# 1998 Hub-5 Conversational Speech Recognition Workshop Overview

Stanley Chen

Carnegie Mellon

*School of Computer Science*
*Carnegie Mellon University*
*Pittsburgh, Pennsylvania 15213*

October 15, 1998

## The Evaluation

- training data
  - Switchboard (160 hours)
    - telephone conversations between North Americans on a preassigned topic
  - CallHome English (20 hours)
    - telephone conversations between a non-native English speaker and family member
- evaluation data
  - 5 min. from 20 calls of both Switchboard and CallHome English

## The Results

word-error rates

| site | overall | SWB | CH | 1997 |
|---|---|---|---|---|
| CU-HTK | 39.5% | 36.7% | 42.2% | 48.7% |
| BBN | 39.7% | 38.4% | 40.9% | 44.9% |
| Dragon | 41.3% | 40.0% | 42.6% | 48.9% |
| BU | 43.6% | 42.1% | 45.0% | 50.1% |
| AT&T | 45.5% | 43.4% | 47.7% | **** |
| SRI | 49.7% | 48.7% | 50.6% | 50.2% |

## BBN/CU-HTK Architecture

| BBN | CU-HTK |
|---|---|
| 120 hr. SWB, 17 hr. CH | 160 hr. SWB, 17 hr. CH |
| 14 mel cep + power, $\Delta + \Delta\Delta$ | 12 PLP cep + power, $\Delta + \Delta\Delta$ |
| 3k senones $\times$ 80 Gaussians | 9k senones $\times$ 16 Gaussians |
| gender-dep., quinphone | gender-dep., quinphone |
| side-based CMS, CVN, VTLN | side-based CMS, CVN, VTLN |
| 35k voc, 3-gram, SWB + CH + CNN, POS-smoothing | 27k voc, 4-gram, SWB + CH + BN, class-based $n$-gram |
| multi-words, weights on pron. variants | LIMSI '93 dict + TTS |
| confidence: GAM on many features | decision-tree on $N$-best homogeneity |

## BBN/CU-HTK Decoding

| BBN | CU-HTK |
|---|---|
| side-based CMS/CVN, guess gender+VTL using GMM | side-based CMS/CVN |
| decode: fast match; coarse model $\rightarrow$ lattice; detailed model $\rightarrow$ $N$-best; POS-smoothed LM $\rightarrow$ 1-best | decode with GI non-VTLN triphone, trigram LM (Pass 1) |
| | find VTL, gender using GD VTLN triphones; decode (P2) |
| | global MLLR, gen lattices, rescore with 4g+class LM (P3) |
| MLLR mean + var, 4 classes; decode with diag. SAT adapted models | MLLR mean + var; 1,2,4,8 classes; quinphone, SAT (P4–7) |
| rescore $N$-best with diff frame rate, ROVER | combine P3 + P7 output with ROVER (P8) |

## CU-HTK Passes

| pass | WER | description |
|---|---|---|
| P1 | 49.3 | GI non-VTLN triphone, trigram LM |
| P2 | 45.6 | GD VTLN triphone, trigram LM |
| P3 | 42.6 | global MLLR, 4-gram+class LM |
| P4 | 40.9 | SAT quinphone, 1-class MLLR |
| P5 | 40.5 | SAT quinphone, 2-class MLLR |
| P6 | 40.4 | SAT quinphone, 4-class MLLR |
| P7 | 40.3 | SAT quinphone, 8-class MLLR |
| P8 | 39.5 | ROVER, P3 and P7 |

## BBN/CU-HTK WER Gains

| BBN | CU-HTK |
|---|---|
| CVN (0.7) | CMS (1.0), CVN (1.8) |
| correct likelihood bias in ML-VTLN (0.8) | VTLN (1.0) |
| quinphones (1.3) | quinphones (2.4) |
| | gender-dependent (1.4) |
| diag. SAT (faster than SAT) | SAT (no gain) |
| MLLR var adapt (0.5) | MLLR mean (1.4), var (0.3) |
| ROVER (1.2) | ROVER (0.8) |
| POS-smoothed LM + CNN (0.6) | 4-gram + class $n$-gram (0.5) |
| more params (0.9), 2-frame HMM state min (0.4), tuning (0.6) | block-diag. semitied cov. matrices (0.3) |

## Alternate Evaluation Metrics

- weighted word-error rate (WWER)
  - weight of word is $-\log_2$(unigram frequency)
  - cost(insert) = wgt(hyp word),
    cost(delete) = wgt(ref word),
    cost(sub) = wgt(ref word) + wgt(hyp word)

$$\text{WWER} = \frac{\text{cost of errors}}{\text{sum of weight of words in ref}}$$

  - relative performance of sites stayed the same
  - WER range: 39.5%–49.7%, WWER range: 44.8%–57.3% or 65.5%–84.7%?
  - SRI — when optimized WWER instead of WER, WWER was 0.4% absolute lower, WER was 0.2% absolute higher

## Alternate Evaluation Metrics

information extraction (IE)

- identify names (organization, person, location), dates and times, and numbers (moneys and percentages)
- fixed IE engine, using speech recognized input
- evaluation metrics
  - slot error rate (SER)
  - F-measure = 2 * precision * recall / (precision + recall)

## Information Extraction

WER $\neq$ IE

| site | WER | F-meas. | SER |
|---|---|---|---|
| reference | 0.0% | 0.71 | 47% |
| CU-HTK | 39.5% | 0.43 | 84% |
| BBN | 39.7% | 0.45 | 80% |
| Dragon | 41.3% | 0.45 | 85% |
| BU | 43.6% | 0.41 | 85% |
| AT&T | 45.5% | 0.38 | 86% |
| SRI | 49.7% | 0.39 | 93% |

## Information Extraction

- optimizing for IE
  - no success in optimizing for F-measure
  - SRI lowered SER by 2% absolute when optimizing for word spotting
- how well do WER, WWER, normalized-cross entropy (NCE) of confidence scores correlate with F-measure, SER?
  - best correlation with F-measure: NCE, then WER, then WWER
  - best correlation with SER: WWER, then WER, then NCE

# ROVER

recognizer output voting error reduction

- given output of several speech recognizers
- align hypotheses word-by-word
- choose each word of final hypothesis by voting
  - by count or avg./max confidence score
- combined five systems from last year's Hub-5 eval
  - best indiv sys: 44.9% WER $\Rightarrow$ ROVER: 39.4% WER

## ROVER

results

- AT&T — combine tetraphone and quinphone rescoring results, improve 0.5% absolute WER over best individual model
- BBN — combine quinphone at 80, 100, 125 frames/sec, triphone at 100 frames/sec; 1% absolute WER improvement
- CU-HTK
  - combine output from two different passes; assign confidence to each word using $N$-best homogeneity
  - combine using ROVER with confidences; 0.5%–0.8% absolute WER improvement

## Resegmentation and Transcription of SWB

- MSU is doing super-clean transcription of SWB
- done with half of 240 hours of SWB?
- transcript WER 8% $\rightarrow$ 2.7%
- adapted AM to new transcriptions, 49.8% $\rightarrow$ 47.9% WER