Incremental Speech Recognition

Almost all speech recognizers use a variant of the token-pass algorithm (Young, Russell, and Thornton 1989) and run *time-synchronously*. Thus, the recognizer’s input side is incremental as-is; at runtime, the algorithm keeps a list of partial results that may later be extended to become the overall best hypothesis, or discarded otherwise. Many approaches have dealt with predicting the one hypothesis during Viterbi search that later turns out to be optimal:

  - Search the token history for a prefix that is common to all active tokens, i.e. the undisputed (later called ‘immortal’ by Selfridge et al. (2011)) prefix; the prefix may lag behind significantly, but this strategy does not introduce errors as compared to non-incremental processing.

  - Consider the best-ranked hypothesis at time $t$ only up to $t - \Delta$ (effectively leaving a fixed right context of $\Delta$, something that is similarly used ASR-internally to restrict the search at word boundaries when using phonetic lookahead (Ortmanns and Ney 2000).

  - Use the same strategy as Wachsmuth, Fink, and Sagerer (1998) for phoneme recognition and finds a right context $\Delta$ of 150 ms as being almost as good as
full recognition. (However, such a short lag does not work well enough for word recognition (Baumann, Atterer, and Schlangen 2009; Wachsmuth, Fink, and Sagerer 1998).)


- Toru Imai et al. (2003). “Progressive early decision of speech recognition results by comparing most likely word sequences”. In: Systems and Computers in Japan 34.14, pp. 73–82. ISSN: 1520-684X. DOI: 10.1002/scj.10193:
  - Combine incrementality with two-pass decoding and use a simple smoothing strategy. In their progressive two-pass setup, a bigram-based first-pass decoder’s partial results are unfolded into a lattice which is rescored in a second pass; whenever two consecutive hypotheses of the second pass are identical, all but the \( M \) most recent words are passed on as output.

  - Analyse performance aspects of speech recognition under the assumption that a later revision of hypotheses is always possible: timing of the first occurrence of a word, and when it is finally decided upon, as well as hypothesis evolution in terms of edit overhead. In addition, two methods are introduced that resemble right context as developed by Wachsmuth, Fink, and Sagerer (1998) and generalize the smoothing method by Imai et al. (2000) in order to trade timeliness of hypotheses against hypothesis stability.

  - Use undisputed parts of the lattice (called ‘immortal nodes’) and sentence-finality (as determined by the LM) for SLMs and grammar-based recognition, and, generalizing Baumann, Atterer, and Schlangen 2009, introduces a gradual notion of stability for partial hypotheses which allows to output hypotheses quickly, but allows to notify consumers about the confidence of the system in this hypothesis.

  - Generate stability estimates for individual words of the hypothesis (generalizing Selfridge et al. 2011). Furthermore, they show that machine learning based on “posteriors, age, words, and a search statistic” for the stability estimate results in a better stability/timing trade-off than using hypothesis age alone.


Investigate the merit of using N-best lists in incremental speech recognition. They find that N-best lists can be considerably larger during recognition but that most of the possible gains (in terms of timing) can already be achieved for low values of N.

**Incremental Speech Synthesis**

  - Outlines requirements and desiderata towards incremental speech synthesis. The diphone synthesis that is employed in that system can, however, not be considered state-of-the-art anymore.
  - Use HMM parameter optimization (which conventionally uses global optimization) within overlapping local contexts and show that a lookahead of only a few phones is sufficient to result in almost indistinguishable synthesis quality.
  - Devise a technique for local variance optimization, which could replace (and surpass) global variance optimization (Toda and Tokuda 2007) in incremental settings.
- Timo Baumann and David Schlangen (2012b). “INPRO_iSS: A Component for Just-In-Time incremental Speech Synthesis”. In: Procs. of ACL System Demonstrations. Jeju, Korea:
  - Present a speech synthesis component embedded into the incremental architecture InproTK that allows to modify delivery parameters on the fly, to pre-specify utterance-plans and to select among those, and to extend partial utterances with further material at runtime.
  - Evaluate the quality trade-off involved when realizing sentence-level prosody within only limited contexts and find that about one phrase of lookahead is sufficient to reach a prosodic realization that closely resembles non-incremental prosody assignments.
- Timo Baumann and David Schlangen (2013). “Interactional Adequacy as a Factor
in the Perception of Synthesized Speech”. In: *Proceedings of Speech Synthesis Workshop (SSW8)*, to appear:

- Present a multi-modal evaluation experiment that shows that users not only rate formulations that are enabled by incremental speech synthesis higher, but also judge this system’s synthesis quality as higher, despite the fact that it can only be worse than non-incremental synthesis. This result highlights that pure listening evaluations may be inadequate and that evaluation in context may lead to more relevant judgements.

**Estimation of Dialogue Flow**

- Nigel Ward, Olac Fuentes, and Alejandro Vega (Sept. 2010). “Dialog Prediction for a General Model of Turn-Taking”. In: *Proceedings of Interspeech*. Tokyo, Japan:
  - Outline a method of predicting the likelihood of the current speaker ending their turn within upcoming, exponentially growing time-frames. The implementation, which is based on some prosodic features alone, results in only modest performance.
- 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:
  - Given that incremental speech recognition recognizes words while they are still ongoing, the authors aim to predict the end of the currently ongoing word, as well as the tempo of the next word, in order for a system to speak in synchrony with the user. Performance is in the range of human performance in similar tasks.

**Incremental Processing Architectures**

- David Schlangen et al. (Sept. 2010). “Middleware for Incremental Processing in Conversational Agents”. In: *Proceedings of SigDial 2010*. Tokyo, Japan:
  - Present three implementations of the IU model (Schlangen and Skantze 2009): IPAACA, Jindigo, and InproTK and discuss their relative strengths.
  - Presents more details on Jindigo.
  - Present more details on InproTK.
References


Imai, Toru, Hideki Tanaka, Akio Ando, and Haruo Isono (2003). “Progressive early decision of speech recognition results by comparing most likely word sequences”. In:


