Partial Representations Improve the Prosody of Incremental Speech Synthesis

Timo Baumann • Natural Language Systems Division • Department of Informatics • Universität Hamburg • Germany • baumann@informatik.uni-hamburg.de

Abstract

Humans speak incrementally [1], based on partial utterance specifications that they modify and extend on-the-fly. Conventional synthesizers, in contrast, process top-down and consecutively each layer of abstraction, limiting their applicability in *highly interactive tasks*.

The architecture of our incremental speech synthesizer enables the system to start utterance delivery immediately, based on partial utterance specifications and to flexibly include all information available so far.

In this paper, we investigate incremental processing for HMM state selection and find that considering phrase/utterance *finality-related features just for final words greatly improves the timeliness/quality trade-off* over previous work [2,3].

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

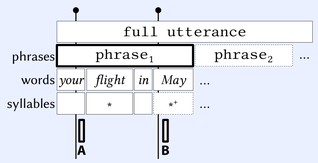
- **conventional** speech synthesis systems are optimized for non-interactive reading tasks
  - full utterances are required as input
  - no changes / extensions / adaptation to ongoing utterance is allowed
  - *ill-suited for highly-dynamic environments*
  - *relatively long utterance-initial delay*
- our **incremental** speech synthesis:
  - starts delivery before the whole utterance has been processed
  - allows to *extend or change delivery* while it is ongoing
  - gives very low latency, and *only little loss in synthesis quality*
- application areas: e.g. dialogue, simultaneous interpreting, sports commentary

Incremental Speech Synthesis **Architecture**

- just-in-time processing:
  - most computations are performed during speech delivery
  - more lookahead on more abstract processing levels

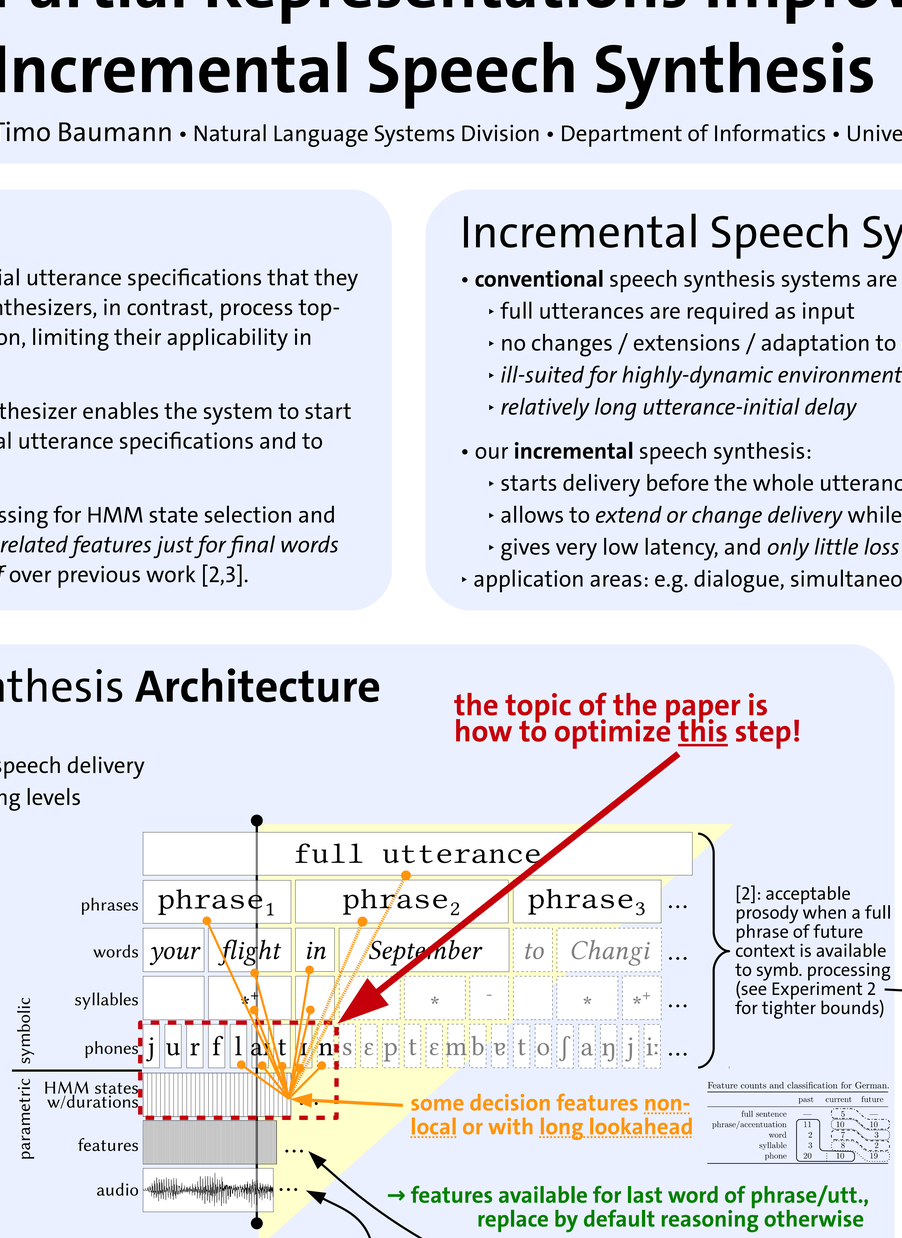
→ triangular processing scheme

- speech synthesis subdivided into symbolic and parametric processing
  - parameter derivation from symbols with decision trees
  - many features are non-local e.g. relating to the full utterance
- we investigate the influence of default reasoning for missing features, *considering the availability of features depending on position in phrase/utterance:*



phrase features likely
A: not available, **B:** available

the topic of the paper is how to optimize this step!



[2]: acceptable prosody when a full phrase of future context is available to symb. processing (see Experiment 2 for tighter bounds)

Feature counts and classification for German.

	past	current	future
full sentence	—	∞	∞
phrase/accentuation	11	10	10
word	2	2	2
syllable	3	8	7
phone	20	10	10

Exp. 1: incremental state selection in isolation

- determine feature usage and infer defaults (for details, see [3])
- use defaults for phrase/utterance-level features unless the word ends the phrase/utterance
- measure pitch and duration RMSE (against non-incr.)

→ phrase/utt-features just on phrase/utt-final words almost as good as requiring them all the time

100 Cent ≙ 1 semitone

setting	root mean f_0 in Cent	squared error duration (ms)
current word (as in [3])	219	26.4
+phrase final	170	25.5
+utterance final	32.5	2.0
current phrase (as in [3])	167	25.4

Exp. 2: combine with limited symbolic processing

- previous experiment uses non-incrementally generated symbolic representation
- here: use method from [2] for incremental symbolic intonation assignments instead

→ cuts down previous lookahead requirement (+2 phrase) down to +1 phrase at little additional cost!

→ for details, see the paper!

Conclusion:

- reduced lookahead requirement down to *word+finality* information for state selection
  → this is *almost word-by-word* synthesis but sounds similar to full-sentence context!
- combined with incremental symbolic processing, this greatly reduces lookahead requirement (however, word-stress and finality must still be determined separately)
- need to validate results in formal listening experiments

Free and Open Source Software!

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

- **inprotk.sf.net** for source code, demos and documentation

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

Implementation details

The presented methods are available in Inpro_iSS [6] which is part of the incremental processing toolkit InproTK [7].

Inpro_iSS is based on MaryTTS [8] and re-uses much of its synthesis-related capabilities, adding to it incremental processing based on the IU framework [9].

References

[1] W. J. Levelt, *Speaking: From intention to Articulation*, MIT Press, 1989.
[2] T. Baumann and D. Schlangen: „Evaluating prosodic processing for incremental speech synthesis", in *Proceedings of Interspeech*, Portland, USA, Sep. 2012.
[3] T. Baumann: „Decision tree usage for incremental parametric speech synthesis", in *Proceedings of ICASSP*, Florence, Italy, May 2014.
[4] H. Kawahara: „Speech representation and transformation using adaptive interpolation of weighted spectrum: Vocoder revisited", in *Proceedings of ICASSP*, 1997.
[5] T. Dutoit, M. Astrinaki, O. Babacan, N. d'Allessandro, 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] 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.
[7] T. Baumann and D. Schlangen: „The InproTK 2012 release," in *Proceedings of SDCTD*, Montréal, Canada, 2012.
[8] 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.
[9] D. Schlangen and G. Skantze: „A general, abstract model of incremental dialogue processing," in *Proceedings of EACL*, Athens, Greece, 2009.

thanks to: **Daimler** und **Benz Stiftung**

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