Interspeech 2018
2-6 September 2018, Hyderabad

Cross-cultural (A)symmetries in Audio-visual Attitude Perception

Hansjörg Mixdorff[1], Albert Rilliard[2], Tan Lee[3], Matthew K. H. Ma[3], Angelika Hönemann[1,4]

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

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

Abstract

This paper evaluates results from a cross-cultural and cross-language experiment series employing short audio-visual utterances produced with varying attitudinal expressions. German and Cantonese-speaking participants freely labeled such utterances in the two languages and assigned to each stimulus a verbal label. Based on the results of the four experiments we were able to establish to what degree the attitudinal frames of reference of the two groups overlap and how they differ. Verbal labels were assessed regarding their emotional content in terms of valence, activation and dominance, and for the linguistic opposition between assertive and interrogative speech act, and hence permit to abstract from the language of the rater and ultimately even abstract from the attitudinal categories used when eliciting the stimuli. Instead we regard each utterance as a data-point in the emotional space. We found that the judgments of the two rater groups agree well with respect to the valence of attitudinal expressions and diverge most as to the perceived activation of the stimulus presenter. Cantonese speaking participants seem to mirror Germans' ratings of German stimuli better than vice versa, which suggests an interesting asymmetry of attitudinal perception. As for the modality of presentation, the audio channel primarily transmits linguistically relevant information regarding the opposition of assertion and interrogation while the visual information signals the emotional content.

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

1. Introduction

In a dialog situation we constantly monitor our interlocutor's behavior and aim to assess his/her intention and attitude while adjusting our own contribution. Hence expressing our attitudes clearly and decoding those of our dialog partner helps us control the dialog and achieve the goals of conversation. If we share the same language and/or cultural background we employ codes belonging to a similar behavioral and value system. However, interactions between partners from different cultures may be compromised by misunderstandings, that is, wrong interpretations of attitudinal expressions. Earlier cross-language studies showed similarities between languages [1][2], but also important culture-specific differences [3].

The present article evaluates 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, audio-visual corpora of attitudinal expressions in German [5] and Cantonese [7] were developed. There are several interesting contrasts between the two languages, such as that between a non-tonal and a tonal language, and a stress-timed and a syllable-timed language, but also contrasts between the cultures, one being Central European and potentially more individualistic, the other one East-Asian and presumably (still) more collective-oriented. Furthermore, to all subjects in our studies the other language is not known.

Subsequently, free-labeling experiments with participants from the two languages were carried out on both of the corpora [6,[7][8][9]. 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 of valence, activation and dominance [10]. Hence we were able to abstract from the raters' languages and compare the responses directly. It was shown that the attitudes essentially cluster in several groups, the members of which share similar properties. There is considerable overlap between different attitudes [8]. For instance, expressions portraying admiration and sincerity are found on the positive side of the spectrum whereas authority, contempt, arrogance, irritation are located on the negative side.

However, we also found that German raters perceived Cantonese performers to be generally less activated and dominant than the German ones [6]. This may be accounted for by the aforementioned more collective-oriented behavior of Asian subjects, who tend to avoid overt voicing of negative emotions and employ a special register of politeness. In contrast, Cantonese-speaking perceivers tend to assign a higher degree of activation to Cantonese stimuli in 12 attitudinal classes than their German-speaking counterparts [9] and therefore seem to be able to decode this information even from the seemingly more restrained presentations. The other four attitudes that were judged more activated by the German subjects have an interrogative mode, indicating that it might have been mistaken for a higher degree of involvement.

10.21437/Interspeech.2018-1373

The results of free-labeling of the 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 the current paper, we summarize and evaluate the findings of our experiment series and draw more general conclusions regarding the inter-language and inter-cultural implications of our findings. We also aim to further abstract from our original paradigm of 16 attitudes and rather treat each stimulus not solely as a result of our elicitation protocol, but rather take it at the face value that our raters eventually assigned it and as a data point in the three-dimensional emotional space. By applying cluster analysis to the complete line-up of results we attempt to determine whether stimulus grouping is more dependent on the perceivers' system of reference or the stimulus proper. We had a total of 20 German- and 10 Cantonese-speaking subjects who produced the stimuli. The six best rated exemplars for each language group per attitude were employed in the perceptual experiments and judged by at least 30 perceivers.

2. Stimuli and Experiment Procedures

Since the design for eliciting the attitudinal expressions has been described in detail in our earlier publications we only provide a condensed summary.

The 16 attitudes are elicited in a dialog between the presenter and the experimenter. The portrayal of each type of attitude is prepared by a dialog meant to immerse the presenter in a suitable communicative situation. Then a short exchange leads up to the target utterances either being "a banana" or "Mary was dancing" in the respective language of the presenter. Short video-clips of the target utterances are extracted and serve as stimuli for the ensuing perceptual tests, after being rated and selected for quality [5]. Due to the better matching of syllable numbers we used the phrase "a banana" (*eine Banane*) for German and "Mary was dancing" (*Mary* 跳緊舞) for Cantonese.

For each of the 16 attitudes the six best-rated samples were selected, yielding 96 auditory-visual (AV) tokens. These were augmented by a subset of the 96 AV stimuli in reduced modality: audio-only (AU) and silent video (VI). The stimuli were presented one by one, and the subject was asked to use a single word to describe the attitude he/she could perceive. The subject was allowed to replay a stimulus as often as necessary. The valid word responses were collected and normalized as described in [6]. In order to abstract from the particular word and language, we classified each response term according to the scheme developed by [9] and [10]. For each term the emotional valence, activation and dominance levels were assigned values of either -1, 0 or +1, 0 being neutral. In addition, a differentiation of assertion versus interrogation was marked. This kind of semantic classification enables us to abstract from the original response terms.

3. Results

3.1. Emotional space

By averaging over the assigned values of valence, activation and dominance 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 the three-dimensional emotional space.

Table 1 lists the average values of all 16 attitudes in the AV case, pooling the results from all four experiments.

It can be seen, for instance, that ADMI was judged most positively in valence, whereas CONT is perceived more negatively than AUTH, and DECL found to have a neutral connotation. ADMI and IRRI are the attitudes showing the highest degree of activation of the talker, whereas IRRI and UNCE mark the extreme ends of the dominance parameter.

Table 1: Sixteen attitudes and respective abbreviations, Positions of sixteen attitudes in the emotional space, pooled results from all four experiments.

attitude	abbrev-iation	valence	activat-ion	domin-ance
admiration	ADMI	.5597	.5708	-.0088
arrogance	ARRO	-.3951	.3642	.2340
authority	AUTH	-.4361	.4141	.3062
contempt	CONT	-.6793	.3697	.2695
neutral statement	DECL	-.0022	.2892	.0353
doubt	DOUB	-.4181	.1401	-.2759
irony	IRON	.1231	.4330	.0352
irritation	IRRI	-.5991	.5441	.3855
obviousness	OBVI	-.2753	.3656	.1247
politeness	POLI	.2009	.3879	.1075
neutral question	QUES	-.2635	-.0517	-.2266
seductiveness	SEDU	,2394	,4539	,0224
sincerity	SINC	,2457	,4668	,0786
surprise	SURP	-,1345	,0318	-,2836
uncertainty	UNCE	-,3981	-,1343	-,3717
walking-on-eggs	WOEG	-,4471	,0120	-,2524

3.2. Inter-language results

We will now examine the similarities and differences between the two rater groups. To that end we step away from the originally intended attitudes of the audio-visual stimuli and rather examine the three emotional dimensions associated with how the perceivers interpreted those stimuli. We calculated means and standard deviations of valence, activation and dominance for each stimulus in our experiments as a function of the rater group. Figure 1 shows three-dimensional displays of each stimulus, AV, AU, and VI in the reference frames of the German (left) and Cantonese (right) speaking judges. The stimulus language is coded in color, German in green, and Cantonese in blue. As can be seen, the values for the German raters appear to be much more scattered and occupy a larger area than those of the Cantonese. It also seems as if the unknown language stimuli are concentrated more than those of the native language. We verify this observation by calculating the mean Euclidian distance of each stimulus' ratings from the

origin of the three-dimensional space as a function of rater and stimulus language:

$$distance = \sqrt{valence^2 + activation^2 + dominance^2}$$

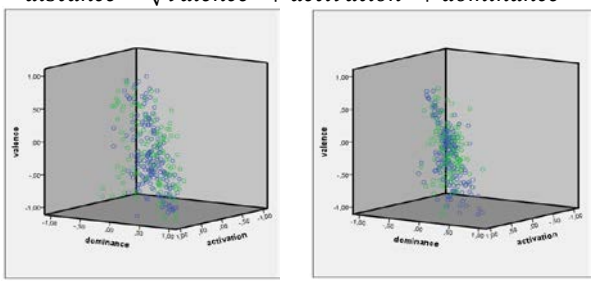

Figure 1: Mean values for all stimuli in the 3D emotional space for German (left) and Cantonese (right) speaking perceivers. German stimuli displayed in green, Cantonese in blue.

This distance indicates how far an expression of attitude is perceived as diverging from neutral.

The mean distance for Germans rating German and Cantonese stimuli is 0.76 (s.d.=0.33) and 0.70 (s.d.=0.34), respectively, whereas it is 0.76 (s.d.=0.29) for Cantonese speakers rating Cantonese and 0.69 (s.d.=0.25) rating German. This suggests that each group employs a larger range of ratings for stimuli from their own language. The higher S.D. for the Germans confirms the wider scattering of data points for this group. The shapes of the scatter plots also indicate that the three dimensions are not independent, at least not for the stimuli that we elicited. Correlation analysis shows a Pearson's r=0.613 between activation and dominance ($p<0.001$) for the German group and r=0.863 for the Cantonese speakers. In contrast, the correlation between valence and activation as well as dominance is not significant.

We now look at the intra-group agreement regarding the ratings for individual stimuli. To this end we calculated the standard deviations of judgments of the members on each stimulus and correlate them to *distance* for that stimulus. There is a moderate negative correlation of -0.563 (Pearson's r, $p < 0.001$) for valence and a weak negative correlation r=-0.296 ($p < 0.001$) for activation for the German raters. For the Cantonese speakers the negative correlation for valence is only -0.155 (Pearson's r, $p < 0.001$) and r=-0.368 ($p < 0.001$) for activation. This indicates that "more extreme" expressions are easier to judge. Now we turn to the agreement between German and Cantonese speaking raters regarding the same stimulus. Figure 2 shows scatter plots for valence, activation and dominance judgments on the AV stimuli.

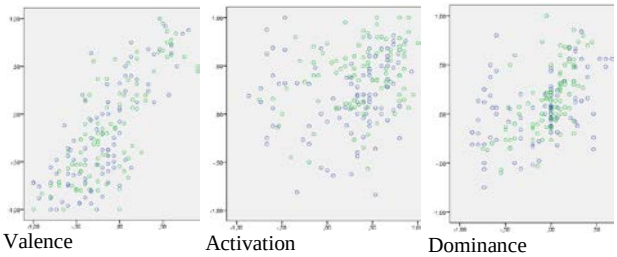

Figure 2: Scatter plots for valence (left), activation (center) and dominance (right). Ratings by Germans on y-axis, Cantonese on the x-axis. German stimuli displayed in green, Cantonese in blue.

We observe a fairly good agreement between the groups as to the valence of the stimuli (Pearson's r=0.773, $p < 0.001$), whereas the values are lower for dominance (r=0.487) and activation (r=0.248). As reported earlier, German raters find it hard to assess the activation levels of the Cantonese presenters [7].

When we subdivide the judgments into those on German and Cantonese stimuli we find an interesting divide. Cantonese listeners seem to be able to mirror the Germans' judgments on German stimuli better than vice versa.

*Table 2: Correlations (Pearson's r) between judgments of German and Cantonese speakers on the stimuli of the other language, as well as split correlations for the two groups on stimuli in their native language (** highly significant, $p < 0.001$).*

	stimulus language			
	German		Cantonese	
dimension	Cant. raters	split corr. Germans	Germ. raters	split corr. Cantonese
valence	0.776**	0.835**	0.768**	0.883**
activation	0.304**	0.674**	0.142 n.s.	0.855**
dominance	0.618**	0.689**	0.403**	0.901**

Table 2 displays the correlations between the German and the Cantonese rater groups on the stimuli of the other language in modality AV. Apparently the Cantonese speakers are better at judging the activation and dominance levels in the German stimuli. This means that while they are able to rate the more restrained Cantonese stimuli they are also to some extent capable of assessing the stimuli from the German group. This may be to do with the generally higher activation levels in the German stimuli - means of 0.394 and 0.375 as judged by German and Cantonese raters – as to those assigned to the Cantonese stimuli – means of 0.186 and 0.222 assigned by German and Cantonese raters, respectively. For better comparison we also calculated split correlations between two randomly selected halves of each rater group on stimuli from their native language (see also Table 2). The intra-group agreement for the Cantonese perceivers is considerably higher. Finally we look at the judgments for reduced modalities. As can be expected, the agreement on audio-only stimuli is lower than for audio-visual stimuli.

*Table 3: Correlations (Pearson's r) between judgments of German and Cantonese speakers on audio-only and silent video stimuli, ** highly significant, $p < 0.001$.*

	stimulus language	
dimension	audio-only	silent video
valence	0.428**	0.822**
activation	-0.069 n.s.	0.377**
dominance	0.288 ($p < 0.026$)	0.430**

Table 3 lists the correlations between the German and Cantonese perceivers on the entirety of audio-only (AU) and silent video (VI) stimuli. For AU, only the valence judgments are moderately correlated whereas the agreement on the VI stimuli in part is even higher than for the audio-visual stimuli. This suggests that listening to the audio can get in the way of the emotional judgment, especially when an unknown language is concerned.

3.3. Clustering of Attitudes

The number of times each original label (obtained during the free labeling task from each of the fours groups of listeners) was used for a stimulus in the bimodal presentations formed

four contingency tables that were analyzed together using a Multiple Factor Analysis (MFA)[13]. This cross-cultural analysis allows for the main dimensions of each table to be grouped according to the lines of the matrix (here, presenting the categories of stimuli). The distribution of expressions obtained that way was submitted to a hierarchical clustering, the output of which is shown in the dendrogram of figure 2. This process allowed defining six clusters (cf. fig. 2 for numbers) based on a criterion of inertia reduction. The description of the clusters in terms of occurrences of labels defining them by each listener group, and in each modality of presentation is described and analyzed hereafter.

The first dimension of the MFA is linked to the linguistic opposition assertion / interrogation: this distinction is major to define the interrogative clusters #1 and #2. These clusters are distinguished by their valence, activation and dominance – all these traits being negative in #2, and neutral in #1. The robustness of audio-visual perception is shown by the lack of confusion in bimodal presentations. The clusters #3 to #6 share an assertive mode: #3 is characterized by dominant and negative traits; the other three having or a neutral or a positive valence. Cluster #6 contains declaration and two assertions performed with courtesy. Cluster #5 is composed of two positive and activated expressions, while the cluster #4 has a positive valence but with a low activation.

Most confusions are observed in mono-modal conditions. Clusters #1 & #2 are best recognized via the audio modality in coherence with their linguistically-related interrogative function. Meanwhile, for audio-only judgments by German listeners presented with Cantonese, confusions are observed: the tonal variations may have had an impact on the perception of interrogative characteristics. Confusion of visual interrogatives arises with SURP being misclassified for its linguistic mode (i.e. described as an assertive), but always in a cluster with high activation (a characteristic of visual SURP), while its valence attribution may be language dependent as both groups presented with German stimuli do mix it with the positive cluster #5 while subjects presented with Cantonese mix it with negative #3. At the opposite of what happen for interrogative expressions, the dominant and negative valence of cluster #3 is better expressed through the visual modality: audio-only expressions show more confusion, and generally toward the neutral cluster #6, but for Cantonese obviousness, which is mixed with the interrogative #2 by both groups of listeners. Interpretations of seduction in monomodal presentations do also lead to confusion: the expression being mixed with either neutral assertions or the negative cluster, without regularities within modality or language group.

4. Discussion and Conclusions

This study concludes a series of German-Cantonese free labeling experiments of audio-visual attitudinal expressions. Comparison of perceptual judgments made by four groups in intra and inter-language situations showed that subjects were able to label the affective content which we then mapped onto the dimensions of valence, dominance and activation dimensions [14]. We found the highest agreement for valence, a primary communicative dimension that is cross-culturally expressed and perceived. Finding reliable acoustic correlates of valence is an interesting challenge, as the most reliable vocal cues found in e.g. [15] are linked to activation. Our silent video and audio-only results suggest that mostly facial cues are responsible for the valence ratings. Notably, a difference in the perceived activation between German and Cantonese productions was found and coherently annotated by listeners with both language backgrounds. In this context an important difference between the language groups is that Cantonese perceivers seem to be better at judging activation and dominance levels in German stimuli than vice versa. This indicates that on the one hand side they are able to pick up subtle signals in the more restrained presentations of their compatriots and on the other hand to some extent able to judge the portrayals of the Westerners. This discrepancy might be in part explained by regular exposure to Western people and culture living in a cosmopolitan environment like Hong Kong. In contrast, the German participants will probably not have been equally exposed to Asian culture. Our analysis also showed an interesting commonality between the two language groups: They seem to assign a larger area of the 3D perceptual emotional space to stimuli from their own language. This is complemented by the observation that 'more extreme' stimuli are rated more unanimously.

The free labeling approach also allowed us to observe another dimension of meaning, intricately linked to the prosodic dimension of speech: the opposition between assertive and interrogative speech acts. This forms the main dimension of the MFA analysis, and is shared across the four groups – i.e. for both German and Cantonese languages and listeners. The underlying (prosodic) information is primarily linked to the audio modality, a finding that supports its linguistic dimension [16][17] and opposes it to the more affective and socially related dimensions of activation, valence and dominance, mainly organized along the visual modality.

In conclusion it has to be stated that neither the rater groups nor the groups of presenters in our experiments are perfectly matched or balanced with respect to age groups and social backgrounds. The number of stimuli evaluated across all sixteen attitudes is also relatively small. However, our results suggest that the research framework is viable and facilitates interesting comparisons. In future work we will look more closely at the visual cues connected with each of the three emotional dimensions.

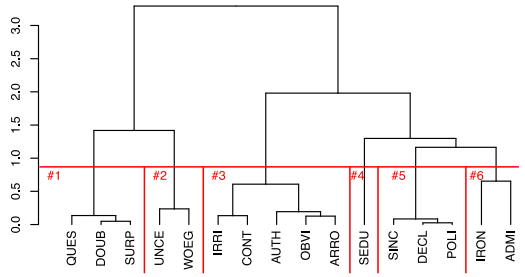

Figure 3: Dendrogram representing the distance between the 16 attitudes in the AV modality, as observed on the dimensions of the MFA, and the split level that separates expressions into six clusters (#numbers), combined for all four experiments.

5. 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).

6. References

[1] T. Shochi, A. Rilliard, V. Aubergé, and D. Erickson, “Intercultural perception of English. French and Japanese social affective prosody”, in S. Hancil (ed.) *The Role of Prosody in Affective Speech*. Linguistic Insights 97. Bern: Peter Lang. AG. Bern. 31-59. 2009.

[2] J.J. Ohala, “The frequency codes underlies the sound symbolic use of voice pitch”, in Hinton. L.. Nichols. J. & Ohala. J. J. (eds.). *Sound symbolism*. Cambridge University Press. Cambridge. 325-347. 1994.

[3] P. Léon, “*Précis de Phonostylistique. Parole et Expressivité*, Paris: Nathan Université, 1993.

[4] A. Rilliard, D. Erickson, T. Shochi, and J.A. de Moraes,, “Social face to face communication - American English attitudinal prosody”, INTERSPEECH 2013. 1648-1652.

[5] A. Hönemann, H. Mixdorff, A. Rilliard “Social attitudes - recordings and evaluation of an audio-visual corpus in German”, Forum Acusticum 2014, Krakow, Poland.

[6] H. Mixdorff, A. Hönemann, and A. Rilliard, “Free Labeling of Audio-visual Attitudinal Expressions in German,” *Proceedings of SST 2016*, Sydney, Australien, 2016.

[7] H. Mixdorff, A. Hönemann, A. Rilliard, T. Lee, and Matthew K.H. Ma, “Cross-Language Perception of Audio-visual Attitudinal Expressions,” Proceedings of AVSP 2017, Stockholm, Sweden.

[8] H. Mixdorff, A. Hönemann, A. Rilliard, T. Lee, and Matthew K.H. Ma, “Audio-Visual Expressions of Attitude: How many different attitudes can perceivers decode?” Speech Communication, Volume 95, December 2017, Pages 114-126.

[9] T. Lee, M.K.H. Ma, A. Rilliard, H. Mixdorff, A. Hönemann, “Free Labeling of Audio-visual Attitudinal Expressions in Cantonese”, accepted for presentation at SpeechProsody 2018, Poznan, Poland.

[10] G. Schauenburg, J. Ambrasat, T. Schröder, C. von Scheve, and M. Conrad, “Emotional connotations of words related to authority and community,” *Behavior Research Methods, 47*, 720-735, 2015.

[11] T. Schröder, J. Hoey, and K. B. Rogers, “Modeling dynamic identities and uncertainty in social interaction: Bayesian affect control theory,” In *American Sociological Review*, vol. 81, issue 4, 2016.

[12] F. Husson, S. Lê, J. Pages, J., “Exploratory multivariate analysis by example using R”. London: Chapman & Hall, 2011.

[13] Bécue-Bertaut, M., & Pagès, J., “Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data”. Computational Statistics & Data Analysis, 52(6), 3255-3268, 2008.

[14] Osgood, C. E., May, W. H., & Miron, M. S., “Cross-cultural universals of affective meaning (Vol. 1)”. University of Illinois Press, 1975.

[15] Goudbeek, M., and K. Scherer. "Beyond arousal: Valence and potency/control cues in the vocal expression of emotion." The Journal of the Acoustical Society of America 128.3 (2010): 1322-1336.

[16] Wichmann, A., "The attitudinal effects of prosody, and how they relate to emotion." ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion. 2000.

[17] Moraes,J. A., A. Rilliard, “Illocution, attitudes and prosody: A multimodal analysis”. In T. Raso & H. Melo (Eds.) Spoken Corpora and Linguistic Studies, John Benjamins Publishing Company, pp.233-270, 2014,