Project Proposal for the Daimler and Benz PostDoc Award

A Concurrent, Distributed, and Incremental Spoken Dialogue Architecture with a First Application to Prosody

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

15. Oktober 2013

Current-day spoken dialogue systems are tedious to interact with (Ward et al. 2005). Their naturalness and (measurable) quality of interaction can be improved through incremental (step-by-step) processing schemes that enable dialogue systems to *interact continuously* (Baumann 2013). However, incremental models have not yet adequately addressed the challenge of joint decision making and optimization of hypotheses across the multitude of components within a modularized system *in real-time*, mostly because their data-flows follow *simple pipeline* approaches. Ad-hoc integration of modules fails completely for distributed systems which are preferred in robotics, for research systems, and in mobile applications. This shortcoming impedes incremental spoken dialogue systems to leverage their full potential.

This project proposes to design and implement an architecture for concurrent, distributed incremental processing and knowledge representation for spoken dialogue in which components share their understanding and collaborate on the *emergence* of desirable dialogue behaviour. The architecture will be applied to (limited) spoken dialogue domains. Prosody and timing are key issues to successful interaction and control dialogue flow, regardless of its content. Thus, the project will focus on the interaction between speakers on the prosodic level.

1 Background

Spoken dialogue systems (SDSs) interact with human interlocutors through spoken language. Current systems are limited in naturalness and can only be used in task-based environments, where humans accept a tedious interaction in order to reach a concrete goal (like reserving a ticket). More advanced systems could also be used in *conversational* settings, such as (ordered by complexity) teaching, entertaining, counseling, or even more empathetic conversation, which are beyond the capacity of current systems. I aim to tackle those limitations in the interaction of current systems, which stem from their architecture and the associated mode of processing.

Due to the complexity of the task, and for psycholinguistic and cognitive plausibility (Levelt 1989), SDSs are highly modular systems, with distinct, specialized modules for different sub-tasks. In conventional systems, the interface between modules is very narrow, with full speaker *turns* being the unit of interaction, and information flow being limited to a classical *pipeline* that only passes on the immediate results. Thus, after recognizing and then understanding an interlocutor's full turn, and then taking dialogue context into account, the best system action is determined, after which spoken output for that action is generated and then synthesized, with each step only being informed about the preceding step's output. Such systems are limited to act only *after* a user turn is over (resulting in unnatural and disturbing pauses in turn-taking), are unable to react to listener feedback while delivering their own turns, and, foremost, ignore the linguistic insight that dialogue is best seen as a complex system (Larsen-Freeman and Cameron 2008). As modules cannot access the system's information in full, decisions are taken locally and separatedly on input and output sides. For example, processing of speech input and output is maximally apart in a pipeline, rendering prosodic interaction impossible.

In my doctoral studies (Baumann 2013), I have worked on *incremental* dialogue processing, where the unit of the turn is broken up into units as small as possible at the respective level of abstraction. This increase in the *granularity* of processing also increases the responsiveness of system interactions and builds on the fact (and allows to model) that dialogue is a collaborative process (Clark 1996) in which interaction behaviour such as turn-taking *emerges* as a result of the behaviour of both the speaker *and* the interlocutor (Thórisson 2008). In the architecture underlying my work (Schlangen and Skantze 2009) information is interchanged in the form of minimal units of information (IUs) which are interlinked to form a network that evolves over time, reflecting the accumulation of understanding in the system. Apart from implementing incremental processing, the IU network somewhat opens up the black-box approach of conventional systems, as all the knowledge of the system is available via links in the IU network.

The model as implemented so far (Baumann and Schlangen 2012) focused on incremental processing only and does not remove the principled limitations of the pipeline architecture: it lacks systematic ways for modules to influence other modules apart from the one that immediately follows in the pipeline. While data is readable in the network, manipulation of data across module boundaries is so far ad-hoc, non-generic and error-prone. It has become clear that further advancements in dialogue behaviour require more elaborate and collaborative inter-module communication. However, multi-directional communication radically complicates the architecture, especially given the modular and incremental processing paradigm, in which modules process concurrently and as autonomously as possible, changes to hypotheses are abundant, and there is pressure to act in real-time. Investigating these problems and finding solutions for them is the target for the present project.

2 Project goal

The goal of the project is to re-formulate the previously developed architecture and toolkit for incremental spoken dialogue processing as a distributed, object-based database management system that is tailored towards the needs of an incremental spoken dialogue system.

Specifically, the architecture will implement *transactions* to allow a module to alter the network of incremental units as a whole (instead of only the units it created itself), without interfering with other modules or loosing consistency. Transaction priorities and update notifications will have to be managed as they result from real-time pressure (e. g. changing how an ongoing utterance is being spoken is possible only up to a certain time in advance) and this pressure should be handled gracefully by the architecture and its components (e. g. by slowing down, or hesitating, or performing some other 'cover-up' as necessary), relating this work to the area of anytime algorithms (Dean and Boddy 1988). Finally, transactions may fail and the system will have to handle such cases. Ideally, the system should not have a central arbiter but instead will be implemented using distributed database management (known as 'NoSQL' databases; Cattell 2011). Distributed databases cannot support all ACID properties of transactions, and ideal trade-offs for the incremental use-case have to be found.

The system will support to share and synchronize 'ownership' of data among modules and thus the collaboration of multiple processing modules to produce emergent behaviour, begging the question of the 'correct' middle-ground between modularity and tight coupling in spoken dialogue processing, which will be investigated as part of the project, in the domain of *incremental prosody modelling*.

This project proposes an 'evolutionary' approach that extends my previous work, namely the software toolkit InproTK which is being used at several institutions world-wide and, even though it is free and open-source software, is highly linked to my person. Supporting and extending it is part of my longterm plans. As InproTK also is the foundation of the envisaged system, some resources will be devoted to maintaining and advancing its codebase. The main research focus, however, will lie on theoretical and architectural considerations and thus be driven by informatics considerations (i. e. the formulation as a database management task). The 'measurable' deliverables of this project will, however, be stated as proof-of-concept systems that deal with prosodic interaction between user and system, as detailed in the following section.

3 Deliverables

In order to simplify measurement of success and to provide clear targets for the project, two applications are proposed that show-case the relevant aspects of the architecture to be conceived.

Synchronous, prosodically motivated co-completion The first prototype combines incremental speech recognition, dialog flow estimation, and incremental speech synthesis with *an incremental prosody model*, in order to build a system that is able to speak in synchrony with a user (similarly to synchronous reading tasks; Cummins 2002), while trying to closely mimic the user's speech (tempo, prosody, possibly accentuation details). The system will build up higher-level structure just-in-time (prosodic modelling, as well as the analyzed output from ASR) and use this structure for estimating/predicting upcoming speech. The successful interplay between prosody recognition and production requires the system architecture to support the aforementioned transaction management in the system.

A much simpler proof-of-concept for a co-completing system has been given in (Baumann and Schlangen 2011), which, however, ignored prosody and used word-by-word speech synthesis. The proposed project will build on that system but significantly extend it by including incremental and adaptive speech synthesis and real-time prosody adaptation.

Speaking in synchrony is rather a good technology test than a useful application; the demonstrator to be described next will feature a more useful capability, by extending over the present demonstrator. Furthermore, a structured, fully incremental just-in-time prosody model will be useful for speech-to-speech translation (Bangalore et al. 2012) and attempts will be made to apply it to this task in collaboration with AT&T.

Prosodically integrated feedback placement The real-time capabilities implemented by the first deliverable are useful for advanced synchronized behaviours, such as precisely aligned feedback utterances. The incremental prosody model will be put to use to select/predict plausible back-channelling opportunities (identified by backchannel-inviting cues; Gravano and Hirschberg 2009) and the system will continuously monitor its behaviour, both wen giving as well as when receiving back-channel feedback.

We will test the resulting back-channelling behaviour both in corpus experiments and in a small user study. While it is well known that humans coordinate their feedback with the interlocutor I would like to investigate whether precisely aligned back-channel feedback is also rated as natural and helpful in *human-machine* interaction.

4 Further research directions

A natural extension to feedback placement is the 'reverse' task of fighting for the floor (the right to talk). While at first this may seem like uncooperative behaviour, turn-fights are indeed highly coordinated behaviour that is also managed prosodically (Schegloff 2000). A system for turn-fights (which can be useful, for example if a computer needs to convey important information to a very talkative user) is a natural successor of the demonstrators mentioned above.

The architecture will include provisions for multi-modal processing, supporting information fusion and fission, as well as concurrent processing. Actual inclusion of vision or other modalities must be left to future work, though.

One of my longterm goals is to use the incremental buildup of linguistic structure for *real-time manipulations* to the user's speech and to use systematic manipulations to subjects' speech for dialogue research, similar in spirit to manipulations of text chats in the DynDial project (Healey et al. 2003; Purver et al. 2009). However, I believe that such a system is far beyond the scope of the current project. Time permitting, I want to start to work on *manipulating* copy-synthesized speech in ways that are impossible to achieve for plain, surface-based voice-morphing (Ye and Young 2006), for example by altering pitch excursions for certain types of accentuations only, altering vowel/consonant proportions, inserting or suppressing material that is spoken, and the like.

References

Bangalore, Srinivas et al. (June 2012). "Real-time Incremental Speech-to-Speech Translation of Dialogs". In: *Proceedings of NAACL-HTL 2012*. Montréal, Canada, pp. 437–445. URL: http://www.aclweb.org/anthology/N12-1048.

Baumann, Timo (May 2013). "Incremental Spoken Dialogue Processing: Architecture and Lower-level Components". PhD thesis. Universität Bielefeld, Germany. URL: http://nbn-resolving.org/urn:nbn:de:hbz:361-25819101.

Baumann, Timo and David Schlangen (2011). "Predicting the Micro-Timing of User Input for an Incremental Spoken Dialogue System that Completes a User's Ongoing Turn". In: *Proceedings of SigDial 2011*. Portland, USA.

– (2012). "The InproTK 2012 Release". In: *Proceedings of SDCTD*. Montréal, Canada.

Cattell, Rick (May 2011). "Scalable SQL and NoSQL data stores". In: *SIGMOD Rec.* 39.4, pp. 12–27. ISSN: 0163-5808. DOI: 10.1145/1978915.1978919.

Clark, Herbert H. (1996). *Using Language*. Cambridge University Press. ISBN: 978-0521567459.

Cummins, Fred (2002). "On synchronous speech". In: *Acoustic Research Letters Online* 3.1, pp. 7–11. ISSN: 1529-7853.

Dean, Thomas and Mark Boddy (1988). "An Analysis of Time-Dependent Planning". In: *Proceedings of AAAI-88*. AAAI. Cambridge, USA, pp. 49–54.

Gravano, A. and J. Hirschberg (2009). "Backchannel-inviting cues in task-oriented dialogue". In: *Proceedings of Interspeech*. Vol. 2009, pp. 1019–1022.

Healey, Patrick et al. (2003). "Experimenting with clarification in dialogue". In: *Proceedings of the 25th Annual Meeting of the Cognitive Science Society*, pp. 539–544.

Larsen-Freeman, Diane and Lynne Cameron (2008). *Complex Systems and Applied Linguistics*. Oxford University Press.

Levelt, William J.M. (1989). *Speaking: From Intention to Articulation*. Mit Pr.

Purver, Matthew et al. (Sept. 2009). "Split Utterances in Dialogue: a Corpus Study". In: *Proceedings of SIGdial*. London, UK, pp. 262–271.

Schegloff, Emanuel A (2000). "Overlapping talk and the organization of turn-taking for conversation". In: *Language in society* 29.1, pp. 1–63.

Schlangen, David and Gabriel Skantze (2009). "A General, Abstract Model of Incremental Dialogue Processing". In: *Proceedings of the EACL*. Athens, Greece, pp. 710–718.

Thórisson, Kristinn R. (2008). "Modeling Multimodal Communication as a Complex System". In: *Modeling Communication with Robots and Virtual Humans*. Ed. by Ipke Wachsmuth and Günther Knoblich. Vol. 4930. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 143–168. ISBN: 978-3-540-79036-5. DOI: 10.1007/978-3-540-79037-2_8.

Ward, Nigel G. et al. (2005). "Root causes of lost time and user stress in a simple dialog system". In: *INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005*. ISCA, pp. 1565–1568. URL: http://www.isca-speech.org/archive/interspeech_2005/i05_1565.html.

Ye, Hui and Steven Young (2006). "Quality-enhanced voice morphing using maximum likelihood transformations". In: *Audio, Speech, and Language Processing, IEEE Transactions on* 14.4, pp. 1301–1312. ISSN: 1558-7916. DOI: 10.1109/TSA.2005.860839.