DEVELOPING A CORPUS OF AUDIO-VISUAL ATTITUDINAL EXPRESSIONS IN HINDI

Hansjörg Mixdorff[1], Navneet Nayan[2], Albert Rilliard[3], Preeti Rao[4], Debashis Ghosh[2]

[1]Berliner Hochschule für Technik [2]IIT Roorkee, India [3]Université Paris-Saclay, France [4]IIT Bombay, India
hmixdorff@bht-berlin.de, nnayan@ec.iitr.ac.in, Albert.Rilliard@limsi.fr, prao@ee.iitb.ac.in,

ABSTRACT

This paper presents a corpus of attitudinal expressions in Hindi and their evaluation by native Hindi speaking raters. The paradigm is adapted from Rilliard et al. (2013), and forms part of an intercultural endeavour aimed at studying prosodic and facial expressions of social affect across languages. Our corpus includes a total of 16 attitudes, such as arrogance, surprise, politeness etc. portrayed by 19 speakers (10f, 9m) of which three experts selected the best four males and females and the better of two turns for subsequent analysis. A follow-up experiment with more participants employed a total of 512 stimuli, 256 hereof full audio-visual stimuli, 128 audio-only and 128 silent video stimuli. Results indicate higher ratings for emotionally loaded attitudes such as irritation and doubt as compared to, for instance, irony or seductiveness. Reduced modality stimuli were rated more poorly.
Keywords: Attitudes, Hindi, audio-visual perception

1. INTRODUCTION

Human communication always has a social goal. Information about e.g. the mental state, emotions, mood or attitudes of the speaker and listener is passed during the dialog. The affective state is influenced, for instance, by the situation or role of the dialog partners. Mutual understanding of the social intention between communication partners should not be difficult as long as they grow up in the same or at least a similar culture. Interaction between partners from different cultures sometimes leads to wrong interpretations of the social expression. It has been shown that the verbal and non-verbal expressions depend on the culture in which we grow up. A study by Shochi et al. investigated twelve social attitudes e.g. surprise, irritation, command-authority for prosodic effects in British English, French and Japanese [1]. 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 for mutual understanding between different cultures.

The primary goal of the work reported in the current paper was the recording of a Hindi speech corpus of social attitudes for an intercultural comparison. It follows the experimental design developed by Rillard et al. [2], discussed in Section 2. In earlier works we collected corpora of attitudinal expressions in German [3] and Cantonese [4] following the same experimental design. These corpora were judged by Hindi speaking subjects [5]. Now the corresponding Hindi attitudinal data were gathered.

This paper presents the process of data collection and a perception study exploring the credibility of the 16 social attitudes (see Table 1 for complete listing), e.g. doubt, surprise and politeness of 19, later 8 Hindi speaking subjects. Section 2 presents the experimental design and the social and linguistic criteria of situations employed to elicit the attitudes from our subjects. In the same section the technical setting for the recordings, as well as the process of the recordings with dialog examples are discussed. The perceptual evaluation of the social attitudes by native Hindi raters is described in section 3. Section 4 presents the results of the evaluated expressions which includes analyses with regards to speaker and attitude. Section 5 concludes this paper with discussion and conclusions.

2. EXPERIMENT SETUP

2.1 Situational attitudes

16 Attitudes are performed by 19 speakers (10f, 9m) which are elicited through a short dialog between each speaker and the experimenter. The dialogs led to two target utterances: “mairee naach rahee thee” (मैरी नाच रही थी) (engl. Mary was dancing) and “ek kela” (एक केला) (engl. a banana). In the remainder of this article we will refer to the English translations to identify the respective phrase. For each expression of attitude, a test dialog was executed in order to prepare the speakers. This dialog was designed according to different social situations differing in the following social and linguistic aspects (A => speaker, B => experimenter):

- Type of speech act: propositional / social attitude

- Hierarchical distance between speaker A and speaker B: A > = < B[1]
- Social distance between speaker A and speaker B: 1-friend, 2-known, 3-unknown[1]
- Valence of speech act: positive / negative
- A dominates B: yes / no

The social situations conceived are adapted from Rillard et al. [2]. As examples detailed descriptions of three of the social situations are presented in the following:

Admiration (ADMI): A and B are almost the same age and know each other well. Both love Samosas, and talk about the delicious ones they had the day before at Prakash's restaurant. The scene is at a coffee shop.

Irritation (IRRI): A and B are almost the same age and know each other. A is sitting next to B. Suddenly B starts smoking and A is very angry; he wants him/her to stop, expressing his irritation toward speaker B. The scene is in a public place.

Surprise (SURP): A and B are friends and the same age. A did not know that B can sing well. One day, B has A listen to his beautiful voice. The scene is at a friend's home.

2.2 Technical setting and recording process

We used a Nikon D850 HD camera set up 1m away from the speaker. Additionally, the acoustic signal was recorded through a dynamic AKG D5 microphone placed under the chin of speaker. A laptop positioned in front of the speaker displays the description of each social situation, the related test dialog and target dialog permitting the speakers to prepare for the ensuing task. The experimenter performed each situation with the subject in order to immerse him/her in the context for the attitude. After a short break the target dialog containing the target phrase followed immediately. The complete sequence of dialogs was recorded twice. An example test and target dialog for the attitude "irritation" (IRRI) could run as follows:

A: Sorry, but don't smoke please.
B: Okay, okay…
A: Don't smoke, please!

B: What did Mary do last night?
A: **Mary was dancing**. I already told you three times (are you deaf?)

In the test dialog A is angry because B starts to smoke and A wants to stop him. The sentence in bold is the target utterances which A should speak with an irritated expression. The sentences in brackets are only for familiarizing the speaker with the task, and are not uttered The main dialog ends with the target utterance "Mary was dancing".

The target phrases were cut out of the session video. We recorded a total of 19 speakers x 16 attitudes x 2 phrases x 2 trials=1216 target phrases performed by 19 speakers. Participants were students of IIT Roorkee and compensated for their time.

3 Evaluation

Before we can subject the data to further analysis or employ them in perceptual tests we need to ensure the quality of presentation. To this end we first had the complete corpus of 1216 files judged by three experts who ranked each video clip on a scale from 9 (convincing) to 1 (implausible), given the intended attitude. The experts were familiar with the elicitation paradigm and supposed to scrutinize the data accordingly. In order to facilitate the tiring task we created four batches of 240 stimuli and one of 256.

In all experiments the stimuli were presented via a survey on the PsyToolkit server[4][5]. After an introduction to the experiment design, listing all 16 attitudes in English as well as Hindi, the stimuli were presented in a randomized order, first displaying for 900 ms a screen with the intended attitude in English and Hindi and a fixation cross in the location where the video was going to appear, then the video for its respective duration was shown. Subsequently the viewers were able to input their judgment via a graphical interface by clicking on a line between "very well" corresponding to a score of 9 and "not at all" corresponding to 1. A progress bar was shown for orientation.

The experts viewed the full audio-visual (A/V) stimuli and helped us identify the four best male and female speakers, as well as the better turn of the two we had collected.

Hence we yielded 8 x 32=256 utterances for the ensuing main experiment with a larger number of participants. In addition to the 256 full A/V stimuli, from these we created 128 silent video (V) and 128 audio-only (A) stimuli, selecting the better one of "Mary was dancing" or "a banana". Hence the experiment employed a total of 512 stimuli, the quality of which was judged by 32 Hindi speaking raters (23m, 9f). Due to the large number of

[1] Notions of distance, cf. Spencer-Oatey(1996)

utterances to be evaluated, we created a PsyToolkit survey with four sets of 128 stimuli each, while ensuring that the reduced A and V stimuli were not in the same set as the corresponding A/V stimuli. This means that each set was rated by eight judges.

After hearing/watching the stimulus only once raters had to judge the plausibility of each attitude during a time window of ten seconds. We collected 4096 scores, 32 of which (0.8%) were invalid due to time-outs. We hence included in total 2034 valid judgments for A/V stimuli 1016 for A, and 1014 for V type stimuli. The experiment took on average 14 minutes to complete.

4 Results

The first analysis presented in this section concerns the judgments on a scale from one to nine assigned by the three expert judges. Statistical tests were performed using the SPSS statistical toolkit [1]. In the second part we look at the judgments of the larger participant group on the reduced set.

4.1. Expert Ratings

The raters generally judged the female speakers (score average/s.d. 7.32/1.95) better than the male speakers (7.06/2.10), or in other words, of the eight best rated speakers six were females. Interestingly, the distance "between the genders" is greater for certain attitudes such as ADMI or WOEG than others like DOUB or IRRI.

In Table 1, columns three and four list means and s.d. for all 16 attitudes. The judgments of the raters indicated a slightly better performance of the speakers at the second trial (average of 7.23) than the first one (7.16). The reason could be that the speakers knew the task after the first trial and on that account, it was easier to prepare the expressions. For the ensuing larger study, we selected for each attitude/target phrase the repetition that was judged better. We did not find any significant difference between the target phrase "a banana" and "Mary was dancing".

As witnessed in earlier studies, complex attitudes like IRON and SEDU were judged more poorly than coping strategies with strong emotions such as IRRI, SURP, CONT and DOUB.

In contrast, POLI which is actually a socially conditioned register also ranges among the highest scorers. In our German data this was the opposite. Obviously politeness exhibits special features in Indian society which are not as strong in the West or just represents a "normal way of speaking".

Interestingly, also the neutral question (QUES) received rather low ratings. We assume that the facial expression together with the interrogative mode often counteracted the intended "neutrality" and rather conveyed doubt or suspicion.

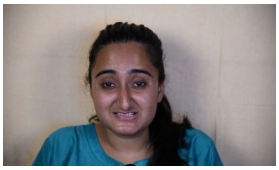
There is a strong negative correlation between the mean score for each stimulus (Pearson's R of -.765, $p < .001$) and its standard deviation, indicating that the raters agreed more on highly rated stimuli. There is also a weak correlation between the reaction time and the score (Pearson's R of -.147, $p < .001$).

4.2. Ratings on subcorpus of best speakers

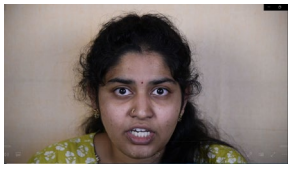
As explained above we selected a subcorpus of the best four male and best four female speakers and had them judged by a larger group of participants. Figure 1 presents snapshots of examples of different attitudes of these speakers and their mean scores and s.d. In general, results among the "non-experts" follow similar tendencies as those of the experts. If we correlate their mean stimulus-wise judgments with those of the experts on the 256 full AV stimuli that both experiments had in common, we yield a Pearson's R of .385 ($p < .001$).

Attitude	abbrev-iation	mean	standard deviation
admiration	ADMI	7.35	2.18
arrogance	ARRO	7.42	1.81
authority	AUTH	7.38	1.62
contempt	CONT	7.25	2.00
neutral statement	DECL	7.30	1.78
doubt	DOUB	7.62	2.05
irony	IRON	6.60	2.03
irritation	IRRI	7.63	2.07
obviousness	OBVI	7.10	1.96
politeness	POLI	7.46	1.57
neutral question	QUES	6.50	2.37
seductiveness	SEDU	6.78	2.29
sincerity	SINC	6.95	2.02
surprise	SURP	7.42	2.11
uncertainty	UNCE	7.30	1.93
walking-on-eggs	WOEG	7.06	2.11

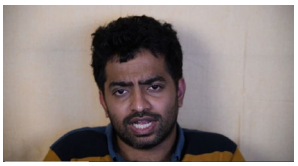
Table 1: List of attitudes, abbreviations used in the paper and mean score and s.d. found in the expert study for 1216 stimuli.



S05 IRRI, 7.33/1.88



S03 SURP, 7.23/1.91

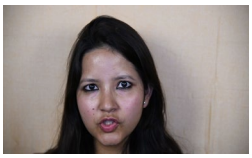


S06 DOUB, 7.17/1.97

S09 ADMI, 6.57/2.25

S08 SINC, 7.01/2.03

S15 SEDU, 7.19/1.91



S16 POLI, 7.15/1.92

S17 UNCE, 7.12/2.05

Figure 1: Examples from the four best male and female speakers performing various attitudes, mean and s.d. of scores.

Attitude	AV mean/s.d.	V mean/s.d.	A mean/s.d.
ADMI	7.39/1.75	7.22/1.85	6.23/2.27
ARRO	7.63/1.55	7.41/1.90	6.75/2.33
AUTH	7.42/1.86	6.65/2.27	6.75/2.49
CONT	7.43/1.86	6.64/2.30	5.48/2.50
DECL	7.30/1.75	6.86/1.95	7.52/1.57
DOUB	7.84/1.55	7.37/1.69	7.61/1.45
IRON	7.01/1.80	6.26/2.19	6.59/1.96
IRRI	7.79/1.59	7.34/1.95	6.41/2.61
OBVI	7.00/2.13	6.62/2.30	7.28/1.70
POLI	7.29/1.74	7.06/1.72	7.33/1.72
QUES	6.71/2.32	5.85/2.15	6.73/1.72
SEDU	6.43/2.34	5.78/2.52	6.03/2.50
SINC	7.33/1.70	7.17/1.85	7.33/1.72
SURP	8.09/1.28	7.80/1.54	8.06/1.21
UNCE	7.64/1.58	7.16/1.91	7.28/1.70
WOEG	6.76/2.33	6.42/2.25	6.63/2.17

Table 2: List of attitudes. mean score and standard deviation as a function of the attitude and modality.

Noticeably, the non-experts extend their range of scores much more towards the lower end. Hence their mean score across the whole data set is almost one point below that of the experts (7.32 vs. 8.12, for AV stimuli) Since we look at many more judges their s.d.s are much higher at 1.89 vs. .68 of the experts.

The most important difference between the two experiments besides the smaller number of stimuli presented is the presence of full A/V as well as V and A stimuli. On average full A/V stimuli were judged better (mean/s.d. of 7.32/1.89) than V (6.86/2.10) and A (6.88/2.10). This is confirmed by independent samples Kruskal-Wallis test between AV and A or V, respectively. Table 2 lists mean scores and standard deviations for each attitude depending on the modality. Interestingly, for the so-called "neutral" attitudes DECL and QUES the audio-only mode yields better ratings than the full AV, indicating the contribution of the acoustic channel, but also DOUB and SURP seem to be conveyed well acoustically. On the video-only data, IRRI and ARRO, two strongly negative attitudes, surpass audio-only by far and show the importance of the visual channel.

5 Discussion and Conclusions

This study presented a recently developed Hindi corpus of attitudinal expressions and first results of its analysis. Starting from an initial set of 19 speakers and 1216 video clips, we first selected the best four male and female speakers for subsequent analysis. We found that the females were the better performers, especially at expressing ADMI.

On the remaining data set we performed a rating test of full AV as well as reduced A and V versions. Our results are in line with those of earlier studies on different languages [3][4]. Reduced stimuli are generally judged more poorly than the full ones, with a few exceptions. This result shows that both the acoustic signal and the visual cues are vital in the portrayal of attitudes.

Future work will entail perception tests with viewers from different cultures. Work is in progress on a free labeling task with German participants. We also plan analysis of acoustic data as well as visual cues. Although we already observed many kinds of head movements, the famous Indian "head wobble" or "wag" occurs very seldom. Of the 256 stimuli only three exhibit this head movement. They all pertain to the UNCE condition.

6 Acknowledgements

This work was supported by DFG grant Mi625-32 to Mixdorff. Thanks go to Priya G for video editing.

7 References

[1] Shochi, T., Rilliard, A., Aubergé, V. & Erickson, D. 2009. 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.

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

[3] Hönemann, A., Mixdorff, H., Rilliard, A. 2014. Social Attitudes - Recordings and Evaluation of an audio-visual Corpus in German. Forum Acusticum 2014, Krakow, Poland.

[4] Lee, T., Ma, M., Rilliard, A., Mixdorff, H., Hönemann, A. 2018. Free Labeling of Audio-visual Attitudinal Expressions in Cantonese. Proceedings of SpeechProsody 2018, Poznan, Poland.

[5] Mixdorff, H., Ghosh, D., Rilliard, A., Hönemann, A. 2020. Perception of Audio-visual Expressions in German and Cantonese by Native Speakers of Hindi. Speech Prosody 2020, Tokyo, Japan.

[6] Stoet, G. 2010. PsyToolkit - A software package for programming psychological experiments using Linux. *Behavior Research Methods, 42(4)*, 1096-1104.

[7] Stoet, G. 2017. PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology, 44(1)*, 24-31.

[8] IBM SPSS , URL https://www.ibm.com/products/spss-statistics, 2022.