Free Labeling of Audio-visual Attitudinal Expressions in Cantonese

Tan Lee1, Matthew K. H. Ma1, Albert Rilliard2, Hansjörg Mixdorff3, Angelika Hönemann3,4

1 The Chinese University of Hong Kong, Hong Kong
2 LIMSI, CNRS, Université Paris-Saclay, France & Federal University of Rio de Janeiro, Brazil
3 Department of Computer Science and Media, Beuth University Berlin, Germany
4 Faculty of Linguistics & Literary Studies, University of Bielefeld, Germany

tanlee@ee.cuhk.edu.hk, khma@ee.cuhk.edu.hk; Albert.Rilliard@limsi.fr
mixdorff@bht-berlin.de; ahoenemann@techfak.uni-bielefeld.de

Abstract

This paper reports results from a free labeling experiment employing short audio-visual utterances of Cantonese produced with varying attitudinal expressions. It is part of a series of such experiments with a cross-language setting between German and Cantonese. Cantonese-speaking perceivers were asked to specify a single word that best described these stimuli, which were presented in audio-visual, audio-only, and video-only modalities. The resulting terms were analyzed with respect to the emotional dimensions of valence, activation and dominance, as well as the linguistic dimension of assertion/interrogation. The analysis results are compared with the outcomes from similar experiments employing German stimuli with Cantonese perceivers, as well as German perceivers assessing both German and Cantonese stimuli. It is found that Cantonese perceivers judge the Cantonese stimuli as more activated than German perceives do. The valence judgments agree relatively well, however, “polite” stimuli were judged less positively by Cantonese perceivers. Generally speaking, valence judgments are mostly influenced by the stimuli whereas activation and dominance judgments depend more on the perceiver group.

Index Terms: social attitudes, auditory-visual speech, free labeling

1. Introduction

In speech communication we often express our attitudes regarding propositions being made by the interlocutor. Hence it is an important skill which we employ seamlessly for controlling the dialogue and achieving the goals of conversation. Dialog partners sharing the same language and/or culture are at an advantage as they make use of codes belonging to the similar behavioral and value systems. In contrast, interactions between partners from different cultures may lead to misunderstandings and wrong interpretations of attitudinal expressions. Earlier cross-language studies [1] showed similarities across languages, such as the use of low pitch to signal dominance [2], and culture-specific uses of prosodic parameters [3]. The study of verbal and non-verbal social affects in a cross-cultural context has attracted great research interests, in order to back and foster increasingly frequent cultural exchange in a globalized world.

The present study is part of a series of cross-lingual studies on multi-modal attitudinal speech following the framework of [4]. In [4], Rilliard et al. presented a paradigm of eliciting, recording and evaluating spoken utterances that express different social affects. A total of 16 types of attitudes, including arrogance, politeness, doubt and irritation, are defined with designated communication goals and social contexts. Based on this paradigm, an audio-visual corpus of German attitudinal expressions was developed [5].

Free-labeling experiments employing native German subjects were carried out on the corpus [6]. The subjects were asked to freely specify one single word to describe the perceived social attitude for each presented stimulus, which could be audio-visual, audio-only or video-only. The collected response words were normalized and analyzed in the three-dimensional emotional space. It was shown that the attitudes essentially cluster in several groups, the members of which share similar properties. Attitudes such as admiration and sincerity are found on the positive side of the spectrum whereas authority, contempt, arrogance, irritation and irony are located on the negative side.

The above mono-lingual study has been extended to a few cross-lingual and cross-cultural studies that involve native Cantonese presenters and perceivers. Cantonese is one of the major Chinese dialects spoken by about 80 million people in Southern China and many overseas Chinese communities. It is known of having significant variation of which share similar properties.

In [7], we described a Cantonese attitudinal speech corpus that parallels the German one and used it to carry out a cross-language free-labeling experiment with German-speaking perceivers. Attitudinal stimuli in the Cantonese corpus (cross-language perception) were perceived to be generally less activated and dominant than those in the German one (mono-lingual perception) [6]. This may be attributed to a more collective-oriented behavior in Asian subjects, who tend to avoid overt voicing of negative emotions and employ a special register of politeness. The results of free-labeling of German corpus by Cantonese perceivers were reported in [8]. It was shown that the judgments of Cantonese and German perceivers regarding valence and dominance are strongly correlated, despite the fact that their locations in the emotional space are slightly different.

In this paper, we present and analyze the results of an intra-Cantonese free-labeling experiment, in comparison to the findings from the cross-language studies. Specifically, we wish to examine how Cantonese perceivers judge Cantonese stimuli and whether the differences we found in the German group – that is, less activation and dominance assigned to the
Cantonese stimuli as opposed to German stimuli are replicated. By applying cluster analysis we attempt to determine whether stimulus grouping is more dependent on the perceivers' system of reference or the stimulus proper.

2. Stimuli and Experiment Procedures

The preparation of Cantonese attitudinal expressions followed the protocol and procedures developed and described in [3] and [4]. Ten native speakers of Cantonese (four males and six females) participated in the recording. All of them were students at the Chinese University of Hong Kong (CUHK), and aged between 18 and 24. Audio-visual recordings were made in a sound-treated room. The 16 attitudes were portrayed using two target phrases: (1) 香蕉 (Transcription: hoeng1 ziu1; English translation: “a banana”), and (2) 玛丽跳舞 (Transcription: Mary tiu3 gan2 mou5; English translation: “Mary was dancing”). As a result, there were 320 audio-visual tokens recorded from the ten speakers. In order to select the best exemplar for each attitude, four native Cantonese perceivers were asked to perform perceptual rating on each recorded token in the full auditory-visual modality. The rating indicates the credibility of an intended attitude on a scale from 1 to 9, with a score of 9 meaning “very convincing” and 1 “completely unconvincing”.

The free-labeling experiment was carried out with the target phrase “玛丽 跳舞。Mary was dancing”, so that the results could be comparable to the preceding investigation with German perceivers [7]. For each of the 16 attitudes, there are 10 recorded tokens, from which we selected 6 with the best perceptual ratings. This leads to a total of 96 stimuli, which could be used for perceptual test in three different modalities: audio-visual (AV), audio-only (AU) and video-only (VI). We created two different sets of test stimuli, each containing 48 AV, 24 AU and 24 VI stimuli. While both sets of stimuli cover all of the 96 selected tokens, each token appears exactly once in a specific modality in each set.

Thirty native Cantonese-speaking subjects participated in the free-labeling experiment. They include 15 males and 15 females, all being undergraduate or postgraduate students (aged 18 – 25) at CUHK. Each subject was asked to rate 96 attitudes, all being the full set of the 117. Semantically equivalent words were collapsed, e.g., 不快, 不高兴, 不爽, 心情不好, 發脾气. An explicit emotion was always normalized to 不高兴 (unhappy); 憂慮, 擔憂, 擔心 were normalized to 擔心 (worry).

The term 平和 (gentle) was the most frequently used label (N=284), followed by 陳述 (statement) (N=234), 不耐煩 (impatient) (N=227), and 高興 (glad) (N=224). For the VI condition, the same group of most common items are observed, with their order slightly changed, i.e., 高興 (glad) ranks the first (N=103) while 陳述 (statement) takes the fourth place (N=57). For the AU condition, 不耐煩 (impatient) and 高興 (glad) are not among the most frequent labels. The top four responses are 陳述 (statement) (N=64), 疑惑 (wondering) (N=55), 平和 (gentle) (N=55) and 疑問 (doubt) (N=49).

Table 1: Sixteen attitudes and respective abbreviations, Positions of sixteen attitudes in the emotional space for Cantonese AV stimuli assessed by HK participants.

<table>
<thead>
<tr>
<th>attitude</th>
<th>abbrevation</th>
<th>valence</th>
<th>activation</th>
<th>dominance</th>
</tr>
</thead>
<tbody>
<tr>
<td>admiration</td>
<td>ADMI</td>
<td>.59</td>
<td>.71</td>
<td>.01</td>
</tr>
<tr>
<td>arrogance</td>
<td>ARRO</td>
<td>-.58</td>
<td>.53</td>
<td>.27</td>
</tr>
<tr>
<td>authority</td>
<td>AUTH</td>
<td>-.48</td>
<td>.51</td>
<td>.21</td>
</tr>
<tr>
<td>contempt</td>
<td>CONT</td>
<td>-.71</td>
<td>.40</td>
<td>.28</td>
</tr>
<tr>
<td>neutral statement</td>
<td>DECL</td>
<td>.02</td>
<td>.46</td>
<td>.02</td>
</tr>
<tr>
<td>doubt</td>
<td>DOUB</td>
<td>-.34</td>
<td>-.14</td>
<td>-.48</td>
</tr>
<tr>
<td>irony</td>
<td>IRON</td>
<td>.06</td>
<td>.44</td>
<td>.09</td>
</tr>
<tr>
<td>irritation</td>
<td>IRRI</td>
<td>-.69</td>
<td>.50</td>
<td>.33</td>
</tr>
<tr>
<td>obviousness</td>
<td>OBVI</td>
<td>-.38</td>
<td>.48</td>
<td>.18</td>
</tr>
<tr>
<td>politeness</td>
<td>POLI</td>
<td>.13</td>
<td>.48</td>
<td>-.01</td>
</tr>
<tr>
<td>neutral question</td>
<td>QUES</td>
<td>-.17</td>
<td>-.59</td>
<td>-.73</td>
</tr>
<tr>
<td>seductiveness</td>
<td>SEDU</td>
<td>-.10</td>
<td>.33</td>
<td>-.11</td>
</tr>
<tr>
<td>sincerity</td>
<td>SINC</td>
<td>.28</td>
<td>.54</td>
<td>-.01</td>
</tr>
<tr>
<td>surprise</td>
<td>SURP</td>
<td>-.17</td>
<td>-.64</td>
<td>-.72</td>
</tr>
<tr>
<td>uncertainty</td>
<td>UNCE</td>
<td>-.14</td>
<td>-.41</td>
<td>-.61</td>
</tr>
<tr>
<td>walking-on-eggs</td>
<td>WOEG</td>
<td>-.37</td>
<td>-.06</td>
<td>-.52</td>
</tr>
</tbody>
</table>

The normalized response terms were classified according to the scheme developed by [9] and [10]. The valence, activation and dominance levels of each normalized label were determined by the experimenters, on a three-point scale, ranging from negative/weak (-1), neutral (0) up to positive/strong (+1). In addition, a differentiation of assertion versus interrogation was marked. By this kind of semantic classification we are able to assess the emotional and linguistic weight of each term with respect to its frequency for a given attitude, hence abstracting from the original term.
4. Results

4.1. Emotional space

By averaging over the assigned values of valence, activation and dominance levels for all normalized response terms to the stimuli pertaining to a certain type of attitude, we obtain the coordinates of this attitude’s location in a three-dimensional emotional space. Table 1 lists the average values of all 16 attitudes in the AV case. It can be seen, for instance, that ADMI was judged most positively in valence, whereas CONT is perceived more negatively than AUTH, and DECL and POLI are found to be neutral connotations.

4.2. Inter-language results

At this point it is interesting to compare the results of the present study with the mean judgments of German perceivers on the same stimuli as reported in [7]. In terms of valence, the mean judgments of the two groups on the Cantonese stimuli show a very high correlation of $r=.89$ ($p < .001$).

Cantonese-speaking perceivers tend to assign a higher degree of activation to Cantonese stimuli in 12 attitudinal classes than their German-speaking counterparts (mean value of .44 vs .17). The four attitudes that were judged less activated by Cantonese perceivers are QUES, SURP, DOUB and UNCE (mean value of -.45 vs .22). This is probably related to the fact that German perceivers interpret the interrogative character of these attitudes as a higher degree of activation. As a consequence, the correlation on the mean activation judgments between groups is not significant. As for the dominance level, there is a moderate correlation between mean judgments by the Cantonese and the German group on the Cantonese stimuli ($r=.70$, $p < .003$).

Figure 1 shows a 3D graph in which the positions of some of the attitudes judged by Cantonese perceivers and German perceivers are displayed for comparison. In [7] we found that on the whole German perceivers judged the Cantonese expressions slightly more negatively than the German perceivers (mean valence of -.22 vs -.16), and also less activated or dominant (mean activation of .19 vs .40, and mean dominance .06 vs .14).

4.3. Clustering of Attitudes

The labels collected on AV stimuli from both groups of perceivers were organized in a matrix counting the number of occurrences for each attitude in each modality for all perceivers. The matrix has 48 rows (16 attitudes in 3 modalities) x (91 + 104) columns (91 / 104 labels used in the AV condition by Cantonese and German perceivers). The matrix was subject to a Multiple Factor Analysis (MFA), using the FactoMineR package [11]. Results obtained on the mono-modal stimuli and for the four semantic scales were mapped to three dimensions of the factor analysis. The repartition of the 16 expressions on these three dimensions was used to group them into coherent clusters, using a hierarchical clustering algorithm. The result of clustering for the AV modality can be depicted by a tree structure as shown in Figure 2. Stimuli presented in other modalities were later associated to a cluster according to their distance from each cluster’s center, in the multidimensional space of the MFA. This tree could be cut according to a criterion of inertia gain, giving four clusters. Hereafter is a description of the expressions pertaining to each of these clusters, and their “definitions” in terms of labels and semantic categories attributed by perceivers of both groups.

Cluster 1 is composed of ADMI, DECL, IRON, POLI, SEDU and SINC, and labeled by both groups as positive assertion without showing dominance and with neutral activation; Cantonese perceivers also use features of positive assertion. The labels that describe the cluster are, in the order of decreasing strength of association with the cluster, for Cantonese perceivers “高興 (glad), 稱道 (statement), 確定 (confirm), 平和 (gentle), 平靜 (calm), 正常 (normal), 興奮 (excited), 嘲笑 (fling in teeth), 友善 (friendly), 平常 (normal), 取笑 (make fun of)”, and for German perceivers “freundlich (friendly), fröhlich (happy), glücklich (happy), erfreut (pleased), aufrichtig (sincere), zustimmend (approving), anmisiert (amused), aufmunternd (encouraging), liebevoll (affectionate)”. Only a few mono-modal audio stimuli were labeled in a similar way (IRON and POLI), while all the visual stimuli were included into the same cluster, plus the visual performance of WOEG (showing it carries positive traits in its visual performance).
Cluster 2 is based on ARRO, AUTH, IRRI, OBVI, described by both perceiver groups as negative assertion of dominance—Cantonese and German groups diverge as for their activation level—Cantonese group rating them as highly activated, while German indicating low activation. The labels used to describe the cluster are, for Cantonese perceivers “不耐煩 (impatient), 不高興 (unhappy), 煩惱 (annoyed), 憤怒 (anger), 煩躁 (tревetful), 不屑 (disdain), 冷漠 (indifferent), 失望 (disappointed), 強調 (emphasizing), 反感 (disgust)”, and for German perceivers “generell (annoyed), neutral (neutral), wütend (angry), gelangweilt (bored), arrogant (arrogant), beleidigt (offended), mahnend (admonishing), erklärend (explaining)”. The corresponding mono-modal stimuli are all classified into the same cluster except OBVI. The audio expressions of ADMI, SEDU and SINC are also found in this cluster, plus the visual performance of SURP.

Cluster 3 contains UNCE and WOEG, which can be described as submissive expressions with low activation; Cantonese perceivers also use an interrogative feature and a neutral valence, while German assign a negative valence. The labels used to describe the cluster are, for Cantonese perceivers “不肯定 (not sure), 猶豫 (hesitate), 沒有自信 (lack self-confidence), 疑惑 (wonderingly), 悲傷 (sad), 害怕 (fear), 說謊 (lie), 坦白 (frank), 為難 (struggling), 不好意思 (sorry)”, and for German perceivers “unsicher (uncertain), nachdenklich (pensive), traurig (sad), zweifeln (doubtful), verwirrt (confused), zögernd (hesitant), zurückhaltend (reserved), abfällig (derogatory), vschlagend (suggestive), müde (tired)”. For mono-modal conditions, the two audio stimuli are classified into the same cluster, while only the visual UNCE goes to it (WOEG being inside cluster 1). Audio performances of OBVI are also grouped in cluster 3.

Cluster 4 is composed of the expressions of DOUB, QUES and SURP, which are described as submissive and interrogatives. Cantonese perceives suggested a low activation. The labels used to describe the cluster are, for Cantonese perceivers “疑問 (doubt), 疑惑 (wonderingly), 懷疑 (doubt), 驚訝 (surprised), 錯愕 (stunned), 不相信 (do not believe), 奇怪 (feel strange)”, and for German perceivers “fragend (questioning), erstaunt (astonished), zweifeln (doubting), überrascht (surprised), erschrocken (frightened), interessiert (interested), vorwurfsvoll (reproachful), entsetzt (appalled), verärgert (angry), wütend (angry)”. Mono-modal conditions group the same expressions except that the visual SURP goes to cluster 2.

5. Discussion and Conclusions

This study examined how Cantonese-speaking perceivers interpret Cantonese stimuli with attitudinal connotations. It is the last in a series of experiments in which we had German- and Cantonese-speaking presenters enact sixteen different attitudes following the same protocol. Subsequently we had perceivers from both languages perform a free labeling task on stimuli in the two languages. Stimulus sets contained audio-only, video-only and full AV presentations.

By clustering analysis in the emotional space, four groups of attitudes were identified. They are interpreted as (1) positive assertions, (2) negative and dominant assertions (imposition on the hearer), (3) submissive expressions with low activation, (4) interrogations with submissive and negative features. These groups are perceived as such, under the same experimental conditions, by both Cantonese and German perceivers. The main feature that separates the prosodic expressions appears to be linked to the sentence’s mode (assertion vs. interrogation). This is very clear for the group of native perceivers: German subjects do label cluster 3 as submissive and low activation only, as they unlikely make a linguistic interpretation of the content. The second factor to separate the clusters is valence, opposing clusters 1 and 2—the second one having a strong imposition feature (+ dominant).

The two groups of perceivers diverge mostly in their attribution of the activation feature, which is even opposed for cluster 2; this agrees with the findings of correlation analysis, where activation receives the lowest scores. The labeling of cluster 3 diverges: correspondence analysis for each group individually show that Cantonese perceivers group UNCE and WOEG with interrogations (QUES, DOUB, SURP), while German find them more assertive and having primarily a negative valence (grouped with ARRO and CONT).

![Figure 2 dendrogram representing the distance between the 16 attitudes in the AV modality, as observed on the first 3 dimensions of the MFA, and the cutting level (red line).](image)

On average Cantonese perceivers perceive Cantonese stimuli as less activated and dominant than German perceivers do (mean activation .22 vs .37 (-.11 vs -.04)), and also as slightly less positive (mean valence of -.19 vs -.12). In terms of valence, the mean judgments of the two groups on Cantonese stimuli show a very high correlation of r=.89 (p < .001). It is even higher when compared with the judgments on German stimuli (r=.91, p < .001) [8]. As for dominance level, there is a moderate correlation between mean judgments by Cantonese and German perceivers on the Cantonese stimuli (r=.70, p < .003). It is lower than for judgments on the German stimuli (r=.84, p < .001). The modality of presentation seems to be related to the types of attitude that are clustered: visual information being more centered on interaction, and audio being more centered on linguistic opposition. This shall be reproduced on a wider and more systematically balanced data.

6. Acknowledgements

This work was supported by the DFG (Deutsche Forschungsgemeinschaft) grant Mi 625-27 funding a visit of Hansjörg Mixdorff to the Chinese University of Hong Kong, and partially supported by the General Research Fund of Hong Kong Research Grants Council (Ref No.: 14227216).
7. References


