Anatomy of an extremely fast LVCSR decoder

George Saon, Daniel Povey and Geoffrey Zweig

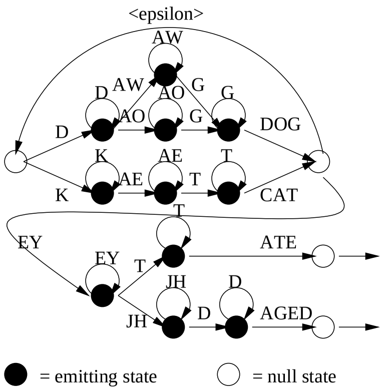
IBM T.J. Watson Research Center
phone (914)-945-2985, email saon@watson.ibm.com

Abstract

We report in detail the decoding strategy that we used for the past two Darpa Rich Transcription evaluations (RT'03 and RT'04) which is based on finite state automata (FSA). We discuss the format of the static decoding graphs, the particulars of our Viterbi implementation, the lattice generation and the likelihood evaluation. Experimental results are given on the EARS database (English conversational telephone speech) with emphasis on our faster than real-time system.

Static decoding graphs

- They are *acceptors* (instead of transducers)
- Arcs in graph have three different types of labels:
  - *leaf* labels (context-dependent output distributions),
  - *word* labels and
  - *epsilon* labels (e.g. due to LM back-off states).
- Two different types of states:
  - *emitting* states for which all incoming arcs are labeled by the *same* leaf and
  - *null* states which have incoming arcs labeled by words or epsilon.

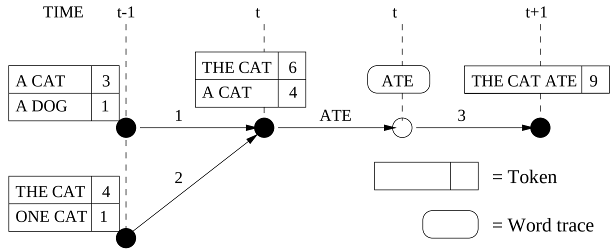


Viterbi search speed-ups

- *Graph memory layout*: graph stored as a linear array of arcs sorted by origin state
- *Successor look-up table*: maps static to dynamic state indices
- *Running beam pruning*: pruning based on current maximum score estimate

Lattice generation

Keep track of the N-best *distinct* word sequences arriving at every state

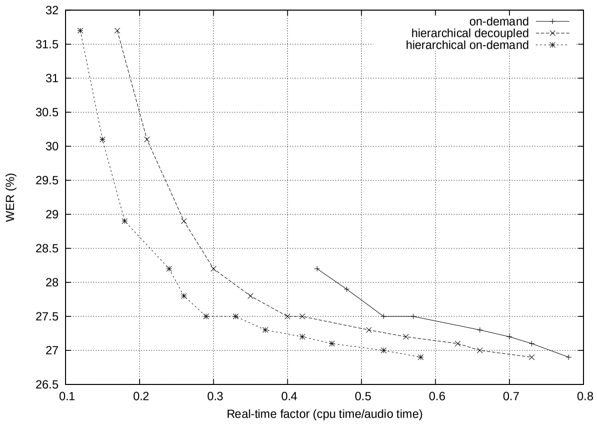


N-best degree	2	5	10
Lattice link density	29.4	451.0	1709.7

	RT03	DEV04	RT04
Speaker-adapted decoding	17.4%	14.5%	16.4%
LM rescoring + consensus	16.1%	13.0%	15.2%

Likelihood computation

- Hierarchical decoupled
- On-demand
- Hierarchical on-demand



Experimental setup (1xRT system)

EARS 2004 evaluation submission in the one times real-time (or 1xRT) category. Two-pass decoding scheme with three adaptation passes inbetween (VTLN, FMLLR, MLLR).

- Decoding graph statistics:

	SI	SA
Phonetic context	±2	±3
Number of leaves	7.9K	21.5K
Number of words	32.9K	32.9K
Number of n-grams	3.9M	4.2M
Number of states	18.5M	26.7M
Number of arcs	44.5M	68.7M

- Search statistics:

	SI	SA
Word error rate	28.7%	19.0%
Search errors	2.2%	0.3%
Run-time factor	0.14xRT	0.55xRT
Likelihood/search ratio	60/40	55/45
Avg. Gaussians/frame	7.5K	43.5K
Max. states/frame	5.0K	15.0K