudes) in different language have been studied in by Fujisaki & Hirose (1993), Aubergé (1998), Mac et al. (2010), Gu et al. (2011) and Lu et al. (2012). In this work, 19 social affects, which are commonly encountered in daily conversation, were selected. Table 1 shows the 19 social affects and their abbreviations, grouped in three categories.

<table>
<thead>
<tr>
<th>Social affects and abbreviation</th>
<th>Attitudes</th>
<th>Social parameters</th>
<th>Social context</th>
</tr>
</thead>
</table>

Table 1: Classification of social affects and their abbreviation

2.3 Corpus recording

One native Mandarin female from Shaanxi province of China took part in the recording. She is teacher of French as a foreign language in a Chinese college, and speaks unmarked standard Mandarin Chinese. The recording was conducted in a sound proof room at GIPSA-Lab in Grenoble, France, both in video and audio modalities. To make the attitudinal expressions consistent, the sentences sharing the same attitude were recorded in one session after the speaker had understood and had got familiar with the situational context of the given affect. 19 social affects were conveyed one by one in the same way. Another native Chinese from the same area of China as the speaker was also present during the recording to supervise the performance of the speaker.

3. Native perceptual validation

3.1 Description of the experiment

To test the validity of the attitudinal speech corpus and to look into the perception and the recognition of attitudes, we designed this perceptual experiment with a sub-corpus of 21 utterances conveying the 19 social affects, i.e. 399 stimuli. The listening subjects were composed of 30 native Mandarin Chinese, from different areas of China: 15 males and 15 females with an average age of 25.2 years. They’re almost all postgraduate students or PhD students in Grenoble, France (except one male subject who works as computer programmer in an IT company in Grenoble), and none of them reported any listening and understanding disorder.

All 399 target stimuli were presented to the subjects through headphones in a quiet room and were introduced by a presentation of the experiment and a description of each social affect with examples of situations in which such social affects can happen. The listeners had the written instructions in their native language at their disposal during the experiment. They listened only one time each stimulus and had to choose the perceived attitude amongst the 19 proposed labels, written in Chinese. The presentation order of the stimuli was randomized for each subject.

3.2 Analysis and results

An analysis of variance (completely randomized three-factorial design) was carried out on the data. The three fixed factors were the subjects’ gender (G, 2 levels), the presented attitudes (A, 19 levels) and the sentences length (L, 4 levels). Each cell of this design contained at least 60 observations. The significance level was set at 0.01. Table 2 shows the results of the analysis of variance for each factor.

The factors “Attitude”, “Length” and the interaction between “Attitude” and “Length” have significant effect; “Attitude” has the highest observed strength of effect (η²). Factors “Gender” and “Length” are significant at the 1% level, but does only explain a small part of the variance observed.

<table>
<thead>
<tr>
<th></th>
<th>Sum Sq</th>
<th>Df</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
<th>η²</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>253.43</td>
<td>18</td>
<td>86.2218</td>
<td>0.0000</td>
<td>0.693</td>
</tr>
<tr>
<td>G</td>
<td>19.97</td>
<td>1</td>
<td>12.0648</td>
<td>0.0005</td>
<td>0.005</td>
</tr>
<tr>
<td>L</td>
<td>16.19</td>
<td>3</td>
<td>33.0582</td>
<td>0.0000</td>
<td>0.044</td>
</tr>
<tr>
<td>A*G</td>
<td>5.57</td>
<td>18</td>
<td>1.8960</td>
<td>0.0122</td>
<td>0.015</td>
</tr>
<tr>
<td>A*L</td>
<td>80.30</td>
<td>54</td>
<td>9.1061</td>
<td>0.0000</td>
<td>0.220</td>
</tr>
<tr>
<td>G*L</td>
<td>0.13</td>
<td>3</td>
<td>0.2556</td>
<td>0.8574</td>
<td>0.000</td>
</tr>
<tr>
<td>A<em>G</em>L</td>
<td>7.87</td>
<td>54</td>
<td>0.8929</td>
<td>0.6958</td>
<td>0.021</td>
</tr>
</tbody>
</table>

Table 2: ANOVA’s results – significant effects in bold

Through the mean recognition rate of 19 social affects and the mean recognition rate of social affects distinguished by stimulus’s length and gender presented in figure 1, it is observed that for native Chinese listeners, almost all of the social affects were recognized above chance level, except “confidence” and they can be classified in the decreasing order (cf. Figure 1, top). The identification of social affects varies with the stimuli’s length: according to the confusion matrix of attitudes by length, there is a clear separation between the 1-syllable stimuli and the longer ones. The 1-syllable stimuli received lower recognition scores while the 4-syllable stimuli received the highest (the 9-and 2-syllable stimuli are just under the 4-syllable ones). The graph of the
mean recognition rate for social affect by length (figure 1, bottom) shows that “infant-directed speech” and “irritation” don’t follow this trend. For “infant-directed speech”, the 1 and 2-syllable stimuli were better recognized than the 4 and 9-syllable ones (who were mixed up with “seduction”). For “irritation”, the 2-syllable stimuli were not well perceived, in comparison with other lengths, and were confused with “declaration” and “confidence”.

Figure 1: Recognition rate for the 19 social affects: rate per attitude (top), detailed per stimuli’s length (bottom)

4. Cross-cultural perception

4.1 Description of experiment
This perception test was aimed to study French and Vietnamese listeners’ perception of these Chinese social affects and to measure a potential interaction between attitudes and tones. A sub-corpus of 16 utterances was selected with a systematic variation of tones values and location. There are no morpho-syntactic nor semantic variations in the sub-corpus. 15 French (6 females, average age of 33 years) and 15 Vietnamese (8 females, average age of 27 years) took part in the experiment. All of them work or study in Grenoble, France. None of the 30 subjects reported any listening disorder. The test’s paradigm was the same as for the first experiment.

4.2 Analysis and results
An analysis of variance (completely randomized three-factorial design) was carried out on the data. The three fixed factors were the presented attitudes (A, 19 levels), the sequence of tones (T, 16 levels) and the native language of subjects (L, 2 levels). The significance level was set at 0.01. Table 3 shows the general results of the analysis of variance for each factor. The factors “Attitude”, “Tones sequence” and the interaction between “Attitude” and “Tones” show significant effects. Both “Attitude” and the “Attitude & Tones” interaction have the most important effect size (cf. the η² column of Table 2), and thus are the most influencing factor on listeners’ answers.

Table 3: Global ANOVA results – significant effects in bold

<table>
<thead>
<tr>
<th></th>
<th>Sum Sq.</th>
<th>Df</th>
<th>F value</th>
<th>P</th>
<th>η²</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>81.85</td>
<td>18</td>
<td>35.8</td>
<td>0.000</td>
<td>0.070</td>
</tr>
<tr>
<td>L</td>
<td>0.11</td>
<td>1</td>
<td>0.8</td>
<td>0.362</td>
<td>0.000</td>
</tr>
<tr>
<td>T</td>
<td>4.61</td>
<td>15</td>
<td>2.4</td>
<td>0.001</td>
<td>0.004</td>
</tr>
<tr>
<td>A*L</td>
<td>10.36</td>
<td>18</td>
<td>4.5</td>
<td>0.000</td>
<td>0.009</td>
</tr>
<tr>
<td>A*T</td>
<td>68.53</td>
<td>270</td>
<td>2.0</td>
<td>0.000</td>
<td>0.060</td>
</tr>
<tr>
<td>L*T</td>
<td>3.11</td>
<td>15</td>
<td>1.6</td>
<td>0.058</td>
<td>0.003</td>
</tr>
<tr>
<td>A<em>L</em>T</td>
<td>37.72</td>
<td>270</td>
<td>1.1</td>
<td>0.131</td>
<td>0.034</td>
</tr>
</tbody>
</table>

Two separated ANOVAs on French and on Vietnamese subjects were run (table 4). Results show that the effect of “Tones” is significant for French subjects while it is not significant for Vietnamese subjects (cf. mean results on fig. 2), although there is a significant interaction between “Attitude” & “Tones”.

Table 4: Separated ANOVAs by language – significant effects in bold

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>French</th>
<th>Vietnamese</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>F</td>
<td>p</td>
</tr>
<tr>
<td>Attitude</td>
<td>18</td>
<td>20.2</td>
<td>0.000</td>
</tr>
<tr>
<td>Tones</td>
<td>15</td>
<td>2.4</td>
<td>0.002</td>
</tr>
<tr>
<td>Attitude*Tones</td>
<td>270</td>
<td>1.4</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Figure 2: Mean recognition rate for each tone sequence, per language background

Figure 3: Mean recognition for the 19 social affects, per language background
Figure 3 shows the mean recognition of the 19 social affects by French and Vietnamese subjects. Almost all of the social affects were identified above chance, except “contempt”, “irony” and “confidence” for French and “irony” for Vietnamese subjects.

5. Discussion and conclusions

A comparison of results obtained in both experiments allows to analyse three aspects: the mean recognition rate of the 19 social affects; the attractiveness of individual social affects and the different clustering made by native and non-native subjects.

5.1 Native and non-native results

In the intra-cultural test, almost all of the social affects were recognized over chance, except “confidence”, and “Declaration” was the best recognized attitude. In the inter-cultural test, almost all of the social affects were identified above chance, except “contempt” “irony” and “confidence” for French and “irony” for Vietnamese. “Declaration” is the best-recognized attitude by French, while it is “disappointment” for Vietnamese. Native listeners received higher recognition scores than non-native listeners. For “seduction” and “authority”, French listeners show the highest recognition scores, and for “confidence”, Vietnamese listeners did the best (cf. figure 1 (top) and figure 3). Concerning the less identified attitudes in this audio modality, they have been supposed to rely strongly on the visual modality (Shochi, 2008). Hence, another multimodal perceptual experiment will be carried out to investigate how the 19 audio-visual prosodic attitudes will be perceived by native and non-native subjects.

Analysis of the cross-cultural experiment also showed that the tones have some influences on the perception of several social affects and that the tonal effect is more important for French subjects than for Vietnamese ones. As it was commonly accepted that there are cross-cultural similarity in the uses of F0 to signal affect, intention, or emotion (Ohala, 1994), in order to validate the findings, our future work will focus on the acoustic analysis of the social affects, with an emphasis on the F0 contour of tones which is the primary acoustic parameter for Mandarin tones (Allard et al., 2006).

5.2 Attractivity of Chinese social affects

The attractiveness of attitudes – the sum of all confusions attributed to a given attitude (cf. fig. 4) – shows some interesting results. For native listeners, the attitude attracting most answers is “declaration”, which is mainly used when judges cannot identify any attitude. This result is coherent to common behaviors of perceiving his language (de Moraes et al., 2010; Diaferia, 2002; Mac et al., 2010; Shochi et al., 2009). Moreover, recognizing a perceived stimulus amongst 19 attitudinal labels is a cognitively complex task. Thus, choosing “declaration” is a way to avoid false or uncertain answers without specifying any information about attitude. French and Vietnamese listeners show, to a lesser degree, the same preference for “declaration”, but with quite clear second choice: “question” for Vietnamese and “obviousness” for French judges. “Irony” was not well recognized by Vietnamese judges, nor did it attract any attitude.

5.3 Clustering of Chinese social affects

In order to measure the perceptive distances between each stimulus and to identify the higher perceptual categories for Chinese, French and Vietnamese subjects, as well as the perceptual differences between the three groups, a hierarchical clustering analysis was run on the dispersion matrix. Distances were expressed as the correlation (r) between rows (1-r is used as the distance). From these perceived distances, a hierarchical clustering algorithm was applied, which allowed the observation of the main clusters of attitudes for each language group (cf. figure 5). The three groups clustered the attitudes almost in the same way and all have separated the attitudes in two subsets: a subset of “assertive” attitudes (represented by “declaration”) and a subset of “interrogative” attitudes (represented by “question” and “doubt”). Meanwhile, in observing closely the clustering, we found that French and Vietnamese listeners have grouped the perceived social affects in the same eight clusters - that differ to some extend from the seven groups made by Chinese subjects. This result is contrary to our hypothesis in which there should be more similarities between Chinese and Vietnamese listeners in respect to cognitive processing of social affects. An evaluation of the classification of the concepts of Chinese and French social affects will be carried out in order to measure the cognitive distances between the attitudinal concepts and propose a cognitive clustering of social affects in daily life.
Figure 5: Hierarchical clustering of perceived social affects, based on R complete grouping criterion. The grouping done by Chinese subjects is shown on top, by French subjects in middle, by Vietnamese subjects on the bottom.

6. Acknowledgements

The corpus couldn’t have been recorded without the technical assistance of C. Savariaux and L. Granjon.

7. References


Can the tones influence the acoustic perception of the Vietnamese attitudes by French listeners? Some evidences for global vs. local processing of prosody

Dang-Khoa MAC1,2, Véronique AUBERGÉ2, Albert RILLIARD3, Eric CASTELLI1

1 International Research Institute MICA, HUST-CNRS/UMI 2954 Grenoble INP, Hanoi, Vietnam; 2 Laboratory of Informatics of Grenoble (LIG), CNRS, France; 3 LIMSI-CNRS, Orsay, France
dang-khoa.mac@mica.edu.vn, eric.castelli@mica.edu.vn, veronique.auberge@imag.fr, albert.rilliard@limsi.fr

Abstract

Attitudes or social affects are strongly implied in the interaction processing, and specifically into the socio-cultural aspects of language. The prosody has been shown as a main vector for expressing attitudes in different languages. In tonal language, the lexical access function is also implemented by the parameters of prosody. This paper presents a study of attitudinal expressions in Vietnamese, a tonal language, under the light of cross-cultural perception. Sixteen Vietnamese attitudes, performed on sentences including with tonal variation, were used in a perception experiment with French listeners. The result of French subjects on the utterances with tones and non-tone allow us to explore the influences of tones on the different Vietnamese attitudes in non-tonal language speakers.

Keywords: attitude; social affect; tone; global prosodic patterns; cross-cultural perception.

1. Introduction

The attitudes, and more generally the social affects, are an important part of the face-to-face interaction and are linked to the language through the socio-culture. These expressions are clearly social: they carry the intentions and points of view of the speaker (e.g. surprise, confirmation, etc.) and can give the social context on the interaction (e.g. intimacy, politeness). When the speaker does not express any attitude in his speech act (in the case of a declaration or a “simple” question), she/he expresses that she/he has no opinion on this utterance or that she/he does not want or cannot express any attitude (Aubergé, 2002).

Even if many such social affects are universal in their values or in their prosodic forms, some prosodic implementation and even some attitudinal values are specific to the culture and the language (Scherer et al., 2001; Shochi et al., 2007). Anyway, the attitudes are built inside each culture and language, and they are acquired by children inside their culture or learned by the learners of second language (Shochi et al., 2010). The understanding of this phenomenon may benefit from cross-cultural studies (Scherer et al., 2001; Shochi et al., 2010).

The attitudes or social affects are supposed to be involved into voluntary cognitive controls, whereas emotions are involuntary controls (Aubergé, 2002). The prosody has been shown as a main vector for expressing attitudes in different languages (Wichmann, 2000; Aubergé, 2002). The “classical” prosodic parameters (F0, intensity, timing), are strongly implied in the expression of attitudes (Fónagy, 1983; Wichmann, 2000; Aubergé, 2002). Campbell & Mokhtari (2003) proposed the voice quality as a 4th dimension of prosody; it has been also shown as a fundamental parameter for emotions (Banse & Scherer, 1996; Audibert, 2005) and is used in some attitudes (Shochi et al., 2007). Many different functions are implemented by prosody by using the same acoustic parameters (F0, intensity, timing and voice quality).

In tonal languages such as Vietnamese, a part of the lexical access function is implemented by F0. The Vietnamese language has 6 tones: level (1), falling (2), broken (3), curve (4), rising (5) and drop (6) as shown in Figure 1. Tone 5b and 6b correspond to tone 5 and 6 on a syllable ended by a stop consonant. Moreover the Vietnamese tonal system can employ some changes of voice quality, with the F0 variations, with co-occurrence of glottalization during the production of tone 3 and tone 6. Tone 3 is accompanied with a harsh voice quality due to a glottal stop (or a rapid series of glottal stops) around the middle of the vowel. Tone 6 has the same kind of harsh voice quality as tone 3; however, it is distinguished by dropping very sharply and it is almost immediately cut off by a strong glottal stop (Do et al., 1998).

The domain of the tonal function is the syllable, which represent a local domain of variation compared to the length of a complete utterance. The attitudinal function concerns the utterance unit, and the prosody of attitude can be described as a global contour related to the utterance (Aubergé, 2002). Modification of F0 values due to either the global attitudinal function or the “local” tonal function seems to be clearly differentiated by native tonal language speakers, but the question of the perceptive
processing of such functional variations by speakers of a non-tonal language, could inform on the cognitive mechanisms of this social signal.

This work aims at exploring possible perturbations by the tonal system on the perception of Vietnamese attitudes by French speakers (i.e. a non-tonal language): will they be able to perceptively extract and separate, from the same acoustic parameters, the tonal values from the attitudinal information? That is to process the lexical access function, attached to word domain, within the attitude function, attached to the whole utterance domain, but morphologically implemented by prominences. May the local tonal variation interfere with the decoding by speakers of a non-tonal language of the utterance-length variations of attitudinal prosody? How does behave such local vs. global cues, described by Gestalt theories of prosodic morphology (Aubergé, 2002).

To answer this question, in this paper, after presenting the construction of the corpus of Vietnamese attitudes, we describe the perceptual experiment of attitudes with tones variation designed for French listeners. The perception results are analyzed, and compared with previous results (Mac et al., 2010) of attitudinal perception on non-tonal (tone 1) Vietnamese utterances by French listeners. The results allow us to answer the question of whether the non-tonal language listeners are able to extract and separate a tone’s lexical F0 value from the attitudinal information. This paper concludes with some discussions and perspectives.

2. Corpus

2.1 Vietnamese attitude corpus

Based on research on some attitudes studied in Vietnamese (Le, 1989) and in other languages (Diaférie, 2002; Shochi, 2008; Rilliard et al., 2009), 16 attitudes have been selected for Vietnamese in our corpus (Table 1).

<table>
<thead>
<tr>
<th>Tone sequence</th>
<th>Utterance in Vietnamese</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>1_1</td>
<td>anh ta</td>
<td>him</td>
</tr>
<tr>
<td>2_1</td>
<td>người ta</td>
<td>them</td>
</tr>
<tr>
<td>3_1</td>
<td>đa xong</td>
<td>finished</td>
</tr>
<tr>
<td>4_1</td>
<td>thuy tinh</td>
<td>glass</td>
</tr>
<tr>
<td>5_1</td>
<td>chung ta</td>
<td>us</td>
</tr>
<tr>
<td>6_1</td>
<td>chỉ ta</td>
<td>her</td>
</tr>
<tr>
<td>5b_1</td>
<td>học ta</td>
<td>hectare</td>
</tr>
<tr>
<td>6b_1</td>
<td>top ca</td>
<td>choral</td>
</tr>
<tr>
<td>1_2</td>
<td>rau cần</td>
<td>celeri</td>
</tr>
<tr>
<td>1_3</td>
<td>dây kem</td>
<td>steel wire</td>
</tr>
<tr>
<td>1_4</td>
<td>cây cảnh</td>
<td>home plant</td>
</tr>
<tr>
<td>1_5</td>
<td>y tá</td>
<td>male nurse</td>
</tr>
<tr>
<td>1_6</td>
<td>danh bù</td>
<td>year book</td>
</tr>
<tr>
<td>1_5b</td>
<td>công tác</td>
<td>mission</td>
</tr>
<tr>
<td>1_6b</td>
<td>sa mac</td>
<td>desert</td>
</tr>
<tr>
<td>4_1_1</td>
<td>bay muroi ba</td>
<td>73</td>
</tr>
<tr>
<td>1_5_1</td>
<td>hai chúng ta</td>
<td>both of us</td>
</tr>
<tr>
<td>6_5b_3</td>
<td>Họp tác xã</td>
<td>cooperation</td>
</tr>
<tr>
<td>1_4_6</td>
<td>em bảo chi</td>
<td>you tell me</td>
</tr>
</tbody>
</table>

Table 2: Sub-set of tonal variation for 2 and 3 syllables length

3. The perception protocol

The perception experiment was carried out to study the influence of Vietnamese tones at varied location on the perception of the 16 Vietnamese attitudes. Twenty French listeners who have no experience with the Vietnamese language took the experiment. The perception test was carried out in a quiet room, using a high-quality headset at a comfortable hearing level. The program interface gave the label and the explanation of the 16 attitudes (in the native language of the listener). No listener expressed any
difficulty in understanding the concepts of these 16 attitudes. All subjects listened to each stimulus only once. After each stimulus, they were asked to indicate the perceived attitude among the 16 attitudes and to indicate the intensity of its expressiveness on a scale ranging from “hardly perceptible” (encoded as 1) to “very marked” (encoded as 100). The score 0 was assigned to the 15 other attitudes.

4. Result analysis
A cross-cultural perceptual experiment has already been performed with Vietnamese attitudes on utterances using only the “neutral” (flat) tone (Mac et al., 2010). This experiment was carried out to have a reference of the non-native perception of attitudes, without tonal variations. For the comparison with the Vietnamese listener’s performances, cf. Mac et al. (2010b)

4.1 Effect of factors
The results of the perception test were first analysed with a repeated measure ANOVA, in order to evaluate the relative effect of the tones and their position on the listener’s perceptual responses. First, the ANOVA of neutral tone sentences (Table 3) show a main effect of the presented attitudes for both Vietnamese and French listeners without a sentence length effect. This result confirms the choice of the 2- vs. 3-syllable length utterances for the experiment on the tonal sentences.

<table>
<thead>
<tr>
<th></th>
<th>Vietnamese</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td>df</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>F</td>
<td>47.804</td>
<td>33.100</td>
</tr>
<tr>
<td>p</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Attitude</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentence Length</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>F</td>
<td>3.735</td>
<td>1.655</td>
</tr>
<tr>
<td>p</td>
<td>0.024</td>
<td>0.191</td>
</tr>
<tr>
<td>Attitude*Sent Len</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>F</td>
<td>3.542</td>
<td>3.007</td>
</tr>
<tr>
<td>p</td>
<td>0.000</td>
<td>0.007</td>
</tr>
</tbody>
</table>

Table 3: Output of ANOVA (on the percentage of attitude recognition) for Vietnamese and French subjects and phrase with tone. Significant effects at the 1% level are set in bold face

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attitude</td>
<td>15</td>
<td>15.790</td>
<td>.000</td>
</tr>
<tr>
<td>Tone</td>
<td>7</td>
<td>1.582</td>
<td>.136</td>
</tr>
<tr>
<td>TonePosition</td>
<td>2</td>
<td>8.301</td>
<td>.000</td>
</tr>
<tr>
<td>Attitude * Tone</td>
<td>105</td>
<td>1.976</td>
<td>.000</td>
</tr>
<tr>
<td>Attitude * TonePosition</td>
<td>30</td>
<td>2.064</td>
<td>.001</td>
</tr>
<tr>
<td>Tone * TonePosition</td>
<td>6</td>
<td>2.519</td>
<td>.020</td>
</tr>
<tr>
<td>Attitude * Tone * TonePosition</td>
<td>90</td>
<td>3.528</td>
<td>.000</td>
</tr>
</tbody>
</table>

Table 4: Output of ANOVA (on the percentage of attitude recognition and level of confidence rating) for French subjects and phrase with tone. Significant effects at the 1% level are set in bold face

For the perception of French subjects on tonal sentences, the ANOVA results (Table 4) show that attitude has a significant effect on perception. There are also significant effects of the interactions between attitude, tones and tone positions. The tone has no significant effect on the perception result. However, the interaction between attitudes and tones is significant. That creates the appearance of the perturbation by tone prosody of some salient cues that are decisive information for some given patterns of attitudes. It must be further verified if it happens only when the local cues can be acoustically confused with salient cues of another global pattern. However the global confusions between attitudes are not changed by tones (see Figure 4 and 5).

4.2 Tones vs. Non-tone structures

4.2.1. Attitude identification
Figure 2 shows the mean recognition rate (in %) for French listeners with 8 representations of Vietnamese tones. The attitude recognition results for the French listeners on the tone variable sentences are not so different. This is verified by the ANOVA result: globally, the tone variation has no effect on attitude perception. That means the non-native listeners can separate the (local) tonal effects and the (global) attitudinal effects.

Figure 2: Mean recognition rate (%) for French listeners for each of the 8 tones presented

Figure 3 shows the perception differences between Vietnamese and French subjects. Globally, most attitudes were recognized above chance level, and native listeners have higher recognition scores than non-native French listeners (except the case of EXn), averaging for tone variation or for the neutral tone sub-corpus. It has to be noted that for French subjects, the neutral tone utterances are better recognized than the non-neutral tone utterances, except the cases of SCO and POL.

Figure 3: Recognition rate of 16 attitudes on non-tone and tonal sentences with Vietnamese and French listeners. The dash line: chance level
4.2.2. Attitude confusion
Figures 4 and 5 show the confusions matrices between attitudes for the varied tone sub-corpus and for the neutral tone sub-corpus. The two most clear results are: (1) on varied tone stimuli, the mean degree of confusion increases; (2) the confusions share the same tendencies in both sub-corpus. Only one new confusion between DOU and INT (that are conceptually close) appears quite clearly for the varied tone stimuli. It means that the local perturbation by tones increases the complexity of the processing of global cues, but does not imply a re-organization, nor a clear misunderstanding by perturbing salient local cues. This result needs to be further explained by studying the similarity in prosodic characteristics of Vietnamese tones and attitudes through the French prosodic patterns.

5. Conclusion
This work aims at studying the cross-cultural perception of Vietnamese social affects, a tonal language where a “neutral tone” can be used. The question of the prosodic influence of the local cues of tones on the global processing of attitudinal prosody can be asked. Some attitudinal stimuli with varied tones were presented to French listeners, who have no experience with lexical tone processing. The main experimental result is that the French listeners can globally separate the tone (local) processing from the attitude (global) processing. The tone processing can be considered as an increased cognitive load for French listeners that reinforces the degree of confusions between attitudes. However, interactions between the tone type, the tone location, and the attitude value indicate that the local cues of tones and the salient cues of global patterns (Aubergé, 2002) could be confused, but depending on the coinciding morphologies of the global and local patterns Thus these results need to be verified by further appropriate acoustic analysis to find out the acoustical parameters that lead to the perception of these social affects.

6. References


### 7. Appendix

Figure 6: Recognition rate per attitude for each tone (1, 2, 3, 4, 5, 6, 5b and 6b) located at the first (top) and the last (bottom) syllable of the sentences. Others syllables in sentences bear the neutral tone (tone 1)
A entonação e a força ilocucionária como pistas da atitude do locutor em atos de fala diretivos

Horácio dos Santos QUEIROZ
Universidade Federal de Minas Gerais
Av. Antônio Carlos, 6627 - Belo Horizonte/MG - Brasil
hsqueiroz@yahoo.com.br

Resumo

O presente estudo relaciona aspectos entonacionais e pragmáticos na expressão de atitudes do locutor a partir do exame do comportamento da curva de \( f_0 \) na produção de atos de fala diretivos do português brasileiro da região metropolitana de Belo Horizonte, Minas Gerais, quais sejam, o pedido, a súplica e a ordem. No nível pragmático, o trabalho explora a Teoria dos Atos de Fala, comparando-se características entonacionais desses atos diretivos à noção de força ilocucionária, sobretudo aos critérios operacionais que a alternam (Vanderveken, 1991). Os resultados demonstram que diferentes estratégias entonacionais estão relacionadas à interpretação do modo de realização efetivo do ato de fala, e que as atitudes do locutor podem ser inferidas, pelo menos em um primeiro instante, com base nas operações que modificam a força ilocucionária.

Keywords: entonação; força ilocucionária; atitude do locutor; atos de fala diretivos.

1. Introdução

A súplica, o pedido e a ordem são atos diretivos que se materializam na comunicação através da forma sintática das sentenças imperativas. A distinção entre esses modos de realização linguística no Português do Brasil (PB) é feita principalmente através da entonação (Rizzo, 1981; Moraes, 1984; Bodolay, 2009; Colamarco, 2009; Queiroz, 2011). Apesar de a literatura concordar sobre o assunto, há certa carência de estudos que desenvolvam mais detalhadamente aspectos referentes à lógica ilocucionária. Por conseguinte, o objetivo do presente estudo é, por um lado, caracterizar padrões entonacionais do português brasileiro para os atos diretivos com modos de realização de pedido, súplica e ordem, mas igualmente privilegiar, com base na Teoria dos Atos de Fala (TAF), aspectos pragmáticos passíveis de envolvê-los, como uma alternativa que possa auxiliar nos estudos voltados para o pragmático. As operações que alteram a força ilocucionária fornecem indícios de relações existentes entre configurações entonacionais específicas e modos de realizações específicos.

2. Entonação e força ilocucionária

De maneira geral, a entonação pode ser considerada como um dos mecanismos utilizados na distinção tipológica de atos de fala. Todo ato de fala pressupõe uma força ilocucionária, um conteúdo proposicional e suas condições de sucesso e satisfação subjacentes à lógica ilocucionária. No entanto, a força ilocucionária mostra-se como um elemento intimamente associado à interpretação do ato de fala, pois é a principal responsável por determinar o modo de realização efetivo do ato de fala (Vanderveken, 1990-91), deduzido com base no ‘vigor’ de sua força ilocucionária, que possui graus variáveis numa mesma dimensão do propósito ilocucionário (Searle, 1995). Alterando-se a força ilocucionária, altera-se necessariamente o modo de realização do ato de fala. Se, por um lado, a entonação é um dos elementos empregados na distinção de atos de fala, por outro, os critérios operacionais que alteram a força ilocucionária fornecem indícios de relações existentes entre configurações entonacionais específicas e modos de realizações específicos.

2.1 Operações que alteram a força ilocucionária

Pela lógica ilocucionária, as operações que alteram a força ilocucionária (Vanderveken, 1991) se resumem em seis e somente seis: (i) restrição o modo de realização do ponto ilocucionário, pela imposição de um modo de realização especial; ii) adicionar um novo conteúdo proposicional particular; iii) acrescentar novas condições preparatórias; iv) acrescentar novas condições de sinceridade; v) v) aumentar ou diminuir o grau de intensidade das condições de sinceridade. O ponto ilocionário é o principal componente da força ilocucionária e determina a direção de ajuste, no caso dos diretivos: fazer o mundo corresponder às palavras; a cada diretivo é imposto um modo de realização especial, com características prosódicas distintas (padrão melódico, duração, amplitudes das variações). A condição de conteúdo proposicional da força ilocucionária é determinada pelo ponto ilocionário, cujo propósito nos diretivos é sempre levar o alocutário a realizar uma ação futura; é basicamente analisável pela boa formação e consistência sintáticas. A condição preparatória dos diretivos consiste nas pressuposições que o locutor faz sobre a situação e seu interlocutor; o locutor pressupõe (ou toma como verdade) que o alocutário seja capaz de realizar a ação futura e que este possa recusar ou não satisfaê-la; as condições preparatórias adicionais estão relacionadas ao acréscimo de elementos adicionais que transcendem de alguma forma a característica autorreferencial do ato de fala diretivo (desejo do locutor), como transmitir um desejo ou vontade; a força ilocucionária é modificada quando há condições de sinceridade adicionais, revelando uma atitude particular, como um desejo somado a uma insatisfação. As condições de sinceridade definem os modos de realização do ponto com diferentes forças ilocucionárias e diferentes
graus de intensidade, por exemplo, quem suplica expressa (estado psicológico expresso) com mais vigor o seu desejo do que quem pede. Em suma, essas seis operações lógicas colocam em xeque a força ilocucionária. Ao investigador cabe estabelecer relações existentes entre os tipos de diretivos e as especificidades entonacionais a fim de identificar o modo efetivo do ato de fala.

3. Métodos

3.1 Corpus


3.2 Coleta de dados

Para coleta dos dados, foram elaboradas dez sentenças imperativas de base, contendo de quatro a sete sílabas, para posteriormente serem proferidas como os atos de fala diretivos propostos. Aos informantes foram explicados os objetivos da pesquisa e o que se pretendia dos atores, de maneira simples e objetiva. A estratégia foi elaborar situações hipotéticas em que a súplica, o pedido e a ordem ocorressem com auxílio de esboços que generalizassem as situações hipotéticas, como abaixo:

<table>
<thead>
<tr>
<th>Súplica</th>
<th>Pedido</th>
<th>Ordem</th>
</tr>
</thead>
<tbody>
<tr>
<td>L depende de A</td>
<td>L não depende de A</td>
<td>A depende de L</td>
</tr>
</tbody>
</table>

Figura 1: Relação Hierárquica Situacional

Quem suplica deseja muito algo e depende daquele a quem dirige seu desejo. Aquele que pede deseja algo e está numa situação de igualdade relativa com o alocutário. Aquele que ordena não depende do outro, ao contrário, está numa posição de autoridade. Para as gravações, foram automatizadas três apresentações distintas de slides, contendo as sentenças a serem proferidas conforme o tipo de diretivo. Ressalta-se que não foi dado ou sugestionado qualquer padrão melódico que pudesse servir de modelo, a fim de que os informantes não mecanizassem padrões melódicos, ficando a critério do conhecimento internalizado dos atores informantes os padrões a serem reproduzidos. Cada uma das dez sentenças de base foi produzida pelos informantes três vezes, em três etapas distintas, com duas etapas livres, sem nenhuma orientação, e uma etapa orientada por uma situação hipotética (Cf. Queiroz, 2011).

3.3 Caracterização pragmática

A atribuição de rótulos para os pedidos baseou-se na Teoria da Polidez (Brown & Levinson, 1987): i) pedido conciso (PdCon), estratégia de polidez aberta e direta (bald on-record); ii) pedido com polidez positiva (PdPol+), estratégia de polidez aberta e indireta com polidez positiva (positive politeness); iii) pedido autoritário (PdAut), estratégia com ações que ameaçam a imagem (face-threatening acts). A ordem foi considerada como prototípica, em razão da literatura na descrição da entonação de atos de fala do PB (e.g. Moraes, 2011; Bodolay, 2009; Colamarco, 2009), e do português europeu (Falé & Faria, 2007). O mesmo se dá no caso da súplica.

3.4 Caracterização melódica

As sentenças foram analisadas através do software Praat (Boersma & Weenink, 1992-2008). As configurações melódicas foram obtidas pela segmentação dos eventos locais (ou eventos-chave): f0 inicial (f0i); pico de f0 (pf0 ou pf0/ton1, quando coincide com a 1ª sílaba tônica não nuclear); sílaba pretônica (preT), antecede imediatamente a nuclear; sílaba tônica proeminent ou nuclear (TonP).

4. Resultados

4.1 Configuração melódica dos diretivos

4.1.1. Pedido conciso

Figura 2: Padrão do pedido conciso

A configuração da curva de f0 no enunciado “Acende a luz” do pedido conciso apresenta um movimento prenuclêntrico (f0i→preT) ascendente/descendente, com o início em um nível relativo médio e pico de f0 (pf0) sobre a primeira sílaba tônica. A configuração intrassilábica da sílaba nuclear (Tonp) descreve um movimento ascendente, exibindo um alinhamento “tardio” (H*→), localizado na porção final da sílaba [s’lus].
4.1.2. Pedido com polidez positiva

![Figura 3: Padrão do pedido com polidez positiva](image)

No pedido com polidez positiva, o contorno prenuclear $f_0 \rightarrow \text{preT}$ é ascendente/descendente, com o início de $f_0$ em um nível relativo médio e pico de $f_0$ localizado sobre a primeira sílaba tônica do enunciado. A configuração intrassilábica da proeminente ($\text{Tonp}$) descreve também um movimento ascendente/descendente, com o pico alinhado à porção mais inicial da vogal da sílaba proeminente (alinhamento adiantado).

4.1.3. Pedido autoritário

![Figura 34: Padrão do pedido autoritário](image)

O padrão melódico global do pedido autoritário é ascendente/descendente. A $f_0$ inicial está situada num nível relativo médio e a curva melódica descreve um movimento ascendente até atingir o pico de $f_0$, localizado no final da vogal da primeira sílaba tônica ($\text{ton1}$). Após o ponto mais alto da curva de $f_0$, a melodia descreve uma suave descida até o final do enunciado, nível relativo mais baixo de $f_0$, com um padrão intrassilábico descendente por toda extensão da sílaba tônica final [aˈ̃du].

4.1.4. Súplica

A configuração melódica global da súplica do enunciado "Acende a luz" é similar à configuração do pedido com polidez positiva. No entanto, o início ($f_0i$) situa-se em um nível relativo significativamente mais baixo de $f_0$, com um padrão intrassilábico descendente por toda extensão da sílaba tônica final [aˈ̃du].

4.2 Interpretação pragmática

A interpretação pragmática considera a relação entre as características dos diretivos e os seis critérios operacionais que alteram a força ilocucionária. A Tabela 1, na página seguinte, caracteriza os atos diretivos conforme as operações que modificam a força ilocucionária e estão sintetizadas como a seguir.
<table>
<thead>
<tr>
<th>Ato</th>
<th>PolCon</th>
<th>PolPol (+)</th>
<th>PolAut</th>
<th>Súplica</th>
<th>Ordem</th>
</tr>
</thead>
<tbody>
<tr>
<td>i) Restrição do modo de realização</td>
<td>Sim</td>
<td>Sim</td>
<td>Sim</td>
<td>Sim</td>
<td>Sim</td>
</tr>
<tr>
<td>ii) Conteúdo proposicional adicional</td>
<td>Não</td>
<td>Não</td>
<td>Não</td>
<td>Não</td>
<td>Não</td>
</tr>
<tr>
<td>iii) Condição preparatória adicional</td>
<td>Sim/Não</td>
<td>Não</td>
<td>Não</td>
<td>Sim</td>
<td>Sim</td>
</tr>
<tr>
<td>iv) Condição de sinceridade adicional</td>
<td>Não</td>
<td>Não</td>
<td>Sim</td>
<td>Sim</td>
<td>Sim</td>
</tr>
<tr>
<td>v) Grau de intensidade ( condições de sinceridade )</td>
<td>Sim/Não</td>
<td>Não</td>
<td>Não</td>
<td>Sim</td>
<td>Sim</td>
</tr>
</tbody>
</table>

Tabela 1: Síntese da caracterização pragmática

i) **Restrição do Modo de Realização** - A cada um dos pedidos é imposto um modo de realização especial, cinco modos de realização do ponto ilocucionário; cinco maneiras distintas, com características entonacionais também distintas: o padrão melódico, a duração e as amplitudes das variações nos movimentos da curva de f0 são pistas importantes para definição do modo efetivo de realização de cada um dos direitivos. Nos pedidos, as diferenças nas configurações melódicas são claras, sobretudo no que se refere às configurações intrassilábicas sobre a sílaba nuclear (TonP). No caso do pedido com polidez positiva e da súplica, embora possuam semelhanças configuracionais, distinguem-se quanto ao nível do ataque e quanto a duração sobre a sílaba nuclear. De modo análogo, o pedido autoritário e a ordem possuem semelhanças melódicas, mas o registro é mais elevado na ordem por toda extensão dos enunciados analisados; o movimento melódico da nuclear é descendente em ambos, mas a ordem apresenta sobre o evento maior variação de f0, taxa de elocução mais elevada e queda mais abrupta do que no pedido.

ii) ** Conteúdo Proposicional Adicional** - A condição do conteúdo proposicional é determinada pelo ponto ilocucionário: toda força ilocucionária de ponto diretivo tem como condição que o conteúdo proposicional represente o desejo do locutor de uma ação futura do alocutário. Todos os cinco tipos possuem o mesmo conteúdo proposicional, sem acrescentar nenhum conteúdo proposicional novo.

iii) **Condição Preparatória Adicional** - A condição preparatória da força ilocucionária consiste nas pressuposições que o locutor faz sobre a situação e seu interlocutor. Em todos os casos o locutor pressupõe (ou toma como verdade) que o alocutário seja capaz de realizar a ação futura e que o alocutário pode recusar ou não a realizá-la. No caso do pedido com polidez positiva e o pedido autoritário não há índices fortes de condições preparatórias adicionais, embora não seja esta uma interpretação estanque. Já o pedido conciso possibilita outra interpretação: o locutor toma como certo que o alocutário seja capaz de realizar a ação futura, mas adiciona a condição preparatória de que esta ação será benéfica ou favorável ao alocutário, ou pelo menos para si mesmo, pois o pedido conciso pode ser interpretado, dependendo do contexto, como uma “sugestão”. Trata-se de um caso representativo em que a relação entre forma e função não se estabelece de maneira exclusiva, pois o padrão melódico do exemplo do enunciado “Acende a luz” (Figura 2, item 4.1.1) caberia confortavelmente em situações nas quais o locutor adicionasse a condição preparatória adicional de que fosse melhor que a luz estivesse acessa. Nessas situações hipotéticas, a interpretação do enunciado com o padrão do pedido conciso seria preferencialmente algo do tipo “Acende a luz... é melhor”. Aliás, o padrão melódico funcionaria de maneira semelhante para situações que requerem interpretações similares: “Feche a porta... é melhor”, “Vai tomar banho... é melhor”, etc. Tanto na súplica quanto na ordem há a adição de condições preparatórias. No caso da súplica, a força ilocucionária é alterada pela condição preparatória adicional de a ação futura ser favorável, pelo menos, e mais geralmente, para o locutor, que toma como certo que está numa situação de dependência, em termos de relação de forças com o alocutário (item 3.2). Na ordem, o locutor toma como verdade que o alocutário pode recusar ou não a obedecer-lá, mas adiciona a condição preparatória de ser ruim para o alocutário, caso não a obediência, pois a relação de forças é desfavorável ao alocutário.

iv) ** Condição de Sinceridade Adicional** - Pelas condições de sinceridade o locutor expressa (ou manifesta) os estados mentais intencionais do locutor, os quais são dirigidos para, ou acerca de objetos e estados de coisas no mundo (SEARLE, 1995) e revelam certos estados psicológicos do locutor. No caso do pedido com polidez positiva e do pedido conciso não há presença clara de índices que os caracterizem como tendo sido modificados pelo acréscimo de condições de sinceridade adicionais. Nos dois tipos, o locutor manifesta abertamente sua intenção, mas os dois tipos sejam diferentes à luz da teoria da polidez (BROWN & LEVINSON, 1978). O pedido conciso é uma estratégia de polidez aberta e direta. O locutor mostra claramente sua intenção, envolve fazê-lo do modo mais direto possível e não há intuito de neutralizar um dano potencial ou conflito que ponha em perigo a própria imagem ou a do alocutário (face-threatening acts). Trata-se, portanto, de um ato passível de ser ameaçador à face do (e.g. um pedido conciso a uma pessoa que mal se conhece, desvaloriza a face do alocutário, criando um dano potencial, mas também a face positiva do locutor, que pode ser visto como uma pessoa grosseira). No pedido com polidez positiva, o locutor mostra abertamente sua intenção, no entanto, a estratégia de polidez positiva é orientada em direção à face positiva do alocutário. Assim, uso do tipo com polidez positiva a uma pessoa que mal se conhece, ao contrário do pedido conciso, é uma estratégia que valoriza a face do alocutário, pois o locutor demonstra uma atitude mais cortês, mais consideração pela face do interlocutor do que no caso do pedido conciso. Já no caso do pedido
autoria, diferentemente dos outros dois tipos de pedidos, a condição de sinceridade adicional ocorre porque o locutor expressa sua vontade, mas adiciona a condição de não estar satisfeito acerca do estado de coisas. O locutor não tem intenção de neutralizar um dano potencial ou conflito, como no pedido conciso, mas difere deste por valorizar face positiva do locutor, ao passo que desvaloriza a face positiva do alocutário (estratégia com ações que ameaça a imagem), o que pode ser interpretado socialmente como um ato de fala ríspido, grosseiro, autoritário etc. Além da impolidez, o pedido autoritário pode indicar impaciência, irritação ou humor momentâneo do locutor. Na súplica, o locutor expressa intencionalmente um estado mental dirigido ao estado de coisas que desejava que estivessem de outra maneira. O locutor valoriza a imagem do outro, de quem sabe hierarquicamente depender (condições preparatórias adicionais), e, por isso, pode sugerir atitudes de submissão ou auto-humilhação, expressando um estado psicológico que desvaloriza a própria imagem, ao mesmo tempo em que busca valorizar a imagem daquele de quem depende, embora possam indicar outros atributos psicológicos, como impaciência, irritação, insatisfação etc. No caso da ordem, o locutor expressa seu desejo, adicionando a condição de sinceridade de não estar satisfeito com o estado de coisas. O modo psicológico é expresso com uma atitude autoritária, de modo a impor o seu desejo, valendo-se de uma condição de sinceridade adicional.

v) Grau de Intensidade das Condições de Sinceridade - Os estados mentais são expressos com diferentes graus de intensidade (degree of strength), dependendo da força ilocucionária. Para que o ato de fala seja perfeito em quantidade (Grice, 1975), o locutor deve expressar sua posição de modo que o ato não seja nem mais nem menos intenso em relação ao seu propósito. Para os pedidos, o grau de intensidade das condições de sinceridade da força ilocucionária é relativamente o mesmo. O desejo que o alocutário faça a ação futura não sinaliza fortemente para graus mais ou menos intensos de desejo, embora, dependendo do contexto, seja possível estabelecer alguma diferenciação, como no caso do pedido conciso, interpretado como sugestão (e.g. "Aceder a luz... é melhor"), visto que a força ilocucionária da sugestão é derivada da força primitiva diretiva, diminuindo-se o grau de intensidade, pois sugerir é uma tentativa mais branda para que o alocutário faça a ação futura, do que pedir, suplicar ou ordenar. O grau de intensidade das condições de sinceridade da súplica, por sua vez, é mais intenso do que quem pede, porque quem suplica expressa um desejo mais intenso do que quem pede ou sugestiona. A ordem apresenta também características que indicam que o desejo do locutor seja mais intenso do que nos pedidos, o grau de intensidade das condições de sinceridade é geralmente expresso através da entonação, logo, um increase in the degree of strength of the intonation contour serves in general to increase the degree of strength of the sincerity conditions” (Vanderveken, 1991: 119).

5. Conclusão

A entonação fornece informações importantes para definição do modo de realização efetivo dos diretivos analisados, evidenciando que o pedido, a súplica e a ordem não são categorias estanques, como exemplifica o caso do ato de pedir, que pode ser feito, pelo menos no dialeto mineiro, de duas maneiras diferentes, dois modos de realização, mas com a mesma intenção comunicativa, apesar de possuírem forças ilocucionárias diferentes. Aliás, o pedido conciso, sua regularidade (136 ocorrências no total de 300) e sua força ilocucionária revelam modos de organização social, visto que, dependendo do contexto, não seria apropriado dirigí-lo, numa situação formal, a alguém que mal se conhece ou acabou de se conhecer, pois o modo pelo qual as coisas são socialmente organizadas exige, a sua maneira, outro comportamento entonacional. Os resultados demonstram que alguns padrões melódicos são mais difíceis de serem relacionados à expressão de atitudes, como nos caso do pedido conciso e do pedido com polidez positiva, considerados padrões entonacionais mais “neutros”, em comparação com os demais diretivos analisados. Os demais tipos são melodicamente marcados, como a súplica, cujo padrão entonacional pode ser relacionado à atitude de submissão ou auto-humilhação, e a possibilidade da sobreposição de atitudes como insatisfação ou descontentamento com o estado de coisas, podendo ser associado às atitudes impaciência, irritação, à polidez e mesmo outros estados afetivos, como, o humor momentâneo do locutor. Ou ainda pela ordem, em que o locutor impõe sua vontade de maneira autoritária. Enfim, nos casos em que a entonação não contribui como um forte índice das atitudes, estas podem ser interpretadas com base em fatores internos e externos ao sistema linguístico, que incluem aspectos sintáticos, semânticos e pragmáticos, bem como as noções de estado psicológico expresso, de conteúdo proposicional, adição de condições preparatórias e de sinceridade adicionais.

6. Referências


Facial gestures in the expression of prosodic attitudes of Brazilian Portuguese

João Antônio de MORAES1,2, Luma da Silva MIRANDA1, Albert RILLIARD3
1Universidade Federal do Rio de Janeiro (UFRJ), Brazil; 2CNPp; 3LIMSI-CNRS, Orsay, France
jamoares3@gmail.com, lumah.miranda@gmail.com, albert.rilliard@limsi.fr

Abstract
This paper presents the results of (i) an identification test of Brazilian Portuguese prosodic attitudes based on visual cues and (ii) a preliminary analysis of the facial gestures involved in its expression. Eleven attitudes, separated between social and propositional categories, performed by two native Brazilian speakers, were audio-visually recorded and analyzed in terms of Ekman’s Action Units, in order to correlate the speaker’s intention and the objective manifestations of facial expressions. Results show the importance of these gestures for the recognition of attitudes as well as the consistency between the two subjects in their use of facial gestures.

Keywords: visual prosody; attitudes; facial gestures.

1. Introduction
In the last decade, several studies in visual prosody have been undertaken to explore in a given language how audio and visual features combine to express the so-called intonational meaning, either properly linguistic (Kendon, 2004; Wollermann & Schröder, 2009; Wollermann et al., 2012) or attitudinal (Rilliard et al., 2009; Tanaka et al., 2010). It seems rather obvious that in face-to-face interactions, attitudes are expressed and perceived within a multimodal paradigm, integrating audio and visual elements (Barkhuysen et al., 2007). However, while several studies on prosodic attitudes have been carried out, most of them have analyzed the acoustic modality only. The multimodal approach has yet to be fully explored. The audiovisual expression of attitudinal meanings is, in a large extent, conventionally encoded within a particular culture and a particular language. They are learned by the speaker and are produced during face-to-face communication, which implies that the manifestation of these attitudes may be ambiguous or even not recognized by foreign speakers.

The importance of facial gestures for the recognition of prosodic attitudes in Brazilian Portuguese (BP) was shown in a previous study (Moraes et al., 2010). In this study, after presenting the main results of these identification tests, we will focus on the description of the gestures involved.

2. Method
2.1 Corpus
A semantically neutral declarative sentence: “Roberta dançava.” (“Roberta was dancing.”) was produced by two BP speakers with eleven different attitudes. These attitudes were grouped in two categories: (i) propositional attitudes, that refer to speaker’s attitudes towards the propositional content of the sentence and (ii) social attitudes, which represent the speaker’s attitudes towards its interlocutor. Five propositional attitudes were performed: doubt (DOU), irony (IRO), incredulity (INC), obviousness (OBV) and surprise (SUR); and six social attitudes: arrogance (ARR), authority (AUT), contempt (CON), irritation (IRR), politeness (POL) and seduction (SED). A “neutral” attitude was also produced, characterized by the absence of any special affect. Each of the 12 attitudes was performed in assertive mode.

2.2 Perceptual validation
The stimuli have been presented in three modalities (audio-only, visual-only, audio-visual) to 29 native BP listeners who had to recognize, in a forced-choice paradigm, the performed attitudes, among the possible attitudes in a given category, propositional or social. Each attitudinal label was completed by a longer description, in order to ease its identification by the listeners.

Each stimulus was played/showed twice on each run. Subjects had to give their answers by selecting on a slider the relative intensity of the perceived attitude. The scale ranged from “barely marked attitude” to “very marked attitude”.

2.3 Description of facial gestures
To describe the facial movements present in the expressions of attitudes a simplified version of the Facial Action Coding System (FACS) proposed by Ekman and colleagues (2002) was adopted. An Action Unit (AU) is defined as a muscular activity that produces momentary changes in the facial features in various areas of the speakers’ face. The facial topography is divided into two principal areas. The first area is the upper face which affects the eyebrows, forehead, and eyelids; the second area is the lower face, which includes movements such as up/down, horizontal and oblique motions of the head, the shoulder and/or the jaw. Using Ekman’s system of facial mapping, the following 15 Action Units were selected for our analysis:

(a) Eyebrow raiser (Inner + Outer brow raiser) AU 1+2
(b) Eyebrow lowerer AU 4
(c) Lid tightener AU 7
(d) Upper lid raiser AU5
(e) Blink AU 45
(f) Lip corner depressor AU 15
(g) Lip corner puller AU 12
(h) Upper lip raiser AU 10
(i) Jaw drop AU 26
(j) Cheek raiser AU 6  
(k) Up and down head movement AU 85  
(l) Right and left head movement AU 51+52  
(m) One side tilt movement AU 55/56  
(n) Head up AU 53  
(o) Shoulder shrug AU 82

Three researchers separately analyzed each video, marking the emergence of AU’s related to the upper face, lower face and head positions based on appearance changes according to the FACS Manual (Ekman et al., 2002), and reached a consensus in case of disagreement; the intensity of the appearance change was not scored.

3. Results

3.1 Identification test

The results of perceptual recognition tests (Moraes et al., 2010, 2011) indicated that the overall recognition rate increases when both audio and visual channels were combined, but when we have access to only one channel, the visual one is generally more effective for the recognition of the speaker's attitude than the audio channel. There is, nevertheless, a significant difference: while for propositional attitudes the performance of each channel separately is relatively close, for social attitudes the difference is striking (Figure 1).

![Figure 1: Mean intensity rating in each modality, for both types of attitudes](image)

By looking at the identification of each propositional attitude (Figure 2), it is clear that for most attitudes the visual information prevailed (although the observed difference was not very pronounced); with the exception of incredulity, in which audio information was dominant, each of these channels was itself actually very effective, showing a recognition rate of the speaker’s intention far above the simple chance.

For social attitudes (Figure 3), however, the presence of visual information is crucial. In some attitudes such as arrogance, contempt and authority the audio information is poorly recognized, near the chance level. This is probably due to the fact that among social attitudes there are not prosodic patterns clearly distinct, as occurs with propositional attitudes (Moraes et al., 2011).

![Figure 2: Mean intensity rating for the identification of propositional attitudes. Results for audio-only in pink (1st column), visual-only in blue (2nd col.) and audio-visual in brown (3rd col.), both speakers](image)

<table>
<thead>
<tr>
<th>percep. produc.</th>
<th>DOU</th>
<th>OBV</th>
<th>INC</th>
<th>IRO</th>
<th>NEU</th>
<th>SUR</th>
</tr>
</thead>
<tbody>
<tr>
<td>doubt</td>
<td>44</td>
<td>4</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>obviousness</td>
<td>2</td>
<td>47</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>incredulity</td>
<td>10</td>
<td>1</td>
<td>33</td>
<td>10</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>irony</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td>46</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>neutral</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>55</td>
<td>0</td>
</tr>
<tr>
<td>surprise</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>49</td>
</tr>
</tbody>
</table>

Table 1: Confusion matrix of visual stimuli in propositional attitudes (both speakers)

If we examine the confusion matrices concerning these attitudes, it can be seen that there were few confusions between the production and the perception, indicating that the gestures were sufficiently distinct in general. Among propositional attitudes (Table 1) confusions based on visual information are rather rare: they occur basically between incredulity vs. doubt and between incredulity vs. irony, in both directions, what can be explained by the fact that these attitudes are semantically close. Interestingly, the visual recognition of incredulity was offset by the audio channel, which received better scores.

Social attitudes were also generally well recognized visually, although in a somewhat less effective way, with confusions for arrogance, interpreted as contempt (quite similar attitudes) and politeness, interpreted as neutral.
3.2 Facial gestures

The preliminary findings in this study disclosed discrete categories formed by the AU’s for the facial expressions of each attitude performed; each attitude was distinguished from the others by the set of its AUs (table 3 in appendix). On average 3.8 AUs were employed in the expression of each attitude, with virtually no difference between the number of gestures present in propositional (3.9) and in social (3.7) attitudes. It is noteworthy that the male subject has used on average more AUs (4.4) than the female one (3.2). Although some attitudes were occasionally conveyed using different strategies between the subjects (irritation, for instance, has no AUs in common between the subjects), the overall similarity of the gestures employed by them is striking, which can be verified by a simple visual inspection of selected photos put side by side (Figures 4 to 15), which illustrate the attitudes expressed by each subject (for a closer and more effective examination, the attached video set can be seen).

Table 2: Confusion matrix of visual stimuli in social attitudes (both speakers)

<table>
<thead>
<tr>
<th>percep. produc.</th>
<th>AR</th>
<th>AU</th>
<th>SED</th>
<th>CON</th>
<th>IRR</th>
<th>NEU</th>
<th>POL</th>
</tr>
</thead>
<tbody>
<tr>
<td>arrogance</td>
<td>29</td>
<td>3</td>
<td>0</td>
<td>25</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>authority</td>
<td>8</td>
<td>43</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>seduction</td>
<td>0</td>
<td>0</td>
<td>40</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>contempt</td>
<td>9</td>
<td>1</td>
<td>48</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>irritation</td>
<td>6</td>
<td>7</td>
<td>0</td>
<td>6</td>
<td>36</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>neutral</td>
<td>5</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>45</td>
<td>2</td>
</tr>
<tr>
<td>politeness</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>23</td>
<td>27</td>
</tr>
</tbody>
</table>

3.3 Propositional attitudes
3.4 Social attitudes

Some AUs, such as (a) AU 1+2 and (b) AU 4 (eyebrow movements) and the four different head movements here considered are rather frequent and very productive in discriminating attitude groups. Others, on the contrary, have an occasional, limited participation, such as (d) upper lid raiser, (i) jaw drop, (j) cheek raiser and (o) shoulder shrug, and are frequently associated to specific attitudes. Thus (d) AU 5 (upper lid raiser) is typical of surprise, and so is (i) AU 26 (jaw drop); (n) AU 53 (Head up) denotes arrogance, and (o) shoulder shrug correlates with obviousness (and also contempt, for the male subject). It is worth noting that (h) AU 10 (upper lip raiser) and in a large extent, (e) AU 45 (blink), are basically dedicated to the expression of social attitudes. Interestingly, the (f) lip corner depressor (AU 15), which appears in four different attitudes, was used only by the male subject, while the (h) upper lip raiser (AU 10) was used only by the female subject; their use in the attitudes of arrogance and contempt seems to suggest that they are individual (or may be gender) gestural variants in the expression of the same set of attitudes.

It can be observed, finally, that pairs of attitudes that are semantically close, such as arrogance/contempt, or politeness/seduction were expressed by a similar set of gestures: they were distinguished from each other by a small number of AUs.

On the other hand, semantically distant propositional attitudes, such as incredulity and obviousness, can be also visually quite similar, which did not prevent them from being clearly identified visually, probably due to the presence of the distinctive shoulder shrug in obviousness, and the difference in head orientation.

4. Conclusions

The results of this study confirm that listeners rely upon the visual channel to better understand what attitudes a speaker is communicating in face-to-face speech, and the facial mapping here undertaken provides a preliminary framework for identifying and interpreting which facial features communicate which particular attitude.

Because of the limited size of this study, the results are not yet conclusive. Additional research with a greater number of native Brazilian Portuguese speakers will be required to confirm the accuracy of these findings and to address other Brazilian Portuguese prosodic attitudes.

5. References


### 6. Appendix

<table>
<thead>
<tr>
<th>Attitudes</th>
<th>AUs</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
<th>g</th>
<th>h</th>
<th>i</th>
<th>j</th>
<th>k</th>
<th>l</th>
<th>m</th>
<th>n</th>
<th>o</th>
</tr>
</thead>
<tbody>
<tr>
<td>doubt</td>
<td>XY</td>
<td>XY</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>XY</td>
<td>XY</td>
<td>XY</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>irony</td>
<td>Y</td>
<td>XY</td>
<td>XY</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>XY</td>
<td>XY</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>incredulity</td>
<td>Y</td>
<td>Y</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>XY</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>obviousness</td>
<td>Y</td>
<td>Y</td>
<td></td>
<td></td>
<td>XY</td>
<td>XY</td>
<td>XY</td>
<td>XY</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>surprise</td>
<td>XY</td>
<td>XY</td>
<td>XY</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>arrogance</td>
<td>XY</td>
<td>XY</td>
<td>X</td>
<td>Y</td>
<td>X</td>
<td></td>
<td>X</td>
<td>Y</td>
<td>XY</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>authority</td>
<td>Y</td>
<td>Y</td>
<td></td>
<td></td>
<td>XY</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>contempt</td>
<td>XY</td>
<td>X</td>
<td>Y</td>
<td>Y</td>
<td>X</td>
<td></td>
<td>Y</td>
<td>XY</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>irritation</td>
<td>Y</td>
<td>Y</td>
<td></td>
<td></td>
<td>Y</td>
<td>X</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>politeness</td>
<td>Y</td>
<td></td>
<td>XY</td>
<td></td>
<td></td>
<td></td>
<td>XY</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>seduction</td>
<td>Y</td>
<td></td>
<td>XY</td>
<td>XY</td>
<td></td>
<td></td>
<td>Y</td>
<td>XY</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: AUs in propositional (in red) and social (in blue) attitudes for female (X) and male (Y) speakers; the letters (a) to (o) correspond to the 15 AUs listed in 2.3
Acoustic analysis of a corpus of Brazilian Portuguese attitudes

João Antônio de MORAES¹, Albert RILLIARD², Donna ERICKSON³, Takaaki SHOCHI⁴
¹Federal University of Rio de Janeiro (UFRJ), Brazil; ²LIMSI-CNRS, Orsay, France; ³Showa Music University, Japan; ⁴Bordeaux 3University, France
jamoares3@gmail.com, rilliard@limsi.fr, ericksondonna2000@gmail.com, shochi38@gmail.com

Abstract

This paper presents the prosodic analysis of a corpus of Brazilian Portuguese attitudes. Attitudes are separated between the social and propositional categories, and performed either with an assertive or an interrogative modality. Previous studies show the particular relevance of prosodic cues for propositional attitudes, while visual cues are more relevant for social ones. This paper shows that this greater relevance of prosody for propositional attitudes is also observed on the prosodic parameters’ variations – and enhance particularly the clearly different and prototypical F0 contours that distinguished such expressions.

Keywords: prosodic attitudes; Brazilian Portuguese.

1. Introduction

The expression of a speaker’s opinion, belief and knowledge to his interlocutor is partly performed through the use of prosodic attitudinal expressions (Wichmann, 2000). The use of such prosodic strategies constitutes an important part of the speaker’s engagement in his speech (Daneš, 1994) and may contribute for an important part of the semantic content of utterances. For example, a sentence produced with an ironic tone of voice will certainly not carry the same meaning than a more its more neutrally performed counterpart. Such prosodic attitudes differ from emotional expressions in that they are voluntarily produced during the interaction, in a given social setting where the attitudes are conventionally encoded for a language and a culture, and may vary with them (Rilliard et al., 2009).

Typologies of attitudinal expressions vary with authors and their points of interest (e.g. Martins-Baltar, 1977; Gu et al., 2011). The present study is based on a separation between two categories of attitudes (already used by Martins-Baltar, 1997 and Fónagy et al.,1984): propositional and social attitudes. The propositional ones address the propositional content of the sentence (e.g. doubt, obviousness, irony), while social ones refer to the interpersonal relationship between the speaker and the receiver (e.g. politeness, irritation, arrogance). Wichmann (2000) proposes a similar distinction between what she calls propositional and behavioural categories of attitudes.

This study describes the prosodic analysis of a corpus of such attitudes in Brazilian Portuguese (BP). The attitudes have been perceptually validated in previous studies (Moraes et al., 2010, 2011), and the present paper will focus on the prosodic parameters relevant to such a perception. After describing the corpus of BP attitudes, the process of prosodic analysis is detailed, and the main results observed on the corpus are given.

2. Method

The set of attitudes used in this study is based on the distinction between propositional and social attitudes introduced above, with a supplementary distinction between the assertive or interrogative modes of the carrying sentences.

The attitudes described here are the following:

**Assertive mode:**
- **Social:** arrogance (ARR), authority (AUT), contempt (CONT), irritation (IRR), politeness (POL) and seduction (SED);
- **Propositional:** doubt (DOU), irony (IRO), incredulity (INC), obviousness (OBV) and surprise (SUR).

**Interrogative mode:**
- **Social:** arrogance (ARR), authority (AUT), contempt (CONT), irritation (IRR), politeness (POL) and seduction (SED);
- **Propositional:** confirmation (CONF), incredulity (INC), rhetoricty (RET) and surprise (SUR).

The labels and the number of attitudes vary according to the sentence’s mode (11 attitudes for assertion, 10 for interrogation), as some attitudes are incompatible with some modes (e.g. obviousness with interrogation).

<table>
<thead>
<tr>
<th>BP sentence</th>
<th>L.</th>
<th>Stress</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ta</td>
<td>1</td>
<td>oxytone</td>
<td>OK</td>
</tr>
<tr>
<td>Vaidançaar</td>
<td>3</td>
<td>oxytone</td>
<td>(s)he is going to dance</td>
</tr>
<tr>
<td>Dançava</td>
<td>3</td>
<td>paroxytone</td>
<td>he/she danced</td>
</tr>
<tr>
<td>Roberta vai dançar</td>
<td>6</td>
<td>oxytone</td>
<td>Roberta is going to dance</td>
</tr>
<tr>
<td>Roberta dançava</td>
<td>6</td>
<td>paroxytone</td>
<td>Roberta danced</td>
</tr>
</tbody>
</table>

Table 1: Sentences used for the attitudes, with their length (L., in syllables), the position of the lexical stress on their last word and an English translation

All attitudes were performed by two native BP speakers (a female and a male), on a set of five sentences from 1- to 6-syllable long and with varying lexical stress position (cf. Table 1). The sentences don’t have any particular meaning in relation to the attitudes nor the modes. Their performances were audio-visually recorded.

---

Heliana Mello, Massimo Pettorino, Tommaso Raso (edited by), *Proceedings of the VIIth GSCP International Conference: Speech and Corpora*
using high quality equipment.

These attitudes (including the neutral assertive and interrogative sentences), performed on the 5 sentences, were recorded (in the audio and video modality) in three repetitions by the two speakers, resulting in 690 stimuli. The recordings were phonetically aligned by hand, using Praat (Boersma & Weenink, 2011).

3. Perceptual validation

In order to assess the pertinence of the speakers’ performances, one repetition of each attitude from the last sentence of Table 1 were chosen in order to perform perception tests, separately for the assertive and interrogative modes. These attitudes have been presented in three modalities (audio-only, visual-only, audio-visual) to native BP listeners who had to recognize the performed attitudes, among the possible attitudes in a given mode and category (propositional or social). The perception results are fully described in Moraes et al. (2010, 2011), and provide a validation of the pertinence of the above-described prosodic parameters. Figure 1 presents the mean recognition scores obtained by each attitude, in each three modality, for the two modes and for propositional or social attitudes.

The most important result that was learned from these perception tests concerns the relative importance of visual and audio modality to the recognition of the two categories. While the visual cues clearly outperformed the audio cues for the recognition of social attitudes, it seems that audio cues are generally more important than the visual ones for the propositional attitudes (mostly for propositional interrogatives).

This primary use of audio cues for signalling information relating to the propositional content of utterances rather than information relating to the interpersonal relationship during a face-to-face interaction is interesting and led us to a complete analysis of the prosodic variation of this attitudinal

corpus.

4. Prosodic analysis

From each of the 690 stimuli, the following prosodic parameters were extracted: the fundamental frequency (F0, expressed in semitones), the intensity (in dB), and the phonemic duration expressed in z-score, following Campbell (1993) method. Both F0 and intensity were measured on each vowel, at three points (at 10, 50, 90% of the vowel’s length).

4.1 Means values over sentences

As it has been claimed by e.g. Gu et al. (2011), the mean distribution of pitch over sentences already gives indication on the type of attitude: a high or low pitch – regarding to the speaker’s mean laryngeal frequency, constitute a first kind of indices.

<table>
<thead>
<tr>
<th>Category</th>
<th>Spk.</th>
<th>F0 (st)</th>
<th>Z-duration</th>
<th>Intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Propositional</td>
<td>F</td>
<td>93.4 (5.1)</td>
<td>0.39 (1.31)</td>
<td>68.9 (6.5)</td>
</tr>
<tr>
<td>Social</td>
<td>F</td>
<td>92.1 (3.4)</td>
<td>0.25 (0.61)</td>
<td>68.8 (6.2)</td>
</tr>
<tr>
<td>Propositional</td>
<td>M</td>
<td>83.8 (6.3)</td>
<td>0.38 (1.26)</td>
<td>68.5 (4.9)</td>
</tr>
<tr>
<td>Social</td>
<td>M</td>
<td>81.6 (4.7)</td>
<td>0.24 (0.68)</td>
<td>68.8</td>
</tr>
</tbody>
</table>

Table 2: Mean (standard deviation) of F0, Z-duration and intensity observed for each category of attitude, and for each speaker (Female & Male).

Figure 2 (in appendix) presents the distributions of F0 values for each attitude over all sentences. In each category of attitude, different patterns of distributions are observed: attitudes with high mean pitch and wide distribution (e.g. CONF and DOU), attitudes with a low and flat pitch (e.g. INC), etc. – supporting the above hypothesis.

A comparison of propositional and social attitudes shows a tendency to a wider distribution of the measured parameters in the case of the former, supporting the perceptual result: a higher importance of prosodic cue for these attitudes (cf. Table 2). This is mainly marked for F0 and Z-duration parameters.

4.2 Prototypical contours

The means and distributions of prosodic parameters can hardly distinguish between a complex set of attitudes. The evolution of these parameters across time and with respect to the carrying sentence’s morphosyntactic structure shall also play a role. To assess such an importance of prosodic contours, they have also been inspected for all three prosodic parameters (cf. Figures 3 and 4 in appendix for the F0 contours in the interrogative mode).

Interestingly, the shapes of contours for propositional attitudes are characteristically different for each one, while the shapes of social attitudes tend to be more similar. For example, for the 1-syllable long sentence (first columns in Figures 3 and 4) propositional attitudes show a large diversity of contours (rising,
falling, flat-rising…), while the contours observed for social attitude are all rising – with small differences of pitch mean. The increase of sentences length shows the evolution of the global contours’ shapes that tend to conserve a similar shape, whatever the length (under some constraints of minimal length). Such an observation is in line with Morlec et al. (2001) principle of “prosodic movement expansion” shown on French prosodic attitudes.

Visual inspection also shows the influence of the linguistic constraints of prosody on the global contours of attitudes: a main difference between Morlec et al. (2001) description and BP attitude is linked with the importance and varying position of lexical stress in BP. Whereas lexical stress in French always occurs at the final syllable, the described corpus proposes a systematic variation of oxytone and paroxytone words at the end of sentences. So, the two 3- and 6-syllable long sentences (respectively oxytone at the 2nd and 4th columns, and paroxytone at the 3rd and 5th columns of Figures 3 and 4) have a different morphosyntactic constraint that imposes a varying position of the main F0 peak and lengthening. This is especially clear for the CONF attitude (2nd line on the figures), where the final slope occurs on the stressed syllable – for both speakers. Such a phenomenon can also be seen for other attitudes. The other parts of the contours remain similar across sentences.

For the segmental duration, similar phenomena are observed. Figure 5 (in appendix) shows the large lengthening of the stressed syllables observed for IRON (5th row) that differ completely from the strategies used by this speaker to perform the others propositional attitudes. In a similar fashion, social attitudes’ duration patterns tend to be more comparable across attitudes.

5. Discussion & conclusion

This paper has presented a prosodic analysis of the variation induced by attitudual expressions into the prosodic parameters of a set of BP sentences. These modifications affect the speaker’s mean register, pitch range and rhythm. To rate the efficiency of mean prosodic patterns to convey attitudinal expressions would require perceptual tests based on a gating paradigm to check whether e.g. a high start followed by a slope at the beginning of a sentence will be systematically perceived as an expression of rhetoric question (cf. Shochi et al., 2009, for such an experiment on Japanese attitudes).

The modifications also affect the sentences’ prosodic contours. Prototypical strategies have been observed for each propositional attitude, and are reproduced in a similar fashion over speakers for several attitudes – but not for all. The CONF attitude show a rise until the last stressed syllable for the female speaker, while the male speaker tend to produce a high plateau, but both make a steep slope on the stressed syllable. This shows that several communication strategies may coexist in a same language, with common grounds. This variation may be accounted for by gender differences, but more investigation is required to confirm this hypothesis. Preceding perception results have also shown such inter-speakers differences, with higher performances obtained by either the female or the male speaker on several attitudes. To describe the possibility of strategic variations inside a given attitudinal expression would require a larger set of speakers to be recorded and analysed.

6. Acknowledgements

This work was partly supported by the French ANR PADE project, and by a Brazilian CNPq grant.

7. References

8. Appendix

Figure 2: Dispersion of the F0 values measured for the female and male speakers on the propositional (left) and the social (right) attitudes, for both assertive and interrogative modes.

Figure 3: F0 contours (mean of 3 repetitions in black, standard deviation in gray) for the 5 interrogative sentences (in columns) with the 4 propositional attitudes plus the neutral interrogation (first row), as performed by the femalespeaker.
Figure 4: F0 contours (mean of 3 repetitions in black, standard deviation in gray) for the 5 interrogative sentences (in columns) with the 6 social attitudes, plus the neutral interrogation (first row), as performed by the female speaker.

Figure 5: Z-duration contours (mean of 3 repetitions, with standard deviation) for the 5 assertive sentences (in columns) with the 5 propositional attitudes (left, in rows) plus the neutral declaration (first row), and for the 6 social attitudes (right panel, in rows) plus the neutral declaration (first row), as performed by the male speaker.
Developmental perception of polite & impolite non-verbal behaviours in Japanese

Albert RILLIARD¹, Takaaki SHOCHI², Donna ERICKSON³, João Antônio de MORAES⁴
¹LIMSI-CNRS, Orsay, France; ²University Bordeaux 3, France; ³Showa Music University, Japan; ⁴Federal University of Rio de Janeiro (UFRJ), Brazil
albert.rilliard@limsi.fr, Takaaki.Shochi@u-bordeaux3.fr, ericksondonna2000@gmail.com, jamaores3@gmail.com

Abstract

This paper uses a corpus containing a set of prosodic attitudes, encoded by Japanese culture and language to express politeness and impoliteness. Three expressions of politeness are used: courtesy-politeness, sincerity-politeness and kyoshuku. A neutral declarative expression and an impolite attitude of arrogance complete the set of expressions. The question addressed here is twofold: first can young Japanese children perceive the expressive differences conveyed by these 5 attitudes in a way similar to that of native adults? And second, can a pair comparison paradigm be used with young (6 to 10 years) children still unfamiliar with the written language? Results show the progression with age of children’s perceptual spaces towards adults’ perception. The perceptions of audio and visual modalities are also compared.

Keywords: prosodic attitudes; Japanese; (im)politeness; perceptual development.

1. Introduction

During face-to-face interactions, speakers are involved in their speech (cf. Daneš, 1994: 253). They convey their message through their lexical and syntactic choices, as well as through gestures, facial expressions and prosodic variations. Emotional expressions may be seen as the most typical example of involvement in speech – as they are always part of an utterance. Such a continuous variation of emotional phenomenon is described by Russell & Barrett (1999) as “core affects” – elementary affective feelings always present and continuously varying. They separate such affects from “prototypical emotional episodes”, which are the rare instances of full-blown basic emotions, conceptualized in language through lexical items. Such prototypical emotional episodes correspond to so-called “emotions”. Their descriptions and the labels naming them may be refined hierarchically up to fine conceptual differences (cf. Golan et al., 2006 for such a list of labels). As any conceptually constructed object, such emotions may be described by scripts, able to capture subtle differences and similarities across cultures (Wierzbicka, 1986; Russell, 1991).

Widden & Russell (2003) describe the acquisition process leading to a diversification of the use of such emotional labels by children of 2 up to 5 years old: children master on average 1 emotional label at the age of 2, and 6 at the age of 5. In a similar fashion, the classification of emotions proposed by Zinck & Newen (2008) postulates an increase of affect types’ complexity with the cognitive development of children, and their different physiological stages. The most complex affects of their classification are coined “secondary cognitive emotions” (such as shame or pride); they are related to cultural norms, and demand an experience of social relationships. Such kinds of affects are strongly linked to the culture and the language in which there are conceptualized. Such “social affects”, as well as other kinds of expressive behaviour such as irony or politeness, are described by Wichmann (2000) as attitudinal expressions – because they allow the speaker to express his attitude towards what he says or towards his interlocutor in a given interaction context. Such attitudes are part of the speakers’ communicative strategies, and to be efficient, they must observe linguistic and cultural norms.

This work is based on a corpus that contains a set of such prosodic attitudes, typical of the Japanese language and culture (cf. Shochi et al., 2009a). A subset of this corpus, grouping 5 attitudes of politeness or impoliteness has been selected. Three politeness expressions are used: courtesy politeness (PO), sincerity politeness (SIN) and a typically Japanese expression of kyoshuku¹ (KYO). A neutral declarative expression (DC) and an impolite expression of arrogance (AR) complete this set. Detailed definitions may be found in Shochi et al. (2009b). This work aims at measuring on one hand if young Japanese children perceive the expressive differences encoded by these 5 attitudes in a fashion similar to that of native adults; and on the other hand if a pair comparison paradigm can be successfully applied with groups of children still not accustomed to written language.

2. Methodology

Shochi et al. (2009b) measured the ability of children (who could read Japanese) to judge the degree of politeness of these five prosodic attitudes. Judgements were made on a politeness scale (ranging from “impolite” to “polite”), with neutral in the middle). The results acknowledge the position of arrogance on one hand and courtesy politeness and sincerity politeness on the other hand to each end of the scale. Meanwhile, both declarative and kyoshuku expressions were placed close to

¹ This Japanese word, without English equivalent is described by Sadanobu as “a mixture of suffering ashamedness and embarrassment, [which] comes from the speaker’s consciousness of the fact that his/her utterance of request imposes a burden to the hearer” (2004: 34).
“neutral”. This result was interpreted as a strong hint of the multidimensional nature of such expressions, which cannot be constrained on this one-dimensional polite-impolite scale. Another limitation of this preceding study is linked to the mandatory use of written instructions describing complex concepts such as politeness. Children below 9 years old do not have sufficient reading skills to adequately perform such a task. But oral presentation of these concepts may also raise difficulties with younger children.

The work of Romney et al. (2000) and Moore et al. (2002) propose an experimental methodology allowing a precise evaluation of the multidimensional nature of a semantic domain, and an evaluation of the cross-cultural variation of this structure. Their work mostly applies to lexical entries (e.g. emotional or kinship terms), but their methodology may be applied also to the comparison of prosodic expressions. Most interestingly, the statistical methods developed by these authors allow a precise (and quantified) comparison of the specificities of different groups of subjects, as well as the quantification of the amount of shared knowledge between groups. Graphical representations of the main dimensions that structure the semantic space under investigation is an additional interest of this approach. The methodology is based on the evaluation by subjects of a perceived distance between stimuli. Such an evaluation of distances between pairs of stimuli is quite straightforward to explain, and does not require complex conceptual definitions or understanding. Such a pair-comparison paradigm was thus selected to test young children at different ages, and compare their results to native adults’.

2.1 Stimuli
Stimuli presented to subjects consist of the 5 attitudes, spoken on one sentence-type by a native Japanese speaker, a teacher of Japanese as a foreign language, trained to play such expressions in front of students. The same sentence, which has no connotation linked with any (im)politeness attitude, is used to produce all the attitudes. The speaker’s performance was recorded with a high quality microphone and a DV camera in a sound proof booth; individual sentences were segmented by hand.

The prosodic and behavioural performances of the speaker show some characteristic differences between the five attitudes that are summarized here. Expressions of sincerity politeness and kyoshoku have a faster mean syllabic rhythm with a more flat F0 contour (especially for kyoshoku) than the others. They show a limited F0 and intensity register, around the speaker’s mean. Courtesy politeness and arrogance, like declaration, show a comparatively wider F0 and intensity slope over the sentence; the F0 slope is more pronounced and linearly decreases in the case of courtesy politeness. The voice quality of each of the five attitudes can be heard as clearly different. Declaration and courtesy politeness use modal voice, while sincerity politeness uses a breathy phonation, which softens the speaker’s voice. Kyoshoku is performed with a characteristic tense, creaky voice. For arrogance, the speaker uses a nasalized phonation (cf. d’Alessandro, 2006, for a description of voice quality).

The facial expressions linked with these five attitudes vary, although very little specific information is shown for declaration (such a “lack” of information may well be typical of such a neutral expression). Courtesy politeness and sincerity politeness show a similar slight rising of the brow with a small movement up and down of the head. Arrogance and kyoshoku are much more marked: while expressing arrogance, the speaker turns his head to his left and raises his brow. For the kyoshoku attitude, the speaker makes a grimace mimicking suffering with a strong frown, wrinkling his nose, and shutting his eyes, and then makes a pronounced bowing.

2.2 Subjects
The 96 subjects of this experiment, all native Japanese speakers, are grouped into four groups of age level.

- 40 adults (AD: 28 females; mean age of 21.6)
- 19 children attending the 4th grade classes (4th grade: 9 females; mean age of 9.5)
- 19 children attending the 2nd grade classes (2nd grade: 13 female; mean age of 7.4)
- 18 children attending the 1st grade classes (1st grade: 11 female; mean age of 6.1)

The adults group is seen as the reference of competent native speakers. The performance of each children groups of growing age will be compared to this adult group.

2.3 Experimental paradigm
All pairs, composed of two different attitudes amongst the 5, are presented to subjects in a random order. Pairs of stimuli are presented in different modalities: audio-only (A), visual-only (V) and audio-visual (AV). The presentation order of these modalities is balanced amongst subjects: half of them took the A modality first, then V and finally AV, while the second half took V first, then A, followed by AV. For each pair, subjects have to judge the perceived difference between the two performances, on a 1 to 9 scale. A pair is only presented once to a subject.

3. Results
Results are analysed following the methodology described by Romney et al. (2000). Details may be found there and in references herein; more specific references will be made for specific points. Statistical methods are tuned for a measure of similarity, thus the obtained judgements of distances are expressed as similarity scores by taking 10 minus the obtained distance score for each pair of different attitudes, and a 10 for the pairs of identical attitudes (not presented to listeners). A 5x5 similarity matrix is obtained for each subject in each modality – with a row containing the similarity scores for an attitude toward each of the possible 5 attitudes. These
matrices are stacked for all subjects and modalities, giving a 1440x5 large matrix X (based on 5 attitudes x 3 modalities x 96 subjects).

3.1 Perceptual distribution of attitudes

A correspondence analysis (CA) is then applied to X, with the row scores standardized using the Kumbasar et al. (1994) method, in order to neutralize possible differences in the use of the judgement scale by subjects. The results of the CA give a cloud of points for each attitude as perceived by each subject in each modality, in the 4-dimensional space of the CA. The first two dimensions of the CA explain more than 70% of the observed variance, thus in subsequent graphs only the first two dimensions will be used to display that attitudes’ distributions. Individual points obtained for each subject are regrouped according to the 4 age groups, and the mean position of attitudes is displayed, surrounded by a 97.5% confidence interval ellipsis, for each group and each modality. Figure 1 presents the observed dispersion of attitudes in each modality, comparing differences between age groups.

The main tendency observed on these graphs is the overall similarity of the attitudes’ distributions over age groups – and to a lesser degree, over modalities. The 2 politeness expressions are on the top right corner, close to declaration, while arrogance is situated on the very left part of the plots. These four attitudes are more or less linearly distributed on an oblique dimension going from arrogance to politeness. This dimension, and the placement of attitudes on it across modalities and age groups, is close to the “impolite-polite” dimension observed by Shochi et al. (2009b). By contrast, the kyoshuku expression is situated on the bottom right corner (in all modalities), not at all on the same “impolite-polite” dimension. However, if we consider the data from the viewpoint of a one-dimensional paradigm, then kyoshuku becomes situated on the same orthogonal line somewhere between polite and impolite, giving this attitude a position close to those obtained by Shochi et al. (2009b). This result confirms the similarity of the tasks performed by subjects in this experiment and in the preceding one. Moreover, it shows that in an open evaluation test as this one, kyoshuku is clearly differentiated by all listeners from other expressions of politeness.

A detailed observation still shows clear differences, between modalities as well as between age groups. The AV modality takes up the wider space, while the audio-only one defines a more restrained one – but with clear distinctions between each attitude (for adults). The visual-only modality, which occupies a space quiet similar to the AV one, only makes differences between three sets of attitudes: kyoshuku, arrogance, and a cluster grouping declaration with the two politeness expressions. Differences between age groups show a progressive extension of the space taken up by the 5 attitudes, from the more limited one with the 1st grade group, expanding with age toward that of the adult’s.

Figure 1: position of each attitude in the 2 first dimensions of the CA, plotted separately for each modality (from the top: A, V, AV)
Figure 2: Coloured points represent the mean placement for the four age groups, all modalities averaged; grey points represent the mean placement for the three modalities (averaged for age). Ellipses correspond to the 97.5% confidence limit from the means.

### 3.2 Quantification of observed variations

The differences subjectively described above may be quantified to obtain a measure of the differences between groups. Romney et al. (2000) propose to use the set of Euclidean distances between points representing the 5 attitudes in the 4 dimensions of the CA to compare the shapes of perceptive structures obtained for each subject (i.e. the shape of the distribution of the 5 attitudes obtained by the CA). The distances between each pair of attitudes compose a vector. The vectors obtained from each subject and modality are then compared via a correlation measure (cf. Rao & Suryawanshi, 1996), resulting in a 288x288 correlation matrix (96 subjects x 3 modalities). The square root of these correlations is used as a measure of the “shared knowledge” between two subjects, or two groups of subjects, by taking the mean of square root correlations (details on this point can be found in Romney et al., 2000). A principal component analysis (PCA) was run on this correlation matrix, in order to observe the differences of shape captured by the paradigm, across groups of modalities and groups of ages.

Figure 2 shows the results of the PCA: the place of subjects in the PCA represents the similarity of their perceptual dispersion of attitudes. These positions are averaged either by groups of the same modality (the grey dots indicating the A, V and AV modalities, surrounded by 97.5% confidence ellipses), or averaged by age groups (the coloured dots indicating the 1st, 2nd, 4th grades and adult groups, surrounded by 97.5% confidence ellipses). It is clear from this figure that the audio and visual modalities constitute the factors introducing most variance in the subjects’ answers, and that the audio-visual presentation, the least, suggesting that perception of attitudes is enhanced by information from both modalities. Moreover, the progressive evolution with age of these perceptive structures toward the adults’ reference is clear – 1st grade subjects showing a maximally different perceptual shape from that of the adults’.

The mean square roots of observed correlations within- and between-age groups (cf. table 1) indicate the average shared knowledge among subjects of that category. The progressive increase of this shared knowledge with age is a clear indication of the acquisition by native children of the proposed attitudinal expressions, from about 6 to 10 years old. This result reinforces the idea of a progressive construction by children of cultural conceptual spaces with age – particularly in the case of such prosodic social affects.

<table>
<thead>
<tr>
<th></th>
<th>1st</th>
<th>2nd</th>
<th>4th</th>
<th>Adults</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>.55</td>
<td>.44</td>
<td>.45</td>
<td>.60</td>
</tr>
<tr>
<td>2nd</td>
<td>.44</td>
<td>.58</td>
<td>.55</td>
<td>.59</td>
</tr>
<tr>
<td>4th</td>
<td>.45</td>
<td>.55</td>
<td>.63</td>
<td>.65</td>
</tr>
<tr>
<td>Adults</td>
<td>.60</td>
<td>.59</td>
<td>.65</td>
<td>.71</td>
</tr>
</tbody>
</table>

Table 1: square roots of the correlations obtained from age groups; within group correlations (in bold) and between groups correlations are given.
4. Conclusions

Since results obtained from this perception test corroborate previous findings, it can be assumed that they validate the use of such a pair comparison paradigm for at least three purposes: testing young children with none or few reading skills, investigating the multidimensional distribution of prosodic expressive performances, and measuring the evolution with age of children’s understanding of their mother language’s social affective strategies.

Such kinds of perception tests are still difficult to run with the youngest group of 1st grade children. Additional information on the children’s understanding of the stimuli is also important. For example, Shochi et al. (2009b) asked subjects questions about the interlocutor, specifically, what kind of interlocutor may be addressed in that way? Informal discussions with subjects during this test gave interesting answers: 1st grade children described the kyoshoku expression as “crying”, while 2nd grade children perceived it as “suffering” – a description closer to Sadanobu’s (2004) definition. Similar, more accurate descriptions by older children were also observed for arrogance, described by 1st grade children as a “sleeping person”, while 2nd grade children thought the speaker was “sulking”. Such informal descriptions may also be an interesting path to follow in order to acquire a deeper understanding of children’s developing capabilities in their social relationships. The main drawback of such experiments is their complexity.

5. Acknowledgements

This work was supported by ANR JCJC 2010 PADE project. It was performed in collaboration with the Yoji school, Japan, and the Showa Music University, Japan.

6. References


