SU Detection for RT-03f at Cambridge University

Marcus Tomalin, Sue Tranter, Phil Woodland
& the CU-HTK STT Team

13th November

Cambridge University Engineering Department

Presentation Overview

- Overview of the CUED CTS SU-Detection System.
- The Prosodic Feature Model.
- The Slash Unit Language Models.
- The Decoder.
- Key Results.
- Scoring Tools.
- Training Data and SU %Err.
- Conclusions and Future Plans.

CTS SU-Detection System Overview

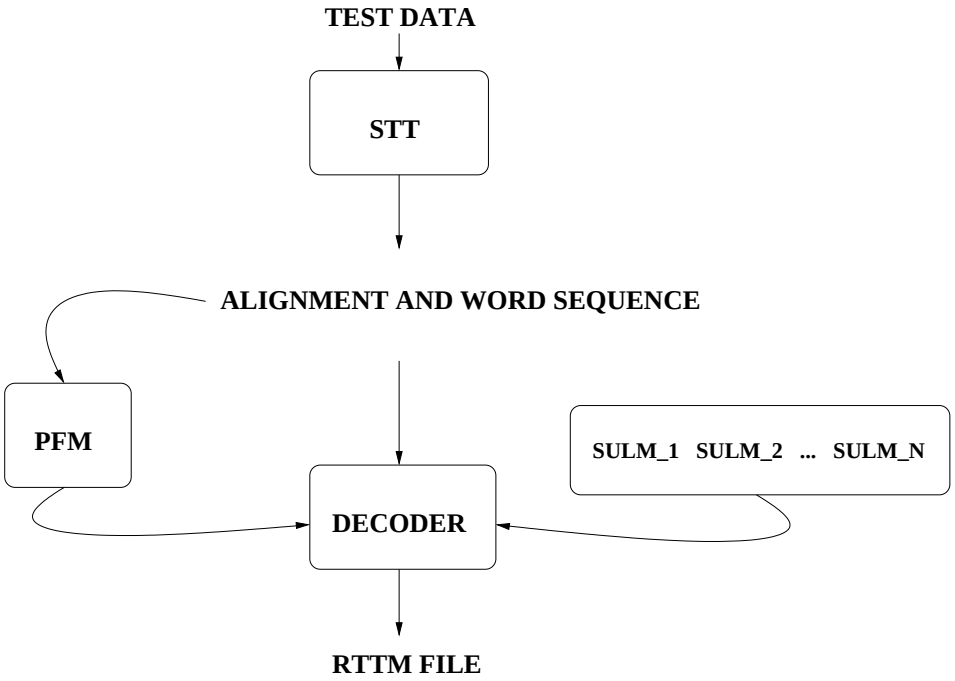


Figure 1: SU-Detection System

STT Output

CU-HTK CTS STT 187×RT System for RT-03s Eval:

- Automatic Segmentation
- Multi-pass System
- MPE Training
- HLDA Transforms
- SAT models
- SPron models
- Adaptation and System Combination

For details see:

Woodland et al. 'CU-HTK STT System for RT-03', Rich Transcription Workshop May 2003

CU-HTK CTS STT 187×RT system output (with optionally deletable tokens retained) used as input to MDE system.

The Prosodic Feature Model

The Prosodic Features (PFs):

Prosodic Feature	Description
Pause_Length	the pause length at the end of the word
Duration	the duration from the previous pause
Avg_F0_L	the mean of the good F0 values† in left window
Avg_F0_R	the mean of the good F0 values in right window
Avg_F0_ratio	Avg_F0_L / Avg_F0_R
Cnt_F0_L	the number of good F0s in left window
Cnt_F0_R	the number of good F0s in right window
Eng_L	the RMS energy in left window
Eng_R	the RMS energy in right window
Eng_ratio	Eng_L / Eng_R

†: 50Hz $\leq$ good F0 values $\leq$ 400Hz

The Prosodic Feature Model

Five SU sub-types defined:

- **SU_S**: statement SU boundary
- **SU_Q**: question SU boundary
- **SU_I**: incomplete SU boundary
- **SU_B**: backchannel SU boundary
- **SU_N**: no SU boundary

Steps in the PFM construction process:

- Convert training data into word sequences.
- Classify each word into one of the above SU sub-types.
- Obtain forced alignments for words in each segment.
- Extract PF info using word start/end times.
- Cross-Validation.
- Construct CART decision tree using PFs and SU sub-type classification.

The Prosodic Feature Model

Training Data	Num PFM Vecs	Num Tree Nodes
LDC train-simple-pilot	27,825	N/A
LDC train-dryrun	12,124	N/A
LDC train-batch1-meteer40 data	94,765	N/A
LDC train-1st-third data	152,737	N/A
LDC train-2nd-third data	80,683	N/A
LDC train-3rd-third data	232,067	N/A
all LDC data	600,201	380 (153 terminal)
SRI+ meteer-mapped V5 data	152,737	336 (170 terminal)
all training data	**752,938**	**397 (183 terminal)**

The Slash Unit Language Models

Insert the required SU token after every word in the training data:

> < s > OKAY **SU_S** ARE WE READY **SU_Q** I THINK WE SHOULD GIVE **SU_I** OKAY **SU_S** ... < /s >

Various SULMs built using standard LM tools:

- N-gram SULMs (i.e., tg = 3gram, fg = 4gram).
- Class-based SULMs (i.e., cl40-tg = 40 class tg).
- Interpolated SULMs (i.e., tg*cl40-tg = interpolated tg and cl40-tg).
- Perplexities (PPs) calculated using the dev03f test data.
- Interpolation Weights (IWs) calculated using the dev03f test data.

The Slash Unit Language Models

Two different types of stream information for SULM interpolation:

- ST_T: obtain stream info for all tokens in training data.
- ST_S: obtain stream info only for SU tokens in training data.

ST_T and ST_S give different PPs and IWs.

Interpolating a **tg**, a **cl40-tg** and a **cl40-fg**:

Stream Type	Tok PP	SU PP	IWs	SU Err
ST_T	106	N/A	∼0.7, ∼0.2, ∼0.1	46.15
ST_S	N/A	6.6	∼0.5, ∼0.2, ∼0.3	**45.88**

- PFM and SULMs trained on all LDC and meteer-mapped V5 data.
- The decoder used posterior decoding.
- Systems tested using dev03f test data.
- Scores obtained using su-eval-v15.pl with the '-w -W -t 1.00' settings.

The Slash Unit Language Models

Some SULM results for the dev03f test set using su-eval-v15.pl:

System	SU PP	IWs	%Del	%Ins	%Err
pfm+tg	7.3	N/A	32.0	16.4	48.4
pfm+fg	7.7	N/A	33.6	15.8	49.4
pfm+cl40-tg	7.6	N/A	33.5	17.3	50.8
pfm+cl40-fg	7.9	N/A	28.9	26.9	55.8
pfm+(tg*cl40-tg)	6.7	∼0.5, ∼0.5	31.1	14.8	45.9
pfm+(tg*cl40-fg)	6.7	∼0.6, ∼0.4	30.3	16.2	46.5
pfm+(tg*cl40-tg*cl40-fg)	6.6	∼0.5, ∼0.2, ∼0.3	31.8	14.1	**45.9**

- All SULMs were trained using LDC and meteer-mapped V5 training data.
- The PFM was trained using LDC and meteer-mapped V5 training data.
- The decoder used posterior decoding
- Systems tested using dev03f test data.
- Scores obtained using su-eval-v15.pl with the '-w -W -t 1.00' settings.

The Decoder

The SU Decoder: lattice-based combination of the PFM and SULM scores.

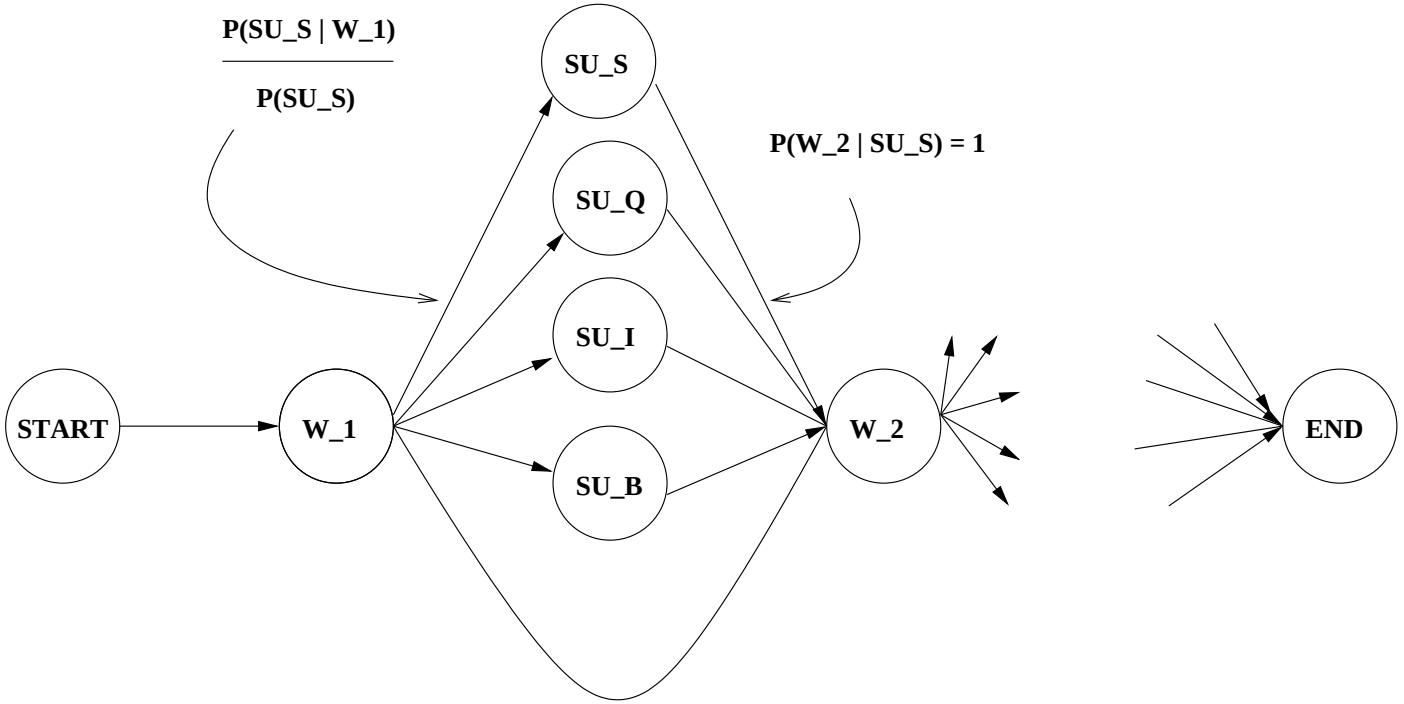


Figure 2: Initial SU Decoder lattice

The Decoder

Comparing two decoding strategies:

- VITERBI-1-BEST
  - Expand initial lattices using SULM.
  - Select hypothesis with highest likelihood.

- POSTDEC-1-BEST
  - Expand initial lattices using SULM.
  - Estimate word-level posterior probs.
  - Sum the posteriors of the SU subtypes.
  - Generate confusion network.
  - Select hypothesis with highest posterior prob.

The Decoder

Su-Detection System:

- PFM
- Interpolated tg, cl40-tg and cl40-fg SULM
- acoustic scale factor = 2.0
- grammar scale factor = 1.0
- insertion penalty = 0.0

Experimental Set-up:

- Training data: LDC and meteer-mapped V5 data
- Test Data: dev03f test set
- Scores obtained using su-eval-v15.pl with the '-w -W -t 1.00' settings.

Decoding Method	%**Del**	%**Ins**	%**Err**
VITERBI-1-BEST	31.36	15.09	46.45
POSTDEC-1-BEST	31.75	14.12	**45.88**

Key Results: Dec02-Oct03

Three CTS SU-detection Systems:

- **Dec02-Sys**: simple rule-based system used for Dec 2002 dryrun.
- **Post-RT-03s-Sys**:
  - TB3 data (c.90 hrs).
  - Side-based forced alignments (i.e., no segment info in training data)
  - PFM (1456 nodes [729 terminal]), 10 prosodic features.
  - SULM (bg).
- **RT-03f-Sys**:
  - LDC data and meteer-mapped V5 data (c.40 hrs).
  - Segment info in training data used when generating forced alignments.
  - PFM (397 nodes [183 terminal]), 10 prosodic features.
  - Interpolated SULMs (tg, cl40-tg, cl40-fg).
  - IWs obtained from SU stream info.
  - Posterior decoding.

Key CTS Results: Dec02-Oct03

System	%Del	%Ins	%Err
Dec02-Sys	58.30	19.00	77.30
Post-RT-03s-Sys	45.60	16.99	62.59
RT-03f-Sys	31.75	14.12	**45.88**

All systems were tested using the dev03f test set.
All scores obtained using su-eval-v15.pl with the '-w -W -t 1.00' settings.

Key CTS Results: Dec02-Oct03

The Ref condition task:

- Ref files segmented automatically.
- Missing dictionary entries added manually.
- Word times converted back to word times in Ref files.

System	%Err (Dev03f)	%Err (Eval03f)
RT-03f-Sys Sys	45.88	46.04
RT-03f-Sys Ref	34.86	34.59

All scores obtained using su-eval-v15.pl with the '-w -W -t 1.00' settings.

System	%Err (Dev03f)	%Err (Eval03f)
RT-03f-Sys Sys	49.52	50.29
RT-03f-Sys Ref	34.96	34.62

All scores obtained using rteval-v2.3.pl.

Scoring Tools

- su-eval-v12.pl and rteval-v2.3.pl used for system development.
- su-eval-v15.pl and rteval-v2.3.pl used to score RT-03f eval submissions.

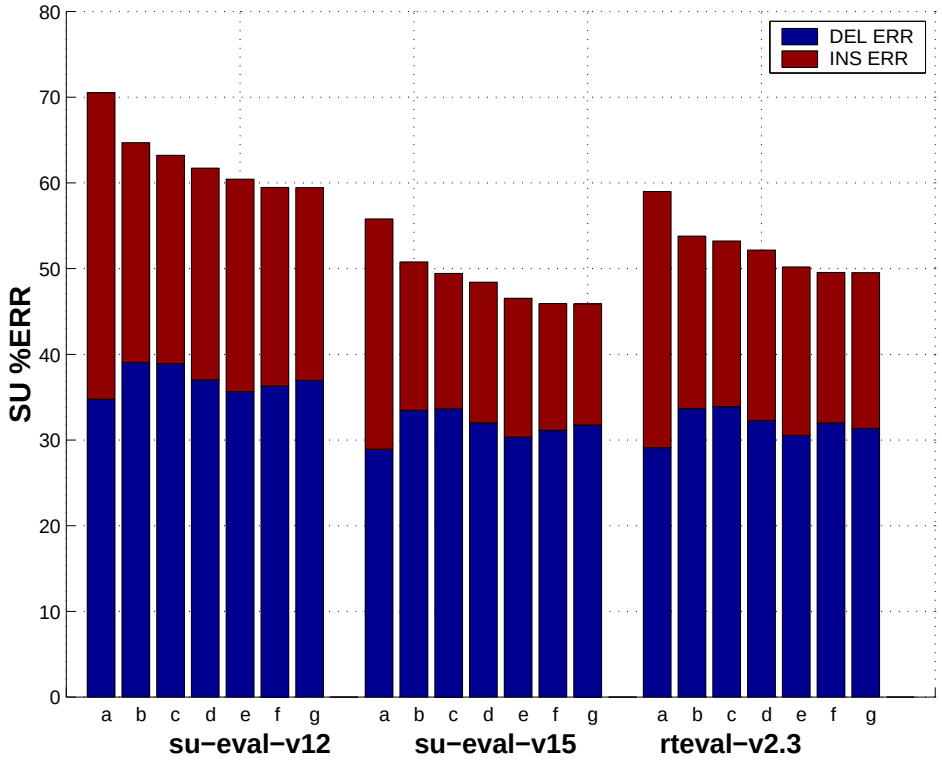
Results obtained for the following systems:

Code	System
a	pfm+(tg*cl40-fg)
b	pfm+(tg*cl40-tg)
c	pfm+fg
d	pfm+tg
e	pfm+(tg*cl40-fg)
f	pfm+(tg*cl40-tg)
g	pfm+(tg*cl40-tg*cl40-fg)

All systems used posterior decoding and scores obtained for dev03f test data.

Scoring Tools

Comparison of scoring tools for different systems:



Basic trends similar; DEL counts closer than INS counts for most recent versions of tools.

Training Data and SU %Err

CTS training data:

- (1) LDC train-1st-third data (c.10 hrs).
- (2) LDC train-2nd-third data (c.6 hrs).
- (3) LDC train-3rd-third data (c.15 hrs).
- (4) SRI+ meteer-mapped V5 data (c.9 hrs).

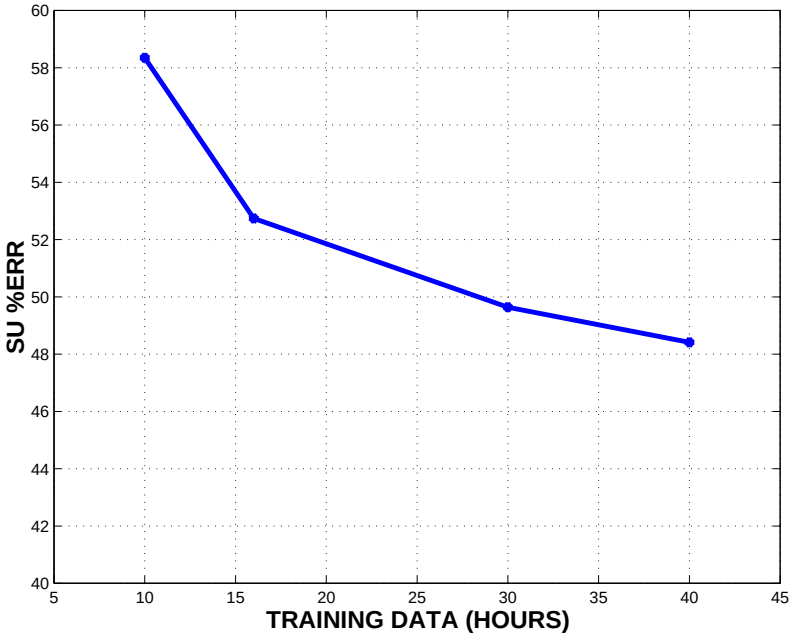
Exploring the cumulative effect of training data on SU %Err rate:

- Build PFM and tg SULM using training data set number (1).
- Obtain results for the dev03f test set.

1. Add next training data set (i.e., cumulative increase in training data).
2. Rebuild PFM and tg SULM.
3. Obtain results for the dev03f test set.
4. Stop if training data set number = (4), else goto 1.

Training Data and SU %Err

The SU %Err rate falls as amount of training data increases:



SU %Err falls at a rate of c.0.25 % (abs) per hour of training data

Conclusions

- Scoring tools still unstable and they have not yet converged.
- SU %Err for CTS task reduced from 62.59 to 45.88 since May 03.
- Task-specific training data reduces SU %Err at rate of 0.25% (abs) per hour.
- Interpolating SULMs reduces SU %Err (c.2.5% abs).
- Calculating IWs using SU stream info reduces SU %Err (c.0.3% abs).
- Posterior decoding strategy reduces SU %Err (c.0.6% abs).

Future Plans

- Continue to provide feedback concerning tools, task definitions etc.
- Develop BN system.
- Explore system combination strategies.
- Develop PFMs (i.e., experiment with other kinds of features).
- Use syntactic parser as post-processing stage (work in progress).
- Consider impact of STT performance upon the SU detection task.