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 utterances 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

Implementation details
The proposed methods are available in Inpro _SS [6] which is part of the incremental processing toolkit InproTK [7].

Inpro _SS 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

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

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

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!

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
Further references:


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


→ 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.


→ 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.


→ 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