MDE Tasks and Results

EARS RT-04 Fall Meeting
Nov 7, 2004

Jonathan Fiscus, Audrey Le, Greg Sanders

Outline

- NIST's tasks
- Motivation for MetaData Extraction (MDE)
- Overview of the MDE tasks and how scored
- Results of this MDE evaluation

Major NIST Tasks

- Establish the evaluation plan
- Maintain a web site for each evaluation
- Work with LDC to produce test datasets
- Provide tools for data production/validation
- Provide metadata scoring tools
  - md-eval.pl
- Score the results and report them
  - Identify progress

EARS Objective

Powerful speech-to-text technology

Input: Human-human speech (broadcasts, conversations)

Output: Rich transcript (words + metadata) accurate enough for

Machines to detect, extract, summarize, translate

Humans to read & understand easily

- Speakers labeled
- Capitalization and punctuation
- Disfluencies removed

Rich Transcription
(Broadcast News Example)

Traditional ASR Output

```
tonight this thursday big
pressure on the clinton
administration to do something
more effective about the latest
killing in yugoslavia airline
passengers and outrageous
behavior at thirty thousand
feet what can an airline do and
now that el nino is virtually
gone there is la nina to worry
about from a. b. c. news world
headquarters in new york this
is world news tonight with
peter jennings good evening
```

Enriched ASR Output

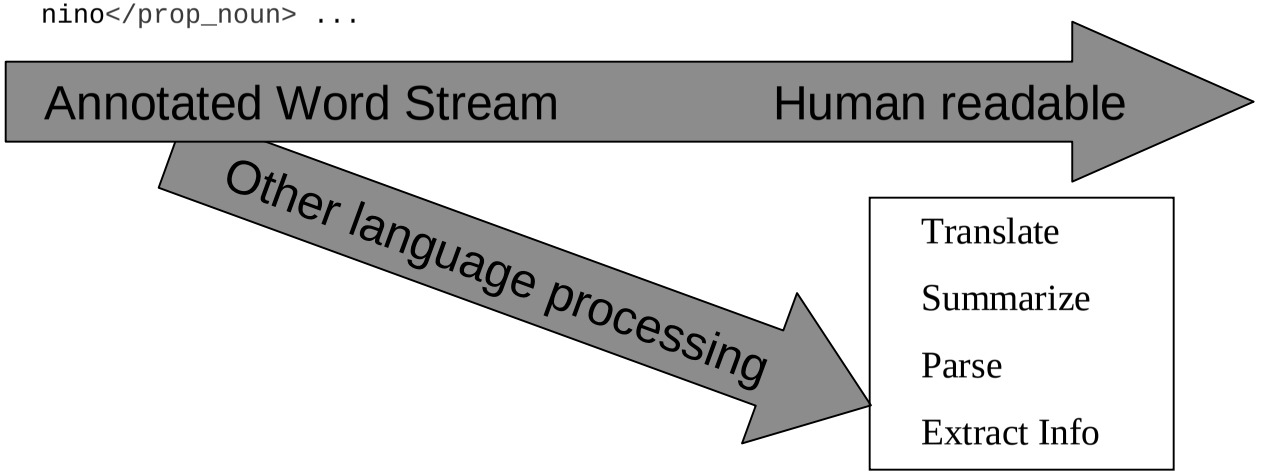
```
<speaker name="Peter Jennings">
<SU type=stmt> tonight this
<prop_noun> thursday
</prop_noun><phrase-bound> big
pressure on the
<prop_noun>clinton </prop_noun>
administration to do something
about the latest killing in
<prop_noun>yugoslavia</prop_noun>
</SU> <SU type=stmt>airline
passengers and outrageous
behavior at <numex
val=30,000>thirty
thousand</numex>feet</SU> <SU
type=question>what can an airline
do</SU> <SU type=stmt>and now
that <prop_noun>el
nino</prop_noun> ...
```

Enable Transcript That Is More Readable

Peter Jennings: Tonight this Thursday, big pressure on the Clinton administration to do something about the latest killing in about the latest killing in Yugoslavia. Airline passengers and outrageous behavior at 30,000 feet. What can an airline do? And now that El Nino is virtually gone, there is La Nina to worry about.

Announcer: From ABC News World Headquarters in New York, this is World News Tonight with Peter Jennings.

Peter Jennings: Good evening.



Metadata Extraction (MDE) tasks

- Structural Metadata Extraction tasks
  - SU Boundary Detection (SUBD)
  - Edit Word Detection (EWD)
  - Filler Word Detection (FWD)
  - Interruption-Point Detection (IPD)
- Diarization Metadata Extraction task
  - "Who spoke when"

Participants

- CU -- Cambridge University
- ICSI -- International Computer Science Institute
- SRI -- SRI International
- UW -- University of Washington (Seattle)
- Brown -- Brown University
- MITLL -- MIT Lincoln Laboratory
- IBM -- International Business Machines
- sriplus -- collaborative effort (SRI, ICSI, UW)

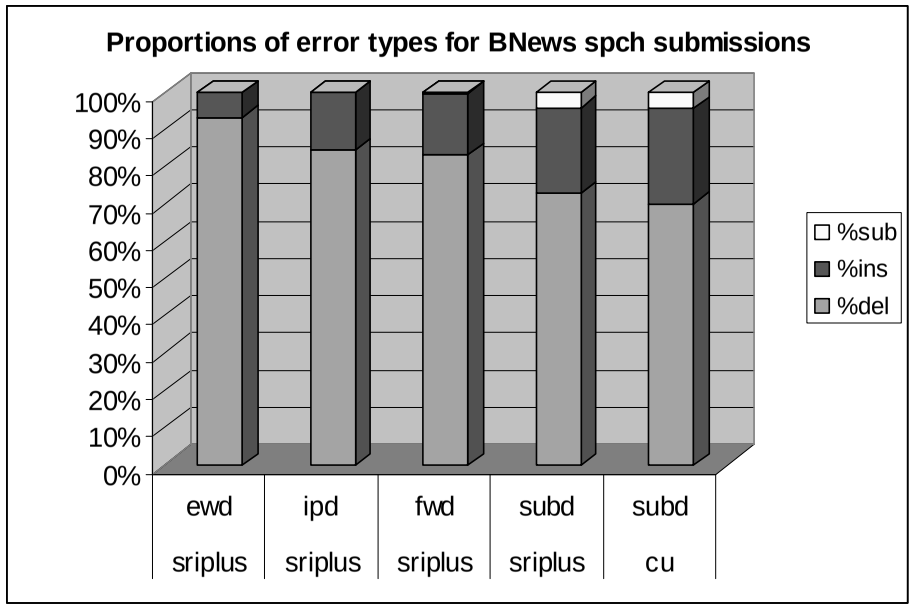
Metric for MDE Detection Tasks

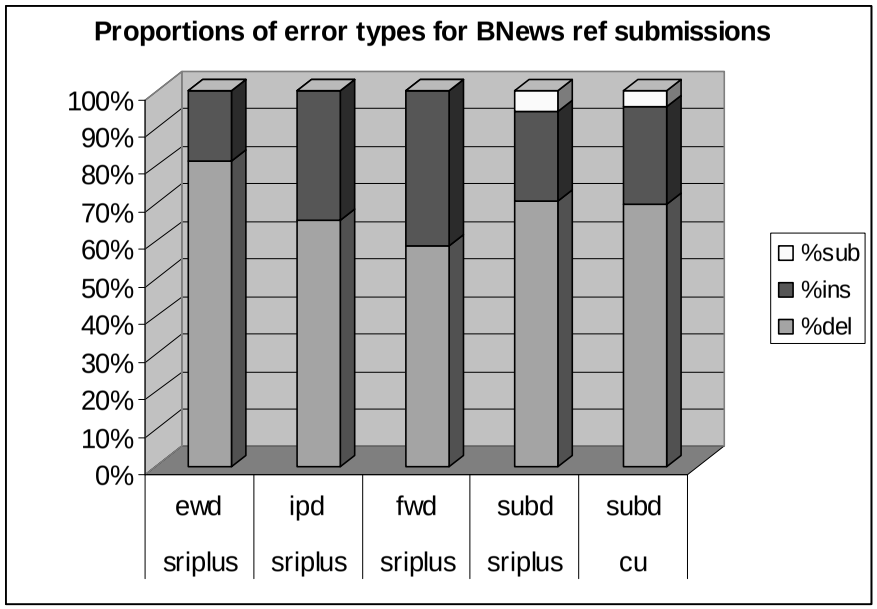
- Somewhat parallel to the scoring of STT
- Metric amounts to

  (Number of Detection Errors) / (Actual number of whatever)

- Detection errors are of two types
  - Miss (e.g., did not find an SU boundary that actually exists)
    - Also known as a “deletion” error
  - False-alarm (e.g., found an SU boundary where none exists)
    - Also known as an “insertion” error
- Details of the formulas are in the Eval Plan
  - Which is in your notebook.

“Miss” Errors Somewhat Dominant in Structural Metadata Extraction





(In diarization, in contrast, substitution errors dominate overwhelmingly)

SU Boundary Detection (SUBD)

- An SU is a unit of speech that expresses a separate thought or idea
  - Does not necessarily correspond to a complete sentence in written language
- Task was to find boundaries between the SUs, in effect to find the last word in each, and the type of each SU.
- This information can be used to add capitalization and punctuation.

Four subtypes of SUs were evaluated

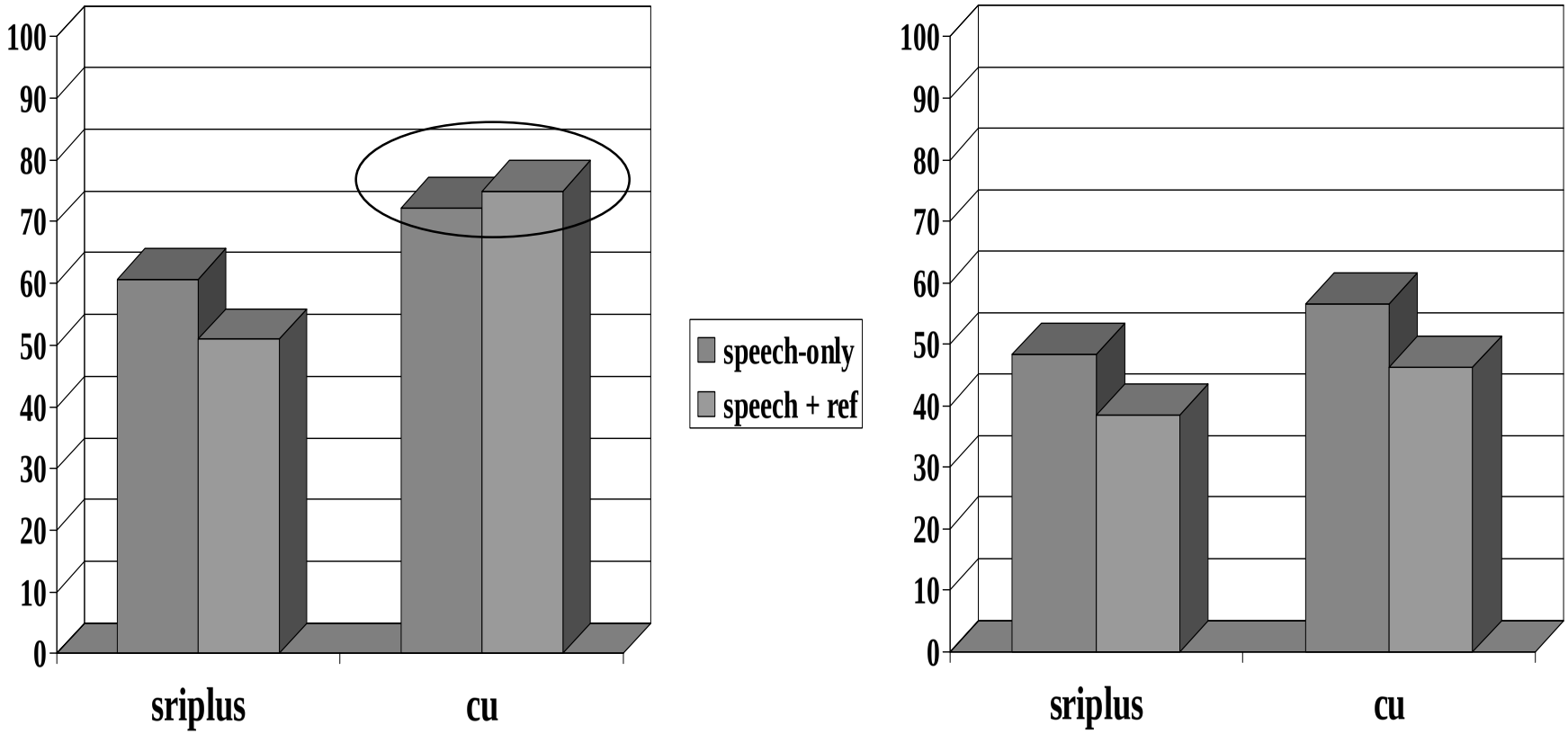
- Statement
  - `We ate dinner.`
- Backchannel (an acknowledgement to the other speaker)
  - `Uh-huh.`
- Question
  - `Good dinner?`
- Incomplete
  - `We went to the` **(*other speaker interrupts*)**

Task included identifying the subtype -- wrong on

about 3 percent of SUs for BNews

about 11 percent for Telephone Conversations

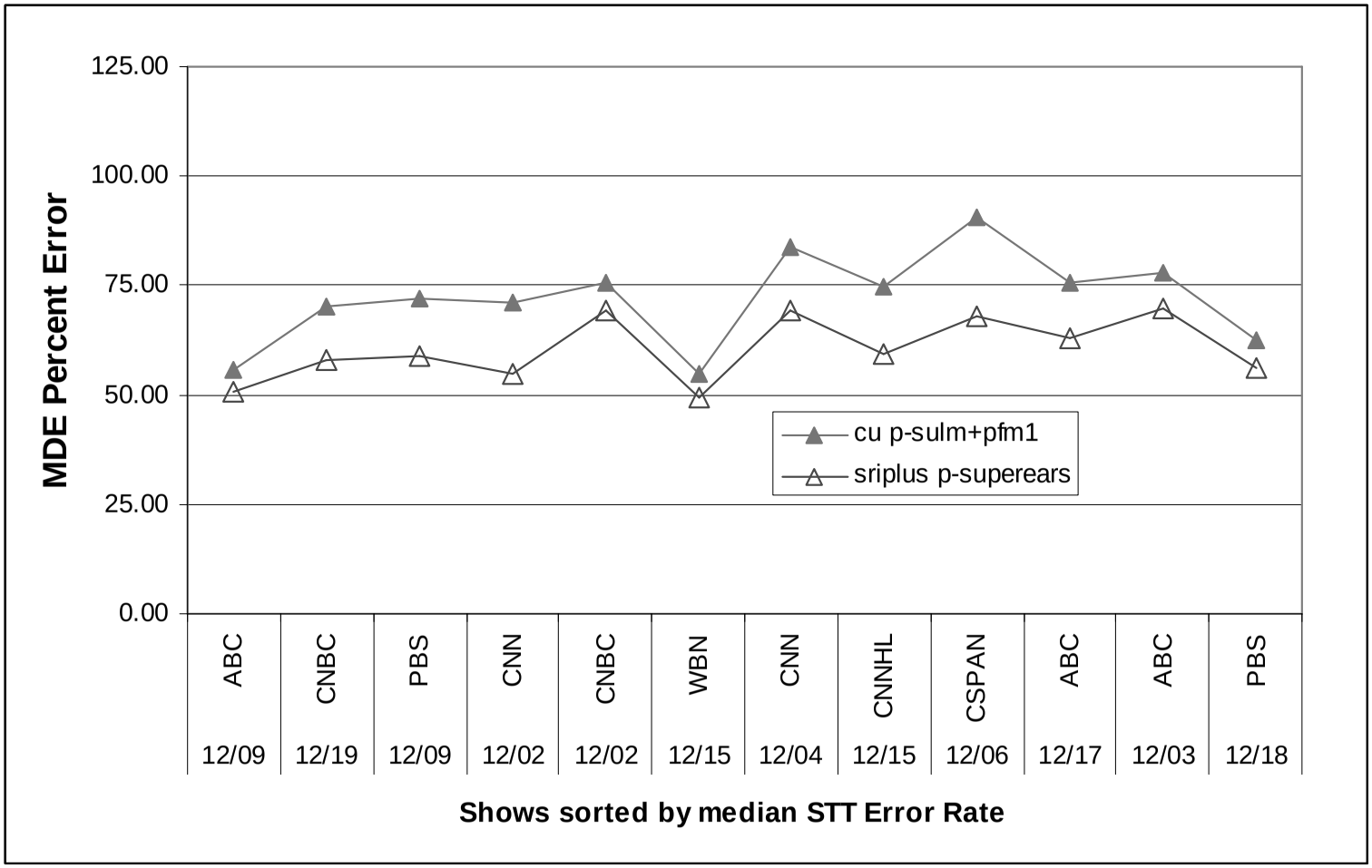
SU Boundary Detection



Broadcast News

Conversational Telephone Speech

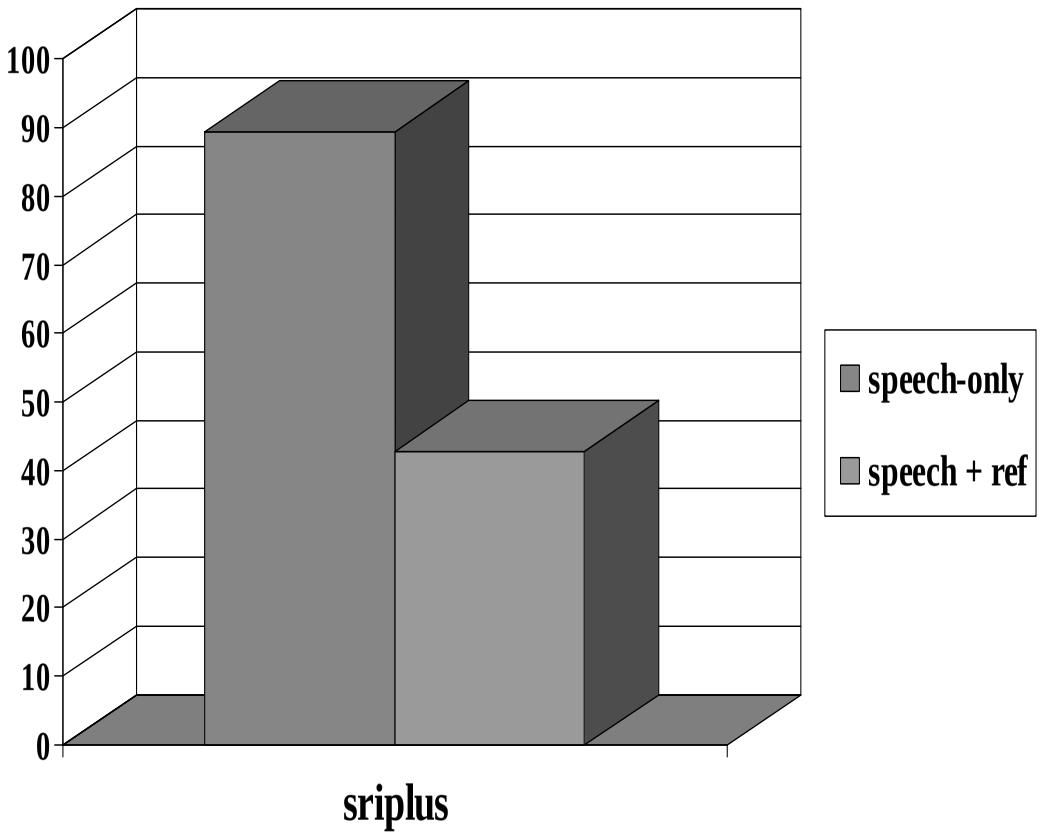
SU Boundary Detection Performance Not Very Correlated with STT Performance



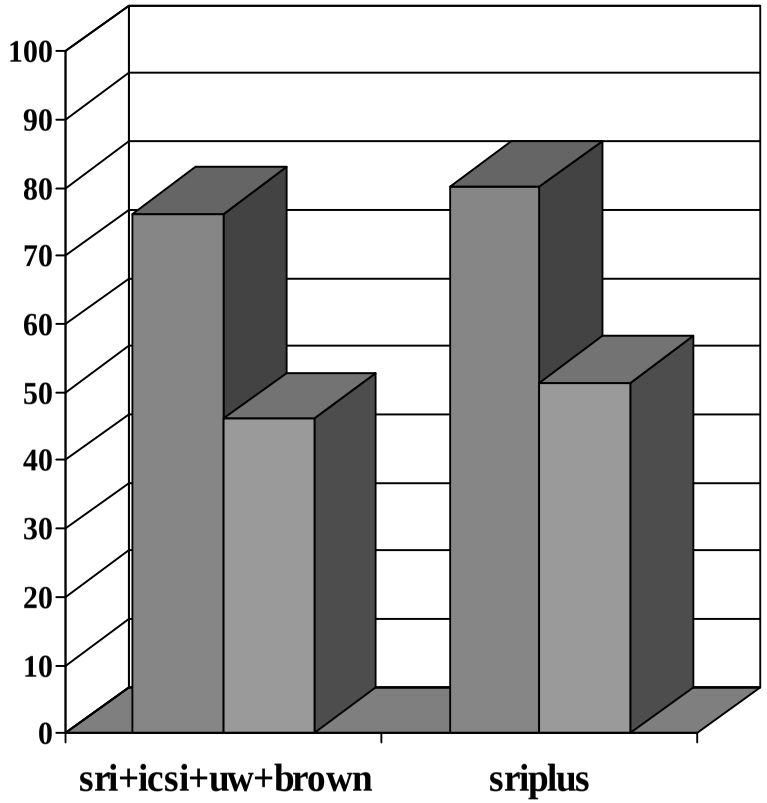
Edit-Word Detection (EWD)

- Identify the words that could be deleted in a "cleaned-up" transcript -- words revised/repeated/abandoned
- Three kinds of edits -- edit-words shown in brackets.
  - Revision -- *original version and revision are related*
    - `He is the [forty-first] I mean forty-third president`
  - Repetition
    - `We [went to] went to New York`
  - Restart -- *original version is abandoned*
    - `You know [they always end up] sometimes they change their mind`

Edit Word Detection



Broadcast News

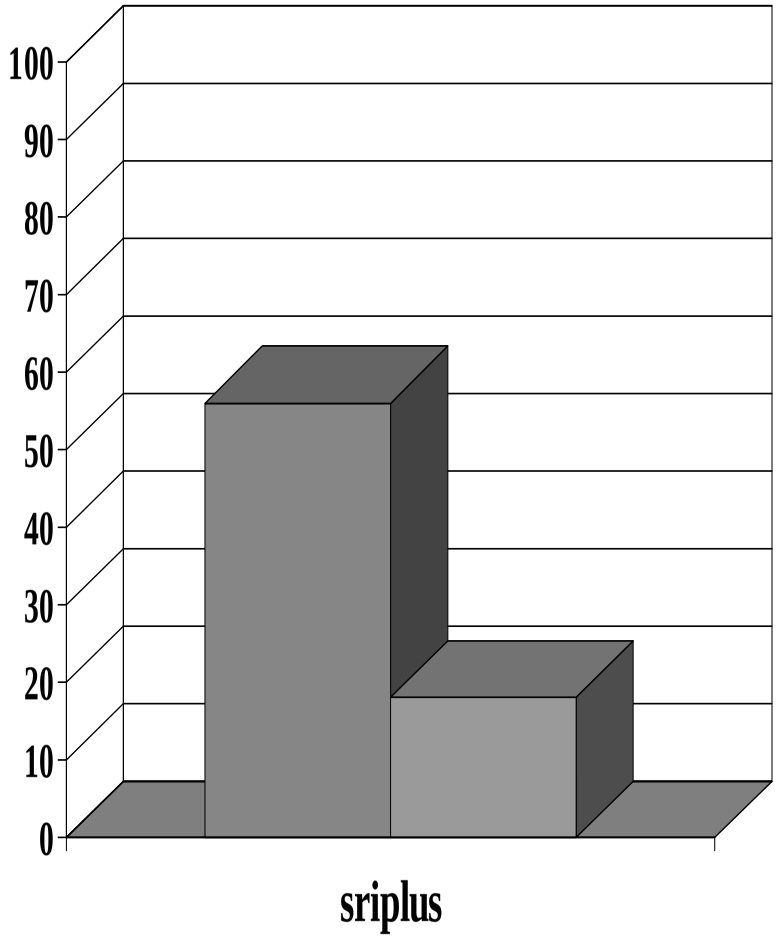


Conversational Telephone Speech

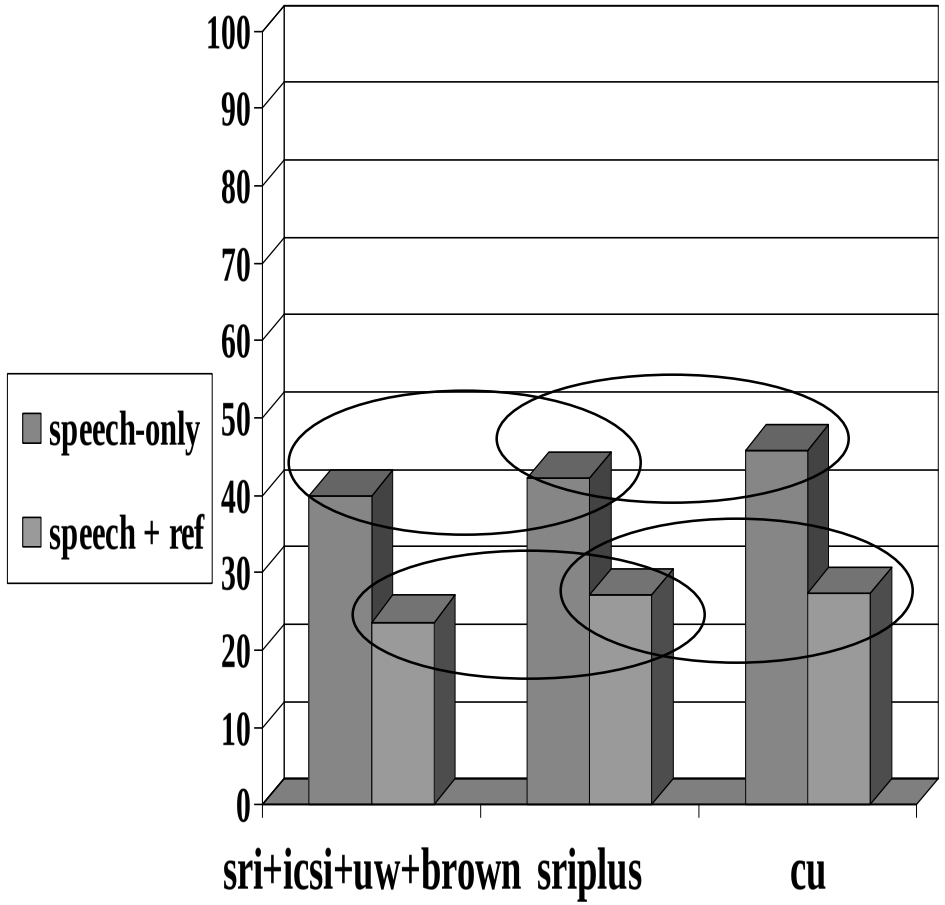
Filler-Word Detection (FWD)

- Four subtypes of fillers
  - Pause fillers, such as **uh**
  - Explicit editing terms, such as **I meant**
    - These occur in an edit disfluency
  - Discourse markers,
    - **Actually** he was **like** gross
  - Aside/parenthetical (*we are not evaluating these*)
- Task included identifying the subtype
  - But subtype was correct 99.9 % of the time

Filler-word Detection



Broadcast News

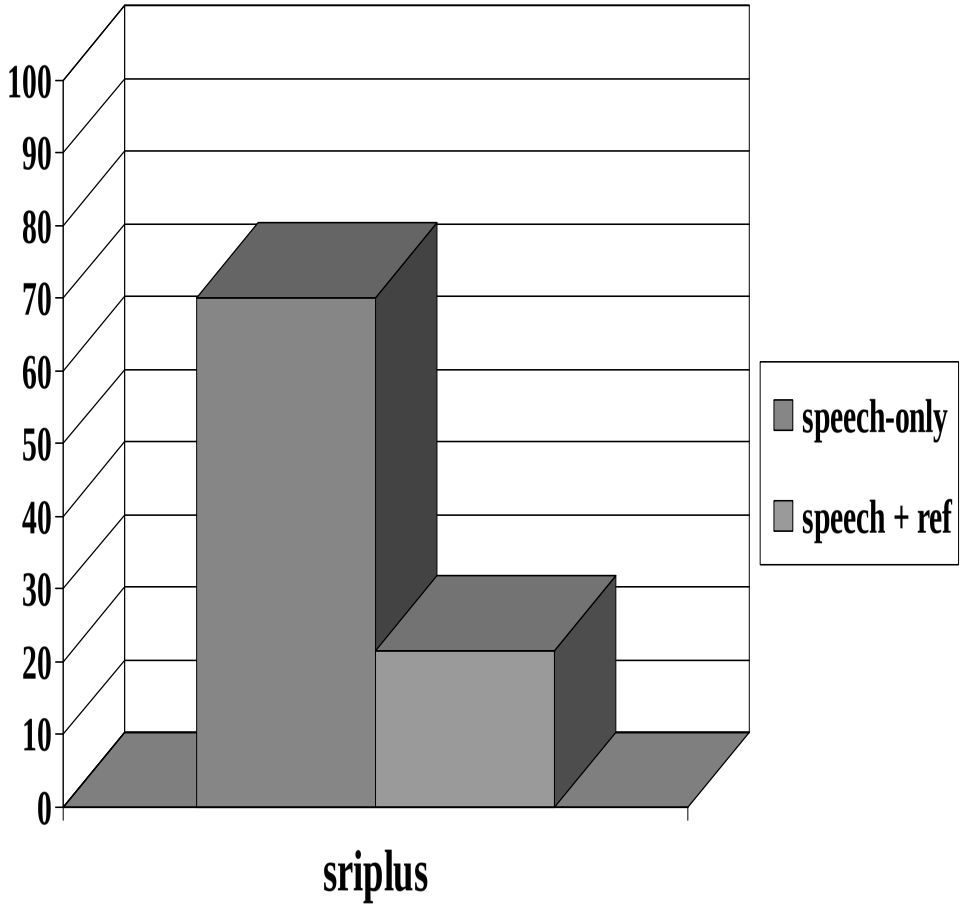


Conversational Telephone Speech

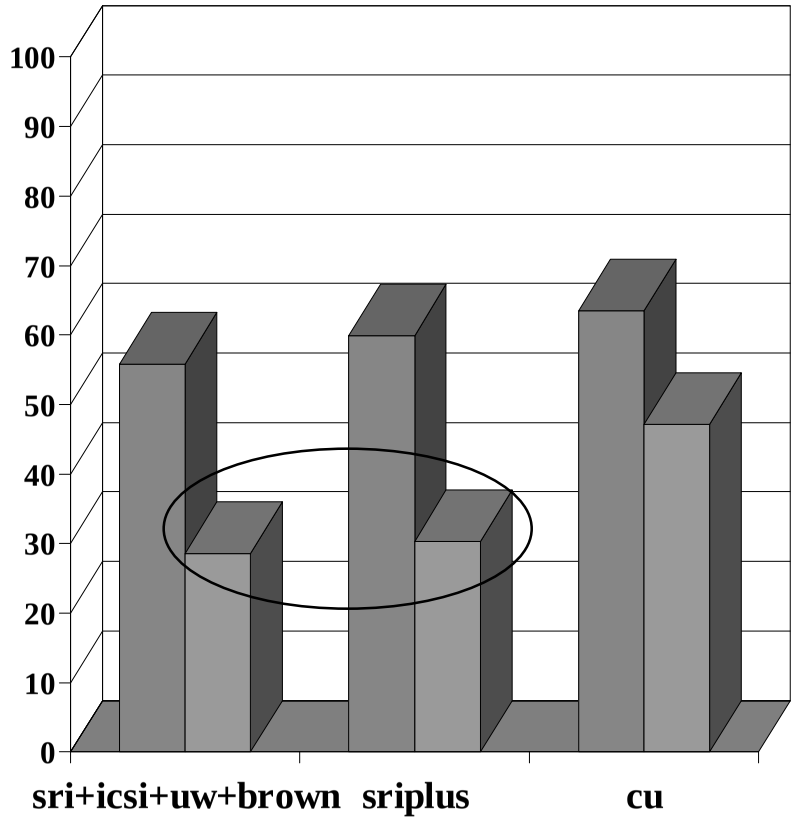
Interruption-Point Detection (IPD)

- Point where fluent speech is interrupted
  - Immediately after the last edit word in an edit disfluency
  - Immediately before each filler disfluency

Interruption Point Detection

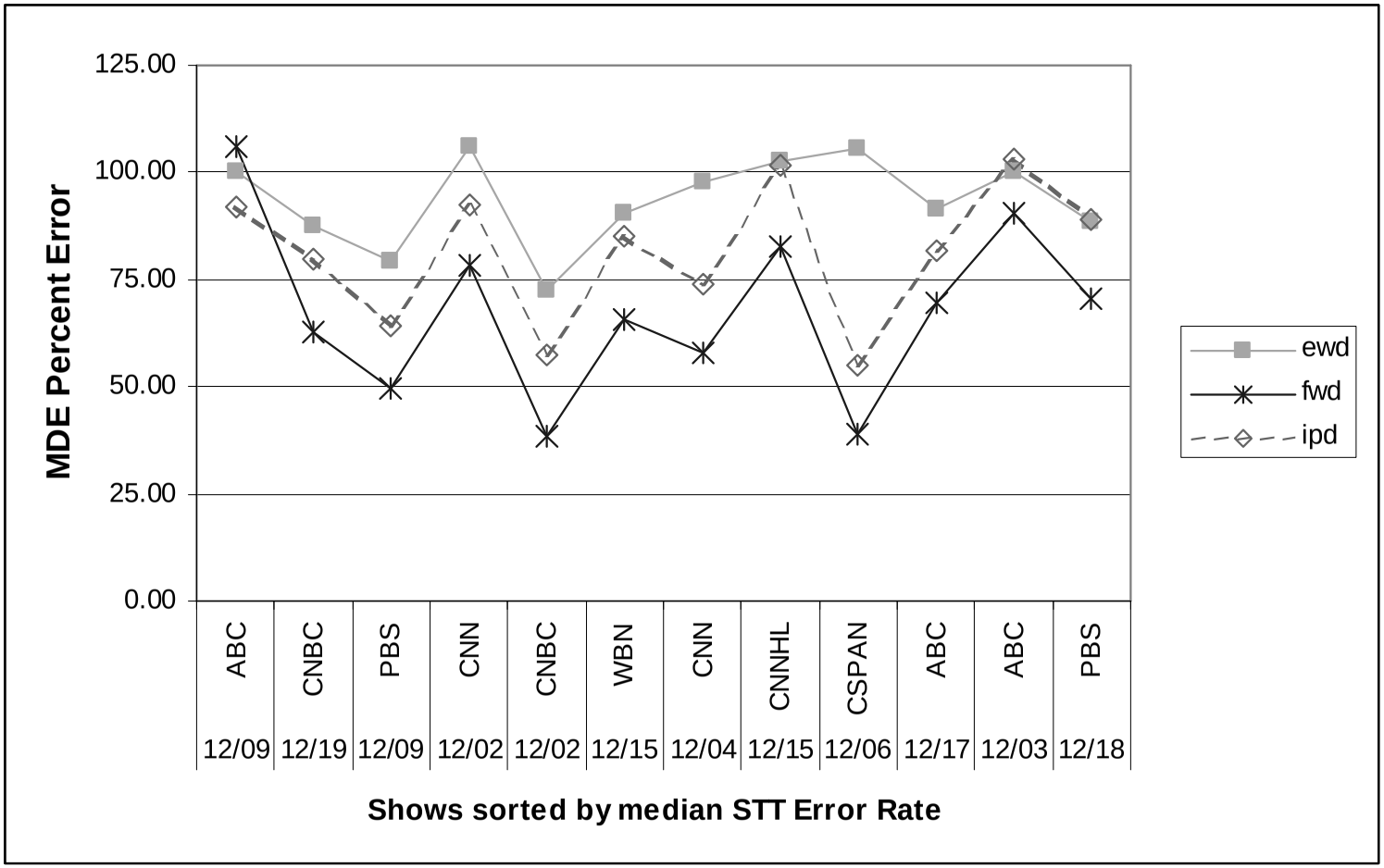


Broadcast News



Conversational
Telephone Speech

Disfluency Detection Performance Not Correlated with STT Performance



Speaker Diarization

- The system first segments the audio into regions where the speaker does not seem to change.
- The system then clusters those segments that seem to be by the same speaker.
- Final result:
  - system knows how many speakers it thinks there are
  - system knows when it think each was speaking
- Task is performed on BNews only
- The system also assigns a speaker-type (a sex) to each speaker. Scored as if speaker type were speaker ID.
  - Very low error rate at identifying sex of adult speakers
  - Data set is about 29% female and 71% male (by time)

Metric for Speaker Diarization

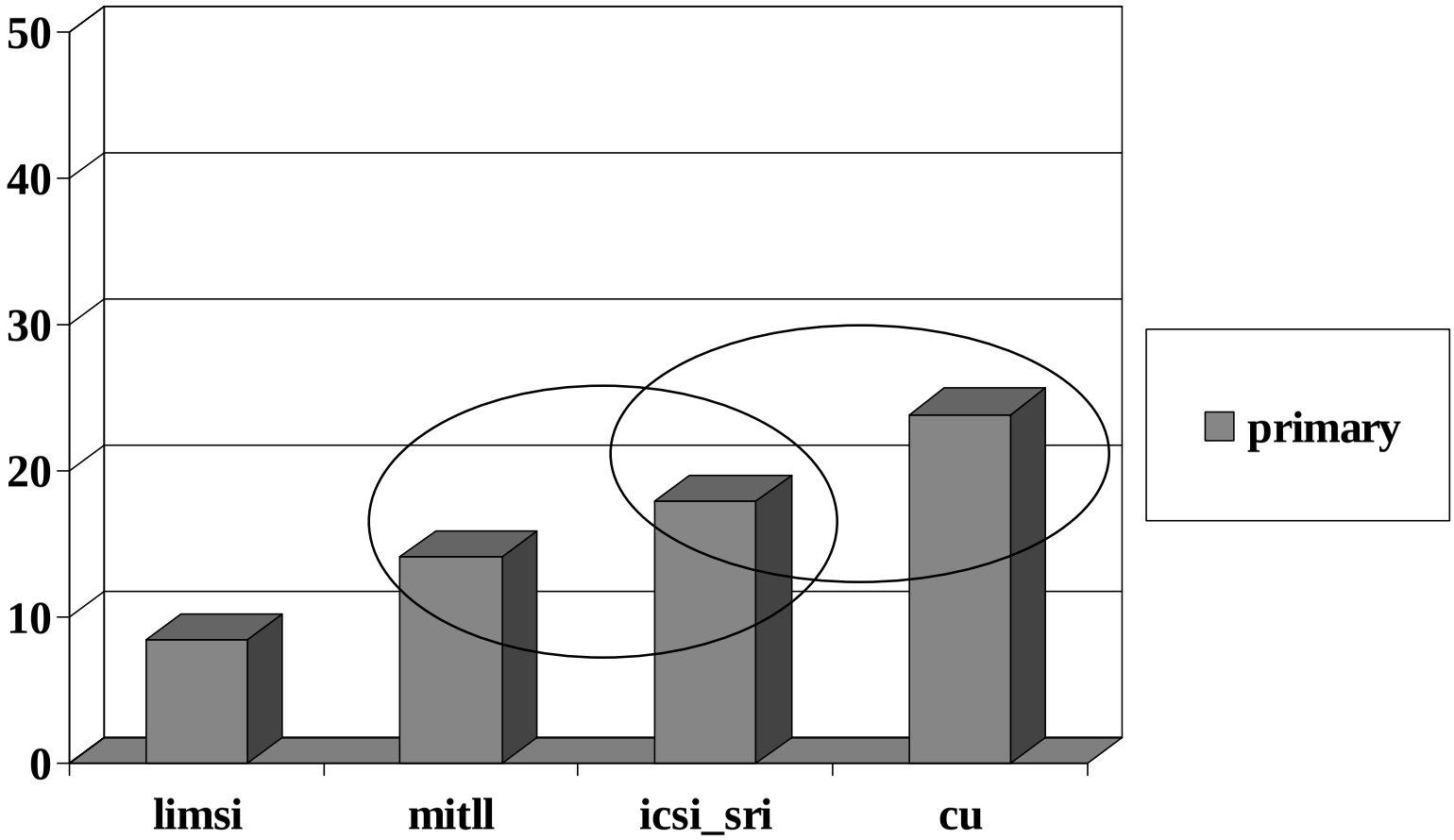
- Simplified form of the metric:

  (Speaker Diarization Error Time) / (Total Actual Speech)

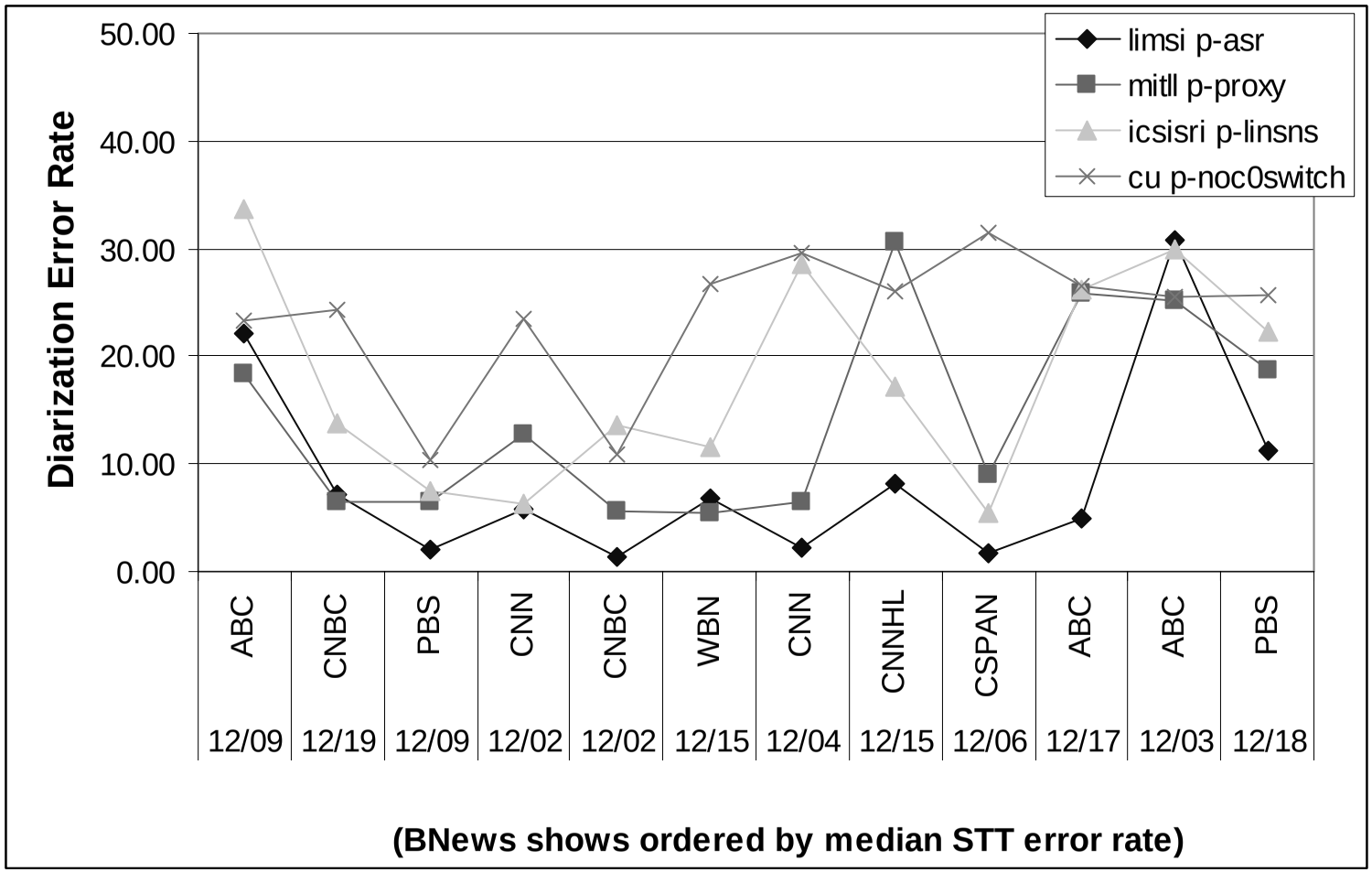
- The Error Time can be decomposed into three parts:
  - Speaker time that is attributed to the wrong speaker
  - Missed speaker time
  - False alarm speaker time

    .

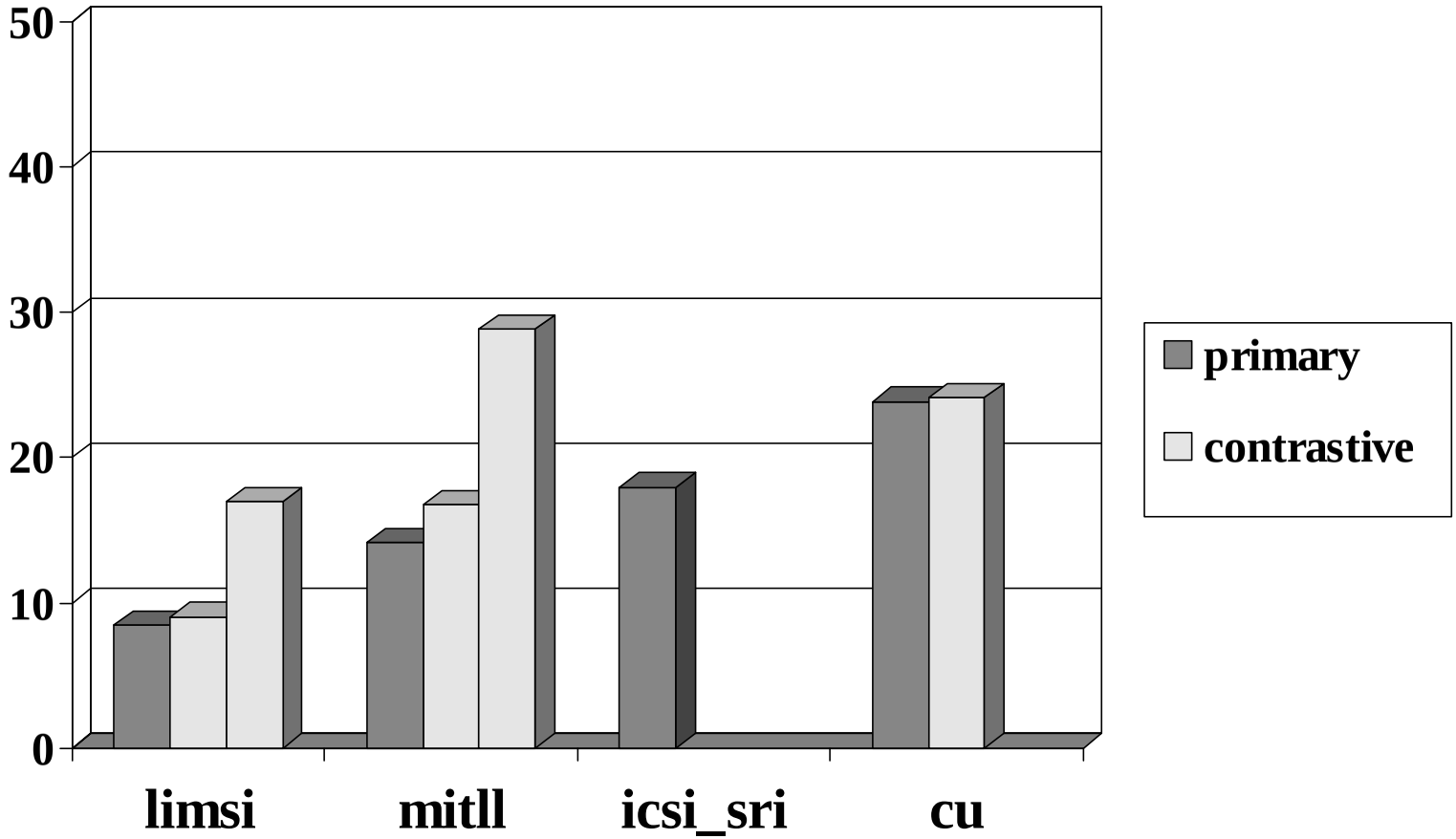
Speaker Diarization on BNews



Speaker Diarization by system, by show



Speaker Diarization on BNews



Summary

- Results of the Diarization and Structural MDE evaluations show benefits of collaborative research
- Structural MDE efforts are newer and less mature
  - Not clear how to make valid comparisons with results of previous structural MDE evaluations
  - Exciting possibilities for further research
- Diarization performance has improved
  - At the clustering stage, researchers are using specialized models for particular types of speakers/bandwidth
- Research represented here is beginning to exploit synergy/feedback between the various kinds of tasks (STT, diarization, and structural MDE)