## Position paper in defense of the doctoral dissertation

# Incremental Spoken Dialogue Processing: Architecture and Lower-level Components

Timo Baumann

This position paper sets into context the results of my thesis, namely that *incremental spoken dialogue processing is technically feasible and successful at enabling more natural interaction*. Furthermore, the thesis provides the low-level *building blocks* incremental speech recognition, synthesis and dialog-flow estimation.

### Proposition 1: Incremental Processing Challenges the Paradigm of Spoken Dialogue Systems

Incremental processing challenges conventional spoken dialogue systems in two ways: the predominant processing paradigm and the predominant interaction paradigm.

Regarding the *processing paradigm*, the thesis shows that incremental spoken dialogue processing can lead to systems that outperform the conventional *full-turn processing* approach in multiple ways, at least in some small example domains. The chosen domains, however, differ from the main application areas of conventional SDSs, and – more importantly – only cover very small dialogue state spaces. Thus, for the foreseeable future, incremental and non-incremental processing will have to co-exist and cooperate in applied systems. Questions that arise are that of combining non-incremental and incremental components (e. g. Baumann et al. 2013), and the integration into the commercial eco-system which mostly uses VoiceXML, SRGS, and SSML as interface languages (which are unsuitable for incremental processing in their current states).

Regarding the *interaction paradigm*, the thesis investigates several issues, such as enabling faster turn-taking, utterance collaboration and utterance co-completion as examples of system-initiative at turn-taking, and direct 'steering' by speech, which all

challenge the conventional *ping pong* paradigm of interaction, leading to more overlap and floor negotiation instead. It should be discussed whether this more assertive floor management is advantageous in full systems, whether it is accepted by users (or some groups of users), and in which cases. Finally, multi-modal systems such as Google Voice Search already employ incremental speech recognition and display partial results. It will be interesting to discuss whether (and how) such immediate cross-modal feedback (e. g. on recognition errors) alters interactions and how users might creatively use system prediction capabilities (e. g. by skipping the realization of predicted parts of utterances as enabled by iTap for typing; Nowlan et al. 2001).

### Proposition 2: Incremental Spoken Dialogue Processing Opens New Ways for Dialogue and Interaction Research

Speech technology has become an influencing factor on speech research itself (Boersma 2002) and dialogue technology has been employed for dialogue research, e. g. to systematically investigate clarification strategies in the *DiET toolkit* (Healey et al. 2003).

Incremental spoken dialogue processing can help to advance research on aspects of *spoken* interaction by engaging a subject in a (limited domain) dialogue where incremental processing could e. g. enable the system to systematically alter turn-taking timing to test the effect of eager or sluggish turn-taking on the interlocutor.

A second, more advanced endeavour would be to combine incremental recognition and (re-)synthesis with voice-morphing technology to construct a system that alters a user's speech in (almost) real-time. This would allow to research the individual contribution of specific aspects of speech on *human-human* spoken dialogue in the spirit of the DiET/DynDial project. Modifications could encompass (in order of growing complexity) pitch deviations/excursions to alter syllable stress, vowel/consonant ratios, tempo and timing, or even introduction/suppression of speech material.

### Proposition 3: A Joint Speech Input and Speech Output Component – Re-Modelling the Dialogue System Architecture

Modular dialogue systems typically handle incoming and outgoing speech in separate components and (in industrial settings) use different interface languages. Some attempts to sharing processing models (based on HMMs) exist (Strecha et al. 2009).

More interesting would be the question how a joint speech input and output component would change a system's modularization, and the overall processing model of the system. A combined speech component (for both input and output) could completely encapsulate all detailed timing issues and would greatly help to reduce the complexity

in linguistic and planning/management modules, while at the same time enabling 'precise' timing (e. g. of back-channels). The close coupling of speech input and output processing also provides for *reflexive* behaviour (such as feedback utterances) without further higher-level intervention. Finally, where mirroring and mimicking require collaboration of separate modules (in the case of speech input and output, these modules are maximally far apart in conventional systems), such behaviour becomes trivial if input and output on one level is handled by a joint component.

Of course, a purely speech-based component (without deep linguistic insight) could only have a sketchy 'default' notion of proper timing/loudness/colouring production and turn-taking behaviour based on the interlocutor's (or other) speech. Thus, an appropriate interface for timing decisions should be devised that supports both underspecified and detailed intervention of higher-level processors on the resulting behaviour.

## References


Baumann, Timo, Maike Paetzel, Philipp Schlesinger, and Wolfgang Menzel (2013). "Using Affordances to Shape the Interaction in a Hybrid Spoken Dialogue System". In: *Proceedings of ESSV 2013*. Ed. by Petra Wagner. TUDpress, pp. 12–19.

Boersma, P. (2002). "Praat, a system for doing phonetics by computer". In: *Glot international* 5.9/10, pp. 341–345.

Healey, Patrick, Matthew Purver, James King, Jonathan Ginzburg, and Greg Mills (2003). "Experimenting with clarification in dialogue". In: *Proceedings of the 25th annual meeting of the cognitive science society*. Citeseer, pp. 539–544.

Nowlan, Steven, Ali Ebrahimi, David Richard Whaley, Pierre Demartines, Sreeram Balakrishnan, and Sheridan Rawlins (Mar. 20, 2001). "Data entry apparatus having a limited number of character keys and method". Pat. US 6,204,848.

Strecha, Guntram, Matthias Wolff, Frank Duckhorn, Sören Wittenberg, and Constanze Tschöpe (2009). "The HMM synthesis algorithm of an embedded unified speech recognizer and synthesizer". In: *Proceedings of Interspeech*.