Design of the CMU Sphinx-4 Decoder

Paul Lamere[1], Philip Kwok[1], William Walker[1],
Evandro Gouvêa[2], Rita Singh[2], Bhiksha Raj[3], Peter Wolf[3]

[1]Sun Microsystems Laboratories, [2]Carnegie Mellon University, [3]Mitsubishi Electric Research Laboratories, USA

Summary: Sphinx-4 is an open source HMM-based speech recognition system written in the Java™ programming language. The design of the Sphinx-4 decoder incorporates several new features in response to current demands on HMM-based large vocabulary systems. New design aspects include graph construction for multilevel parallel decoding with multiple feature streams without the use of compound HMMs, the incorporation of a generalized search algorithm that subsumes Viterbi decoding as a special case, token stack decoding for efficient maintenance of multiple paths during search, design of a generalized language HMM graph from grammars and language models of multiple standard formats, that can toggle between a flat search structure and tree search structures.

Description

- Speaker independent
- Large vocabulary
- Continuous speech
- Support true 3-gram, FST and BNF grammars
- Written entirely in the Java™ programming language
- Open source – BSD style license
- Based upon Sphinx-3 and Sphinx-3.3 developed at Carnegie Mellon University (CMU)
- Collaboration between:
  - Sun Microsystems Laboratories
  - Carnegie Mellon University
  - Mitsubishi Electric Research Laboratories (MERL)

Goals

- Highly flexible recognizer
- Speed and accuracy equal to or exceeding Sphinx-3
- Give the world an open source state-of-the-art system for
  - Application development
  - Speech research

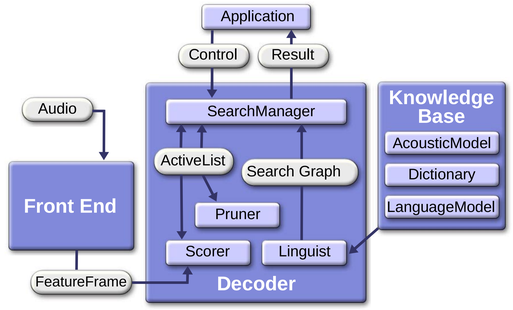
Learn More

- Visit the Sphinx-4 project website at:
  cmusphinx.sourceforge.net
- Or contact the team at:
  cmusphinx-contacts@lists.sourceforge.net

The Java™ Programming Language

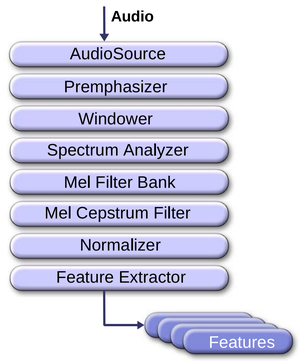
- Object-Oriented
  - Decoupled modules
  - Pluggable modules
- High performance
- Garbage collection
- Multi-threaded language
- Rich set of standard libraries
- Write Once, Run Anywhere™
- Large developer base

Sphinx-4 Architecture

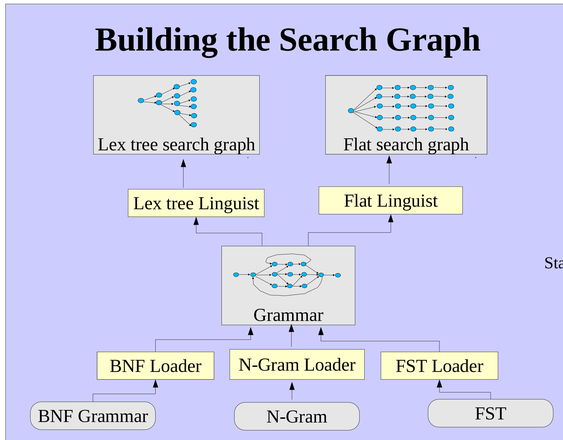


The Front End

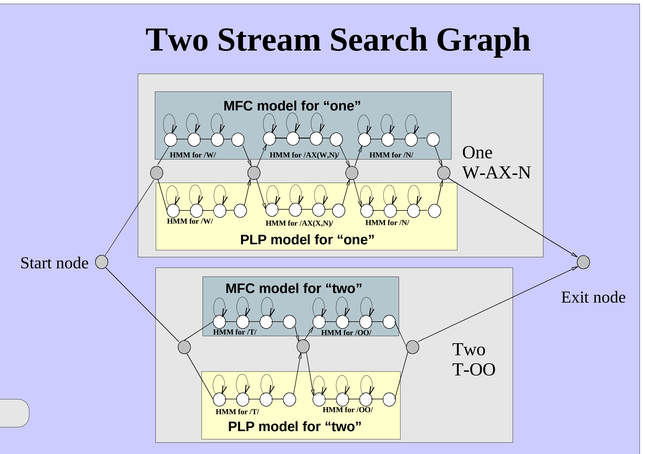
- Transforms speech into MFCC features
- Set of signal processing filters
- Configurable



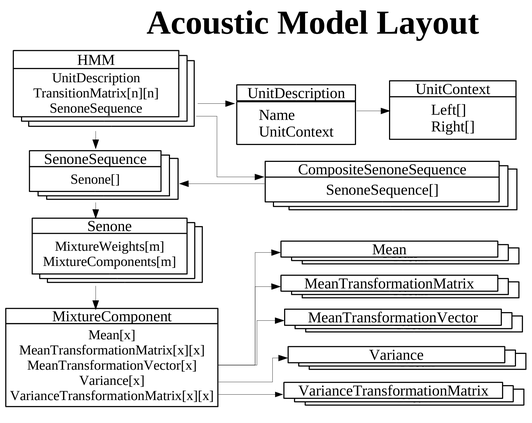
Building the Search Graph



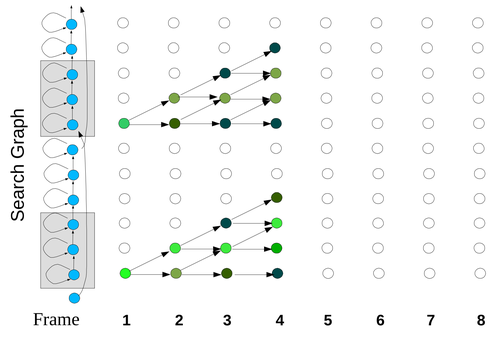
Two Stream Search Graph



Acoustic Model Layout



The Search Trellis



Tests

- **TI46 –** 11 word dictionary, isolated speech recognition, uses TI-46 data from the LDC
- **TIDIGITS** - 11 word dictionary, continuous speech recognition, uses TIDIGITS data from the LDC
- **AN4** - 105 highly confusable words and letters, continuous speech recognition, 3-gram language models, from CMU
- **RM1** - 1,000 word vocabulary, N-gram language models, ARPA's Resource Management database
- **HUB4 –** 64,000 word vocabulary, 3-gram language models, pre-segmented **F0** (baseline/broadcast speech), from LDC 1999 HUB-4 Broadcast News evaluation test material

Performance Comparison

Test	Word Error Rate (%)		Speed (X RT)	
	Sphinx-3	Sphinx-4	Sphinx-3	Sphinx-4
TI46	0.67	0.17	0.41	0.28
TIDIGITS	1.10	0.32	0.79	0.22
AN4	6.23	6.21	3.70	2.28
RM1 1-gram	15.67	14.77	1.88	> 10[†]
RM1 2-gram	2.10	2.10	2.00	> 10[†]
RM1 3-gram	0.90[‡]	0.99[◦]	1.74	> 10[†]

[†]Note that RM1 tests have not been optimized for speed
[‡]Note that Sphinx-3 implements a pseudo-3-gram, not a pure 3-gram
[◦] Not optimized for accuracy

Discussion

- Performs well in terms of speed (RT) and accuracy (WER) for smaller tasks
- Performs well in terms of accuracy (WER) for medium-sized tasks
- Not yet optimized for speed for medium-sized or larger tasks
- Evaluation of the medium vocabulary tasks (RM1) and the large vocabulary tasks (HUB4) is ongoing
- Performance data collected on a Dual CPU, 750 MHz UltraSPARC® III processors with 1024 MB memory

Current and Future Work

- Optimize medium vocabulary tasks - improve speed to meet or exceed Sphinx-3
- Complete and optimize large vocabulary tasks
- Create Sphinx-4 trainer to generate Sphinx-4 acoustic models
- Support word lattices
- Support confidence scoring
- Support Java Speech API (JSAPI)