Free Labeling of Audio-visual Attitudinal Expressions in German

Hansjörg Mixdorff[1], *Angelika Hönemann*[2], *Albert Rilliard*[3]

[1] Department of Computer Science and Media, Beuth University Berlin, Germany
[2] Faculty of Linguistics & Literary Studies, University of Bielefeld, Germany
[3] LIMSI-CNRS, Orsay, France

mixdorff@bht-berlin.de, ahoenemann@techfak.uni-bielefeld.de, Albert.Rilliard@limsi.fr

Abstract

This paper presents results from a free labeling experiment employing short audio-visual utterances of German produced with varying attitudinal expressions. Raters were asked to freely specify one single word to describe these. Words were classified with respect to emotional dimensions of valence, activation and dominance, as well as assertion/interrogation. As regards modality, video-supported stimuli yielded significantly higher dominance levels than audio-only ones. The main dimensions separating expressions are assertive vs. interrogation, valence, and dominance. The illocutionary strength is associated with the perceived activation, and primarily linked to the visual channel, while sentence mode is primarily conveyed by acoustic cues.

Index Terms: social attitudes, free labeling

1. Introduction

When two talkers converse they always convey information above and beyond pure linguistics, e.g. their mental state, emotions, mood or attitudes. This affective state is influenced, for instance, by the situation or roles of the dialog partners in the social hierarchy. People who share the same language or culture are therefore 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. A study investigated twelve social attitudes e.g. surprise, irritation, command-authority for prosodic effects in the languages British English, French and Japanese [1]. They found similarities across these languages, but also some culture-specific uses of prosodic parameters. The similarities may be explained within the framework of a theory such as the frequency code [2] which proposes the use of pitch level as a marker inverse to dominance. Other codes have been proposed [3] that may refine the predicted use of pitch for communicative purposes. Conversely, culture-specific uses have been documented [4]. 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 current work is based on the framework developed by [5] in which attitudes are characterized by a situational description of between whom and where they occur. Recordings also concern the visual channel, as facial gestures are known to be a vital part of attitudinal expressions [6].

Attitudes such as arrogance, politeness, doubt or irritation - see Table 1 for abbreviations henceforth used in this paper - were elicited 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 [7]. We then examined the acoustic-prosodic properties of the data and determined the respective differences between types of attitudes [8]. Finally, we ran an identification study in which we asked subjects to choose from a set of five labels the one they deemed most appropriate [9].

Both latter 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. Due to the experiences with the identification study we suspected that offering raters a set sub-group of labels introduces a strong bias. Therefore we decided to follow the approach by [10]. In this study, raters were free to select a single word, either a noun or adjective that best fit their impression of the attitudinal expression. Different from [10], we also included audio-only and video-only examples to test for differences in the modalities. Our methodology for evaluating the results is also slightly different.

2. Perception Study

We selected the stimuli for the study based on our earlier results regarding the performance of the speakers. Eventually stimuli of the best 15 speakers were included. Of these we chose those examples that had been rated best for a given attitude, yielding 6 stimuli for each attitude. In our previous work we found that there was no significant difference of the target utterances in the judgment of the raters therefore we decided to use only the utterance ‘Eine Banane’ to reduce the amount of 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 80 stimuli each.

Further developing the design adopted in [10], a presentation software was developed that included audio-visual, audio-only and video-only stimuli. 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. The ultimate task was to describe each of the stimuli with a single word, either a noun or adjective. As mentioned earlier, every subject had to rate 80 examples (48 AV, 16 AU, 16 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 students (30 male, 5 female) 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 2732 labels: 1631 for AV, 546 for AU and 595 for VI presented stimuli. Analysis of written expressions showed quite a variation of terms used, yielding a total number of 647 different tokens. Despite the instruction to use just a single word, some subjects had entered two or even a whole phrase to describe their impressions. Oftentimes 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 one, e.g. "akzeptierend" (*accepting*), "bestätigend" (*confirming)* and "zustimmend" (*approving*) were collapsed to the more frequent term "zustimmend". The term "fragend" was the most frequently chosen term (N=292), followed by "genervt" (*irritated*, N=149) and überzeugt (*convinced*, N=115) of the 127 non-neutrally perceived terms. Following are the top three expressions across the attitudes depending on the modality: AV: fragend (questioning), N=171 genervt (irritated), N=103, zweifelnd (doubting), N=74, AU: fragend, N=67, erstaunt (surprised), N=34, gelangweilt (bored), N=27, VI: fragend, N=54, entschlossen (determined), N=27, genervt, N=27. There were a small percentage of terms (2.43 %) 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 117 terms.

In order to further pull apart the linguistic and affective content of each normalized term and become more independent of the respective word identity for the ensuing analysis, we classified them following the scheme developed by [11][12][13]. In principle, we analyzed each term in the three-dimensional space of valence, activation and dominance and added to these the linguistic dimension of statement vs. interrogation. We restricted this classification to three possible values: negative, neutral and positive for valence and – , 0 and + for activation and dominance. Such a semantic classification [O] permits the analysis of the emotional and linguistic weight of each term with respect to its frequency of occurrence for a given rater and attitude without being tied to the original term. For a term such as "genervt" (*irritated*), for instance, we assigned negative valence, +activation and +dominance. Depending on the stimulus used to elicit the term in the given case we assigned the sentence mode, here statement. In order to calculate the position assumed by each attitude in the three-dimensional emotional space we mapped the three values onto a scale from -1 to +1 and averaged over all ratings for that given attitude.

4. Results of Analysis

Based on the frequency and semantic values of labels assigned to each attitude we yielded centers of gravity in the emotional space for each attitude. Table 1 lists the positions of all 16 attitudes in the emotional space for audio-visual stimuli. We can see, for instance, that CONT is judged more negatively than AUTH, while POLI has an almost neutral connotation. Based on these results we also compared the impact of reduced modalities on the assessment of attitudes. The result presented in

Figure *1* only concern the subset of utterances presented audio-visually, audio-only as well as video-only.

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

attitude	abbrev-iation	valence	activat-ion	domin-ance
admiration	ADMI	.5347	.7030	-.0198
arrogance	ARRO	-.2885	.4423	.4615
authority	AUTH	-.4078	.3398	.4369
contempt	CONT	-.6700	.6100	.6000
neutral statement	DECL	-.1456	.0194	.0000
doubt	DOUB	-.3462	.5096	-.3173
irony	IRON	.1053	.6737	.0842
irritation	IRRI	-.7767	.7961	.6893
obviousness	OBVI	-.3529	.5294	.3431
politeness	POLI	.0577	.1250	.1923
neutral question	QUES	-.1471	.0294	-.0294
seductiveness	SEDU	.6600	.6000	.1600
sincerity	SINC	.3564	.4455	.2277
surprise	SURP	-.0385	.4904	-.1731
uncertainty	UNCE	-.3725	.0686	-.2157
walking-on-eggs	WOEG	-.6117	-.0097	-.1553

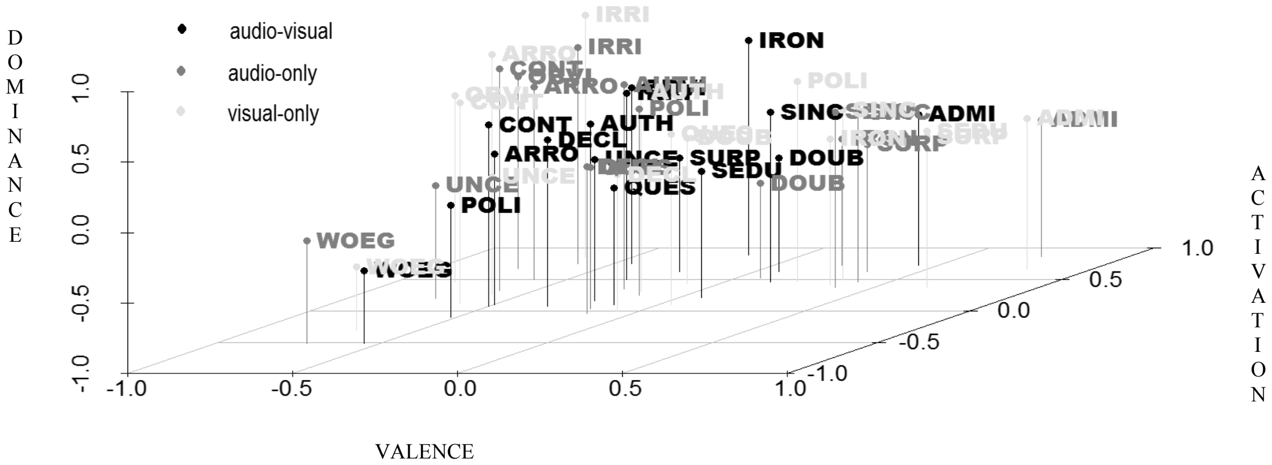

Figure 1: Position of attitudes in the emotional space, subset presented audio-visually, audio-only and video-only.

We will discuss the differences between modalities in detail later in this paper.

Normalized labels were organized in a contingency table with the presented stimuli's 16 categories, in each of the three presentation modalities in rows (i.e. rows present the expressive behavior of speakers, ordered by presentation modality), and the 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) [14]. The CA was run on the results obtained on the audio-visual modalities only with audio-only and video-only results used as supplementary individuals. The semantic classification of labels according to their valence, dominance, activation and linguistic mode were used as supplementary variables. An elbow criterion indicates to keep the first four dimensions (which explain 56% of the variance) of the CA for further analysis,. Table 2 reports 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. The second dimension relates to valence, while the third is linked to +activated labels. The fourth is related to expression with -activation.

The first dimension of the CA is mainly built on the expressions of AUTH, CONT, IRRI (on the assertive and +dominant side) and of DOUB, SURP, UNCE and the interrogative and –dominant side. Dimension 2 contrasts SEDU and ADMI (labeled with positive valence labels) to the others. The third dimension – linked to +activated labels – separates IRRI and CONT from the more neutrally perceived expressions of POLI and DECL. The fourth sets apart SEDU and WOEG - being -activated expressions - from OBVI, IRON and IRRI as non-minus activated expressions (but not necessarily +activated).

Table 2: Coordinates and cos^2 (cos^2 are multiplied by 100 and rounded for convenience) of the supplementary semantic categories of labels, on the first 4 dimensions of the CA.

coord.		Dim. 1	Dim. 2	Dim. 3	Dim. 4
Val.	neg	0.2	-0.3	0.3	0.0
	neu	0.0	-0.2	-0.4	0.1
	pos	-0.3	0.7	0.1	-0.1
Activ.	-	-0.1	-0.3	-0.4	0.4
	0	0.0	0.0	-0.4	0.0
	+	0.0	0.1	0.4	-0.1
Dom.	-	-0.5	-0.4	-0.1	0.1
	0	-0.2	0.2	-0.2	0.0
	+	0.6	-0.1	0.4	0.0
Ling.	Ass.	0.4	0.3	0.0	0.0
	Int.	-1.0	-0.8	0.0	0.1

cos^2		Dim. 1	Dim. 2	Dim. 3	Dim. 4
Val.	neg	9	**43**	24	0
	neu	1	10	**60**	2
	pos	12	**80**	1	2
Activ.	-	2	14	15	**16**
	0	1	1	**64**	0
	+	0	7	**79**	5
Dom.	-	**42**	26	1	1
	0	24	25	28	1
	+	**62**	3	28	0
Ling.	Ass.	**55**	33	0	1
	Int.	**55**	33	0	1

In order to reach a better representation of the multi-dimensional spread of expressions, we applied a hierarchical clustering on the distribution of rows obtained with the first four dimensions of the CA (cf. [14]). Results of this clustering summarize the observed spread of expressions in a 7-cluster solution. In the following we list these clusters with the most frequent labels (in decreasing order of importance, English translations given in italics), semantic connotation, and primarily associated attitudes and their modalities:

Cluster #1: fragend (*asking*), zweifelnd (*doubting*), erstaunt (*astounded*), überrascht (*surprised*), unwissend (*unknowing*); interrogation, –dominant ; DOUB (AU, AV, VI), QUES (AU, AV), SURP (AU, AV), UNCE (AU, AV, VI)

Cluster #2: fragend, unsicher (*unsure*), zurückhaltend (*restrained*), ängstlich (*afraid*), verachtend (*contemptuous*), enttäuscht (*disappointed*); interrogation, -activation,
-dominance and 0 valence; WOEG (AU, AV, VI)

Cluster #3: amüsiert (*amused*), erfreut (*pleased*), ironisch (*ironic*), fröhlich (*cheerful*), begeistert (*zealous*), schwärmend (*enthusiastic*), erzählend (*narrating*), erleichtert (*relieved*), verwirrt (*confused*); positive valence, +activation, assertion, and 0 dominance; SEDU (AU, AV)

Cluster #4: erregt (*aroused*), verführerisch (*seductive*), geheimnisvoll (*mysterious*), begeistert (*excited*), sinnlich (*sensual*), amüsiert, fröhlich, freundlich (friendly); positive valence, 0 dominance, assertion and +activation; ADMI (AU, AV, VI), IRON (AU, AV, VI), SEDU (VI), SURP (VI)

Cluster #5: arrogant (*arrogant*), überzeugt (*convinced*), offensichtlich (*obvious*), zustimmend (affirmative), abfällig (*condescending*), erschrocken (*scared*); assertion,+dominance; ARRO (AV, VI)

Cluster #6: neutral, feststellend (*ascertaining*), gelangweilt (*bored*), bestimmend (*determining*), berichtend (*reporting*); 0 activation, neutral valence, assertion, and 0 dominance; CONT (AV, VI), IRRI (AV, VI), OBVI (AU, AV, VI)

Cluster #7: genervt (*irritated*), aggressiv (*aggressive*), wütend (*furious*), verärgert (*angry*), entschlossen (*determined*), fordernd (*demanding*), autoritär (*authoritarian*); +dominance, negative valence, +activation, assertion; ARRO (AU), AUTH (AU, AV, VI), CONT (AU), DECL (AU, AV, VI), IRRI (AU), POLI (AU, AV, VI), QUES (VI), SINC (AU, AV, VI)

5. Discussion and Conclusions

Audio-only presentations of IRRI, ARRO and CONT (semantic impositions of the speaker on the interlocutor) did not pertain to the same cluster as their audio-visual counterparts. Video-only presentations of SEDU, SURP and QUES, and both mono-modal presentations of IRON are also judged differently than their respective audio-visual versions. Audio-only ARRO and CONT are judged less dominant than audio-visual presentations; audio-only performances of IRRI are perceived as less negative. In these three cases a reduction of perceived illocutionary strength is observed in the AU modality as compared to VI and AV presentations. Video-only SURP and QUES are not understood as interrogations – thus the absence of the acoustic channel makes it more difficult to decode the linguistic meaning (difficult, but not impossible, as video-only DOUB is classified correctly). VI-only SEDU conveys an interrogative meaning, but lacks the joyful and sexually-oriented aspects that are conveyed when the audio modality is present.

The description of IRON by the listeners is interesting, as it is a complex construct: it is described by [15] as a meaning contrasted by prosodic means. For English there does not seem to be one single reliable strategy to express irony. Rather, prosody is used to create a contrast where IRON is to be conveyed. It seems the contrast lies in a mismatch between modalities: Ironic single-modality performances are perceived as dominant (i.e. expressing one type of dominant expression that, however, is not ironic), whereas their combination conveys a positive valence. Irony emerges from this contrast between the interpretations of the two modalities, like a contradiction between modalities (cf. also [16]). Another noticeable fact from the interpretation of irony by German listeners is in contrast with similar tests in French and German [12][17]: the French linked it to obviousness, thus lacking the positive character observed here, and the Japanese rated irony solely as negative, something to be avoided in a conversation. This positive evaluation of irony also contrasts with the results obtained in the categorical perception test, where the "ironic" label was mixed with more negative labels, while the same performance are judged here positively: this bias between perceptual protocols may illustrate a difference between tests that focus on predefined concepts (categorical recognition) vs. tests that allow judging prosodic performances (like this free labeling one).

In contrast to [9], attitudes WOEG, ARRO and SEDU now occupy separate clusters indicating an unambiguous assignment of the labels. In contrast, in our previous experiment no attitude formed its own cluster and e.g. the positive attitude SEDU was mixed with negative attitudes such as ARRO, AUTH and CONT. SEDU and WOEG yielded very low recognition scores. In the previous work SURP and DOUB built one cluster but the same cluster also included POLI, OBVI and CONT which was not plausible. As can be seen in the current clustering SURP, DOUB are joined with the other interrogative attitudes UNCE and QUES. This suggests that free descriptions of attitudes yield more plausible classifications.

6. 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] Gussenhoven. C., *The Phonology of Tone and Intonation*, Cambridge: Cambridge University Press. 2004.

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

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

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

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

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

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

[10] 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.

[11] http://tschroeder.eu/weblog/?page_id=2, accessed on 5 February 2016.

[12] 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.

[13] Schröder, T., Hoey, J., & Rogers, K. B. (in print). Modeling dynamic identities and uncertainty in social interaction: Bayesian affect control theory. *American Sociological Review*.

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

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

[16] González-Fuente, S., Escandell-Vidal, V. & Prieto, P., "Gestural codas pave the way to the understanding of verbal irony," *Journal of Pragmatics*, 90, 26-47, 2015.

[17] M. Guerry, A. Rilliard, D. Erickson, T. Shochi, "Perception of prosodic social affects in Japanese: a free-labeling study", In Proc. Speech Prosody 2016, Boston, 811-815, 2016.