7 - 12 SEPTEMBER 2014, KRAKÓW
FORUM ACUSTICUM

Social Attitudes - Recordings and Evaluation of an audio-visual Corpus in German

Angelika Hönemann
University Bielefeld, Bielefeld, Germany

Hansjörg Mixdorff
Beuth University of Applied Science, Berlin, Germany

Albert Rilliard
LIMSI-CNRS, Orsay, France

Summary
This paper presents an experimental design for recording different social face-to-face expressions of native German speakers and the evaluation of these performances by native German raters. The paradigm is adapted from Rilliard et al. (2013) and hence our work forms part of an intercultural endeavor aimed at studying prosodic and facial expressions of social affect across languages. Our speech corpus includes a total of 16 attitudes, such as arrogance, surprise, politeness, irony, doubt or irritation, portrayed by 10 speakers (4f, 6m). The attitudes were elicited by performing dialogs between the speaker and the experimenter. Each condition was recorded twice. A total of 800 audio-visual recordings were evaluated by 30 raters (8f, 22m) who judged the quality of the intended expressions on a scale of 1 (implausible) to 9 (convincing). We yielded 4294 valid judgments. The first three most convincing expressions for the AV recordings are doubt, irritation and surprise; the most implausible are walking on eggs, politeness and seductiveness. This ranking is similar to what was observed for American English, with the notable exception of irony, which was assigned better ratings for the German speakers.

PACS no. 43.72.+q

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, to some extend, 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 the languages British English, French and Japanese [8]. They found similarities across these languages, but also some culture-specific uses of prosodic parameters. The similarities may be explained under theory like the frequency code [5] – a code phylogenetically derived that (roughly) proposes the use of pitch level as a marker of dominance. Other codes have been proposed [1] that may refine the predicted use of fundamental frequency for communicative purposes. Conversely, culture-specific uses have been documented [2]. 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 German speech corpus of social attitudes for an intercultural comparison. It follows the experimental design developed by Rilliard et al. [7]. This paper presents the process of data collection and a perception study exploring the credibility of the 16 social attitudes, e.g. doubt, surprise and politeness of ten native German subjects. Section 2 presents the experimental

Table 1: Short terms of 16 social attitudes, the meaning of them and the situational settings (see 2.1.)

DECL	QUES	UNCE	SURP	DOUB	OBVI	ADMI	IRON
Decl. sentence	Question	Uncertainty	Surprise	Doubt	Obviousness	Admiration	Irony
Prop., A=B, 2	Prop., A=B, 2	Prop., A=B, 2	Prop., A=B, 2	Prop./Soc., A=B, 2, Neg., yes	Prop./Soc., A=B, 2, Neg., yes	Prop./Soc., A=B, 1, Neg., no	Prop./Soc., A=B, 1, Pos./Neg., yes
SEDU	AUTH	IRRI	ARRO	CONT	POLI	SINC	WOEG
Seduction	Authority	Irritation	Arrogance	Contempt	Politeness	Sincerity	"Walking on eggs"
Soc., A=B, 1, Pos./Neg., yes	Soc., A>B, 3, Pos./Neg., yes	Soc., A=B, 2, Neg., yes	Soc., A>B, 2, Neg., yes	Soc., A>B, 2, Neg., yes	Soc., A=B, 3, Pos., no	Soc., A<B, 2, Pos., no	Soc., A<B, 2, Pos., yes

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 German 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

Corpora in French, American English and Japanese exist which are drew up with the same experimental design. Because of the intercultural comparability we used also for the German data collection the same experimental conditions.

2.1. Situational attitudes

16 Attitudes are performed by 10 speakers (4f, 6m) which are elicited through short dialogs between each speaker and the experimenter. The use of dialogs to elicit prosodic attitudes is intended to avoid problems with the definition of these affects, a problem particularly acute as soon as translation of a given concept is used for cross-cultural comparison [10], and to allow the investigation of culture-specific affect that are not defined in some language (cf. the case of Japanese's *kyoshuku*, which typical situational occurrences similar to the "walking on eggs" attitude [7]). The dialogs led to two target utterances: "Marie tanzte" (eng. Marie was dancing) and "Banane" (eng. Banana). 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 [3]
- Hierarchical distance between speaker A and speaker B: A > = < B[2]
- Social distance between speaker A and speaker B: 1-friend, 2-know, 3-unknown[2]
- Valence of speech act: positive / negative
- A dominates B: yes / no

The social situations conceived are described by Rilliard et al. [7]. Attitudes being considered refer to the social settings and can be found in Table 1. 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 French cuisine, and talk about the very delicious food they had the day before at a famous French 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 Sony FX1000 HD Camera set up 1.5m away from the speaker. Additionally, the acoustic signal was recorded through a MCE 86 S II microphone from Beyerdynamic placed at a distance of 0.3m from the speaker. A laptop positioned in front of the speaker shows the description of each social situation, the related test dialog and target dialog permitting the speakers to prepare themselves for the next task. The

[2]On the notions of distance, cf. Spencer-Oatey (1996)

experimenter performed each situation with the subject in order to immerse him/herself in the context of the attitude. After a short break the test dialog was executed. The target dialog containing the target sentences followed immediately. The complete sequence of dialogs was recorded twice. An example test and target dialog for the attitude "irritation" (IRRI) is shown (the social situation is described above):

A: Tschuldige, aber bitte rauche nicht.
(Excuse me, but don't smoke please.)
B: Ok, ich weiß, ich weiß…
(Ok, I know I know…)
A: **Rauche nicht, bitte!**
(Don't smoke, please!)

B: Was hat Marie letzte Nacht gemacht?
(What was Mary doing last night?)
A: **Marie tanzte** (Ich habe es dir schon dreimal gesagt. Bist du taub oder was?)
(Mary was dancing. I already told you three times before. Are you deaf or what?)

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

Because of the difficulty performing the target utterance "Banane" in a particular mood a supporting image was presented. Figure 1 shows an example image and the associated dialog for the attitude "irritation".

B: Was möchtest du haben?
(What would you like?)
A: **Eine Banane** (Ich habe es dir schon dreimal gesagt.)
(A Banana. I already told you three times before.)

Figure 1. Image and dialog for attitude "Irritation"

The target sentences had to be cut out of the session video. We recorded a total of 640 target sentences performed by 10 speakers.

3. Evaluation

The quality of the intended expressions was judged by 30 German raters (8f, 22m) by ranking them on a scale from 1 (implausible) to 9 (convincing). Due to the large number of utterances to be evaluated we created five sets of 160 stimuli. For comparing the perceived effect of reduced modalities, a subset of 80 recordings was presented in audio only and silent video condition. Each contained data from six of the speakers but only either AV, A or V stimuli, respectively. Therefore a total of 800 audio-visual recordings were evaluated. The recordings were presented via an experimental software developed using the LiveCode Framework [11].

In order to examine the influences of modality on the judgment of the rater, silent and non-silent videos have been presented the recordings of the audio-visual stimuli from the first and second trail of the recordings $AV_{(1+2)}$ and the silent video from the first trail $V_{(1)}$ were played in a random sequence from the same speaker. A second part of the evaluation contained the audio-only stimuli $A_{(1)}$ from the first trail of the recordings in a random sequence by the same speaker. After hearing/watching the stimulus only once raters had to judge the plausibility of each attitude during a time window of ten seconds.

We included in total 4294 valid judgment thereof 416 (9.69%) for the $A_{(1)}$ stimulus, 496 (11.55%) for the $V_{(1)}$ stimulus and for the $AV_{(1+2)}$ stimulus 1496 (35.17%) / 1864 (43.60%) judgments.

4. Results

The first analyses presented in this section concern the raw judgments on a scale from one to nine. In addition, a statistical test was performed by applying a randomized four-factor ANOVA test using the R software [11]. To that effect the judgments of each rater were standardized calculating their z-score value.

4.1. Speaker analysis

One interesting result was that the raters judged the male speakers (04, 05, 06, 08, 09, 10) better than the female speakers (01, 02, 03, 07) in each modality, independently of the raters' gender. As an example, for the AV-modality the males were judged on average (s.d.) with 6.32 (2.15) and the females with 5.47 (2.48) in the first trial. The other modalities were judged in a similar fashion.

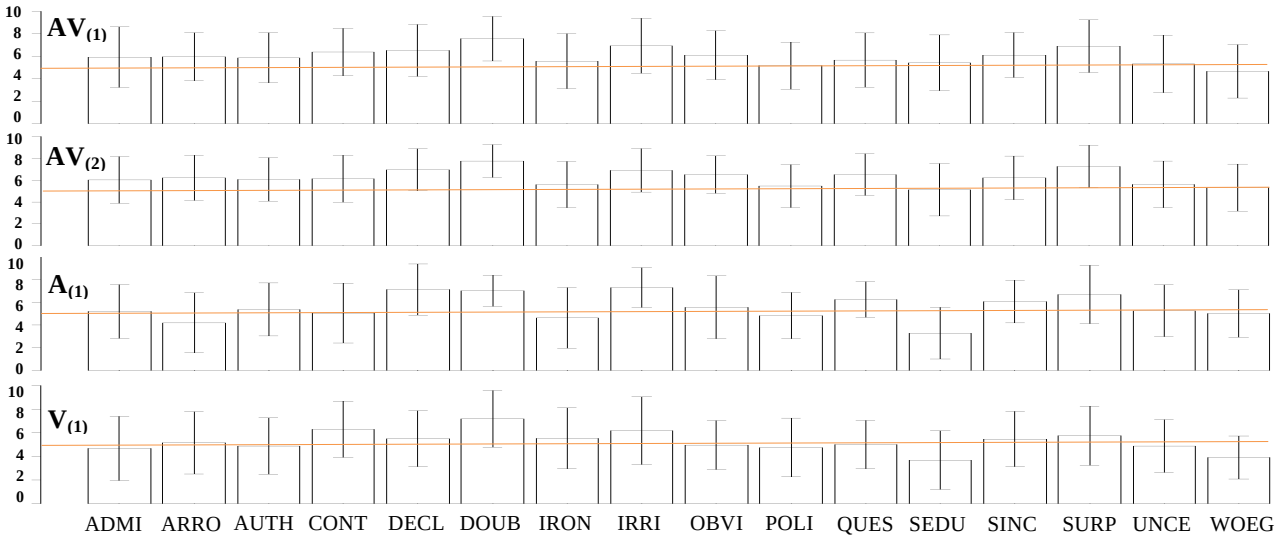

Figure 2: Mean judgment and s.d. of each attitude and for each modality $AV_{(1)}$, $AV_{(2)}$, $A_{(1)}$ and $V_{(1)}$of the 30 rater's

The judgments of the raters indicated a better performance of the speakers at the second trial in almost all cases of AV presentations (.05, $p<0.01$). In addition the standard deviation is lower for the speakers' performances at the second trial. The reason could be that the speakers knew the tasks after the first trial and on that account it was easier to prepare the expressions.

Table 2: Mean and s.d. of the judgment for the 10 speakers over the modalities $AV_{(1)}$, $AV_{(2)}$, $A_{(1)}$ and $V_{(1)}$

SP	$AV_{(1)}$ mean/sd	$AV_{(2)}$ mean/sd	$A_{(1)}$ mean/sd	$V_{(1)}$ mean/sd
01	6.61/2.35	7.17/1.84	4.93 /2.31	5.51/3.05
02	4.72/2.64	5.77/2.23	6.30/2.56	4.82/2.80
03	5.34/2.62	5.50/2.11	4.53/2.25	5.10/2.24
04	6.40/2.36	7.24/1.83	6.75/1.93	7.43/1.99
05	5.54/2.20	5.85/1.93	4.65/2.98	5.03/2.19
06	6.74/1.91	6.30/1.95	5.90/2.65	6.55/2.11
07	5.20/2.30	5.61/2.17	5.16/2.69	4.81/1.89
08	7.00/1.93	6.59/2.15	5.40/2.39	4.98/2.41
09	5.68/2.16	6.10/2.10	4.95/2.25	4.56/2.53
10	6.61/2.35	6.21/2.28	5.80/2.43	4.80/2.30

There are significant differences between the modalities. It is also significant that in most cases the AV presentation leads to a more reliable detection of the attitudes than presenting audio-only (A) or silent videos (V) (.16, $p<0.01$). Table 2 lists the average judgment and s.d. across the 10 speakers for the modalities $AV_{(1)}$, $AV_{(2)}$, $A_{(1)}$ and $V_{(1)}$.

4.2. Attitude analysis

Figure 2 presents the judgments of each attitude averaged over the 10 speakers and for modalities $AV_{(1)}$, $AV_{(2)}$, $A_{(1)}$ and $V_{(1)}$. It can be clearly seen that the judgment of the $AV_{(1)}$ and $AV_{(2)}$ recordings show no significant differences: listeners do answer reliably. That means the better performances of the speakers, seen in Table 2, did not lead to a better perception of the expressions at the end. The ANOVA test shows that the ratings significantly depend on the speaker and the attitude portrayed (.09, $p<0.01$). We found also significant differences between the audiovisual stimuli and the audio-only stimuli and silent videos (.17, $p<0.01$). The horizontal line in Figure 2 indicates the mid-point of the scale. Whereas all of the $AV_{(1+2)}$ recordings (with the exception of the attitude "walking on eggs" (WOEG) of the first trial) are rated in the upper half of the scale, some of the audio-only and visual-only stimuli are in the lower half of the scale. For example for $A_{(1)}$ stimuli the following means and standard deviations of judgments were calculated for "arrogance" (ARRO) 4.18 (2.64) and for "seductiveness" (SEDU) 3.27 (2.27), and the $V_{(1)}$ recordings for "admiration" were assessed by the raters with 4.68 (2.72) and for "walking on eggs" (WOEG) with 3.91 (1.82). However, there are also exceptions e.g. the attitudes "declarative sentence" (DECL) and "irritation" (IRRI) presented in the audio-only mode were judged better than in the audiovisual modality. Below the three most convincing and most implausible attitudes of non-silent videos across the 10 speakers (average judgment/s.d.): "doubt"$_{(1)}$: 7.56 (1.98), "irritation"$_{(1)}$: 6.93 (2.45)

and "surprise"$_{(2)}$: 7.25 (1.95), "walking on eggs"$_{(1)}$: 4.70 (2.37), "politeness"$_{(1)}$: 5.17 (2.10) and "seduction"$_{(2)}$: 3.69 (2.48). The ANOVA confirms the significance of the scores between the attitudes (.10, $p<0.01$).

We found no influence of the target sentence – either "Banane" or "Marie tanzte" – on the judgments. The average deviation (s.d.) across the attitudes is 0.35 (0.16). In contrast the average deviation (s.d.) between the $AV_{(1)}$ and $A_{(1)}$ is 0.69 (0.35) and between $AV_{(1)}$ and $V_{(1)}$ 0.76 (0.22).

5. Discussion and Conclusions

This paper presented the collection of a corpus of attitudes in German. So far the corpus contains A/V recordings by ten native subjects. Their performances were judged by native perceivers and judged as to the plausibility of the attitudes portrayed. Due to the small number of subjects assessed our results can only be preliminary. However, they permit to evaluate and if necessary discard samples from the database.

We found significant differences in the judgments of full A/V stimuli and the silent video/audio-only stimuli. This result shows that both the acoustic signal and the visual cues collude in the portrayal of attitudes. In some cases there were large differences in the performances of the speakers. Nevertheless all attitudes presented in the AV modality were judged on the upper half of the scale. This means that the mean performance for all the intended attitudes received good scores. We did not find any dependency of ratings on the target sentence "Marie tanzte" and "Banane".

The results of the current study are in line with those of an earlier one on American English [2] despite the fact that we only evaluated 10 speakers. A notable difference concern the expression of irony which received consistently the lowest performance scores for American English, but it is not the case in German. Could it be the case that the prosodic expression in German is more clearly marked in this language [4]. However, our study can be only seen as a first step. We intend to collect more subjects.

Nevertheless the performed attitudes constitute the basis for further analysis and cross-cultural comparisons. Prosodic features such as F0, intensity, voice quality and durations should be examined as correlates of the linguistic and paralinguistic information conveyed by attitudes which lead to a correct understanding between speaker and listener.

Acknowledgement

Great thanks go to the speaker for their performances and to all the students from the Beuth University of Applied Science who took part at the evaluation.

References

[1] Gussenhoven, C., "The Phonology of Tone and Intonation", Cambridge: Cambridge University Press, 2004.

[2] Léon, P.,"Précis de phonostylistique, Parole et expressivité," Paris: Nathan Université, 1993.

[3] Moraes, J. A., "The pitch accents in Brazilian Portuguese: Analysis by synthesis", in Proceedings of Speech Prosody 2008, Campinas, 389–397, 2008.

[4] Niebuhr, O., ""A little more ironic" – Voice quality and segmental reduction differences between sarcastic and neutral utterances, in Proceedings of Speech Prosody, Dublin, 608-612, 2014.

[5] 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.

[6] R Core Team, "R: A language and environment for statistical computing". R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/, 2012.

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

[8] 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

[9] Spencer-Oatey, H., "Reconsidering power and distance," Journal of Pragmatics 26: 1–24, 1996.

[10] Wierzbicka, A., "Defining emotion concepts," Cognitive Science 16: 539–581, 1992.

[11] www. livecode.com