Evaluating Prosodic Processing for Incremental Speech Synthesis

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

Abstract

We analyze the prosodic quality of our incremental speech synthesis component INPRO_iSS [1], which, in incremental processing, only has limited context available.

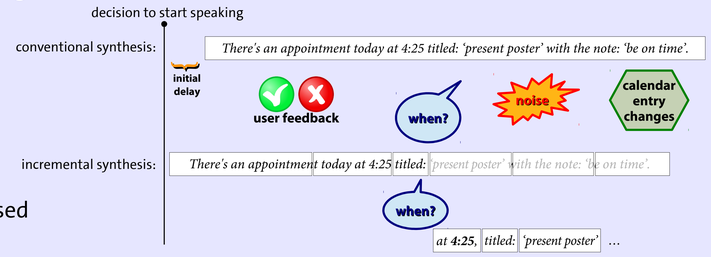
For incremental prosodic assignment, there is a tradeoff between the amount of lookahead vs. the resulting timeliness and quality of the generated prosodic contours.

We found that high quality incremental output can be achieved even with a lookahead of less than one phrase, allowing for timely system reaction.

In our method, we encapsulate a non-incremental processor which is called repeatedly, which proves to be a reliable and simple solution.

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

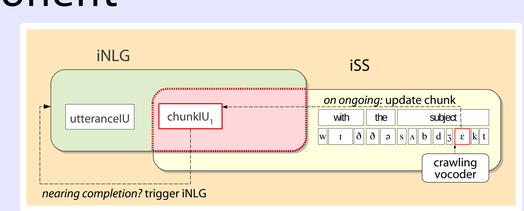
- **conventional** speech synthesis systems are optimized for non-interactive reading tasks
  - full utterances are required as input
  - no changes / extensions / adaptation to ongoing utterance is allowed
  - *ill-suited for highly-dynamic environments*
  - *relatively long utterance-initial delay*

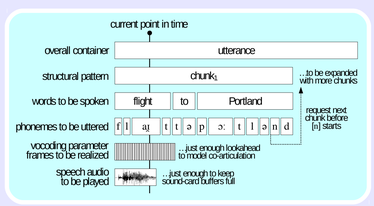


- our **incremental** speech synthesis allows:
  - to start delivery before the whole utterance has been generated and processed
  - to change delivery while it is ongoing
  - gives very low latency, and *only little loss in synthesis quality* (that's the topic of this paper)

Our **Incremental Speech Synthesis** Component

- implemented in INPROTK [2] using MaryTTS [3] based on the IU framework [4]
  - interconnected modules create and extend a network of IUs
  - IUs can trigger updates via a call-back mechanism
- data are produced just-in-time in a triangular processing scheme
- a *crawling vocoder* performs piece-wise HMM optimization [5] and vocoding
  - hardly any lookahead on synthesis level for high responsivity
- as synthesis progresses, updates are sent to iNLG [6] demanding for more chunks
  - other, external events could also lead to the iNLG changing the IU networks
- key question: *when can updates occur at the latest, without deteriorating quality too much?*





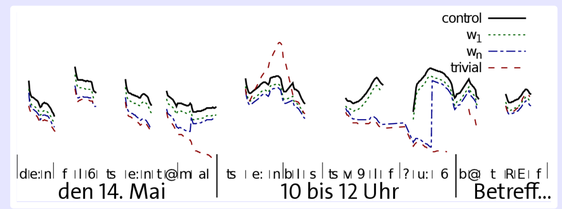
Design Space for Incremental Prosody

- how much history to consider ⟶ question of *left context*
- when to add more words ⟶ question of *lookahead*
- how many words to add at a time ⟶ question of *granularity*

- granularity of *semantic chunks* as generated by our iNLG component [6].
- no need to restrict left context as symbolic processing is very fast
  - key question: *how much lookahead?*

Exemplary Analysis

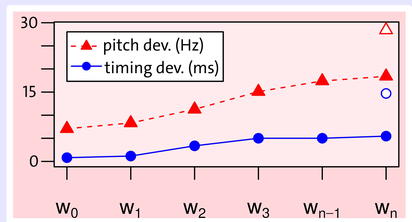
- exemplary plot of pitch curves resulting from various lookahead conditions:
  - strongest deviations at ends of phrases
  - the trivial setting wrongly inserts start-of-sentence intonations



- dis-continuities could be easily smoothed by enforcing a maximum gradient
  - this would further increase RMSE for 'bad' settings

Evaluation

- we compare the incremental prosodic assignment relative to non-incremental assignments
- we measured *phoneme duration deviation* and *pitch deviation* (RMSE)
- measured values are close to *just noticeable differences* (JND) for speech [7,8].



Results:

- next phrase must be appended no later than after the current phrase's first word
- more lookahead only marginally improves the results
- almost reaches just noticeable differences (JND) given in the literature

Open Source!

Our software for incremental dialogue processing is available as open source:

- inprotk.sf.net for the source code and documentation
- www.inpro.tk for more information on the Inpro project

We value your feedback to inprotk-devel@lists.sourceforge.net !

lookahead condition	the next phrase is integrated:	**timing** dev. RMSE (ms)	**pitch** dev. RMSE (Hz)	**pitch** dev. (Hz) 95% quantile
control condition	(non-incremental synthesis)			
w_0	with one full phrase of lookahead	0.81	7.08	10
w_1	after first word of current phrase	1.16	8.32	19
w_2	... after second word	3.37	11.27	27
w_3	... after third word	5.01	15.10	37
w_{n-1}	... one word before end of the phrase	5.01	17.40	46
w_n (w/ left context)	immediately before the next phrase	5.47	18.42	50
w_n (trivial)	phrase-by-phrase, no left context	14.70	28.42	67

References:

[1] T. Baumann and D. Schlangen: „Inpro_iSS: A component for just-in-time incremental speech synthesis," in *Procs. of ACL System Demonstrations*, Jeju, Korea, 2012.
[2] T. Baumann and D. Schlangen: „The InproTK 2012 release," in *Proceedings of SDCTD*, Montréal, Canada, 2012.
[3] M. Schröder and J. Trouvain: „The German Text-to-Speech synthesis system MARY: A tool for research, deveiopment and teaching," *Int. J. of Speech Tech*, **6**(3), 2003.
[4] D. Schlangen and G. Skantze: „A general, abstract model of incremental dialogue processing," in *Proceedings of EACL*, Athens, Greece, 2009.
[5] T. Dutoit, M. Astrinaki, O. Babacan, N. d'Alessandro, and B. Picart: „pHTS for Max/MSP: A streaming architecture for statistical parametric speech synthesis," numediart Research Program on Digital Art Technologies, Tech. Rep. 1, 2011.
[6] H. Buschmeier, T. Baumann, B. Dorsch, S. Kopp and D. Schlangen: „Combining incremental language generation and incremental speech synthesis for adaptive information presentation," in *Proceedings of SigDial*, Seoul, Korea, 2012.
[7] H. Quené: „On the just noticeable difference for tempo in speech," *Journal of Phonetics*, **35**(3), 2007.
[8] S. G. Noteboom: „The prosody of speech: Melody and rhythm," in *The Handbook of Phonetic Sciences*, W. J. Hardcastle and J. Laver, Eds., Oxford: Blackwell, 1997.