Cross-Language Perception of Audio-visual Attitudinal Expressions

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

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

mixdorff@bht-berlin.de, ahoenemann@techfak.uni-bielefeld.de, Albert.Rilliard@limsi.fr, tanlee@ee.cuhk.edu.hk, matthewhang92@gmail.com;

Abstract

This paper presents results from a cross-language free labeling experiment employing short audio-visual utterances of Cantonese produced with varying attitudinal expressions. German perceivers were asked to specify a single word that best described these stimuli, some of which were presented in audio-only and video-only modality. The resulting terms were classified with respect to the emotional dimensions of valence, activation and dominance, as well as the linguistic dimension of assertion/interrogation. We compared the results with the outcomes from a similar experiment employing German stimuli. Most types of attitudes presented in Cantonese were rated less positive and portrayed with less activation than those presented by Germans. Video-supported stimuli yielded significantly higher activation and dominance levels than audio-only ones. The main dimensions separating expressions are assertive vs. interrogation, valence, and dominance. Illocutionary strength is associated with the perceived activation, and primarily linked to the visual channel, while linguistic information is primarily conveyed by acoustic cues, but of course only for the German stimuli.

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

1. Introduction

The expression of attitudes is an important aspect of speech communication as it helps us navigate in human-human dialogs and achieve our conversational aims. Interlocutors who share the same language or culture are conditioned to similar codes, behaviors and even belief systems. In contrast, interaction between partners from different cultures may lead to wrong interpretations of social expressions. Earlier cross-language studies (see, for instance, [1]) found similarities across languages, such as the use of low pitch to signal dominance [2] but also some culture-specific uses of e.g. prosodic parameters [3]. Intercultural comparison of linguistic and paralinguistic effects has enjoyed growing attention as the knowledge about how verbal and non-verbal social affects are expressed in different languages is paramount in a globalized world.

The current work is based on the framework developed by [3] in which attitudes are elicited by immersing the presenters in various communicational situations. Recordings also concern the visual channel, as facial gestures are known to be an important part of attitudinal expressions [4].

Attitudes such as arrogance, politeness, doubt or irritation - see Table 1 for abbreviations henceforth used in this paper - were portrayed by native German presenters through short dialogs which ended in the target sentences ‘Eine Banane’ (engl. *a banana*) or ‘Marie tanzte’ (engl. *Marie was dancing*). Preceding the target dialog a test dialog was performed in order to prepare the speakers and help them immerse themselves in the context of the attitude.

In earlier perception studies we had native German subjects rate the credibility of the expressions portrayed by the first 10 of the speakers [5] and examined the acoustic-prosodic properties of the data [6]. Then we ran an identification study in which we asked subjects to choose from a set of five labels the one they deemed most appropriate [7]. From the outcomes of the identification study we suspected that offering raters a set sub-group of labels introduced a strong bias. Hence, in a follow-up inspired by [8], raters were free to select a single word, either a noun or adjective that best fit their impression of the attitudinal expression [9]. Different from [8], we also included audio-only and video-only examples to test for differences in the modalities. The expressions chosen were normalized and rated regarding their location in the three-dimensional emotional space of valence, activation and dominance.

All studies showed that attitudes essentially cluster in several groups, the members of which share similar properties. On the positive side of the spectrum we find attitudes such as *admiration* and *sincerity*, whereas *authority*, *contempt*, *arrogance*, *irritation* and to a certain degree *irony* gather on the negative side. “Neutral” statements and questions which we initially regarded as a standard are often confounded with their affective partners *politeness* and *surprise*, respectively.

In an effort to extend our work with a cross-language setting, the current study aims to compare the results from the intra-German free labeling experiment to a situation where German listeners are exposed to expressions of attitude portrayed in Cantonese, a language and culture they are typically not familiar with. We wish to examine whether resulting clusters are similar to those formed when stimuli in German are presented or whether there are systematic differences, especially as a function of seeing the presenter’s face.

10.21437/AVSP.2017-23

2. Recording of Stimuli and Perception Study

Following the protocol developed in [3], expressions of attitude were elicited from ten native speakers of Cantonese, four males and six females, all of them students at the Chinese University of Hong Kong, aged between 18 and 24. To that end, English versions of the materials describing the communicative situations for the sixteen attitudes were translated to Cantonese. The participants were recorded in a sound-treated room. Each attitude was recorded once. Eventually the target phrases in the session videos were extracted from the context, yielding 320 tokens for sixteen attitudes and two target phrases by ten presenters. In order to select the best exemplars for each attitude, four native Cantonese listeners rated each phrase in full auditory-visual (AV) modality given the intended attitude on a scale from 1 to 9, a score of 9 being very convincing and 1 completely unconvincing.

Subsequently we selected the stimuli for the ensuing perceptual study with the German participants based on the mean ratings for each of the stimuli. Different from [9] where we had employed the phrase “Eine Banane” for the free labeling, we selected the Cantonese equivalent of “Mary was dancing” (Mary tiu3 gan2 mou5, Mary 跳緊舞) and not “A banana” (hoeng1 ziu1, 香蕉). As Mann-Whitney-U test had not shown any significant differences of the performance ratings as a function of the phrase, we decided that the five-syllable “Mary” utterance better matched the German “Eine Banane” which also contains five syllables.

For creating our stimulus sets we chose those AV examples that had been rated best for a given attitude, yielding 6 stimuli for each attitude, or a total of 96 stimuli. A sub-set of the selected stimuli was added in audio-only and video-only mode. In total we had 96 audio-visual (AV), 48 audio-only (AU) and 48 video-only (VI) samples which we split into two sets of 96 stimuli each.

A warm-up phase was added in which eight stimuli were displayed to familiarize subjects with the range of expressions they were going to rate, however, without asking their assessment. Hence each of the stimuli should be described by a single word, either a noun or adjective. As mentioned earlier, every subject had to rate 96 examples (48 AV, 24 AU, and 24 VI) for the experiment. Warm-up stimuli were presented only in the audio-visual modality and not used in the experiment proper. The rating procedure was allowed to take as long as the subject required. It took between 25 and 45 minutes to complete the task. Subjects were 42 students (31 male, 11 female), aged between 19 and 36 years, of Media Informatics in their second year at the Department of Computer Science and Media at Beuth University Berlin. Participants received course credits in exchange for their time.

3. Normalization and Semantic Analysis of Labels

We collected a total number of 3920 valid labels: 1957 for AV, 979 for AU and 984 for VI presented stimuli. Analysis of written expressions showed quite a variation of terms used, yielding a total number of 728 different tokens. Despite the instruction to use just a single word, some subjects had typed two words or even a whole phrase to describe their impressions. Many two-word terms included an emotional and a linguistic component, such as “genervt fragend” (engl. *asking irritably*). After the correction of typos we normalized the raters’ inputs by collapsing similar words, for instance, such as “Frage” (*question*) and “fragend” (*asking*) onto a single term. We also collapsed semantically equivalent terms onto the more frequent ones.

We excluded one of the male raters due to a multitude of “no idea” responses. Of the remaining 41 perceivers, there were only 16 terms that we were unable to interpret sensibly and hence failed to map onto any of the normalized expressions. These were all single-occurrence tokens that we excluded from further analysis. After consolidating all expressions we yielded 122 different terms.

The term “neutral” was the most frequently chosen (*neutral*, N=421), followed by “freundlich” (*friendly*, N=341), “fragend” (*asking*, N=310), and “genervt” (*irritated*, N=296). Depending on the modality, the same terms appear at the top of the list, only the order slightly changes, and “traurig” (*sad*) reaches the 4th position in audio-only condition.

Following the methodology in [9] we classified all terms according to the scheme developed by [10] and [11]. We located each term with respect to the three-dimensional space of valence, activation and dominance and statement vs. interrogation, respectively, and assigned one of three possible values: negative, neutral and positive for valence, and –, 0 and + for activation and dominance. By performing 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.

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

attitude	abbreviation	valence	activation	dominance
admiration	ADMI	.7163	.4184	-.0284
arrogance	ARRO	-.6143	.0571	.1357
authority	AUTH	-.5248	.3546	.3759
contempt	CONT	-.7194	-.1511	.0576
neutral statement	DECL	.1786	.2357	.0714
doubt	DOUB	-.5200	.2800	-.0333
irony	IRON	.1400	.2400	-.0667
irritation	IRRI	-.5177	.4113	.4255
obviousness	OBVI	-.3860	.1140	.0439
politeness	POLI	.4649	.1842	.0702
neutral question	QUES	-.4362	.1383	.0319
seductiveness	SEDU	-.0108	.0860	-.0215
sincerity	SINC	.0938	.1354	.0417
surprise	SURP	-.2842	.6000	.1158
uncertainty	UNCE	-.6190	-.1429	-.3333
walking-on-eggs	WOEG	-,5192	-,0673	-,1250

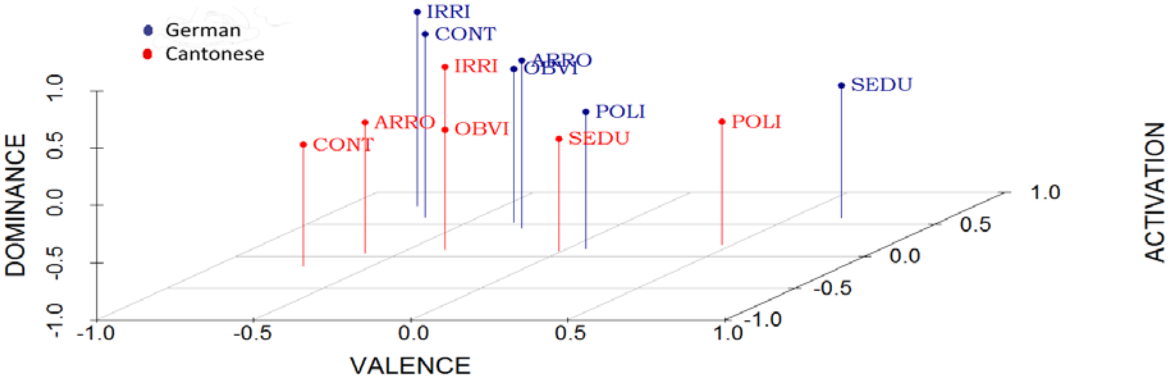

Figure 1: Examples of some of the attitudes that are significantly different for German and Cantonese in the three-dimensional space of valence, activation and dominance, values given for AV stimuli.

4. Results of Analysis

4.1. Emotional space

Based on the frequency and semantic values of labels assigned to each attitude we yielded centers of gravity in the emotional space. Table 1 lists the positions of all 16 attitudes for audio-visual stimuli. We can see, for instance, that CONT is judged more negatively than AUTH, while POLI has a very positive connotation. Figure 1 shows a 3D graph of some of the attitudes for Cantonese stimuli as well as their respective locations found for German stimuli in [9] discussed in the next section.

Based on these results we also compared the impact of reduced modalities on the assessment of attitudes. We will discuss these differences later in this paper when we compare them to those observed for the German stimuli.

The labels obtained from German raters listening to the German performances (cf. [9] for details) are also used in the following analyses, in order to compare the variation of perception of these performances. The same procedure was used to obtain these labels and normalize them, yielding 128 normalized labels for the German performances.

4.2. Comparison with intra-language results

If we compare the results for the Cantonese AV stimuli to the outcome of our earlier study with German AV stimuli [9], we find that overall German perceivers judged the Cantonese performers slightly more negatively than the Germans (mean valence of -.22 vs. -.16), but also less activated or dominant (mean activation of .19 vs. .40, and mean dominance .06 vs. .14).

Mann-Witney U-Test of independent samples confirms that there are significant differences in all three dimensions. The significance is slightly less (p=.005) for valence than for activation and dominance (p=.001). If we look at individual attitudes, almost all perceived mean activation levels are higher for German performers than for the Cantonese ones, CONT (.61 vs. -.15) and SEDU (.60 vs. .09) being the extreme cases. This seems to indicate that the Cantonese performers are perceived as more restrained (see examples for IRRI and SURP in Figure 2).

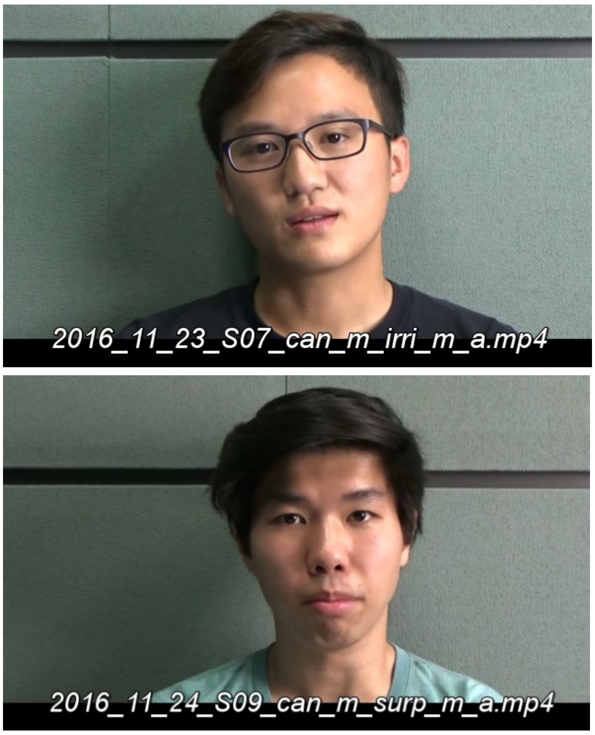
2016_11_23_S07_can_m_irri_m_a.mp4
2016_11_24_S09_can_m_surp_m_a.mp4

Figure 2: Snap shots from Cantonese speaking performers. Top: S07 IRRI, bottom: S09, SURP.

The differences are smaller for dominance, with CONT again being the extreme case (.60 vs. .06). As mentioned before, valence for most attitudes is perceived for Cantonese more negatively, even for the neutral question (mean valence of -.44 vs. -.15). However, the positive attitudes POLI and ADMI also yield better ratings for the Cantonese speakers, namely, .46 and .72 vs. .06 and .53 for the Germans. It is known that in German POLI is not a specific register, but rather a neutral way of speaking whereas Cantonese might have specific devices to express politeness which the Germans perceive as more friendly. As mentioned earlier Figure 1 displays those attitudes that are perceived maximally different in Cantonese and German stimuli in the 3D space. Reduced modality does not change the general picture described above for full AV stimuli, however, audio-only presentations are judged slightly more negatively for both German and Cantonese performers, as well as less activated.

4.3. Clustering of Attitudes

Normalized labels were organized in a contingency table with the presented stimuli's 16 categories, for both languages (German and Cantonese), in each of the three presentation modalities along the 96 rows of the matrix (i.e. rows represent the expressive behavior of speakers, ordered by presentation modality and languages), and the normalized labels assigned in columns (i.e. columns present what was perceived from the stimuli). An analysis of the distribution of these expressive behaviors according to the labels was performed using a correspondence analysis (CA) [12]. The CA was run on the results obtained on the audio-visual (AV) modalities (for both language groups) only, with audio-only (AU) and video-only (VI) results used as supplementary individuals. For this reason, only the 144 normalized labels used to characterize the AV stimuli were kept (resulting from both 104 labels obtained from Cantonese and 112 from German stimuli). The semantic classification of labels according to their valence, dominance, activation and linguistic mode were used as supplementary variables. An elbow criterion was employed to keep the first 8 dimensions which explain about 64% of the variance of the CA for further analysis. Table 2 displays the coordinates and quality of representation ($\cos^2$) of supplementary semantic labels attributed to each labels. This allows us to interpret the abstract dimensions without referring directly to the respective labels collected. The first dimension is mostly linked to the linguistic distinction between assertive and interrogative terms and to the dominance dimension (mostly for submissive labels, that correlate with interrogative labels). The second dimension relates to valence and to dominance—which is also related to the third. The fourth is linked to +activated labels. The fourth is related to expressions with neutral or low activation, while the fifth is linked to high activation.

From the repartition of performances on these dimensions, a hierarchical clustering was performed so to group together expression that shares similar labels. The dendrogram obtained is shown in Figure 3. The first four partitions derived from this tree are described hereafter. In the following we list these clusters with the most frequent labels used to describe them (in decreasing order of importance, English translations given in italics), the semantic connotation significantly linked to them, and the attitudes primarily to them, in the AV modality, as well as in the mono-modal conditions. The assignment of supplementary individual to their respective clusters is performed by calculating the closest center of gravity of each cluster regarding the position of the individual in the CA first eight dimensions:

Table 2: $\cos^2$ (multiplied by 100 and rounded for convenience) of the supplementary variables of the CA, indicating their quality of representation on the first X dimensions. The signs (+/-) indicate the positive or negative localization on the dimensions. Only the first 5 dimensions are detailed, as the others show weak correlations with supplementary variables.

	Dim. 1	Dim. 2	Dim. 3	Dim. 4	Dim. 5
- valence	27/-	37/-	2/+	12/+	4/+
0 valence	3/-	4/-	6/+	27/-	27/-
+ valence	36/+	50/+	8/-	0/+	2/+
- activation	4/-	14/-	9/-	29/+	21/-
0 activation	9/+	0/+	0/+	42/-	15/-
+ activation	1/+	11/+	9/+	2/+	65/+
- dominance	42/-	1/-	16/-	10/+	8/-
0 dominance	20/+	35/+	9/-	13/-	4/-
+ dominance	0+	34/-	37/+	3/+	18/+
Statement	65/+	4/-	8/+	5/+	0/+
Interrogation	65/-	4/+	8/-	5/-	0/+

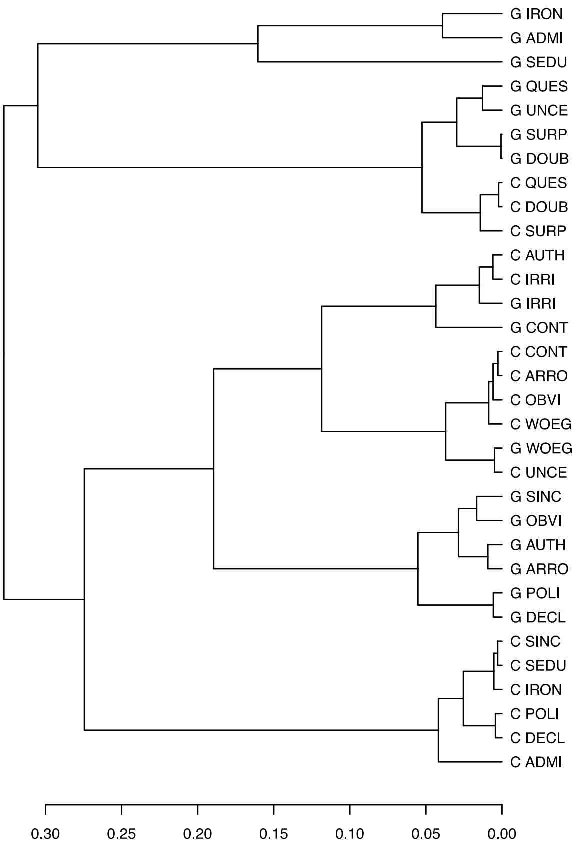

Figure 3: Hierarchical representation of the similarities between audiovisual performances of German (G) and Cantonese (C) speakers, in the 16 situations.

Cluster #1 is composed of the following expressions (modalities given in parenthesis):
for German speakers
DOUB (AV, AU, VI), UNCE (AV, AU, VI), SURP (AV, AU), QUES (AV, AU), and mono-modal ADMI (AU), and WOEG (AU, VI);
for Cantonese speakers
DOUB (AV, AU, VI), QUES (AV, VI), and SURP (AV).

Cluster #1 is mostly described by the following labels: “fragend” (*asking*), “zweifelnd” (*doubtful*), “erstaunt” (*astonished*), “überrascht” (*surprised*), “unsicher” (*insecure*), “unwissend” (*ignorant*), which are linked to the dimensions of Interrogation, -dominance, and neutral valence.

Cluster #2 is the largest cluster and is composed of the following expressions:
for German speakers
ARRO (AV, AU, VI), AUTH (AV, AU VI), CONT (AV, AU, VI), DECL (AV, AU, VI), IRRI (AV, AU, VI), OBVI (AV, AU, VI), POLI (AV, AU, VI), SINC (AV, AU, VI), and WOEG (AV), plus mono-modal IRON (AU, VI) and QUES (VI);
for Cantonese speakers
ARRO (AV, AU, VI), AUTH (AV, AU, VI), CONT (AV, AU, VI), IRRI (AV, AU, VI), OBVI (AV, AU, VI), UNCE (AV, AU, VI), WOEG (AV, AU), plus the mono-modal ADMI (AU), DECL (AU), IRON (AU), QUES (AU), SINC (AU), SURP (AU, VI).
This cluster #2 is described mostly by the labels “genervt” (*irritated*), neutral (*neutral*), “gelangweilt” (*bored*), “überzeugt” (*convinced*), “arrogant” (*arrogant*), “entschlossen” (*determined*), “offensichtlich” (*obvious*), which are linked to the dimensions of +dominance, negative valence, assertion and -activation.

Cluster #3 is composed of expressions only from German speakers:
ADMI (AV), IRON (AV), and SEDU (AV, AU, VI). Cluster #3 is mostly described by the labels “amüsiert” (*amused*), “erregt” (*excited*), “erfreut” (*pleased*), “verführerisch” (*seductive*), “begeistert” (*enthusiastic*) which are related to positive valence, +activation, 0 dominance, and assertion.
Cluster #4 is composed of audio-visual expressions only from the Cantonese speakers:
ADMI (AV, VI), DECL (AV, VI), IRON (AV, VI), SEDU (AV, AU, VI), SINC (AV, VI), POLI (AV, AU, VI), and mono-modal WOEG (VI) — plus mono-modal from German speakers ADMI (VI), and SURP (VI). The cluster is described by labels “freundlich” (*friendly*), “fröhlich” (*cheerful*), “glücklich” *(happy*), “aufrichtig” (*sincere*), “erfreut” (*pleased*), “aufmunternd” (*encouraging*), which are related to the dimensions of +valence, 0 dominance, assertion and 0 activation.

Cluster #1 thus contains interrogative expressions from both language groups, on most modalities for German speakers, while more Cantonese mono-modal performances are misclassified. The link between interrogation and submission proposed by [2] is corroborated by the interpretation of this cluster. Furthermore, the expression of interrogation is probably harder to decode by the Germans since they do not speak Cantonese.

Cluster #2 regroups the greatest number of expressions from both language groups, including a number of misperceived ones, especially in mono-modal presentations. It is mostly composed of expressions imposing the speaker’s will on the interlocutor (AUTH, ARRO, IRRI, CONT), as well as simple declaratives (DECL). Interestingly, mono-modal performances of IRON are also grouped in this cluster, and thus tend to be perceived negatively.

Clusters #3 and #4 are both composed of positive assertions, the first one by German speakers, the second one by Cantonese speakers. Cluster #3 is mostly related to the expression of SEDU by German, and also contains bi-modal expressions of ADMI and IRON. It is interesting to note that German AV performances of IRON are perceived as positive, whereas mono-modal performances pertain to the more negative cluster #2. Cluster #4, containing positive expressions by Cantonese speakers, is mostly described by labels linked to politeness. The main difference between these two clusters is related to the degree of activation.

5. Discussion and Conclusions

Some mono-modal presentations were misclassified by perceivers. This arises, for instance, for ADMI and WOEG being regrouped inside cluster #1 with interrogative meanings. However, AV WOEG was not assigned to the interrogative cluster, but the negative cluster #2, whereas in French and Japanese WOEG shows strong links with uncertainty. Neutral questions performed by Germans and shown in the VI only modality were not recognized, but confused with assertions—underlying the linguistic nature of this sentence mode. Likewise, the Cantonese audio-only versions of QUES were misinterpreted by the German listeners. Most misclassified AU-only productions by Cantonese (ADMI, DECL, IRON, QUES, SINC, SURP) are regrouped under the assertive cluster #2, a fact probably related to the decoding of attitudes via the audio modality in a foreign language.

SURP constitutes a case of synergy between both modalities, which is misclassified when only audio or video are presented. This can also be observed for IRON, the mono-modal productions of which are perceived as rather negative (cluster #2), while the bimodal productions are perceived as positive, for both groups, with more activation for the German version (cluster #3) than the Cantonese one (cluster #4). In this case the two modalities may convey incoherent meanings (see, for instance, [14]) whereas the combination facilitates an understanding of the ironic intention.

The different interpretations (and compositions) of the two positive clusters - the German-speaking one contains SEDU, ADMI and IRON, and is labeled as “amused”; the Cantonese-speaking one contains ADMI, DECL, IRON, SEDU, SINC, POLI, and is labeled as “kind” - show a shift in the activation of the expressions between the two groups of speakers: positive German performances are mostly presented with high activation, and linked to playful behavior, while positive Cantonese performances are grouped with neutral assertions, and related to polite behavior. This is also supported by the fact that video-only performances of ADMI and SURP from German speakers pertain to this cluster #4.

Our analysis shows the importance of the presentation modality for the decoding of attitudinal information. Audio modality plays a primary role in signaling linguistic information like interrogation, when conveyed in the language of the perceiver, whereas visual modality is not constrained to language, but also shows language-specific traits, facilitating an interpretation of politeness, for instance. Future work will examine how Cantonese perceivers interpret attitudes portrayed in German and Cantonese.

6. Acknowledgements

This work was supported by DFG (Deutsche Forschungsgemeinschaft) grant Mi 625-27 funding a visit by Mixdorff to the Chinese University of Hong Kong.

7. References

[1] Shochi. T.. Rilliard. A.. Aubergé. V. & Erickson. D. “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] Ohala. J. J.. “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] Léon, P., “*Précis de Phonostylistique. Parole et Expressivité,* Paris: Nathan Université, 1993.

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

[5] Swerts, M. and Krahmer, E., “Audiovisual prosody and feeling of knowing”, Journal of Memory and Language 53(1): 81-94, 2005.

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

[7] Mixdorff, H., Hönemann, A., Rilliard, A., “Acoustic-prosodic Analysis of Attitudinal Expressions in German.” Proceedings of Interspeech 2015, Dresden, Germany, 2015.

[8] Hönemann, A., Rilliard, A., Mixdorff, H., “Classification of Auditory-Visual Attitudes in German.” FAAVSP 2015, Vienna, Austria, 2015.

[9] Guerry, M., Shochi, T., Rilliard, A., and Erickson, D. “Perception of prosodic social attitudes affects in French: A free-labeling study”, Proceedings of ICPhS 2015, Glasgow, Scotland.

[10] Mixdorff, H., Hönemann, A. and Rilliard, A. (2016): Free Labeling of Audio-visual Attitudinal Expressions in German. SST 2016, Sydney, Australien.

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

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

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

[14] Bryant, G. A., “Prosodic contrasts in ironic speech,” Discourse Processes, 47(7), 545-566, 2010