Facial activity of attitudinal speech in German

Angelika Hönemann[1,2], *Petra Wagner*[1]

[1]Bielefeld University, Bielefeld
[2]Beuth University of Applied Science, Berlin
ahoenemann@beuth-hochschule.de, pwagner@uni-bielefeld.de

Abstract

The current study is a continuation of previous work on how attitudes are expressed in speech. Generally, we found that the availability of facial expression (together with speech) leads to a higher plausibility of an expressed attitude. Therefore, a further analysis of facial cues in the expression of attitudinal states appears useful. For this purpose, we selected a subset of sixteen attitudes portrayed by five performers, who have shown to be most convincing in their portrayal of attitudinal states. The Facial Action Coding System (FACS) is then applied on the output of OpenFace, which automatically detects the presence and the intensity of Action Units using a computer vision algorithm. Based on detected Action Units (Aus), a cluster analysis on one representative frame for each video is carried out. Furthermore, a linear discriminant analysis of the clusters is carried out including the whole dataset. The analysis shows that based on two linear discriminants, no clear separation of the cluster groups is possible. Some of the results for the facial expressions can be explained by their corresponding attitudes' distribution in emotional space. Not surprisingly, we also found that the facial expressions are strongly influenced by their phonetic content, thus, the attitudinal expression related to lip movements may be difficult to detect in speech.

Index Terms: German attitudes, attitudinal activity, visual cues, modalities, FACS

1. Introduction

Attitudes are indispensable for the understanding of human behavior in social communication. Both positively connotated attitudes such as *politeness* and negatively connotated attitudes such as *arrogance* are ubiquitous in everyday social communication. In contrast to emotions, attitudes provide information about the cognitive appraisal of a situation (or an object) [4]. Thus, the appropriateness of attitudes depends on its contextual embedding and its expression and perception is likely to be more complex than the expression of emotions.

The decoding and encoding of attitudinal speech is a field of research that has recently been gaining in interest in speech research [7,10,14,16]. We know that attitudes can be expressed both with the voice and with the help of visual cues such as facial or gestural expression. The correct interpretation of attitudinal expression becomes more difficult with increasing linguistic and cultural distance between interlocutors: [15] investigated twelve attitudes such as *surprise*, *irritation* and *command-authority* with respect to their prosodic characteristics. The result shows some cross-linguistic similarities, but cross-cultural differences.

To this day, we know little about the precise acoustic and visual cues that lead to a correct interpretation of intended attitudinal expression, i.e. which among a large number of potential cues are essential for attitudinal expression. In earlier studies we investigated sixteen attitudes, portrayed by sixteen native speakers of German. These attitudes were analyzed with respect to their individual acoustic expression. We also analyzed their modality-specific recognition by comparing the recognition based on audio-visual, visual-only and audio-only material. Our results show that attitudes that are expressed audio-visually are more plausible or convincing compared to attitudes expressed only in the acoustic or the visual modality. Thus, the lack of both acoustic and visual cues may lead to a misinterpretation of attitudinal expression [13, 8]. We also found that participants often confuse attitudes which are similar with respect to their linguistic structure (interrogative or declarative), or which have a similar allocation of attributes within a multidimensional emotional space with the dimensions of *valence*, *dominance* and *activation*: The attitudes *doubt* and *surprise* tend to be confused because of their inherent interrogative characteristics, while *arrogance*, *sincerity* and *authority* are often taken for a *neutral declarative* expression due to their common lack of emotionality [9].

The analyzed data is a subset of the material used in previous studies and is comprised of those speakers whose attitudinal expressions yielded the most convincing results. As we used an extended data set compared to our previous studies, we first confirm our prior results, that audio-visually presented attitudes are generally more convincing that those presented only via audio or video (section 2). In a next step, we extract visual cues used by the performers in the recordings (section 3) and analyze these facial expressions further (section 4), using a Correspondence Analysis (CA) / Hierarchical Cluster Analysis (HCPC) as well as a Linear Discriminant Analysis (LDA), thereby getting a detailed picture of how the various attitudes correspond to Action Units of facial expression. This is followed by a discussion and conclusion (section 5).

2. Modalities and attitudinal perception

This section describes the analysis of attitudinal display with respect to the role of the visual and acoustic modality. In a rating task sixteen attitudes (Table 2) portrayed by ten performers were judged by 30 participants. The participants were asked to judge how convincingly a target attitude was expressed based on stimuli presented either as audio-only, visual-only or audio-visual portrayals of attitudes, using a 9-point Likert scale [8]. For this purpose, we extended our dataset with six additional performers, whose performances were rated with the same paradigm [8, 14]. For each modality, 412 judgments were analyzed. Figure 1 displays the judgments for the three modalities (audio-visual: *M=6.3, SD=2.4*, visual-only: *M=5.8, SD=2.3*, audio-only: *M=5.9, SD=2.5*).

A Kruskal-Wallis test confirms previous results, that modality of presentation has an impact on how well an attitude is portrayed *(H=23.18, df=2, p<0.001).* The post hoc analysis (Dunn) shows that an audio-visually presented performance leads to a more convincing perception in contrast to an audio-only *(z=-4.61, p<0.001)* or a visual-only *(z=3.49, p<0.001)* presentation. No significant differences between audio-only and visual-only judged stimuli were found *(z=-1.12, p<0.13).*

However, perception also depends on the attitudes displayed. We detected attitudes which are perceived as more convincing when presented in audio-only, e.g. *authority*, or visual-only, e.g. *obviousness* and p*oliteness*. *Arrogance, authority, contempt* and *uncertainty* receive almost equal ratings between stimuli presented audio-visually and visual-only, while the attitudes *walking-on-eggs* and *politeness* are perceived as equally convincing when comparing audio-visual and audio-only presentations.

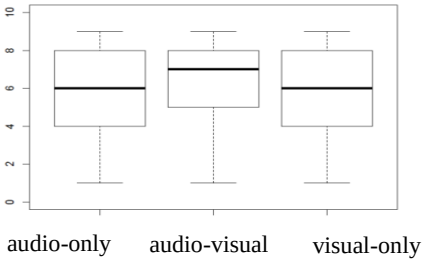

Figure 1: *Judgments (N=412) of audio-only, audio-visual and visual-only*

3. Analysis of visual cues

In order to quantify facial expressions, we applied the Open Source Tool OpenFace which contains three different modules: (1) facial landmarks detection, (2) head pose and eye-gaze estimation, (3) recognition of facial action unit [3]. OpenFace uses the recently proposed Conditional Local Neural Fields (CLNF) for facial landmark detection and tracking. CLNF is an instance of a Constrained Local Model (CLM) [4, 6], that uses more advanced patch experts and optimization functions. The two main components of CLNF are a (1) Point Distribution Model (PDM) which captures landmark shape variations and (2) patch experts which capture local appearance variations of each landmark [2, 3].

3.1 Data corpus

For the visual analysis we selected a subset from our total dataset, consisting of 80 videos presenting sixteen attitudes portrayed by the five performers while they produce the utterance "Marie tanzte". In previous studies, we found that the perception of how convincing an attitude comes across, is independent of the sentence produced [8]. The five performers selected for our study have been shown to produce the most convincing portrayals of the sixteen attitudes investigated. The average convincingness of the performers is listed in Table 1 and the attitudes in Table 2 (descending order). Table 2 also lists the abbreviations for each attitude which will be used henceforth to refer to the various attitudes in question.

The average rating across the various attitudes ranges between 5.2 and 7.7. The attitudes DOUB, IRRI and SURP were most convincing, while IRON, WOEG and SEDU were judged as less plausible. WOEG was the attitude most difficult to perceive. This may be partly explicable by the lack of this attitudinal concept in German culture, which may have caused a less coherent expressive behavior by the performers and a worse recognition by the raters.

Table 1: *Mean and sd. of ratings for the performer presented audio-visual and the utterance 'Marie tanzte'. The bold written performers are analyzed in the current visual study*

perf	Mean	s.d.	Perf	mean	s.d.
S01	**7.38**	**1.86**	S14	6.45	2.32
S08	**7.32**	**1.83**	S11	6.21	2.14
S04	**7.21**	**1.87**	S16	6.06	2.35
S10	**7.09**	**1.68**	S05	6.02	1.94
S06	**7.05**	**1.80**	S07	5.92	2.01
S13	6.53	2.05	S02	5.81	2.27
S18	6.47	2.13	S20	5.71	2.10
S09	6.45	1.89	S03	5.69	2.33

Table 2: *Description and short term of the sixteen attitudes as well as mean and s.d. of ratings for the audio-visual presented stimuli*

Attitude	short term	mean	s.d.
doubt	DOUB	7.68	1.64
Surprise	SURP	7.63	1.55
Irritation	IRRI	7.09	1.99
neutral statement	DECL	6.88	1.91
obviousness	OBVI	6.75	1.81
Contempt	CONT	6.63	2.06
Arrogant	ARRO	6.59	2.10
sincerity	SINC	6.54	1.74
Admiration	ADMI	6.45	2.20
authority	AUTH	6.38	2.08
neutral question	QUES	6.26	2.11
Uncertainty	UNCE	6.17	2.28
politeness	POLI	5.88	1.79
Irony	IRON	5.84	2.26
walking-on-eggs	WOEG	5.37	2.09
Seductiveness	SEDU	5.24	2.43

3.2 Visual cues estimation

Before the extraction of visual features we cut our videos to 41 frames for each video, to get a comparable number of frames. Additionally some videos were corrected in their quality such as brightness and/or contrast to guarantee an optimal detection. OpenFace then uses an internal 3D projection of the facial points (CLNF) to compute head position (translation and orientation) [6, 7]. The blue rectangles in Figure 2 show the 3D representation we extracted from our videos. Furthermore, we extracted 68 facial points as displayed in Figure 2 likewise (blue dots) which are not analyzed in this paper.

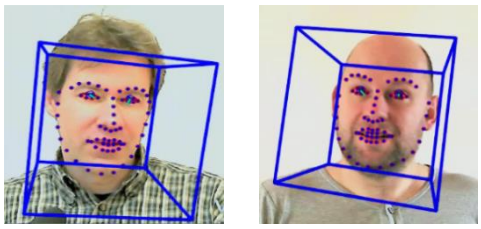

Figure 2: *3D representation of the head as well as detected facial point of two performers*

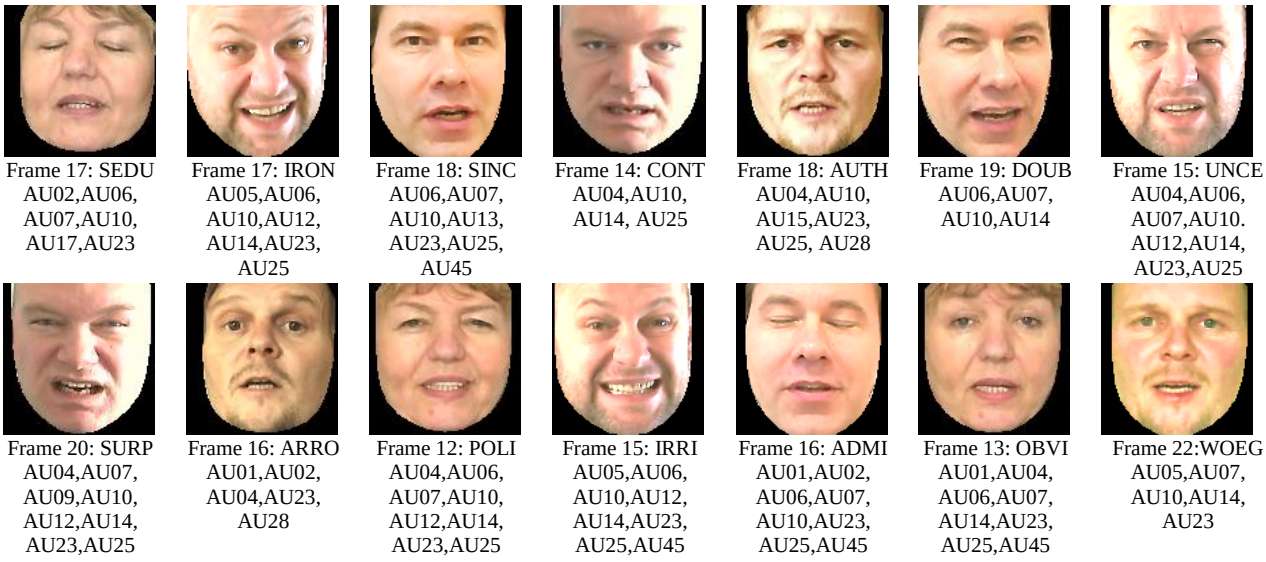

Figure 3: *Aligned faces and detected action units (AU) of the sixteen attitudes SEDU,IRON, SINC, CONT, AUTH, DOUB, UNCE, SURP, ARRO, POLI, IRRI, ADM, OBVII and WOEG (from left to right/ top to bottom) portrayed by five performers. One frame (stressed phoneme /a/) is presented*

The detection of the Action Units is based on the appearance as displayed by Histograms of Oriented Gradients (HOGs) [3] and facial features geometry. The algorithm estimates both the presence as well as the intensity of an Action Unit by using Support Vector Regression (SVR). The cropped faces presented in Figure 3 deliver the base to generate the Histograms of Oriented Gradients by doing a transformation from the detected landmarks to a frontal face representation from a neutral expression. The Action Units are based on their underlying facial expression according to Ekman [5] and are described in Table 3.

Table 3: *Description of each action unit*

Point	Description
AU1	Inner Brow Raisers
AU2	Outer Brow Raiser
AU4	Brow Lowerer
AU5	Upper Lid Raise
AU6	Cheek Raiser
AU7	Lid Tightener
AU9	Nose Wrinkler
AU10	Upper Lip Raiser
AU12	Lip Corner Puller
AU14	Dimpler
AU15	Lip Corner Depressor
AU17	Chin Raiser
AU20	Lip stretcher
AU23	Lip Tightener
AU25	Lips part
AU26	Jaw Drop
AU28	Lip Suck
AU45	Blink

4. Results

The frame extraction for each video yielded a total of 3286 frames which are included in the further analyses. To get a first overview of AUs involved in the facial expression of the individual attitudes, we examined the results of a single frame, taken from the production of the stressed vowel /a/ of each performer and attitude.

4.1 Single frame detection

We selected a single frame during the production of the stressed vowel /a/ as part of the word "t/**a**/nzte" for each performer and attitude, since most facial motions occur at prominent vowels [11, 17]. For various videos, facial expressions and corresponding AUs detected for this frame are presented in Figure 3. The activation intensities across AUs and corresponding attitudes are shown in Figure 4.

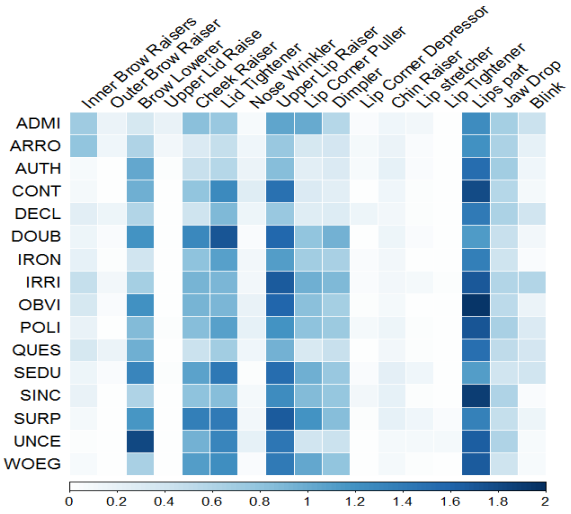

Figure 4: *Intensity of AUs of the 16 attitudes for one frame (stressed phoneme /a/)*

It is not surprising that AU25 and AU26 are detected for all attitudes at a similar activity level as the performers produce an open vowel. AU25 is detected with an overall high intensity (M=1.5) but AU26 with an overall lower intensity

(M=0.5). ADMI and ARRO show the highest intensity for the raising of the inner brows (AU01=0.72/0.79) in contrast to the other attitudes. There is a generally high activation level of the AU04 (M=0.89) in which UNCE shows the highest activity (AU04=1.8). AU06 (M=0.83), AU07 (M=1.03), AU10 (M=1.26), AU12 (M=0.69) and AU14 (M=0.60) are detected for all attitudes, but with unique levels of activation intensity. AUs for which we generally found only low activation are AU15 (M=0.05), AU17 (M=0.14), AU20 (M=0.05) and AU23 (M=0.01).

4.2 Clustering

We carried out a correspondence analysis (CA) with the single frame data described previously as input. The data is described by the first dimension explaining 44.44% and by the second one explaining 22.83% of the variance. The results were then used in a three cluster hierarchical cluster analysis (HCPC). Table 4 gives an overview of the mean activation intensity for each cluster and AU.

Table 4: *Mean of each AU for each cluster*

	CL1	CL2	CL3
AU01	0.068	0.205	0.526
AU02	0.003	0.035	0.152
AU04	1.265	0.895	0.632
AU05	0.012	0.009	0.084
AU06	0.734	1.019	0.493
AU07	1.054	1.180	0.695
AU09	0.212	0.105	0.116
AU10	1.263	1.442	0.879
AU12	0.320	0.906	0.491
AU14	0.329	0.784	0.421
AU15	0.034	0.048	0.076
AU17	0.145	0.148	0.131
AU20	0.054	0.059	0.054
AU23	0.034	0.011	0.000
AU25	1.661	1.540	1.348
AU26	0.626	0.500	0.606
AU45	0.095	0.206	0.346

The first cluster (CL1) contains the attitudes CONT, UNCE and AUTH which show a high activity (mean > 1.0) for AU04, AU07, AU10 and AU25. The second cluster (CL2) comprises the largest number of attitudes (IRON, POLI, OBVI, WOEG, SURP, DOUB, SINC, SEDU, IRRI). This cluster is built by AU06, AU07, AU10 and AU25 (mean > 1.0). The third (CL3) one includes the neutral attitudes QUES, DECL as well as ADMI and ARRO and shows the highest activation intensity for lip part motion AU25 (mean > 1.3).

The overall mean intensity for the cluster CL1 which is dominated (lowest distance to the cluster center) by the attitude CONT is 0.465, for the cluster CL2, dominated by the attitude IRON, it is 0.535, and for the cluster CL3 which is dominated by the attitude QUES it is 0.415. The mean activity across the AUs and cluster is 0.572.

4.3 Linear Discriminant Analysis

The three clusters (henceforth referred to groups GrA, GrB and GrC) are taken from the HCPC analysis to carry out a Linear Discriminant Analysis (LDA) on the whole dataset. The proportionate mean intensity of activation for each group is shown in Figure 5. The overall intensity means for each group are: GrA: 0.367, GrB: 0.431, GrC: 0.338. In comparison to the cluster means of the single frame HCPC analysis (Table 3) there exists no difference in the proportion of activity between the groups, however the overall mean across groups and activity is reduced to 0.379.

The attitudes CONT, UNCE and AUTH of the group GrA show the highest activation level for the 'Brow Lowerer'-activity AU04 (M=1.155) but also for AU09, AU15, AU17, AU23, AU25 and AU26. The attitudes of GrB, dominated by the expression IRON show the highest activation for the 'Lid Tightener'-activity AU07 (M=1.024) and 'Upper Lip Raiser '-activity AU10 (M=1.115) but also for AU06, AU12, AU14, AU15, AU17, AU23, AU25, AU26 and AU45.

GrC which includes the neutral declarative attitudes DECL and QUES show the highest activity of all groups for the 'Inner Brow Raisers'-activity AU01 (M=0.203), but even more activity for AU04, the 'Brow Lowerer'-activation (M=0.633). Further activations are shown in 'Lid Tightener'-activity AU07 (M= 0.696), 'Upper Lip Raiser'-activity AU10 (M=0.803), AU14 and AU25. GrC shows overall lowest intensities for almost all AUs. It is not surprising that AUs evoked by vowel production, e.g. depressing (AU15) and stretching (AU20) of the lips or the tightening of the chin (AU23) show similar intensities across groups. Similarly, only small differences between groups are found for the opening of lips (AU25).

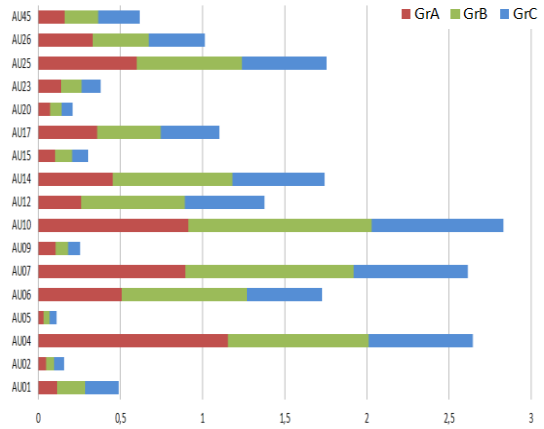

Figure 5: *Mean activation intensity for each group GrA, GrB, GrC*

Figure 6 shows histograms of the linear discriminant coefficient calculated for the LD1 and LD2. It is evident that the separation between groups is not entirely clear. The percentage of separation achieved by each discriminant function is 55% for the first discriminant and 45% for the second discriminant.

We can see from the histogram for the LD1 that GrA and GrB show high overlaps from the first discriminant function, particularly between -1 and 1, while the overlaps with the GrC are less strong. Here the values lie between 0 and 3. The histograms for LD2 show a high overlap between GrB and GrC, even in the region from -1 to 1 while GrA show the highest values around 1 and 2. Therefore groups GrA/GrB and GrC are separated mostly by the first discriminant, while groups GrA and GrB/GrC are separated by the second discriminant. Still, there are considerable overlaps between the groups, illustrated in Figure 7.

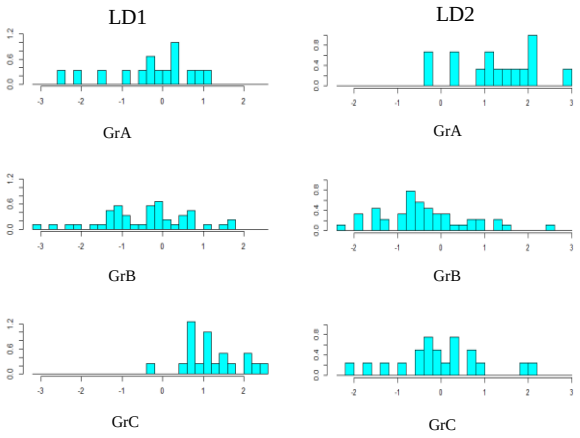

Figure 6: *Histograms of the linear discriminant coefficient calculated each group for the LD1 and LD2*

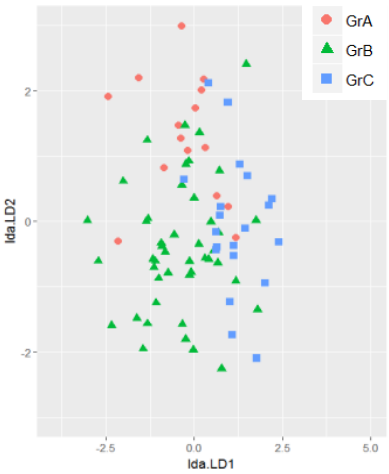

Figure 7: *LDA distribution calculated for each group*

5. Discussion and Conclusions

The result of our analysis on how the convincingness of attitudinal displays depends on the modality of their presentation confirms previous work in that the combined presentation of acoustic and visual cues leads to a better plausibility of their expression. The audio-only as well as the visual only presentations were overall less convincing to the listeners. Thus, our further analyses tried to shed light on the question on *which* facial cues are responsible for the perception of attitudes in speech.

The three clusters we extracted from the HCPC based on the visual cues of a single frame show plausible results. The results can be explained by the attitude's emotional component expressed in a multidimensional space, i.e. their inherent valence, dominance and activation level. The distributions of the attitudes also make sense in the light of an earlier evaluation study, where participants described the sixteen attitudes without predefined labels [12].

The first cluster GrA including CONT, UNCE and AUTH is plausible because these attitudes are expressed with a controlled level of activation. All three attitudes convey a negative valence and both CONT and AUTH express a high degree of dominance. This dominance is expressed through a 'brow lowerer'- and 'nose wrinkle-' activation. The attitude UNCE shows these activations as well but probably rather due to its interrogative, doubtful character. As shown in Figure 3 UNCE can be confused both with CONT and AUTH. Furthermore, CONT, AUTH and UNCE show almost the same ratings for stimuli presented audio-visually or only visually, thus their visual expression seems sufficient for them to be convincing. However, AUTH seems to be largely expressed by the voice, as it received better ratings in its audio-only presentation.

Despite their completely opposite linguistic character (assertion vs. interrogation) DECL and QUES are classified in one cluster. At first glance, this looks surprising but we can assume that these linguistic differences are mostly expressed acoustically, using a falling (DECL) or rising (QUES) fundamental frequency contour, but that both attitudes show similar activation levels [13]. Still, we find that the interrogative character of QUES is accompanied by a higher degree of 'brow lowerer'- activity. ADMI and ARRO, which are classified in the same cluster, have opposite emotional valence and dominance levels: ADMI is commonly perceived as having a positive valence but low dominance, while ARRO has negative valence and high dominance. Still, both attitudes have low activation [12]. So, the low level of activation may be a reason for these attitudes being clustered together, as well as their comparatively low level of convincingness (Table 1). The third cluster GrC shows the least overall activation of all three clusters, possibly reflecting the low activation expressed by their corresponding attitudes.

The highest degree of overall facial activation is detected for the second cluster GrB which consists of the remaining attitudes such as IRON, POLI, WOEG, SURP and IRRI.

Across clusters, little activation is found for the 'outer brow raiser', 'upper lid raiser' and 'lip stretcher'. As the examined single frame examined was extracted from the open, unrounded vowel /a/, this is not surprising and may have caused articulatory needs to overrule facial display of attitudinal expression. The discriminant functions we got from the discriminant analysis based on entire sentence productions do not separate the clusters properly. The greatest overlap is between the GrA and both remaining groups. GrA shares the 'brow lowerer'-activity and the 'blink'-activity with GrC, but most of the lip motions with GrB. Before further detailed analyses are available for the facial expression of attitudes, it is difficult to interpret these results.

It is important to note that the mean intensities of AU activations were very similar between the single frame analysis and those based on entire sentences. The selection of a frame in the vowel of the accented syllable has shown to be a good representative of the dynamic facial movements carried out to transport the expression of a particular attitude. To some degree, this confirms previous findings showing that attitudinal expression concentrates on accented syllables and that speech and co-speech gestures tend to align strongly in accented syllables [11]. However, this may also be a side effect of the phonetic structure of the investigated sentence *"Marie tanzte"*, in which the vowel /a/ occurs twice.

Overall, we shed some light on how attitudinal speech is facially encoded and decoded by native listeners. We detected similarities between some attitudes that are perceived in a similar manner, or share similarities in their distribution in a

multidimensional emotional space. Taken together, our findings are compatible with the assumption that at least when with respect to their facial expression, different attitudinal expressions should be regarded as attitudinal groups rather than uniquely displayed singular attitudes. Not surprisingly, we also found that the facial expressions are strongly influenced by their phonetic content, thus, the attitudinal expression of lip movements may be difficult to detect in speech.

6. References

[1] Baltrusaitis, T., Morency, L.-P. and Robinson, P., Constrained local neural fields for robust facial landmark detection in the wild. In ICCVW, 2013

[2] Baltrusaitis, T., Morency, L.-P. and Robinson, P., Cross-dataset learning and person-specic normalisation for automatic Action Unit detection. In Facial Expression Recognition and Analysis Challenge, in conjunction with FG, 2015.

[3] Baltrušaitis, Tadas, Peter Robinson, and Louis-Philippe Morency. "Openface: an open source facial behavior analysis toolkit." *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, 2016.

[4] Cristinacce D. and Cootes T. Feature detection and tracking with constrained local models. In BMVC, 2006

[5] Ekman P. and Friesen W. V. Manual for the Facial Action Coding System. Palo Alto: Consulting Psychologists Press, 1977.

[6] Felzenszwalb, P. F., Girshick, R. B., McAllester, D. and Ramanan, D. Object Detection with Discriminative Trained Part Based Models. IEEE TPAMI, 32, 2010.

[7] Fazio,R.H and Olson, M.A. Attitudes: Foundations, Functions, and Consequences. M.A. Hogg & J. Cooper. The SAGE Handbook of Social Psychology (pp. 139-160), London: Sage, 2003

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

[9] Hönemann, A., Rilliard A., Mixdorff, H., Classification of Auditory-Visual Attitudes in German, FAAVSP - The 1st Joint Conference on Facial Analysis, Animation and Auditory-Visual Speech Processing. Vienna, Austria 2015

[10] Kim J., Davis, C. The Consistency and Stability of Acoustic and Visual Cues for Different Prosodic Attitudes. Proc. Interspeech 2016, 57-61.

[11] Krahmer, E., and Swerts, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses auditory perception and visual perception. *Journal of Memory and Language, 57*(3), 396–414.

[12] Mixdorff, H., Hönemann, A., Rilliard, A., Free Labeling of Audio-visual Attitudinal Expressions in German. SST 2016, Parramatta City, New South Wales, Australia, 2016

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

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

[15] Rilliard, A., Shochi, T. Martin, JC., Erickson, E., Aubergé V., Multimodal indices to Japanese and French prosodically expressed social affects - Language and speech, 2009

[16] Valstar, M., Girard, J., Almaev, T., McKeown, G., Mehu, M., Pantic, L. Yin, M. and J. Cohn. FERA 2015 - Second Facial Expression Recognition and Analysis Challenge. In IEEE FG, 2015.

[17] Wagner P, Malisz Z, Kopp S. Gesture and Speech in Interaction: An Overview. *Speech Communication*. 2014;57 (Special Iss.):209-232.