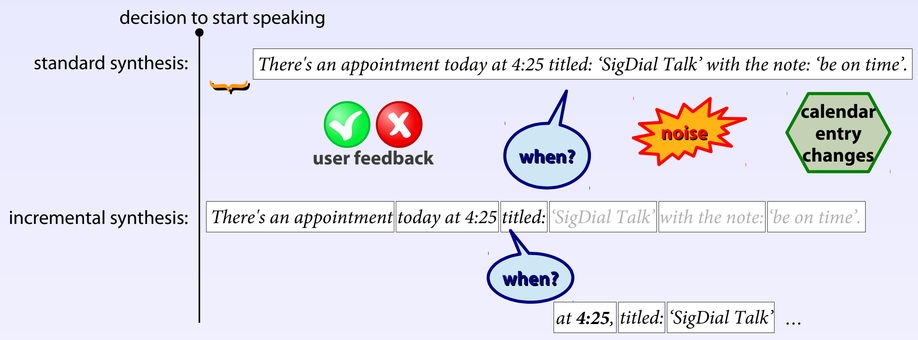
INPRO_iSS: A Component for **Just-In-Time Incremental Speech Synthesis**

Timo Baumann • Natural Language Systems Division • Department of Informatics • University of Hamburg • Germany • baumann@informatik.uni-hamburg.de
David Schlangen • Dialogue Systems Group • Faculty of Linguistics and Literature • Bielefeld University • Germany • david.schlangen@uni-bielefeld.de

Incremental Speech Synthesis: **What is it good for?**

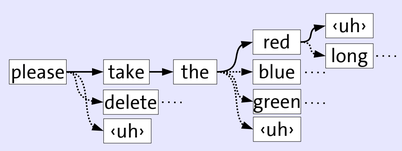
- conventional speech synthesis systems are optimized for non-interactive reading tasks
  - full utterances are required as input
  - no changes / extensions / adaptation to ongoing utterance is allowed
  - *ill-suited for highly-dynamic environments*

- incremental speech synthesis allows:
  - to start delivery before the whole utterance has been generated and processed
  - to change delivery while it is ongoing
  - with very low latency, and only little loss in synthesis quality



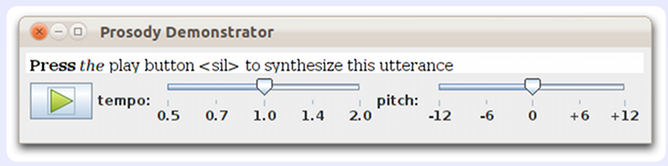
Demo: Switching Utterance Branches

- a tree of different utterance out-comes can be pre-compiled
  - very low latency for highly-interactive environments
  - good prosodic quality as future of the utterance is taken into account
  - branches can also be added on-the-fly
- Demo: click to select utterance branches



Demo: Online Prosodic Adaptation

- HMM synthesis separates duration, pitch and cepstral properties
  - simple to manipulate pitch and tempo
  - immediately before vocoding (*just-in-time principle*)
- Demo: sliders to change tempo and pitch

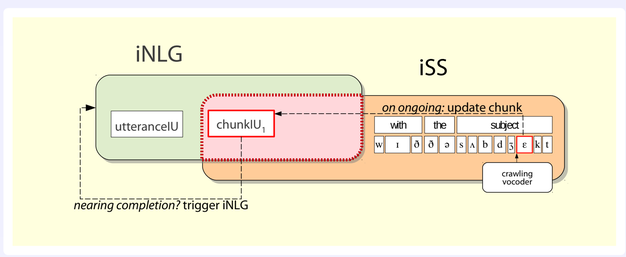


  - (resulting parameter sequence is not optimal w.r.t. original HMM state sequence – but this does not have practical implications)
- a more principled approach (future work):
  - determine HMM state sequence incrementally
  - compute duration/f_0 incrementally from structured representation

Demo: Integration with iNLG

(see also: Buschmeier et al., SIGDial 2012)

- the full iSS component has been integrated with an incremental NLG component that reacts to a noisy channel
  - upon noise: stop after current word, restart phrase after noise ends
  - rated significantly more natural than two baselines (ignore, stop/continue)



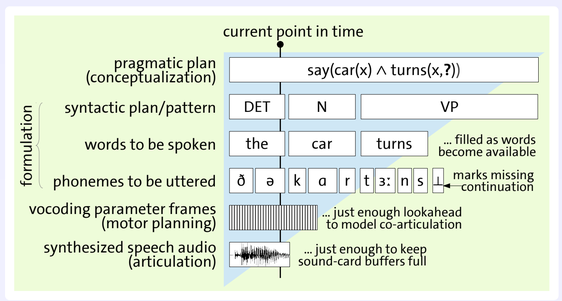
- prosodic quality *similar to non-incremental system* with just one phrase of lookahead
  - details: Baumann and Schlangen (Interspeech 2012)

How can I use it?

- INPRO_iSS is a component of INPROTK:
  - output to speakers, RTP, and other audio streams
  - flexible framework for incremental processing
- clean interface, based on the IU framework:
  - add sub-utterance chunks of words (or individual words)
  - add alternatives and switch between them (see left)
  - adapt words that haven't been realized yet
  - stay informed about delivery via update-listener interface
- relies on MaryTTS for non-incremental processing, voice models, ...
  - does not support out-of-the-box Java-Webstart functionality yet

How does it work?

- uses a triangular processing scheme where data on each level are produced just-in-time and as late as possible
  - enough lookahead on higher levels to produce high-quality prosody
  - hardly any lookahead on synthesis level for high responsivity



- details Buschmeier et al.: „Combining Incremental Language Generation and Incremental Speech Synthesis for Adaptive Information Presentation“ *Proceedings of SIGDial 2012*.

Further Information

This software is available as open-source in **INPROTK**, the incremental dialogue processing toolkit, at **http://inprotk.sourceforge.net** , Background on INPROTK is available at **http://www.inpro.tk** .

Acknowledgements

This work was largely funded by DFG in the Emmy-Noether Proramme and supported through a travel grant by the German Academic Exchange Service.

Further references:

Hendrik Buschmeier, Timo Baumann, Benjamin Dorsch, Stefan Kopp and David Schlangen (2012): "Combining Incremental Language Generation and Incremental Speech Synthesis for Adaptive Information Presentation", in *Proceedings of SigDial 2012*, Seoul, South Korea.

→ discusses in depth the approaches for incremental speech synthesis and incremental natural language generation and their combination in an adaptive, incremental speech output pipeline.

Timo Baumann and David Schlangen (2012): "Evaluating Prosodic Processing for Incremental Speech Synthesis", to appear in *Proceedings of Interspeech 2012*, Portland, USA.

→ discusses an evaluation method suitable for incremental speech synthesis (comparing incremental with non-incremental synthesis) and presents numbers that support our claim that slightly less than one intonation phrase of lookahead is sufficient for high-quality iSS.

Timo Baumann and David Schlangen (2012): "The InproTK 2012 release", in *Proceedings of the SDCTD Workshop*, Montréal, Canada.

→ discusses features and properties of InproTK, our toolkit for incremental spoken dialogue processing.

Timo Baumann 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.

→ a system that incrementally co-completes the user's ongoing speech (i.e., it says the same words as the user, at the same speed, at precisely the same time), which highlights the need for incremental speech synthesis.

Timo Baumann, Okko Buß and David Schlangen (2011): "Evaluation and Optimisation of Incremental Processors", in *Dialogue & Discourse*, **2**(1), Special Issue on Incremental Processing in Dialogue.

→ discusses our model of incremental processing, evaluation methodology and results for incremental speech input processing.

InproTK is open-source and available at http://inprotk.sourceforge.net
More information on the Inpro project is available at http://www.inpro.tk