The LIUM speech transcription system: a CMU Sphinx III-based system for french broadcast news

Paul Deléglise, Yannick Estève, Sylvain Meignier, Teva Merlin
LIUM/CNRS — Université du Maine, Le Mans, France

ABSTRACT

The system used by the LIUM (Science Computer Lab of the University of Maine, France) to participate in ESTER, the french evaluation campaign of broadcast news transcription, is based on the CMU Sphinx 3.3 (fast) decoder. Some tools have been added to different steps of the Sphinx recognition process: segmentation, acoustic model adaptation, word-lattice rescoring.
Several experiments have been conducted on studying the effects of the signal segmentation on the recognition process, on injecting automatically transcribed data into training corpora, or on testing different approaches for acoustic model adaptation.
With very few modifications and a simple MAP acoustic model estimation, Sphinx 3.3 decoder reached a word error rate of 28.2%. The entire system developed by LIUM obtained 23.6% as official word error rate for the ESTER evaluation, and 23.4% as result of post-evaluation experiments.

SYSTEM OVERVIEW

CMU Sphinx 3.3

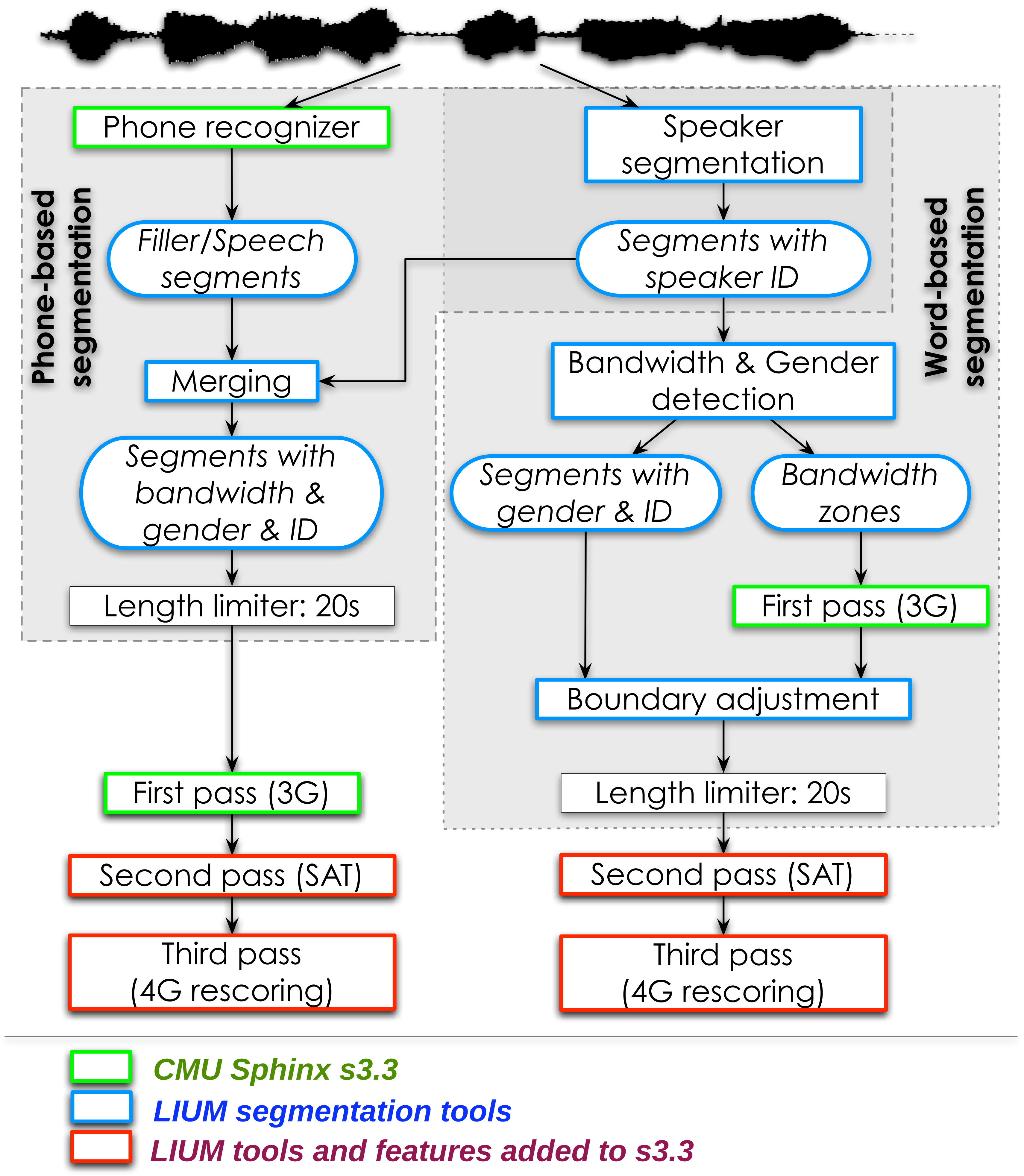
- ✓ A branch from CMU Sphinx III project which is distributed under an *Apache*-like license
- ✓ Uses fully continuous acoustic models with 3 or 5-state left-to-right HMM topologies (phones on context)
- ✓ Bigram or trigram language models can be used
- ✓ Vocabulary size limited to about 65K words

Added features

- ✓ **Speaker Adaptive Training** (SAT) procedure based on CMLLR [Gales1997]: *CMLLR can be computed either on a sentence-by-sentence basis or on a speaker-by-speaker basis*
- ✓ **4-gram lattices rescoring**: *Improvement of a tool (s3_dag) using trigram language models distributed with the last release of the fast decoder, CMU Sphinx 3.5*

Segmentation

- ✓ **2 steps**: speaker segmentation + adjustment of the speaker segment boundaries to speech-based segment boundaries
- ✓ **Speaker segmentation**:
  - ◆ Initial over-segmentation determined according to sliding GLR
  - ◆ BIC-based, bottom-up hierarchical clustering of segments
  - ◆ Viterbi-based adjustment of segment boundaries
- ✓ **Re-adjustment of segment boundaries**: *2 strategies investigated*
  - ◆ "phone-based": speaker segments adjusted to fit filler/speech segments detected by a phone recognizer
  - ◆ "word-based": speaker segments adjusted to fit sentences decoded by first pass transcription



EXPERIMENTAL RESULTS

ESTER: french broadcast news evaluation campaign

Training corpus:

- ✓ *90 hours of audio from 4 radio stations: RFI, France Info, France Inter, RTM, with manual transcription (period: 1998-2003)*
- ✓ *Articles from french newspaper "Le Monde" from 1987 to 2003*

Test corpus:

- ✓ *10 hours of audio from 5 radio stations: the ones from training corpus + Radio Classique (2h from each station, period: 2004)*

Additional data:

- ✓ *1400h of untranscribed audio from various radio stations*

Linguistic resources

- ✓ Lexicon: *about 65K most frequent words with their pronunciation*
- ✓ Out-of-vocabulary word rate: *1.18%*
- ✓ Language models (3g and 4g): *linear interpolations of 3 LMs estimated from manual transcriptions of audio files and newspaper*
- ✓ Discounting method: *Kneser-Ney modified*
- ✓ Number of n-grams:

	1-grams	2-grams	3-grams	4-grams
Trigram model	65.5K	18.4M	25.4M	–
Quadrigram model	65.5K	18.4M	22.2M	19.7M

Acoustic models

- ✓ 35 phones + 4 filler phones (silence, music, breath, long 'e')
- ✓ 5500 tied states, 22 Gaussians per state
- ✓ Trained from 72 hours of broadband and 8 hours of narrrowband
- ✓ Broadband and narrowband models adapted using a MAP procedure to specialize models on gender
- ✓ Gender- and bandwith-dependent models are used to compute CMLLR transformation for each sentence (or each speaker)

Results (word error rate)

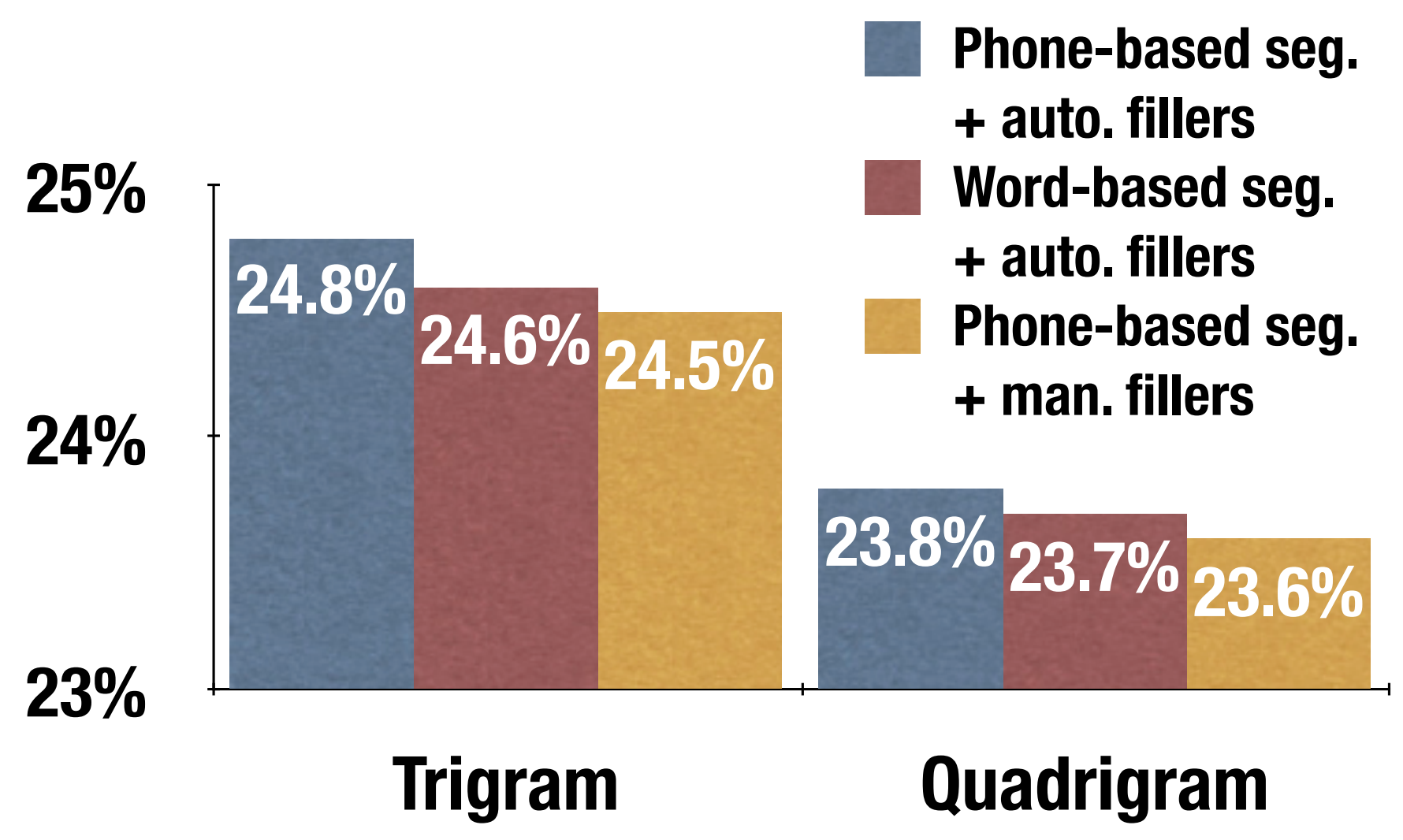
***Sphinx 3.3 baseline:* 28.2%** *(with MAP acoustic model estimation)*
***LIUM system:* 23.4%**

SAT: sentence-by-sentence *vs* speaker-by-speaker

- ✓ Sentence-by-sentence better by 0.4 point

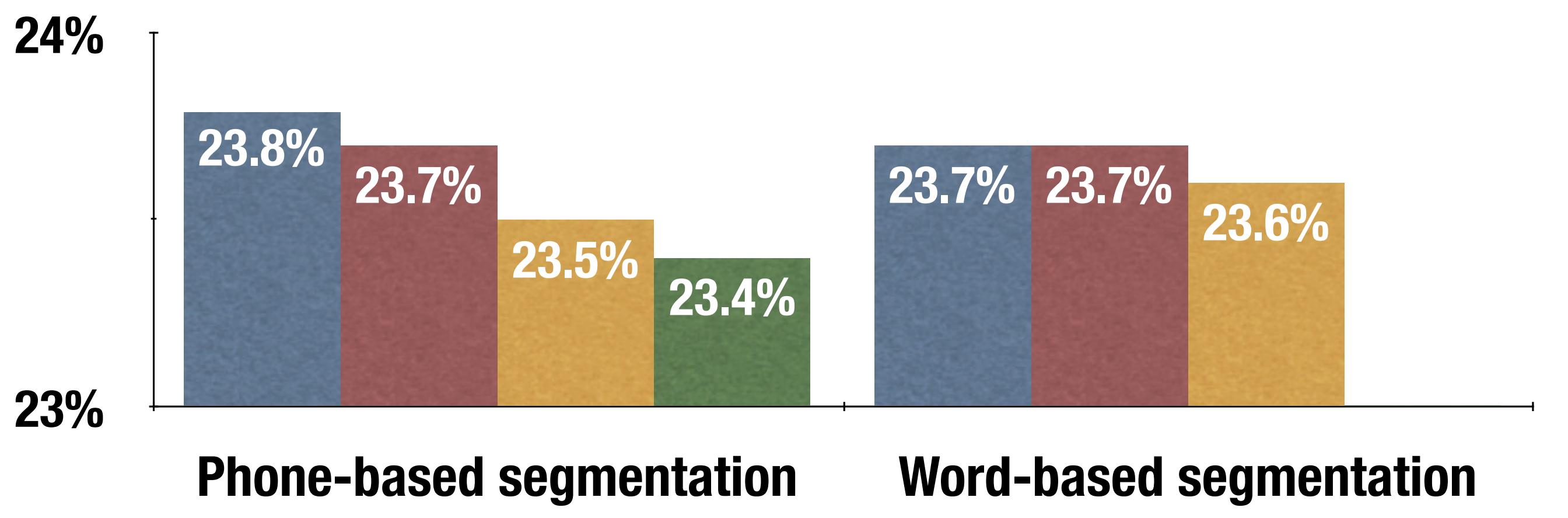
Segmentation and alignment

- ✓ Transcription alignment strategies to train acoustic models: manual or automatic detection of filler words
- ✓ With or without word-lattices rescoring with quadrigram LM



Addition of automatically transcribed data for acoustic training

- ✓ 3 data sets of 25 hours each:
  - ◆ *S1: France Culture, December 2003*
  - ◆ *S2: mixed radio stations, 2003-2004*
  - ◆ *S3: France Culture, September 2004*
- ✓ No data filtering



Added features, segmentation LIUM tools, and some resources available on: http://www-lium.univ-lemans.fr/speechtools/