Research Statement

Timo Baumann, baumann@informatik.uni-hamburg.de

I work on the cross-section of spoken language, computational linguistics, informatics and human-machine interaction. My research is geared towards *interaction management* in spoken dialogue. I'm specifically interested in the role of subtle linguistic and non-linguistic cues (prosody, timing, gaze, …) that control the flow of interaction and are independent of dialogue content. I pursue my interaction research by building computational real-time models of dialogue that can be used as foundations for spoken dialogue systems (SDSs). The current systems' mode of interaction is a major impediment to the acceptability of dialogue systems. I believe that a leap in SDSs design and performance will result from considering more the interactional cues, which will extend the application of SDSs to less clearly-defined tasks that require a larger degree of *conversational competence*. Going beyond the mere improvement of applied SDS research, I hope to advance our knowledge of dialogue as a system, as well as to improve tools that support human-human dialogue, e. g. simultaneous speech-to-speech translation, silent speech interfaces, tools for the visually impaired, etc.

My past work convinces me that *system architecture* is a main driving factor towards more natural SDS interactions. In the past, I have devised an architecture and software toolkit for *fine-granular incremental* spoken dialogue processing. Using the architecture, I have developed and analyzed incremental speech recognition, as well as incremental speech synthesis, and incremental dialogue flow estimation. Using these components, I have been able to build (partial) SDSs that are incremental *from the bottom up* and have been shown to be preferred over non-incremental systems, or even to enable behaviours that are impossible to achieve at all without incremental processing.

Having increased the granularity of processing, I want to turn to the system theoretic full-system perspective of dialogue processing, which I find to be another main impediment of current architectures. Specifically, I believe that dialogue is a *complex system* in which behaviour *emerges* from the collaboration of individual modules, as well as collaboration between the communicating agents in the dialogue. While *decoupling* is a goal in modularization, more advanced interaction of modules (coupling, e. g., speech input and speech output) makes it easier to rely on *attractors* of dialogue as stabilizing factors. I believe that some (soft) notion of *transaction* is required to manage the interplay between components in an incremental complex system.

I believe that *prosody* plays a vital role in everyday conversation (and joint work with N. Ward and A. Vega has shown that the inclusion of prosody into language modelling results in improved speech recognition performance) and that it is still too often ignored in practical systems, possibly because its advantages cannot be exploited in non-incremental, 'sluggish', and pipelined systems, but only if results and analyses become available with very low delays and can be handled across many module boundaries. It is for this reason that I have started to work on *fully incremental prosody models*, based on partial utterance specifications, for speech synthesis. I plan to extend this work towards higher-level incremental prosody modelling (going beyond the mere features to a higher-level model) as part of the proposed project, for both the analytical as well as the synthesis aspects of prosody, ideally combining them in a unified model. Thus, prosody modelling is not only an example problem for the architecture developed in this project, but also a proper research area in itself. While already necessary for prosody, higher abstraction levels (such as semantics and pragmatics) depend on the architectural advancements that are proposed here to an even larger degree.

I believe that fine-granular incrementality in combination with architectures providing for complex interaction hold a large value for conversational competent dialogue behaviours and will – in the long run – help to enable better, more natural spoken dialogue systems, as well as a deepen our understanding of human dialogue itself: my systems could themselves become tools for research, e. g. by allowing to distort human-human dialogue in a controlled way in order to investigate repair mechanisms.