Ph.D. thesis – abstract

Synthesis and Coordination of Speech and Gesture for Virtual Multimodal Agents

Stefan Kopp

Humans intuitively employ speech and gesture in most communicative acts. Thereby, both modalities are closely connected in conveying an intent such that coexpressive elements appear temporally synchronized as well as in semantically and pragmatically coordinated form. Generating lifelike and coherent multimodal utterances for communicative agents therefore includes, first, the synthesis of intelligible verbal and gestural acts per se and, secondly, their mutual coordination and integration into a continous, human-like flow of multimodal behavior. This thesis addresses both problems for the basic levels of speech and gesture production, from declarative specifications of the speech's surface structure and the morphology of accompanying gestures to multimodal utterances of a virtual humanoid agent.

As a starting point, a XML compliant markup language for the representation of multimodal utterances (MURML) is defined. It provides a feature-based notation system for the outer form of hand-arm gestures along with flexbile means of describing verbal parts (including prosodic properties) and cross-modal correspondences at this surface level. Affiliation usually results from gesture and speech expressing concertedly the rhematic element that in speech frequently receives prosodic focus and hence forms the nucleus of the intonation phrase. The increased speech accentuation engenders a tighter coupling to the affiliated gesture whose meaning-bearing phase (stroke) is consistently found to overlap the nucleus. An analysis of empirical and theoretical findings further reveals that humans accomplish this synchrony by means of cross-modal adaptation: Gesture significantly adapts to speech at the boundaries of ballistic movement phases (preparation onset, stroke onset and duration) while in speech only the onset of phonation is influenced. In result, the onsets of both, the gesture phrase and the intonation phrase covary with the position of the verbal affiliate and the gesture stroke, respectively.

For speech synthesis, a system is presented that builds on and extends the capabilities of txt2pho and MBROLA. It controls prosodic parameters like speechrate and intonation and delivers detailed timing information at the phonem level. In addition, it is able to prosodically focus single words by simulating realistic pitch accents (e.g., for emphatic or contrastive stress).

For gesture synthesis, a comprehensive approach to building and executing gesture animations is presented that emphasizes the accurate and relieable reproduction of the prescribed spatiotemporal properties of the stroke phase. The generation system combines high-level gesture planning with low-level motor planning and drives in real-time the anthropomorphic kinematic skeleton of the agent. Gesture planning defines the expressive gesture phase in terms of movement constraints by way of (optionally) selecting a lexicalized gesture template, allocating body parts, expanding movements constraints, resolving deictic references, and applying timing constraints. Motor planning seeks a solution to control movements of the agent's upper limbs that satisfy the imposed constraints.

Following a biologically motivated, functional-anatomical decomposition of motor control, the movements are driven by multiple kinematic controllers (local motor programs) running concurrently and synchronized. Specialized motor control modules for the hands, the wrists, the arms, and the neck instantiate the motor programs and arrange them in controller networks that lay down their potential interdependencies (de-/activation). At execution-time, the motor programs are able to de-/activate themselves as well as other controllers in the network. That way, the application of suitable motion generation techniques to control realistic submovements, i.e., within a limited set of DOFs and for a designated period of time, is coordinated. In particular, a new method is introduced for forming parametric curves that resemble natural wrist trajectories in work space while directly reproducing the required gestural form features.

To integrate speech and gesture production, a hierarchical structure of overt multimodal utterances is assumed that suggests an incremental nature of the overall production process: Single chunks, each expressing a single idea unit by an intonation phrase paired off with a gesture phrase, are subsequently uttered, but produced in an interleaved fashion. A production model is presented that for each chunk extends the classical two-phase, planning-execution procedure by additional phases ("pending", "lurking", and "subsiding") in which the production processes of subsequent chunks interact. That way, cross-modal coordination takes place on two levels of the utterance's structure: Within a chunk, the gesture stroke is synchronized with the verbal affiliate and focus, respectively, to ensure cross-modal coherence. At the inter-chunk level, gesture transitions depend on the, in turn, speech-dependent timing of the successive stroke (co-articulation). Resulting automatically from the self-activation of motor control primitives, they may range from the adoption of intermediate rest positions to direct transitional movements. Likewise, the onset of the – more ballistically executed – intontation phrase is influenced by the gesture's timing (deferred in accord to the anticipated duration of gesture preparation).

Using this production model, complex multimodal utterances comprising multiple chunks have been synthesized in a coherent and continous fashion, with realistic coordination effects between gesture and speech. This is demonstrable in "Max", a virtual humanoid agent that is able to utter himself multimodally and in real-time from given MURML representations.