An audiovisual political speech analysis incorporating eye-tracking and perception data

Stefan Scherer[1], Georg Layher[2], John Kane[3], Heiko Neumann[2], Nick Campbell[3]

[1]Institute for Creative Technologies
University of Southern California
Playa Vista, CA, USA
scherer@ict.usc.edu

[2]Institute of Neural Information Processing
University of Ulm
89069 Ulm, Germany
firstname.lastname@uni-ulm.de

[3]Centre for Language and Commuication Studies
Trinity College Dublin
Dublin 2, Ireland
{kanejo, nick}@tcd.ie

Abstract

We investigate the influence of audiovisual features on the perception of speaking style and performance of politicians, utilizing a large publicly available dataset of German parliament recordings. We conduct a human perception experiment involving eye-tracker data to evaluate human ratings as well as behavior in two separate conditions, i.e. audiovisual and video only. The ratings are evaluated on a five dimensional scale comprising measures of insecurity, monotony, expressiveness, persuasiveness, and overall performance. Further, they are statistically analyzed and put into context in a multimodal feature analysis, involving measures of prosody, voice quality and motion energy. The analysis reveals several statistically significant features, such as pause timing, voice quality measures and motion energy, that highly positively or negatively correlate with certain human ratings of speaking style. Additionally, we compare the gaze behavior of the human subjects to evaluate saliency regions in the multimodal and visual only conditions. The eye-tracking analysis reveals significant changes in the gaze behavior of the human subjects; participants reduce their focus of attention in the audiovisual condition mainly to the region of the face of the politician and scan the upper body, including hands and arms, in the video only condition.

Keywords: political speeches, speaking style, statistical measurement, speech analysis, audiovisual, eye-tracking

1. Introduction

Humans use a variety of apparent communicative features in order to judge social aspects of behavior. Either in direct interactions or during observation in passive roles observers quickly rate other people's personality and trustworthiness utilizing verbal and non-verbal cues (Argyle, 1975; Grammer et al., 2002). For example, in intergroup communication speakers' attitudes are signaled verbally and nonverbally (Gallois and Callan, 1988), while the leadership and affective strength of an actor are judged on the basis of non-verbal cues in situations of listening to announcements and speech (Albright et al., 1988). It has been demonstrated that motion cues are a rich source of communicating nonverbal information which is perceived and reliably interpreted by observers in a context-dependent fashion (Koppenheimer and Grammer, 2010; Grammer et al., 2002). Such judgements can be accomplished even with impoverished visual motion signals, as in point-light displays, still allowing the perception of emotion cues in interpersonal dialogue situations (Clarke et al., 2005). More recently, it has been shown that observers selectively sample information of motion cues in point-light displays using eye-movements with gaze patterns depending on the particular task (Saunders et al., 2010). Taken together, this demonstrates that active communicators and listeners make use of various cues to encode and decode communicative signals and actively search for the presence of specific hints for socially relevant stimuli.

While in the latter reported eye-movement task the analysis focuses on one modality only, it remains unclear whether and how specific eye-movement patterns also vary when multimodal information provides verbal and non-verbal signals. In this study, we present details of an analysis of audio and visual factors and features of political speaking styles that correlate with human perceptual evaluations. We produced several audiovisual features of political speeches from little-known speakers of the German parliament and performed a statistical analysis of eye-tracking data and perceptual ratings from seven naive subjects on this data. We compared audiovisual features related to the perception ratings and analyzed the gaze behavior in both video-only and multimodal conditions (see Section 2.).

The audio features comprise prosodic parameters such as articulation rate, pitch range, voice quality parameters, intensity measures, and speech timing that were subject of the analysis in related work (Rosenberg and Hirschberg, 2005; Strangert and Gustafson, 2008), except for the voice quality parameters. As a basic but nevertheless meaningful visual feature the relative motion energy contained within a sequence of a speaker was used (much in the spirit to consider movement quality analysis as suggested by (Grammer et al., 2002)). Unlike the approach of (Koppenheimer and Grammer, 2010), we refrained from using complex high level visual features, such as the characteristics of an estimated geometric body model or the trajectory of the hand, since most of them are difficult to obtain under unrestricted realistic conditions. Gaze behavior was analyzed using the relative time a subject was fixating a specific body part (in this case the face) of a speaker as an indicator on the influence of that body part on the perception of a speaker's qualities.

The remainder of this paper is organized as follows: in Section 2. the data and the experimental setup are introduced. Section 3. reports key findings in the various statistical tests and Section 4. summarizes the results and presents future work. The paper concludes with a discussion of potential real-world applications of this work.

2. Data and Perception Experiment

The stimuli for the perception study were taken from three individual plenary sessions (i.e. earthquake in Japan (March 17th 2011), adjustments within the organization of the German armed forces (May 27th 2011), and the plagiarism scandal of the defense minister (February 23rd 2011)) of the German parliament[1]. We chose 40 sequences by eight different rather little-known speakers (four female, four male) of 10-20 seconds in length (average: 16.68s, variance: 4.35s). Each speaker is represented in five different sequences (exemplary pictures are shown in Fig. 1). Two separate experimental runs are conducted in two sessions. For each subject one randomly chosen half of the stimuli is presented as video only and the other half audiovisual. In the second session the former video only half is presented audiovisual and vice versa. Within the runs the stimuli's order of presentation is as well randomized. The subject's position is fixed at a distance of 50 cm by a stationary eye-tracker (SMI iView X^{TM}Hi-Speed) to precisely record their gaze direction. The stimuli are presented on a flat 20.1 inch LCD Display (Dell 2001FP). Each stimulus stands still for three seconds in order to ground the subjects. After the stimulus is presented, a dialog appears where the subject has to answer five questions on a five point Likert-scale (from absolutely disagree to absolutely agree) using the mouse. The questions posed to the subject were shown to be reliable in previous studies (Rosenberg and Hirschberg, 2005; Strangert and Gustafson, 2008), namely "The speaker is ..."

- "... insecure."
- "... monotonous."
- "... expressive."
- "... persuasive."
- "... overall a great speaker that is capable of capturing the attention of an audience."

In the presented study we recorded the eye-movements and acquired the subjective speaker ratings of seven subjects (two female, five male; with an average age of 24). Currently, effort is undertaken to record a larger cohort of subjects.

3. Evaluation

In order to evaluate the influence of the speakers' behavior on their rating we conducted multiple statistical tests with two basic foci, i.e. the disparity in the speakers' perception with and without audio using the eye-tracking data, and the influence of acoustic prosodic measures, as well as basic visual features on the perception of the speakers' qualities. The results and evaluations of the investigated foci will be covered in subsections 3.1., 3.2. and 3.3.

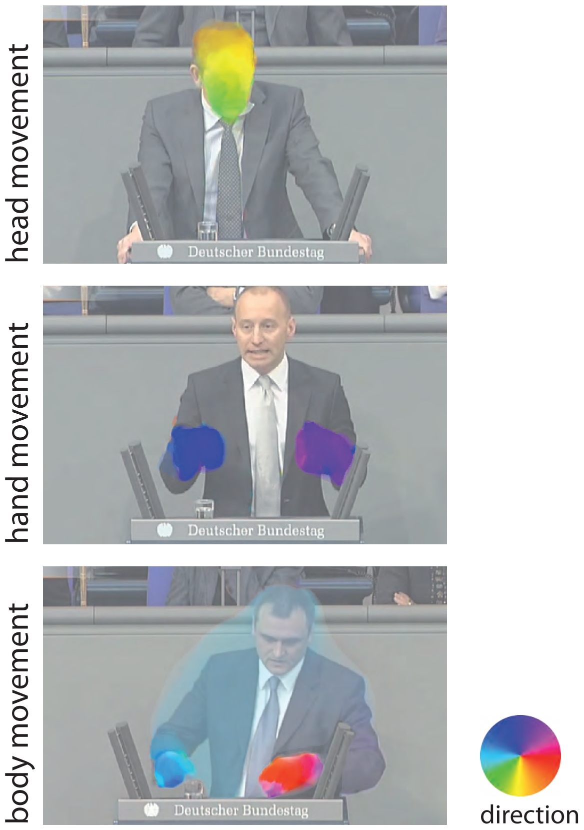

Figure 1: Exemplary pictures of three different speakers contained in the used dataset. Regions with strong optical flow changes are shown in differently colored blobs, where the color encodes the direction of the movement. The reference for the directions is shown in the rainbow circle at the bottom right.

3.1. Audio Measure Evaluation

For the audio evaluation, we extracted a battery of parameters representing different aspects of prosody, including statistics of the fundamental frequency (f_0), intensity, voice quality parameters, and timing related features. Among the extracted features are articulation rate (i.e. number of syllables per second), statistics of f_0 (i.e. minimum, maximum, mean and span) extracted using the ESPS/*waves*+ software package, mean F_1 (i.e. the first formant), average pause time, mean amplitude and normalized amplitude quotient (NAQ) (Alku et al., 2002) and the so called Peak-Slope parameter identifying breathy regions of the speech (Kane and Gobl, 2011). In order to identify their influence on the perception of the speakers' style and quality (with respect to the set of questions posed to the subjects after each segment), we calculated Pearson correlation coefficients $\rho \in [-1, 1]$. The ρ values represent the strength of positive or negative linear correlation of the extracted parameters and the perception of a speaker. Table 1 summarizes the analyzed parameters and the found correlations for all speakers as well as for male and female speakers separately.

It is seen that multiple prosodic parameters have significant negative or positive correlations (marked with * or **)

[1] Freely available at: `http://webtv.bundestag.de/iptv/player/macros/bttv/index.html`

Feature	Group	Ov.	Ins.	Mon.	Exp.	Per.
Mean f_0	ALL	.190	.001	-.202	.189	.149
	M	.525*	-.360	-.558*	.521*	.473*
	F	.320	-.270	-.400	.413	.394
Min f_0	ALL	.250	-.135	-.144	.210	.201
	M	.340	-.294	-.250	.287	.257
	F	.136	-.063	.040	.096	.150
Max f_0	ALL	.266	-.089	-.284	.261	.218
	M	.305	-.214	-.351	.274	.239
	F	.397	-.294	-.461*	.508*	.426
f_0 span	ALL	.224	-.065	-.262	.226	.184
	M	.216	-.137	-.287	.199	.172
	F	.348	-.266	-.453*	.465*	.372
Mean F_1	ALL	.449**	-.285	-.460**	.484**	.473**
	M	.736**	-.800**	-.622**	.712**	.767**
	F	.270	-.118	-.424	.379	.333
Articulation rate	ALL	-.045	.202	-.047	-.026	-.077
	M	-.224	.271	.084	-.214	-.231
	F	.086	.070	-.299	.254	.094
Pause time	ALL	-.556**	.480**	.597**	-.568**	-.525**
	M	-.605**	.510*	.701**	-.643**	-.581**
	F	-.485*	.622**	.262	-.353	-.464*
Mean intensity	ALL	.538**	-.395*	-.621**	.573**	.509**
	M	.793**	-.709**	-.806**	.771**	.736**
	F	.231	-.113	-.347	.275	.226
NAQ	ALL	-.394*	.375*	.395*	-.460**	-.449**
	M	-.421	.432	.354	-.438	-.465*
	F	-.409	.310	.576**	-.580**	-.456*
PeakSlope	ALL	-.475**	.390*	.588**	-.535**	-.453**
	M	-.793**	.685**	.851**	-.792**	-.732**
	F	-.130	.038	.228	-.171	-.104

Table 1: Table showing Pearson's ρ values and significant positive or negative linear correlations between prosodic parameters and subjects' perceptual ratings (PART 1). The correlations are calculated for three groups, i.e. all speakers (ALL), female speakers (F), and male speakers (M). The perceptual ratings are shortened using the following abbreviations: overall (Ov.); insecurity (Ins.); monotony (Mon.); expressiveness (Exp.); and persuasiveness (Per.). Significant correlations are denoted with * ($p < .05$) and ** ($p < .01$). Leading zeros were omitted.

with the perceptual ratings of the naive subjects that rated the short speech segments. The strongest correlations were found for the pause time parameter which highly negatively correlates with the overall rating (i.e. the shorter the time spent for pauses the better the overall rating), and the speakers' expressiveness and persuasiveness. Further, the parameter is positively correlated with insecurity and monotony.

Also the voice quality parameters PeakSlope and NAQ highly correlate with the perceptual ratings. In general, breathy voice qualities (i.e. high NAQ and PeakSlope values) correlate strongly and positively with insecurity and monotony. The other three categories correlate negatively with these parameters, as small NAQ and PeakSlope values indicate more tense voice qualities.

Relatively moderate correlations were found for the f_0 related parameters. Mean f_0 had only slight effects for male speakers and span f_0 as well as max f_0 had small significant correlations for female speakers. This finding might be an effect of the strong speaker dependence of the f_0 parameter.

In (Rosenberg and Hirschberg, 2005) highly significant correlations (i.e. $p < .001$) for f_0 related statistics were found for charismatic speech, which roughly corresponds to our overall rating scores. However, we could not verify those highly significant results as f_0 only rarely significantly correlates with overall ratings. Also speaking rate (i.e. syllables per second) correlated significantly with charisma ($p = .085$). As we did not assume significance at $p < .1$ (as in (Rosenberg and Hirschberg, 2005)) but at $p < .05$, the results are not necessarily comparable. Unfortunately, no correlation values such as Pearson's ρ were given in (Rosenberg and Hirschberg, 2005), with which we could compare our values.

Further, Table 1 shows that gender specific differences are found in the data. The mean intensity of the speech for example only shows significant correlation with ratings for male speakers but none for female speakers, which indicates that a raised intensity in speech only changes the perceptional quality of the speech for male speakers. PeakSlope and mean F_1 also show similar effects. These findings show that it is important to separate the analysis for male and female speakers as the same prosodic parameters might have opposing effects for the different genders. However, no significant change of directional correlation was found for a single parameter.

3.2. Visual Feature Evaluation

As a basic but meaningful visual feature we estimated the optical flow in each of the 40 sequences using the algorithm proposed by (Brox et al., 2004). The results are vector fields of regularized estimates of the spatiotemporal changes resulting in image shifts and deformations (or warpings) of structures in the intensity function over time. At each location the estimated amount (speed) and direction of the velocity of a pixel is encoded. In Fig. 1 exemplary optical flow fields are shown for three different underlying body movements; the different colors resemble the direction of the movement. Using such flow fields $\vec{u}(x, y, t)$ we calculate the motion energy of each image by a frame-wise summation of the length of the vectors and discarding their direction information. This identifies wheather motion occurs in an image and also captures its average strength. Since the speaker is the only person visible in each sequence and the background is almost perfectly constant, the calculated motion energy can be treated as an indicator for the overall intensity of a speaker's motion. Any conceivable body movement causes a change in the motion energy, and thus a direct relationship between the motion energy and the gesticulation of a speaker exists. The motion energy is finally averaged over the whole sequence and used as a measure reflecting the overall motion amplitude a speaker exhibits within one sequence. In more formal terms the operation is denoted by

$$E_\tau(x, y, t) = \sum_{t=0}^{\tau-1} \|\vec{u}(x, y, t)\|/\tau \qquad (1)$$

showing the pixel-wise averaging of motion vector lengths (compare (Bobick and Davis, 2001) for a slightly different approach based on temporal differencing alone). Fig. 2 shows the estimated motion energy over the first 400 frames of two sequences, one with a speaker rated as monotonous (Mean: 3.571, Std.: 0.787), and the other showing a speaker who is perceived as expressive (Mean: 3.714, Std.: 0.488). Almost all rating categories were found to be strongly correlated to the relative motion energy (see Table 2). Merely insecurity showed no correlation in the audiovisual setup. In accordance to the semantic relationship between the rating categories, monotony showed a strong negative correlation to the relative motion energy, whereas overall, expressiveness and persuasiveness were positively correlated. There was a slight tendency to stronger correlations in the visual only than in the audiovisual setup, potentially indicating a higher importance of the motion energy in the absence of audio. Further, we could show a significant difference of the relative motion energy within the tested rating categories. Here, the ratings below 2.5 were treated as disagreement to one of the speaker characterizations, whereas ratings exceeding 3.5 were considered as agreement. Within most of the categories, there is a highly significant difference in the relative motion energy between the sequences which were rated in agreement with a category and their disagreement counterparts (see Fig. 3). As for the correlation, the results show stronger significant differences for the visual only setup, again supporting the assumption of a higher importance of visual features in the absence of audio.

Rating Category	Group	Motion Energy
Overall	**Mute**	.553**
	AV	.368*
Insecure	**Mute**	.360*
	AV	-.170
Monotonous	**Mute**	-.585**
	AV	-.552**
Expressive	**Mute**	.606**
	AV	.489**
Persuasive	**Mute**	.489**
	AV	.368*

Table 2: Table showing Pearson's ρ values and significant positive or negative linear correlations between visual features and subjects' perceptual ratings (PART 1). The correlations are calculated for the audiovisual and the visual only setup. The perceptual ratings are shortened using the following abbreviations: overall (Ov.); insecurity (Ins.); monotony (Mon.); expressiveness (Exp.); and persuasiveness (Per.). Significant correlations are denoted with * (p $<$ .05) and ** (p $<$ .01). Leading zeros were omitted.

3.3. Audiovisual vs. Video-only Evaluation

Beside the speaker ratings, we recorded the gaze direction of the subjects to address the question whether the absence of the acoustic channel influences the kind of visual features used for the judgments of the speakers. During the experiments, the point of regard (POR) in pixel coordinates was measured at a speed of 240 Hz using an SMI iView X$^{\text{TM}}$Hi-Speed eye-tracker. Analyzing the POR over time allows to infer from a rich set of different features, such as the fixation time on a specific target or the characteristics of its trajectory. In the presented study, we compared the relative time a subject is focusing on a speaker's face in the audiovisual and the visual only setup. The fixation time on the face reflects the ratio between facial and other body features (e.g. facial expression vs. upper body pose) used for the judgment of the speaker's performance. In Fig. 4, the relative fixation time of an exemplary subject is shown. The accordance of the POR and the position of the face in an image was verified using the OpenCV implementation of the Viola Jones face detection algorithm[2] (Viola and Jones, 2004). As shown in Table 3, the subjects focused significantly longer (p $<$ .01) on the face in the audiovisual than in the visual only condition. We hypothesize, that in the absence of acoustic input subjects are trying to gather a larger amount of visual features contributing to their judgment. An example for the influence of audio on the fixation on the face is shown in Fig. 4.

[2] In case of a miss, the position of the face was interpolated between neighboring frames in time. The presence of a face in each frame is guaranteed within the data.

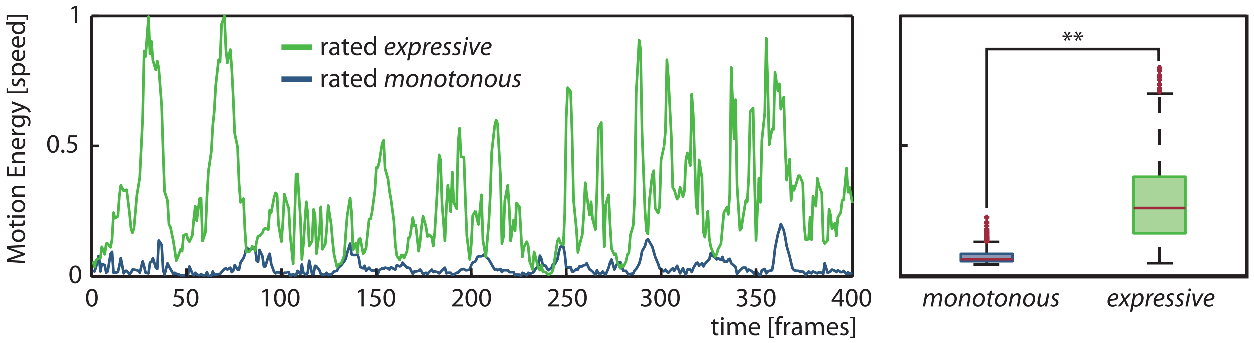

Figure 2: Estimated motion energy over the first 400 frames of two exemplary sequences. The speaker in the first sequence (green) was rated as expressive (Mean: 3.714, Std.: 0.488), the second speaker (blue), in contrast, was clearly perceived as monotonous (Mean: 3.571, Std.: 0.787). The two corresponding motion energy plots almost never overlap and show a strong significant difference (p $<$ .001).

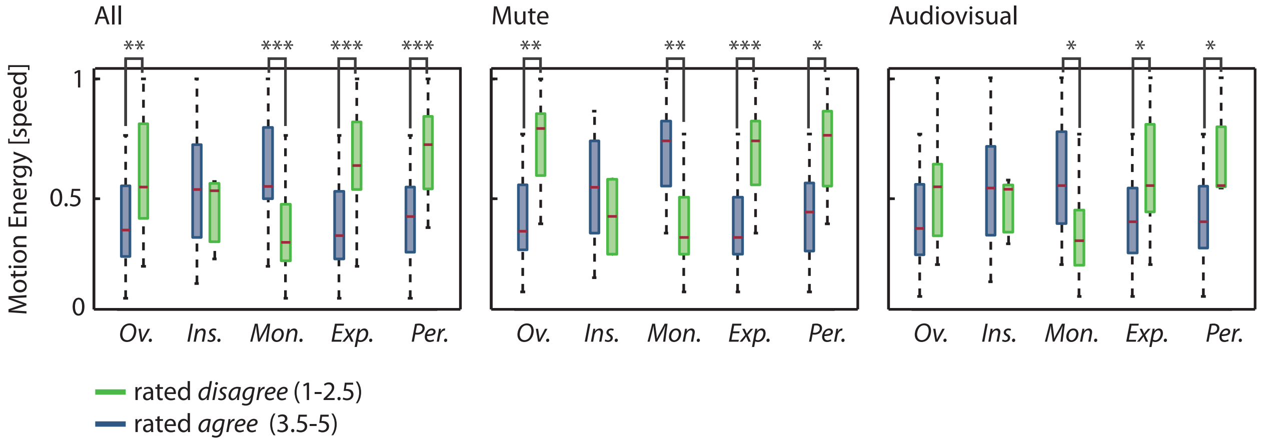

Figure 3: Differences in the relative motion energy within the tested rating categories. Mean ratings in the interval $[1, 2.5)$ were treated as disagreement to a speaker's characterization, whereas ratings in $(3.5, 5]$ were considered as agreement. Analyzing audiovisual and visual only sequences as a whole, we found strong significant differences in all categories, except insecurity. Considered separately, the visual only setup showed a higher number of significant differences. This, in conjunction with the correlation results, indicates a higher importance of visual features in the absence of audio. Significant differences in the Mann-Whitney-U-Test are denoted with * (p $<$ .05) , ** (p $<$ .01) and *** (p $<$ .001) (homoscedasticity was checked using Levene's test). The perceptual ratings are shortened using the following abbreviations: overall (Ov.); insecurity (Ins.); monotony (Mon.); expressiveness (Exp.); and persuasiveness (Per.).

4. Conclusions and Outlook

In the present study, we investigated the impact of combined verbal and non-verbal features in the communication and judgement of target persons in zero acquaintance situations. In particular, we could show various effects with respect to the combined visual and auditory perception of political speeches in comparison with visual-only movie presentations in the same speech scenario. The most prominent effects could be found for the correlations of the five perceptual speaker ratings (i.e. insecurity, monotony, expressiveness, persuasiveness and overall quality) by seven naive subjects and several prosodic parameters and motion energy. Our results thus add new knowledge to the discussion of the role of different verbal and non-verbal cues in communication and social interaction. It has already been shown previously that different cues from speech and non-vocal channels are evaluated in order to judge speakers as friendly or showing solidarity (Gallois and Callan, 1988) while the rapid analysis and judgement of personality is mainly based on visual cues (Albright et al., 1988). The use of multiple channels in communication and the context-dependency of the meaning of received social signals has been emphasized in (Shanker and King, 2002; Grammer et al., 2002). The authors foster the emergence of a paradigm shift from treating communication as a sequential information processing mechanism towards a dynamical system's approach in which communication and interaction is a parallel, tunable, and dynamic process. With our results, we add to the latter by showing that a receiver actively seeks for specific signatures of relevant social signals in the audi-

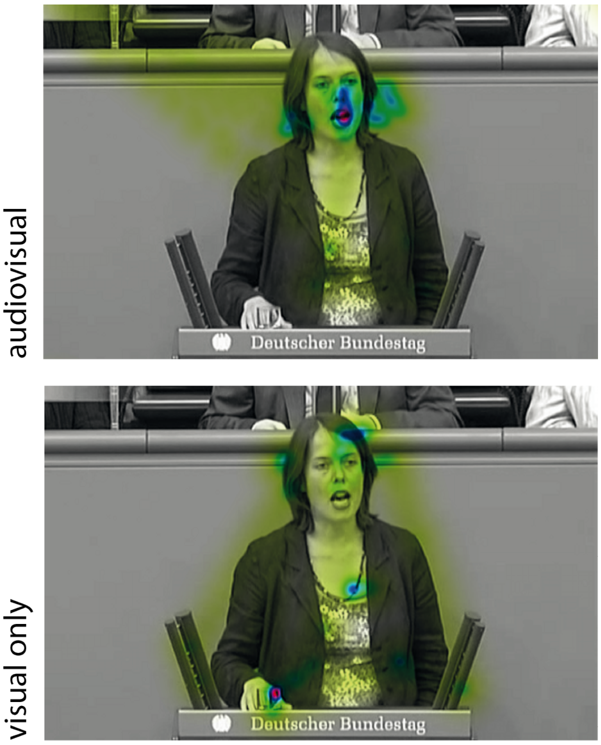

Figure 4: Influence of the acoustic channel on the relative fixation time on the face and limbs of a speaker. The relative fixation time is color encoded (from yellow indicating a very short fixation time, through green and blue to red, indicating a very long fixation time). In the audiovisual case (at the top), the subject mainly focuses on the face of the speaker and just rarely on the upper body or the background of the scene. In contrast, the subject virtually spends the same time fixating the face, the body and the hand of the speaker if no audio is present (at the bottom).

	Group	All speakers	Female	Male
Mean	**Mute**	.77**	.76	.78**
	AV	.80	.78	.82
Std.	**Mute**	.15	.15	.15
	AV	.13	.14	.11
Q_1	**Mute**	.69	.67	.71
	AV	.73	.71	.76
Q_3	**Mute**	.88	.87	.90
	AV	.90	.88	.90

Table 3: Table showing the gazing behavior of the subjects with respect to visual only (i.e. mute) and audiovisual (i.e. AV) conditions. Mean values, standard deviations (Std.), first and third quartiles (Q_1, Q_3) are given as relative time the subject looks at the speakers face within one of the segments. Significant differences in the Wilcoxon sign-rank tests are denoted with ** ($p < .01$) Leading zeros were omitted.

tory and visual stream, depending on the currently available information channels.

In our study the role of the perceiver is rather passive in terms of a communication scenario. The subjects had to watch (and listen to) political speeches of unknown actors, a situation that closely resembles the zero acquaintance scenario in which the perceiver cannot directly interact with the target. Still, as a result of our evaluation of comparison conditions, we demonstrate that the dynamics of body movements, as encoded in the perceived motion, is an important indicator people use for judging the personality of speakers. Similar to the approach suggested by (Grammer et al., 2002) we utilized the optical flow field estimated from the speakers' video sequences and computed the average motion energy pattern thereof. The input scene is actively sampled over time by the observer through actively gazing at specific locations which show high dynamics in the structure and its changes over space and time (Vig et al., 2012). We have demonstrated that the selection of specific locations is not only driven by sensory data, but by the active search for information based on the available input modalities. Thus, top-down expectations and task-dependent information sampling is observed here (Rothkopf et al., 2007; Saunders et al., 2010). Unlike the approach in (Koppenheimer and Grammer, 2010) we do not instantiate geometric movement models but make use of image-based motion energy directly. It should be noted here, that further information could be easily derived from this optical flow pattern since we have access to a more rich repertoire of, e.g., motion direction patterns over time, in turn, allowing to estimated robust spatiotemporal parameters from the input directly. At the moment we kept the analysis as simple as possible to highlight the key statistical dependencies.

In our investigation, we identified that pause time and voice quality parameters (i.e. NAQ and PeakSlope), which have not been used in previous studies, showed the strongest correlations with the ratings. The results for the audio feature correlations can be found in Table 1. Values indicating breathy voice qualities for example correlate strongly with perceived insecurity and monotony of male speakers, which is in line with previous expectations. Further, we were able to identify strong positive correlations between the motion energy feature and the ratings of good overall performance as well as strong expressivity and persuasiveness. In addition, we found strong negative correlations of motion energy with monotony, which corresponds to our previous hypothesis. We found some differences in the strength of correlation between the feature of motion energy and all the speaker ratings in the mute and audiovisual conditions. To be precise, the correlations of motion energy with all ratings is weaker in the audiovisual condition indicating that the raters rely less on visual features if auditory cues are at their disposal. In the case of insecurity the correlation between the motion energy feature and the subject's even switches the algebraic sign from the mute to the audiovisual condition; this in turn indicates that for insecurity auditory features complement perception.

Standard parameters such as statistics of f_0, however, show little to no correlations, which might be an effect of the coarse level of analysis. As already pointed out above, a more fine-grained analysis of both audio and video features might reveal more interesting effects and is subject to future analysis. It should be noted here that all the analysis assumes a dedicated actor that is engaged in his political speech. The mere correlation analysis could be fooled through clownish presentations with large movement and otherwise useless speech content.

The eye-tracking analysis revealed that subjects are more likely to look at the faces of the speakers in the audiovisual condition than in the video only condition (see Table 3 for details). This indicates that the subjects have a less focused gaze in the video only condition and scan the full politician's body for cues. It almost seems like the observers are searching for indications in the whole picture that might help them in judging the speaker without knowing the verbal content of the speech.

The revealed effects and correlations have great potential for future applications such as the automatic classification of the quality of public speeches or the training of speakers. In order to be able to train classifiers the findings need to be confirmed on a larger scaled analysis for more speakers as well as more naive subjects. Currently effort is undertaken to increase the number of investigated features as well as the cohort of human subjects. In addition, the study reveals that information fusion in humans using visual and auditory streams influences the patterns of active seeking and sampling of relevant information in the ambient environment. Further studies are needed which will highlight the dynamic processing and evaluation of features from different sensor streams and their subtleties in different situations of social communication.

Acknowledgement

The authors would like to thank the students Florian Oberlies and Michael Strauß for their programming efforts and the setup of the study. The presented work was developed within the Transregional Collaborative Research Centre SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG). This work was further supported as part of the FASTNET project - Focus on Action in Social Talk: Network Enabling Technology funded by Science Foundation Ireland (SFI) (Grant 09/IN.1/I2631) and by the SFI (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (`http://www.cngl.ie`).

5. References

L. Albright, D.A. Kenny, and T.E. Malloy. 1988. Consensus in personality judgements at zero acquaintance. *J. of Personality and Social Psychology*, 55(3):387–395.

P. Alku, T. Bäckström, and E. Vilkman. 2002. Normalized amplitude quotient for parameterization of the glottal flow. *Journal of the Acoustical Society of America*, 112(2):701–710.

M. Argyle. 1975. *Bodily Communication*. Londoin, Routledge.

A. Bobick and J.W. Davis. 2001. The recognition of human movement using temporal templates. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23:257–267.

T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. 2004. High Accuracy Optical Flow Estimation Based on a Theory for Warping. In Tomás Pajdla, Jiri Matas, Tomás Pajdla, and Jiri Matas, editors, *ECCV (4)*, volume 3024 of *Lecture Notes in Computer Science*, pages 25–36. Springer.

T.J. Clarke, M.F. Bradshaw, D.T. Field, S.E. Hampson, and D. Rose. 2005. The perception of emotion from body movement in point-light displays of interperonal dialogue. *Perception*, 34:1171–1180.

C. Gallois and V.J. Callan. 1988. Communication accomodation and the prototypical speaker: predicting evaluations of status and solidarity. *Language & Communication*, 8(3/4):271–283.

K. Grammer, B. Fink, and L. Renninger. 2002. Dynamic systems and inferential information processing in human communication. *Neuroendocrinology Letters*, 23 (suppl. 4):15–22.

J. Kane and C. Gobl. 2011. Identifying regions of non-modal phonation using features of the wavelet transform. In *Proceedings of Interspeech 2011*, pages 177–180. ISCA.

M. Koppenheimer and K. Grammer. 2010. Motion patterns in political speech and their influence on personality ratings. *J. of Res. in Personality*, 44:374–379.

A. Rosenberg and J. Hirschberg. 2005. Acoustic/prosodic and lexical correlates of charismatic speech. In *Proceedings of Interspeech 2005*, pages 513–516. ISCA.

C.A. Rothkopf, D.H. Ballard, and M.M. Hayhoe. 2007. Task and context determine where you look. *J. of Vision*, 7(14):16:1–20.

D.R. Saunders, D.K. Williamson, and N.F. Troje. 2010. Gaze patterns during perception of direction and gender from biological motion. *J. of Vision*, 10(11):9:1–10.

S.G. Shanker and B.J. King. 2002. The emergence of a new paradigm in ape language research. *Behav. and Brain Sciences*, 25:605–656.

E. Strangert and J. Gustafson. 2008. What makes a good speaker? subject ratings, acoustic measurements and perceptual evaluations. In *Proceedings of Interspeech 2008*, pages 1688–1691. ISCA.

E. Vig, M. Dorr, T. Martinetz, and E. Barth. 2012. Intrinsic dimensionality predicts the saliency of natural dynamic scenes. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, (in press).

P. Viola and M.J. Jones. 2004. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154.