BEYOND VOCAL TRACT ACTIONS: SPEECH PROSODY AND CO-VERBAL GESTURING IN FACE-TO-FACE COMMUNICATION

Bernd J. Kröger[1], Peter Birkholz[1], Emily Kaufmann[2], Christiane Neuschaefer-Rube[1]

[1]Department of Phoniatrics, Pedaudiology, and Communication Disorders, RWTH Aachen University, Aachen, Germany

[2]Human Technology Centre, RWTH Aachen University, Aachen, Germany

{bkroeger, pbirkholz, cneuschaefer}@ukaachen.de, kaufmann.emily@gmail.com

Abstract: A comprehensive approach for describing the functional and behavioral aspects of *communicative actions*, e.g. facial, manual, and vocal tract actions, has been established for face-to-face-communication (Cogn Process 11:187-205, 2010). Within the speech domain, this approach will now be extended in two ways: (i) by introducing *level actions* such as turn, breath group, phrase, stress group, syllable, and sound group level actions, and (ii) by distinguishing between prosodic and segmental *speech actions* such as respiratory, voice quality, tonal, and vocal tract actions. Moreover this paper describes how speech actions are temporally coordinated with *co-verbal (or co-speech) communicative actions* such as manual and facial actions. **Index Terms**: face-to-face communication; communicative action; speech; prosody; vocal tract action; co-verbal action

1 Introduction: The concept of communicative actions

It is recognized that *bodily movements* within the *production* as well as *perception* of communicative actions in face-to-face communication are important for co-speech communicative actions (i.e. facial and manual actions) as well as for speech actions [1]. *Speech actions* (e.g. vocal tract actions such as lip closing or tongue body lowering in the production of /ba/) are accomplished by goal-directed movements of vocal tract effectors (tongue, lips, velum, glottis); manual actions (e.g. pointing) are accomplished by goal-directed movements of the hand–arm systems, and facial expression actions (e.g. smiling) by goal-directed movements of facial effectors such as the corners of the mouth, eyelids, eyebrows, etc. It is hypothesized in [1] that the *functional goal* in the production as well as in the perception of all these different types of communicative actions is *shape formation*, and that these shapes, which are approximated by the actions, are conveyed from speaker to listener via *movements*, which themselves are coded within a time-varying visual and/or a time-varying auditory signal. This paper will show that the concept of communicative actions can be extended to other types of speech actions, such as respiratory, voice quality, and tonal actions (i.e. *prosodic speech actions*), and it will show how speech and co-verbal actions can be temporally coordinated via *level actions*.

2 Types of communicative actions in face-to-face communication

In face-to-face communication a speaker's *turn*, lasting from turn-taking to turn-giving, comprises *speech* and *co-verbal (or co-speech) actions*. Speech actions comprise vocal tract and prosodic actions (i.e. respiratory, voice quality, and tonal actions), and the temporal coordination of these actions is organized by *level actions* (Fig. 1). Level actions comprise turn, breath group, phrase, stress group, syllable and sound group actions. ("Sound group" is used as a synonym for "sub-syllabic constituents" such as syllable onset, center, and offset;

see section 3.1). These level actions also coordinate speech with co-verbal facial and manual actions (Fig. 1).

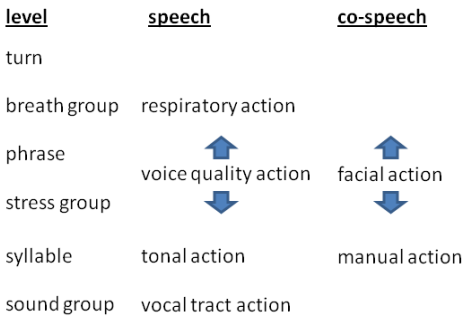

Figure 1 – Hierarchical organization of level actions and their relation to speech and co-speech actions.

The system of level actions is hierarchically organized: Each turn action comprises one or more sequentially ordered breath group actions; each breath group action comprises one or more sequentially ordered phrase actions; and so on over the stress group and its sequences down to the syllable and sub-syllabic actions and their sequences. (For a comprehensive discussion of prosodic units and intonation in general, see [2]). The breath group level is introduced here, since a speaker's turn may "degenerate" into a monologue and thus may comprise more than one breath group. In normal (i.e. dialogue-like) face-to-face communication, a speaker's turn is generally relatively short, and so in most cases a turn can be completed in a single breath group.

The prosodic levels introduced above (Fig. 1) do not necessarily correlate with linguistic levels. A word comprises one or more syllables linguistically, whereas in spontaneous speech, a word may be divided by a stress group or even a phrase. In spontaneous speech – and face-to-face communication is spontaneous – a turn may not necessarily even coincide with a complete or "correct" sentence in a linguistic sense.

Speech actions directly reflect the *movement* of effectors for approximating the spatial, visual and/or auditory target ([1] and [3]). *Level actions* reflect the temporal organization of speech actions. Moreover, movement actions are not necessarily meaning bearing, but they are always *distinctive*, while level actions such as phrase or turn actions convey *meaning* or *communicative intentions* (i.e. the meaning of a word or phrase; or the communicative intention of a whole turn, including speech and co-verbal actions). Thus, while *movement actions* primarily determine the (sensorimotor) *behavioral shape* of communicative actions, level actions determine the shape of communicative actions by arranging the temporal coordination of movement actions. For the dichotomy of sensorimotor behavior and cognitive features of actions, see [1].

3 Speech actions and a preliminary outline of an action-based speech prosody model

Following the ideas of [4], *voicing* is of central importance for speech prosody as well as for the creation of an acoustic-auditory speech signal per se. The voice source signal is the *carrier signal*, which is modulated in order to transfer information. Carrier signal *modulation* results from laryngeal source signal modulation as well as from the modulation of supralaryngeal articulation.

Modulations of the source signal result from four (and possibly more) different mechanisms. In order to tease apart these mechanisms it is necessary to differentiate between slow and rapid modulations. A modulation from syllable to syllable or on a lower (i.e. sub-syllabic) level is referred to as rapid, while a modulation on the phrase level or higher (i.e. phrase, breath group, or turn level) is referred to as slow. The four different mechanisms for carrier signal modulation are as follows. (i) A slow modulation in voice source intensity (e.g. from turn to turn) can be accomplished by a slow modulation of tracheal pressure (respiratory action; Fig. 1). (ii) A slow or rapid modulation in voice quality as well as a rapid modulation in voice source intensity can be accomplished by a modulation of vocal fold rest position (voice quality action). Voice quality actions – which can be superimposed by segmental glottal abduction or adduction actions for realizing voiceless consonants – are coordinated on the phrase level or higher, e.g. in order to adjust the overall voice quality and loudness of voicing during an utterance. But voice quality actions may also sometimes vary on the stress group level, e.g. a change in voice quality which signals the end of an utterance. (iii) A slow or rapid modulation in fundamental frequency (F0) can be accomplished by a modulation of vocal fold tension (tonal action). (iv) A rapid modulation in voice source intensity can be accomplished by segmental glottal abduction or adduction actions, as occurs during the realization of voiceless consonants. These segmental glottal abduction/adduction actions are part of the vocal tract action system [3].

Modulations of supralaryngeal articulation, which lead to an additional carrier signal modulation, result from vocal tract actions such as vocalic, consonantal and velopharyngeal actions (see section 3.1 below). These actions realize segmental changes and thus mainly lead to rapid modulations of the carrier signal. There are two (and possibly more) different mechanisms at work here: (i) The temporal succession of consonantal and vocalic vocal tract actions leads to a rapid modulation of the intensity level of the acoustic speech signal, even if the intensity of the voice source signal is constant. (ii) Vocalic, consonantal, and velopharyngeal vocal tract actions lead to rapid modifications of the formant pattern over time as well as to abrupt onsets and offsets of antiformants (nasality), frication noise (i.e. time intervals of noise for fricatives) and noise bursts (i.e. for plosives).

Thus, the modulation of carrier signal voicing is the cue for information coding. Voicing results from the close linkage of a *respiratory action* with one or more successive *voice quality actions*, which provide a specific vocal fold rest position that enables voicing. Voicing is controlled at the phrase level or higher and includes voice quality adjustments as well as the adjustment of the loudness level (e.g. a soft, normal, or loud voice). However, voice quality need not remain constant over a complete phrase. Moreover, a succession of two or more voice quality actions may occur during a phrase, for example in the transition from normal voice quality at the beginning of a phrase to breathy-laryngealized voice quality which often indicates the end of a phrase, breath group or even a turn [5]. Furthermore, it should be mentioned here that a complete phrase level intonation pattern is created on the syllable level by elementary goal-directed *tonal actions* ([6]; see also section 3.2).

3.1 Vocal tract actions and syllable action

Vocal tract actions comprise vocalic, consonantal, velopharyngeal, and glottal movement actions (see [1, 3, 7]; glottal actions, if labeled as vocal tract actions, are limited to glottal abduction/adduction actions in so far as the goal is to produce voiced or voiceless segments). This concept is based on the task dynamics approach [8, 9, 10] and on articulatory phonology [11, 12, 13, 14]. These actions generally comprise a *movement phase* as well as an optional *target phase* [1, 3, 7]. For example, in the case of consonantal actions, the target phase represents the consonantal closure or near closure time interval, while the movement phase represents the movement of the effector towards that constriction or closure (see Fig. 2). Movement phases can be perceived visually in the special cases of lip, jaw, and possibly

tongue tip actions. However, more generally, vocal tract effector movements are perceived in the auditory domain since the purpose of these movements, which are not visible, is to generate formant pattern transitions (formant trajectories, e.g. for coding the place of articulation). In fast speech, target phases can disappear, especially in the case of lax vowels. In this case, vocalic actions tend to be dramatically reduced in order to shorten the amount of time needed for the whole utterance. Thus, vowel quality (which leads to the phonemic discrimination of vowels) is perceived mainly from the formant movement (i.e. from the movement phase) towards the vocalic target (cf. the concept of formant undershoot [15]).

Since the syllable is the basic unit of articulation, vocal tract actions are coordinated in time by *syllable actions* in cooperation with sub-syllabic sound group actions. A syllable action (i) coordinates the vocalic (i.e. tract-forming) action specifying the syllable center (or syllable core), (ii) coordinates one or more consonantal actions describing syllable initial and/or syllable final consonants or consonant clusters, and (iii) coordinates velopharyngeal and glottal actions, which may overlap in time with the vocalic and consonantal actions (Fig. 2). Velopharyngeal opening actions create nasal consonants; velopharyngeal closing actions produce non-nasal sonorants; and velopharyngeal tight closing actions are needed for the production of obstruents, which requires a pressure built-up in the oral cavity. Glottal opening actions form voiceless consonants; glottal closing actions form voiced sounds; and glottal tight closing actions form the glottal stop [ʔ]. In this syllable-based concept for the temporal coordination of vocal tract actions, only one glottal closing action is allowed to occur within the syllable center, while for voiceless consonants, an additional single glottal abduction action each is allowed in the syllable onset and offset (see Fig. 2). Thus, in our action-based approach, glottal abduction/adduction actions are closely related to syllable constituents such as syllable onset/offset and syllable core rather than segments (phones or phonemes).

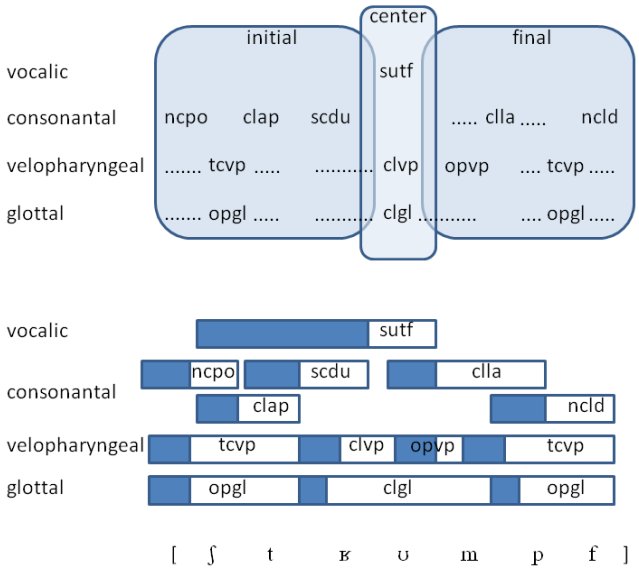

Figure 2 – Organization of vocal tract actions within the complex syllable /ʃtʁʊmpf/ in Standard German ("Strumpf" means 'socks/stockings'). Upper part: organization of actions with respect to initial, center, and final part of syllable (sound groups); Lower part: exact temporal alignment of vocal tract actions; dark blue boxes = movement phase; white boxes = target phase of action. Naming of actions following [3]: sutf = tract-forming short /u/ action; ncpo = postalveolar near-closing action; clap = apical closing action; scdu = dorso-uvular short-closing action (approximant); clla = labial closing action; ncld = labio-dental near-closing action; tcvp = velopharyngeal tight-closing action; clvp = velopharyngeal closing action; opvp = velopharyngeal opening action; opgl = glottal opening action; clgl = glottal closing action.

It becomes apparent from the complex example “Strumpf” (Fig. 2) that vocal tract actions do not necessarily have a one-to-one relationship with segments. It can be seen that in most cases, a segment (vowel or consonant) is composed of more than one vocal tract action. For example, the obstruent /ʃ/ comprises a consonantal near-closing action, a velopharyngeal tight-closing action and a glottal opening action; and the vowel /ʊ/ comprises a vocalic tract-forming action (mainly acting on the tongue body and lips) and a glottal closing action for realizing phonation. However, there are also vocal tract actions which encompass more than one segment. In our example, these are: (i) the glottal closing action which is responsible for voicing in three central segments (i.e. /ʁʊm/); the glottal opening action which is responsible for voicelessness in two segments of the initial consonant cluster (i.e. /ʃt/) as well as in the final consonant cluster (i.e. /pf/); and the labial closing action within the syllable offset (i.e. /mp/).

Over the last decade, *quantitative dynamic models* for describing the movement behavior of elementary vocal tract actions have been developed. These models are capable of fitting natural effector movement data with high levels of accuracy [16, 17, 18]. Following [17], four parameters are sufficient for describing an elementary vocal tract action quantitatively: (i) *Starting time* and (ii) *ending time* for *activation* of the dynamical system which describes a specific vocal tract movement action, (iii) *target position* including *target shape* of the action-executing effector, and (iv) a time constant of the dynamical system which controls how rapid the effector approaches its target. This time constant was labeled inverse *rapidity* in [3]. Low rapidity values indicate fast target approximation, and high values indicate slow approximation. The starting time of an activation of a vocal tract action may occur earlier than that of the movement phase of an vocal tract movement action (cf. Fig. 2), and its ending time may occur earlier than that of the target or even that of the movement phase of an action. This is due to the fact that starting and ending time describe the *activation interval for the whole dynamical system* including high level control of the action. Thus, high level control represents the cortical premotor activation for an action [19]. But in addition to cortical activation, the dynamical system described here also models lower level control of neuromuscular activation. Moreover, our dynamic system for a quantitative description of actions also includes the biomechanical (mass-spring) response function of the effector.

Because consonantal vocal tract actions typically indicate a vocal tract constriction or closure created by contact between effector surfaces (e.g. lower and upper lips) or between the effector and the vocal tract wall (e.g. tongue tip or tongue body and alveolar ridge or hard palate), the quantitative dynamical model [17] was augmented with *virtual targets* in order to include these types of actions. Virtual targets are located beyond the contact area (place of articulation), so introduce a truncation (clipping) within the resulting movement trajectory during the target phase [18]. So, even during the target phase of an action, target approximation still continues. This ongoing process of target approximation quantitatively expresses the increase in contact between effectors or between the effector and the vocal tract wall as it occurs in natural constriction or closure formation. It should be noted here that effector movement velocity is relatively high during the movement phase and relatively low during the target phase (cf. the concepts of acceleration phase and asymptotic approximation phase as introduced for formant trajectories of vocal tract actions in [20]).

In the case of speech, the auditory domain is more important than the visual domain, meaning that speech can be understood unambiguously solely in the auditory domain (e.g. in a phone conversation) but not solely in the visual domain (i.e. even lipreading cannot produce an unambiguous representation of the corresponding speech). For this reason, speech action goals should be defined in the auditory domain in addition to their definition in the spatial or motor domain. Speech acquisition research has shown that speech actions can be defined straightforwardly in the auditory domain *as well as* the motor domain since it is one of the major goals

of early speech acquisition to learn sensorimotor relations [19]. So, the goal of a vocalic action, for example, can be coded on the one hand as static formant pattern and on the other as a vocal tract shape.

Moreover, it should be mentioned that a vocal tract shape must be coded on a *high* motor level (i.e. the coding of cavity shapes rather than of effector positions), since perturbation experiments indicate that speech sound qualities can be produced in spite of unexpected effector perturbations (e.g. of the lower jaw) without any need for additional learning or adaptation (e.g. [21]).

3.2 Tonal actions and phrase level intonation patterns, respiratory and voice quality actions

A phrase level intonation pattern in tone languages (e.g. Mandarin) as well as in stress languages (e.g. English) [2] can be described as a *temporal sequence of syllable-synchronized elementary tonal actions* [6, 22, 23]. The functional goal of each elementary tonal actions is to approximate an *underlying pitch target*. Underlying pitch targets can be dynamic (e.g. rising or falling) or static (e.g. high, middle, or low). The quantitative concept behind the description of elementary tonal actions is compatible with the concept introduced in [17] for vocal tract movement actions.

Moreover, it has been emphasized that the *temporal synchronization* of elementary tonal and vocal tract movement actions is possible with a high degree of accuracy only if both tonal and articulatory events are described in terms of the target approximation model [20].

In the framework of our action based approach [3, 17], which basically describes bodily movements and in which goals such as specific static bodily shapes are introduced in the visual or motor domain, it seems difficult to integrate "underlying static or dynamic pitch targets" (i.e. auditory goals). However, there are two reasons why this works. Firstly, the definition of bodily shapes or targets in our approach also includes dynamic shapes, for example manual actions in sign languages; see the description of "secondary movement actions" [7]. Secondly, the concept of movement (or elementary) actions should not be limited to movements which can be perceived exclusively in the visual domain. Especially in the case of speech it is widely accepted that effector movements can be perceived in the auditory domain (e.g. via formant trajectories).

Particularly for glottal actions, the description of goals in the auditory domain rather than in the spatial or motor domains seems to be much more feasible since very different articulatory solutions (i.e. laryngeal configurations) may exist for the realization of different pitches as well as different voice qualities. In contrast, for many supralaryngeal vocal tract actions, there is little physiological variation in articulatory formation, e.g. in vowel realization or the realization of a consonantal constriction or closure. For example, the same pitch can be realized by different humans using different configurations of muscular activations (e.g. vocalis vs. cricothyroid activity); another example would be a complex voice quality such as a breathy-laryngealized voice, which can be realized using different arytenoid configurations in different speakers.

The goal of a *respiratory action* can also be defined in the auditory domain. There is evidence that tracheal pressure is roughly constant during the production of an utterance and that the pressure level mainly determines the loudness of the utterance (e.g. a loud, normal, or soft voice); Active changes in tracheal pressure initiated by the respiratory system occur only in the case of emphatic stress [24]. Indirect or passive changes of subglottal pressure occur during phrases, but these changes result from changes in glottal aerodynamic resistance and are causes by changes in glottal configurations during the time course of each syllable, stress group, and phrase.

Thus, the goal of a respiratory action is always closely linked to the goal of one or more sequentially ordered *voice quality actions*. The voice quality action (e.g. breathy, normal, or pressed voice) determines the glottal aerodynamic resistance in the center of each syllable and so determines the strength of the appropriate respiratory action needed in order to realize the intended voice loudness. On the basis of these well-coordinated respiratory and voice quality actions, a *continuous voicing* is created, and it is interrupted only by (segmental) glottal abduction/adduction actions (i.e. by vocal tract actions for producing voiceless consonants).

It should be mentioned here that voice quality actions, like co-verbal facial actions (see Fig. 1 and section 4), can be composed of two or more movement actions. For example, in the case of breathy-laryngealized voice quality, a first movement action adjusts the arytenoids towards a whispering triangle, while one or more additional movement actions put the vocal folds into a slack and non-compressive position [5].

4 Co-verbal communicative actions and their temporal coordination with speech

While speech actions are perceived mainly in the auditory domain, co-verbal manual and facial actions are perceived in the visual domain. Thus, the goals of co-verbal movement actions can easily be defined in the spatial domain, and the goal of co-verbal manual and facial actions is shape formation [1].

Facial actions, whether co-verbal or not, can be described on the behavioral level by using the facial action coding system [25, 26, 27]. This system comprises a set of about 50 *facial movement actions* (i.e. facial action units or FACs) which control different parts of the face, e.g. corners of the mouth, eyebrows, eyelids, etc. A *facial expression action* (i.e. a higher-level meaningful or intentional facial action) is composed of facial movement actions which occur more or less synchronously in time. Thus, the temporal coordination of facial movement actions for a facial expression action is much simpler than the temporal coordination of vocal tract movement actions for a meaningful speech action such as a word (which in the simplest behavioral case comprises one syllable; see Fig. 2 and Fig. 3 and the examples given in [7]).

Moreover, a facial expression action – like a voice quality action – in many situations exhibits a *long target phase*, e.g. one that lasts the duration of a whole utterance or turn. In some situations, facial and voice quality actions may change faster, for example on the phrase or stress group level, i.e. from one stressed syllable to the next. Both types of actions are very well suited to conveying information concerning the emotional or affective state of the speaker.

Just as co-verbal facial actions may be related to voice quality actions, co-verbal manual actions may be related to tonal actions. As stated above, a *phrase level intonation pattern* is based on a sequence of syllable synchronized *elementary tonal actions* (see section 3.2). In a comparable way it can be assumed that a *manual gesture phrase*, which is composed of a preparation phase and a nucleus phase, where the nucleus phase comprises a stroke and an optional post-stroke hold phase (cf. [28]), is based on a sequence of *elementary manual movement actions* [1]. It can be assumed that the most meaningful part of a gesture phrase, the stroke phase, is synchronized in time with that syllable within a phrase that exhibits the intonational focus or tonal center of the phrase [29]. The subsequent post-stroke hold phase lasts until the production of the speech phrase is completed (ibid.).

Thus, the temporal coordination of the sequence of gesture phases, which constitute a meaningful gesture phrase, is closely related to syllable production and speech prosodic categories such as intonation patterns.

Moreover it should be noted that each gesture phase comprises not just one or more sequentially ordered manual movement actions but also temporally synchronous elementary manual

movement actions. For example, a pointing action may comprise two more or less simultaneous elementary manual movement actions (i) a hand-shaping action (for stretching the index finger) and (ii) a hand-arm movement action which moves the hand-arm system into the desired position.

5 A cognitive and sensorimotor control approach for face-to-face communication

A preliminary approach for production and/or perception (including comprehension) in face-to-face communicative situations has been outlined in [1]. With respect to the dichotomy of the cognitive function (meaning) and the behavioral form (sensorimotor realization) of actions, two main modules can be distinguished: *cognitive planning and comprehension* vs. *sensorimotor realization and perception*. From the production view, *planning* is on the border between cognition and sensorimotor processing. On the cognitive level we can distinguish between (i) *action planning* (i.e. the selection of *meaningful* actions) and (ii) *motor planning*. Motor planning comprises (a) the specification of all *elementary movement actions* which are needed for the production of a series of *meaningful (or higher-level) actions* and (b) the activation of a coarse temporal scheme for coordinating all elementary movement actions. But even motor planning as defined in [1] is still relatively abstract, since only goals or target shapes (including temporal landmarks for reaching these goals) are planned, while the concrete sensorimotor realization of elementary movement actions including a detailed movement description for all effectors involved is accomplished on a lower sensorimotor level (*motor programming,* following [1]).

Additionally, due to the importance of emotions and affect in face-to-face communication, and due to the influence of emotions and affect on cognitive processes (and vice versa), it will be important in future work to extend our control model as outlined in [1] by introducing an "emotion module" in addition to the "cognition module" introduced above (cf. [30]).

6 Discussion and conclusions

In this paper, our action based approach for describing vocal tract and co-verbal manual and facial actions [1] has been extended in two ways: (i) by including prosodic speech actions such as respiratory, voice quality, and tonal actions and (ii) by introducing the concept of level actions for describing the temporal coordination of vocal tract and prosodic speech actions and coordinating speech and co-verbal manual and facial actions.

This paper emphasizes the fact that elementary vocal tract movement actions as introduced in [17] are analyzed following the same qualitative and quantitative approach as outlined for elementary tonal actions [5]. This allows the formulation of a truly comprehensive quantitative behavioral approach to face-to-face communication.

Some parallels between facial and voice quality actions were drawn, on the behavioral level – i.e. long target phases and temporal organization above the syllable level – as well as on the level of communicative intention, e.g. expressing an emotional state. We hypothesized that a second parallel behavior may exist between tonal and manual actions. These actions are temporally coordinated primarily on the level of the syllable; however, in both cases the temporal serial ordering of tonal and manual actions leads to the phrase level intonation pattern or to a pattern of co-verbal gesturing.

In addition to manual and facial actions, gaze and head actions are important types of co-verbal actions, but the discussion of these types of actions is beyond the scope of this paper and will be a subject of future work.

Acknowledgements This work was supported in part by the German Research Council, project KR 1439/15-1, and in part by EU-COST action 2102.

Literature

[1] 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.

[2] Botinis, A., Granström, B., Möbius, B, "Developments and paradigms in intonation research", Speech Communication 33: 263-296, 2001.

[3] Kröger, B.J., Birkholz, P., "A gesture-based concept for speech movement control in articulatory speech synthesis", In: A. Esposito, M. Faundez-Zanuy, E. Keller, M. Marinaro [Eds], Verbal and Nonverbal Communication Behaviours, LNAI 4775 (Springer, Berlin), 174-189, 2007.

[4] Dogil, G., Möbius, B., "Towards a model of target oriented production of prosody", Proceedings of the European Conference on Speech Communication and Technology (Aalborg, Denmark), Vol. 1, 665-668, 2001.

[5] Klatt, D., Klatt, L., "Analysis, synthesis, and perception of voice quality variations among female and male talkers", Journal of the Acoustical Society of America, 87: 820–857, 1990.

[6] Xu, Y., Xu, C. X., "Phonetic realization of focus in English declarative intonation", Journal of Phonetics 33: 159-197, 2005.

[7] 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, in press. DOI: 10.1007/s12559-010-9071-2

[8] Saltzman, E., Kelso J. A. S., "Skilled actions: A task dynamics approach", Psychological Review 94: 84-106, 1987.

[9] Saltzman, E., Munhall, K. G., „A dynamical approach to gestural patterning in speech production", Ecological Psychology 1: 333-382, 1989.

[10] Saltzman, E., Byrd, D., "Task-dynamics of gestural timing: Phase windows and multifrequency rhythms", Human Movement Science 19: 999-526, 2000.

[11] Browman, C., Goldstein, L., "Articulatory gestures as phonological units", Phonology 6: 201-251, 1989.

[12] Browman, C., Goldstein, L., "Articulatory phonology: An overview", Phonetica 49: 155-180, 1992.

[13] Goldstein, L., Byrd, D., Saltzman, E., "The role of vocal tract action units in understanding the evolution of phonology", In: M. A. Arbib [Ed], Action to language via the mirror neuron system. (Cambridge University Press), 215-249, 2006.

[14] Goldstein, L., Pouplier, M., Chen, L., Saltzman, E, Byrd, D., "Dynamic action units slip in speech production errors", Cognition 103: 386-412, 2007.

[15] Lindblom, B., "Role of articulation in speech perception: Clues from production", Journal of the Acoustical Society of America 99: 1683-1692, 1996.

[16] Ogata, K. and Sonoda, Y., "Reproduction of articulatory behavior based on the parameterization of articulatory movements", Acoustical Science and Technology, 24: 403-405, 2003.

[17] Birkholz, P., Kröger, B. J., Neuschaefer-Rube, C., "Model-based reproduction of articulatory trajectories for consonant-vowel sequences", IEEE Transactions on Audio, Speech, and Language Processing, 19: 1422-1433, 2011.

[18] Birkholz, P., Kröger, B. J., Neuschaefer-Rube, C., "Articulatory synthesis and perception of plosive-vowel syllables with virtual consonant targets", Proceedings of Interspeech 2010 (Makuhari, Japan), 1017-1020, 2010.

[19] Kröger, B. J., Kannampuzha, J., Neuschaefer-Rube, C., "Towards a neurocomputational model of speech production and perception", Speech Communication 51: 793-809, 2009.

[20] Xu, Y., Liu, F., "Tonal alignment, syllable structure and coarticulation: towards an integrated model", Rivista di Linguistica 18: 125-159, 2006.

[21] Kelso, J. A. S., Tuller, B., Vatikiotis-Bateson, E., Fowler, C. A., "Functionally specific articulatory cooperation following jaw perturbations during speech: Evidence for coordinative structures", Journal of Experimental Psychology: Human Perception and Performance 10: 812-832, 1984.
[22] Xu, Y., Wang, Q. E., "Pitch targets and their realization: Evidence from Mandarin Chinese", Speech Communication 33: 319-337, 2001.
[23] Prom-on, S., Xu, Y., Thipakorn B., "Modeling tone and intonation in Mandarin and English as a process of target approximation", Journal of the Acoustical Society of America 125: 405-424, 2009.
[24] Finnegan, E. M., Luschei, E. S., Hoffman, H. T., "Modulations in respiratory and laryngeal activity associated with changes in vocal intensity during speech", Journal of Speech Language and Hearing Research 43: 934-950, 2000.
[25] Ekman, P., Friesen, W. V., "Measuring facial movement", Environmental Psychology and Nonverbal Behavior 1: 56-75, 1976.
[26] Ekman, P., Friesen, W. V., "The Facial Action Coding System (FACS): A Technique for the Measurement of Facial Action", Consulting Psychologists Press. (Palo Alto, CA), 1978.
[27] 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), 203-221, 2007.
[28] Kendon, A., Gesture: Visible Action as Utterance. (Cambridge University Press, New York), 2004.
[29] Kopp, S., Wachsmuth, I., "Synthesizing multimodal utterances for conversational agents", *Journal* of *Computer Animation* and *Virtual Worlds* 15: 39-51, 2004.
[30] Becker, C., Kopp, S., Wachsmuth, I., "Why emotions should be integrated into conversational agents", In: Nishida, T. [Ed.], Engineering Approaches to Conversational Informatics. (John Wiley & Sons, Hoboken, NJ), 49-68, 2007.