Modulation of Cognitive Goals and Sensorimotor Actions in Face-to-Face Communication by Emotional States: The Action-Based Approach

Bernd J. Kröger[1,2]

[1]Neurophonetics Group, Department of Phoniatrics, Pedaudiology, and Communication Disorders, Medical School, RWTH Aachen University, Aachen, Germany
[2]Cognitive Computation and Applications Laboratory, School of Computer Science and Technology, Tianjin University, Tianjin, China
bernd.kroeger@rwth-aachen.de

Abstract. Cognitive goals – i.e. the intention to utter a sentence and to produce co-speech facial and hand-arm gestures – as well as the sensorimotor realization of the intended speech, co-speech facial, and co-speech hand-arm actions are modulated by the emotional state of the speaker. In this review paper it will be illustrated how cognitive goals and sensorimotor speech, co-speech facial, and co-speech hand-arm actions are modulated by emotional states of the speaker, how emotional states are perceived and recognized by interlocutors in the context of face-to-face communication, and which brain regions are responsible for production and perception of emotions in face-to-face communication.

Keywords: face-to-face communication, emotion, speech, facial expression, gesture, sensorimotor action, emotional speech, brain imaging, fMRI

1 Introduction

Speech production comprises the activation of cognitive goals or intentions, their transformation into a (still cognitive) lexical phonological representation and subsequently into a sequence of speech articulator movements and an acoustic speech signal. Thus we can separate the cognitive part of speech production [1, 2] and its sensorimotor part [3-5]. The emotional state of the speaker modulates both, the cognitive part (e.g. by choice of different lexical items, choice of different syntactic structure, etc.) as well as the sensorimotor part of speech production (e.g. type of phonetic realization of the utterance by varying loudness level, speaking rate, intonation, voice quality, etc.) [6-8]. In the case of face-to-face communication in addition the co-speech facial expression and the co-speech hand-arm gestures are modulated by speaker's emotional state at both levels. At the cognitive level different types of facial expressions and different types of hand-arm gestures can be chosen (non-conscious or conscious). At the sensorimotor realization level amplitude and duration of facial as well as hand-arm movements can vary [9, 10].

It is the goal of this paper to discuss these processes from the viewpoint of our action-based approach for production, perception, and acquisition of communicative actions in face-to-face-situations [11], as well as to review literature which identifies the brain regions involved in these processes.

2 The Action-Based Approach for Face-to-Face Communication

The main proposition of our action-based approach is that speech, facial, as well as hand-arm actions can be described by one comprehensive cognitive and sensorimotor concept. On the one hand the intention or meaning of an utterance, its co-speech facial expression and its co-speech hand-arm gestures can be specified in form of a hypermodal cognitive pattern [11]. This cognitive pattern can be seen as the starting point for production as well as the end point of perception. On the other hand during the production process the complex temporarily overlapping sequence of movement actions controlling the speech articulators (lips, tongue, velum glottis etc.), the facial articulators (eye brows, eye lids, cheeks, mouth corners etc.) as well as the articulators or the hand-arm system (position of lower arms and hand, orientation of palm by wrist, and form of hand and fingers; see also Figure 1) needs to be specified. This is done at the motor plan level ([11] and see Fig. 1 as an example for a motor plan).

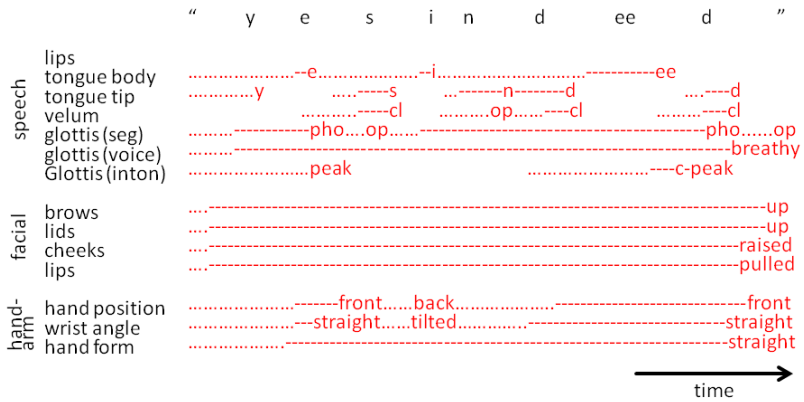

Fig. 1. Example of a motor plan for temporally coordinated speech, facial, and hand-arm movement actions (utterance: "yes indeed"). Movement phases are marked by "…" and hold phases are marked by "---" for each action. Tongue body and lip actions are labeled orthographically by the resulting speech sound. In addition, temporally overlapping velopharyngeal opening (op) and closing (cl) actions and glottal opening (op) and phonation (pho) actions are needed in order to produce proper speech sounds (cf. [18]). In addition, voice quality actions (here for breathy voice) and glottal actions for production of specific intonation contours (peak and central peak actions) are specified. Only one co-speech facial expression is produced here ("happy"&"interest" in combination, cf. [16]), and two co-speech hand-arm actions (front, straight) are produced in this example in order to underline the stressed syllables. The third hand-arm action (back, tilted) can be interpreted as rest position of the hand-arm system in this communication situation. Only actions of left hand-arm system are displayed here because right hand-arm system was at rest in this utterance.

The hierarchical structure of actions is very steep in the case of speech. Here, each utterance can be separated into one or more phrases, each phrase in one or more stress groups (each containing one stressed syllable), and each stress group in one or more syllables [12]. Moreover each syllable comprises a set of temporally overlapping movement actions of speech articulators (Fig. 1). In the case of co-speech hand-arm actions each gesture at least comprises a preparation, stroke, and recovery phase [13, 14]. Moreover each of these phases comprise one or more temporally overlapping

target-directed movement actions of hand-arm articulators controlling (arm and) hand position, wrist angle (leading to palm orientation), and hand shape (Fig. 1). In the case of facial actions the hierarchical structure is less steep. Here, each facial expression is directly realized by a set of temporally overlapping and mainly synchronously occurring facial action units as defined in FACS [15, 16] (see also Fig. 1: four temporally synchronous facial actions). These target-directed movement actions control the shape of parts of the facial structure like form/location of mouth, eye lids, eye brows etc. The movement actions activated at the motor plan level are always target- or goal-directed, where the goal is defined in the auditory domain in the case of speech and in the visual domain in the case of hand-arm gestures as well as in the case of facial expressions.

The realization of movement actions (of articulator movements) should not just be interpreted as the endpoint of the production process but as well as starting point or input perceptual vehicles for comprehending the intentions of a speaker in face-to-face communication [11]. It is an important aspect of our action-based approach, that each movement action can be subdivided in a movement and in and target or hold phase [17, 18]. From the viewpoint of action perception, the movement phase is as important as – or sometimes even more important than – the target or hold phase. Movement phases of speech actions produce (auditory perceivable) formant trajectories, which are important for perception of place of articulation of consonants as well as vowel quality. Movement phases often are important parts of stroke phase in case of co-speech hand-arm actions, and specific parameters of movement phase of facial actions help to identify, whether a facial expression is spontaneously produced or acted [19]. Moreover, long hold phases especially occur in co-speech facial movement actions, in specific phases of co-speech hand-arm gestures, as well as in voice quality actions, because the overall timing of face-to-face communication actions are dominated by speech movement actions, defining the segmental and syllabic structure of an utterance (cf. [17] for a vice-versa situation in sign language production).

3 Emotions Modulate Behavior: The Functional View

Internal and external aspects of emotion can be separated [20]. While the internal aspect of emotion is important for organization of behavior like action selection, attention, learning etc., the external aspect of emotion focuses on production and perception of emotional expressions which is an important aspect in communication and social coordination [ibid., p. 554]. A comprehensive model for emergence of emotions and how emotions influence behavior and perception is given in [21] and summarized in Fig. 2. In this model, the drive system is central. Drives establish the top-level goals of each individual and need to be satisfied by specific behavior or specific actions of the individual. Thus, the intensity of a drive increases again after satisfaction (e.g. degree of hunger, degree of fatigue, degree of loneliness etc.). The intensity of drives on the one hand organizes behavior (e.g. start to eat or drink, try to rest or sleep, start to communicate etc.) and on the other hand indirectly influences the emotional system of the individual. Here it is important to state, that intensity of drives change more slowly, while emotive responses work on a faster time scale (ibid., p. 129). An emotional reaction can be due to a recent external event occurring in the environment of the individual. For example a specific reaction of a communi-

cation partner could make the individual happy. Thus, the emergence of emotions mainly results from high-level processing of perceptual input. A set of specific releasers at a higher level of the perception system is postulated in [21], which on the one hand is capable of releasing specific behaviors, e.g. if a goal is not satisfied or if a stimulus is not desired, but which on the other hand is capable of influencing affective and thus emotional states. The emotion system itself can be subdivided in three stages, i.e. affective appraisal, elicitation and activation of emotions (ibid.). Affective states are mainly driven by the perceptual system and can be defined by specifying values on three scales (i.e. arousal, valence, and stance). Thus, a stimulus may be very arousing (positive arousal), but at the same time very unfavorable (negative valence) and therefore is refused or avoided (negative stance). Each affective state is associated with a specific type of emotion (e.g. "anger" in the case of the example given above; see ibid., p. 135 for more examples) and an emotion is elicited, if the activation of an affective state exceeds specific threshold values. The current emotional state now directly influences the cognitive behavioral system (e.g. selection of behavioral goals) as well as the sensorimotor system (i.e. the explicit temporal motor planning and execution for the actions, which are already defined at the cognitive level).

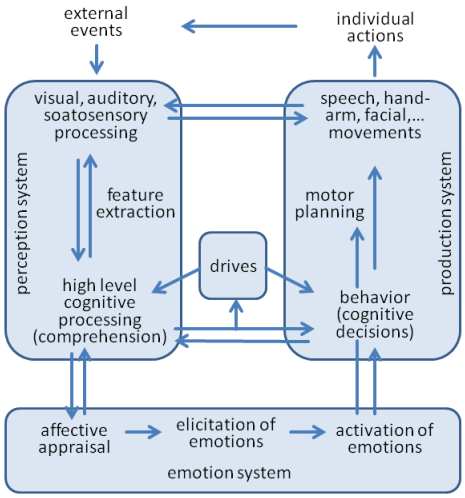

Fig. 2. Overview on interplay of emotion system, production system, perception system, and drives, following [21]. Here, the behavior system (after [21]) is included as cognitive part of production system. Lower levels of the production system comprise motor planning and execution; lower levels of perception end at the feature extraction level (cf. [11]).

Different sets of emotions are proposed in literature. A minimal set of six basic emotions is widely accepted: "anger", "disgust", "fear", "happiness", "sadness", and "surprise" [22]. But especially in the case of face-to-face communication including speech, beside "neutral", often two more emotional states are cited: "boredom" and "stress" (e.g. [8]). These additional states in part might be blends of basic emotions and as well might be biased by cultural influences. Even more emotional categories

can be defined, if emotions are ordered with respect to the 3-dimensional space defined by the dimensions arousal, valence, and stance (see [21], p. 135).

4 Neuroanatomical Correlates

It is widely accepted that the limbic system, especially the amygdala plays a crucial role in learning and (re-)activation of emotions [23]. Input as well as output pathways to and from the amygdala involve the brainstem and connect this "emotion processing center" within the limbic system with cortical and subcortical processing centers for sensory input as well as with cortical processing centers for context evaluation as well as for generation of specific emotionally motivated or emotionally modulated behavior. Thus, in addition cortical cognitive centers are involved in processing of emotions [24, 25]. In accordance, recent studies assume a complex network comprising the amygdala and its synaptic connections with medial prefrontal cortex for emotion regulation [26].

A complex cerebral network is postulated for processing emotionally loaded speech. This network comprises temporal regions such as the posterior and middle part of the superior temporal cortex for acoustic feature extraction, frontal structures including the inferior frontal cortex for emotion identification, as well as parts of the limbic system such as the amygdala [27]. Here, the neural associations towards the limbic system are mainly inhibitory, because the evaluation process of externally perceived emotions should not directly be influenced by the current emotional state of the subject. Emotional reactions are assumed to occur after this evaluation process (i.e. after evaluation of external voice signals produced by interlocutors). Comparable networks are postulated for processing of emotionally loaded facial expressions. Here, beside additional activation patterns within the occipital cortex for a basic perceptual processing of the incoming visual signals, the network comprises the amygdala, as well as frontal and temporal cortical areas [25-28].

Because of its experimental complexity only few studies are known, which investigate brain activities during speech production within face-to-face communication. Here – in contrast to brain imaging studies investigating the processing of emotionally loaded auditory speech or visual facial input (see above) – not the process of emotion evaluation, conveyed within the external signals (emotions of communication partner) is investigated, but the emotional activation of the subject involved in a free conversation with an interlocutor is measured [29]. But interestingly, comparable brain regions are involved in activation of emotions as are known for emotion evaluation. Brain activations of subjects in an emotionally loaded conversation situation appear in frontal cortex (BA 45 and BA 47) as well as within the limbic system. These results are in accordance with the fact that perception and production of speech prosody (i.e. the aspect of speech which conveys most information concerning the emotional state of a speaker) share common neural networks [30]. Here, common network regions (i.e. regions of overlap of activity for producing and/or perceiving speech prosody) were found especially in left inferior frontal gyrus.

5 Modulation of Communicative Actions by Emotion: Towards a Transcription System for all Three Domains

From the idea of defining a unified motor plan structure for movement actions in all three domains of face-to-face communication (speech, facial expressions, hand-arm gestures; see above) we aim at describing face-to-face communication behavior in a form as is already displayed in Fig. 1. Movement and hold phases of facial and hand-arm actions can be identified from video recordings. For identifying speech movement actions we use an articulatory speech resynthesis procedure as is described in [31]. From a first transcription of parts of the eNTERFACE05 audio-visual Emotion Database [32] (speech and co-facial actions) and of the Bielefeld Speech and Gesture Alignment corpus (SaGA) [33] (speech, co-speech facial actions, and co-speech hand-arm actions), we got results which are in accordance with literature (cf. Introduction of this paper and [6-10]). Two points seem to be very obvious using our transcription system (Fig. 1): On the one hand, the type of facial expression as well as the type of voice quality (defined at cognitive levels in our model for each utterance) seems to be selected directly with respect to the emotional state of a speaker and seem to hold over a whole utterance or whole turn of one speaker in face-to-face communication scenarios. On the other hand, speech actions as well as co-speech hand-arm movement actions are selected with respect to the lexical content of the utterance. In the case of these actions, the emotional state of the speaker merely modifies quantitative parameters of these movement actions, e.g. duration, amplitude, and degree of realization of these movement actions, as well as degree of temporal overlap with adjacent movement actions.

6 Conclusions and Future Work

The action-based approach has been developed for describing movement actions in face-to-face communication in a comparable way for speech, facial expressions, and hand-arm gestures. This approach allows a transcription of movement actions occurring in all three domains of face-to-face communication in a similar way. This transcription system is qualitative (i.e. denominating each type of action) as well as quantitative (i.e. annotation of beginning and end of movement phase and hold phase for each action). Thus, this approach allows a qualitative and quantitative evaluation of face-to-face communication e.g. in order to identify differences in cognitive goals and sensorimotor realizations of actions in face-to-face communication. In future work a broader corpus-based analysis of face-to-face communicative actions in different emotional states is planned by using our qualitative-quantitative transcription system.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (Grant No. 61233009)

References

1. Levelt W.J.M., Roelofs, A., Meyer, A.S.: A theory of lexical access in speech production. Behavioral and Brain Sciences 22: 1-75 (1999)
2. Levelt, W.J.M.: Models of word production. Trends in Cognitive Sciences 3: 223-232 (1999)
3. Guenther, F.H.: Cortical interactions underlying the production of speech sounds. Journal of Communication Disorders 39: 350-365 (2006)
4. Guenther, F.H., Ghosh, S.S., Tourville, J.A.: Neural modeling and imaging of the cortical interactions underlying syllable production. Brain and Language 96, 280–301 (2006)
5. Kröger, B.J., Kannampuzha, J., Neuschaefer-Rube, C.: Towards a neurocomputational model of speech production and perception. Speech Communication 51, 793-809 (2009)
6. Halberstadt J.B., Niedenthal P.M. Kushner J.: Resolution of lexical ambiguity by emotional state. Psychological Science 6: 278-282 (1995)
7. Bänziger T., Scherer, K.R.: The role of intonation on emotional expressions. Speech Communication 46: 252-267 (2005)
8. Scherer K.R.: Vocal communication of emotion: A review of research paradigms. Speech Communication 40: 227-256 (2003)
9. Ekman, P., Oster, H.: Facial expressions of emotion. Annual Review of Psychology 30: 527-554 (1979)
10. Castellano, G., Villalba S.D. Camurri A. : Recognising human emotions from body movement and gesture dynamics. In: A. Paiva, R. Prada, R.W. Picard (eds.) Affective Computing and Intelligent Interaction, LNAI 4738. Springer Verlag, Berlin, pp. 71-82 (2007)
11. Kröger, B.J., Kopp, S., Lowit, A.: A model for production, perception, and acquisition of actions in face-to-face communication. Cognitive Processing 11: 187-205 (2010)
12. Kröger, B.J., Birkholz, P., Kaufmann, E., Neuschaefer-Rube, C.: (2011) Beyond vocal tract actions: speech prosody and co-verbal gesturing in face-to-face communication. In: B.J. Kröger, P. Birkholz (eds.) Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2011. TUDpress, Dresden, Germany, pp. 195-204 (2011)
13. Kendon, A.: Gesture: Visible Action as Utterance. Cambridge University Press, New York (2004)
14. Kopp, S., Wachsmuth, I.: Synthesizing multimodal utterances for conversational agents. Journal of Computer Animation and Virtual Worlds 15: 39-51 (2004)
15. Ekman, P., Friesen, W.V.: Facial Action Coding System. Consulting Psychologists Press, Palo Alto, CA (1978)
16. Cohn, J.F., Ambadar, Z., Ekman, P.: Observer-based measurement of facial expression with the facial action coding system. In: J.A. Coan, J.J.B. Allen (eds.) Handbook of Emotion Elicitation and Assessment. Oxford University Press US, New York, pp. 203-221 (2007)
17. Kröger, B.J., Birkholz, P., Kannampuzha, J., Kaufmann, E., Mittelberg, I.: Movements and holds in fluent sentence production of American Sign Language: The action-based approach. Cognitive Computation 3: 449-465 (2011)
18. Kröger, B.J., Birkholz, P.: A gesture-based concept for speech movement control in articulatory speech synthesis. In: Esposito A, Faundez-Zanuy M, Keller E, Marinaro M (eds.) Verbal and Nonverbal Communication Behaviours, LNAI 4775. Springer Verlag, Berlin, pp. 174-189 (2007)

19. Schmidt, K.L., Ambadar, Z., Cohn, J.F., Reed, L.I. (2006) Movement differences between deliberate and spontaneous facial expressions: Zygomaticus major action in smiling. Journal of Nonverbal Behavior 30: 37-52 (2006)
20. Arbib M.A., Fellous, J.M.: Emotions: from brain to robot. Trends in Cognitive Sciences 8: 554-561 (2004)
21. Breazeal, C.: Emotion and sociable humanoid robots. International Journal of Human-Computer Studies 59: 119-155 (2003)
22. Ekman P.: An argument for basic emotions. Cognition and Emotion 6: 169-200 (1992)
23. LeDoux J.E.: Emotion circuits in the brain. Annual Reviews of Neuroscience 23: 155-184 (2000)
24. Lazarus, R.S. Cognition and motivation in emotion. American Psychologist 46: 352-367 (1991)
25. Pessoa, L., Adolphs R.: Emotion processing and the amygdala: from a 'low road' to 'many roads' of evaluating biological significance. Nature Reviews Neuroscience 11: 773-782 (2010)
26. Whalen, P.J., Raila, H., Bennett R., Mattek, A., Brown, A., Taylor, J., van Tieghem, M., Tanner, A., Miner, M., Palme, A.: Neuroscience and facial expressions of emotion: the role of amygdala-prefrontal interactions. Emotion Review 5: 78-83 (2013)
27. Brück, C., Kreifelts, B., Ethofer, T., Wildgruber, D.: Emotional voices: the tone of (true) feelings. In: J. Armony, P. Vuilleumier (eds.) The Cambridge Handbook of Human Affective Neuroscience. Cambridge University Press, New York, pp. 256-285 (2013)
28. Kesler-West, Marilyn.L., Andersen, A.H., Smith, C.D., Avison, M.J., Davis, C.E., Kryscio, R.J., Blonder, L.X.: Neural substrates of facial emotion processing using fMRI. Cognitive Brain Research 11: 213-226 (2001)
29. Mitsuyoshi, S., Monnma, F., Tanaka, Y., Minami, T., Kato, M. Murata, T.: Identifying neural components of emotion in free conversation with fMRI. Defense Science Research Conference and Expo, Singapore 2011. DOI: 10.1109/DSR.2011.6026845, pp. 1-4 (2011)
30. Aziz-Zadeh, L., Sheng, T., Gheytanchi, A.: Common premotor regions for the perception and production of prosody and correlations with empathy and prosodic ability. PLoS ONE 5: e8759. DOI:10.1371/journal.pone.0008759, pp. 1-7 (2010)
31. Bauer, D., Kannampuzha, J., Kröger, B.J.: Articulatory Speech Re-Synthesis: Profiting from natural acoustic speech data. In: Esposito A, Vich R (eds.) Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions, LNAI 5641 (Springer, Berlin), pp. 344-355 (2009)
32. Martin, O., Kotsia, I., Macq, B., Pitas, I.: The eNTERFACE05 Audio-Visual Emotion Database. First IEEE Workshop on Multimedia Database Management, Atlanta, USA. DOI: 10.1109/ICDEW.2006.145 (2006)
33. Lücking, A., Bergmann, K., Hahn, F., Kopp, S., & Rieser, H.: Data-based analysis of speech and gesture: the Bielefeld Speech and Gesture Alignment corpus (SaGA) and its applications. Journal on Multimodal User Interfaces, DOI: 10.1007/s12193-012-0106-8 (2012)