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.

Cambridge University Engineering Department

13th November

Marcus Tomalin, Sue Tranter, Phil Woodland

SU Detection for RT-03f at Cambridge University
Tomalin et al.: SU Detection for RT-03f at Cambridge University

**CTT System Overview**

SULM_1   SULM_2   ...   SULM_N

RTTM FILE

TEST DATA

DECODER

PFL

ALIGNMENT AND WORD SEQUENCE

STT

Figure 1: SU-Detection System

CU-HTK CTS STT 17×RT system for RT-03, Rich Transcription Workshop May 2003

For details see:

- Adaptation and System Combination
- SPn models
- SAT models
- HLA Transforms
- MPE Training
- Multi-pass System
- Automatic Segmentation

CU-HTK CTS STT 17×RT System for RT-03 Eval:

**STT Output**

Tomalin et al.: SU Detection for RT-03f at Cambridge University

RT-03f Workshop, 13th November 2003

For details see: Woodland et al., CU-HTK STT System for RT-03, Rich Transcription Workshop May 2003

RT system output (with optionally deletable tokens retained) used as input to MDE system.
### Prosodic Feature Model

**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.
- Construct CART decision tree using PFs and SU sub-type classification.
- Cross-validation.

### Prosodic Features (PFs):

<table>
<thead>
<tr>
<th>Description</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pause Length</td>
<td>the pause length at the end of the word</td>
</tr>
<tr>
<td>Duration</td>
<td>the duration from the previous pause</td>
</tr>
<tr>
<td>Avg F0 L</td>
<td>the mean of the good F0 values in left window</td>
</tr>
<tr>
<td>Avg F0 R</td>
<td>the mean of the good F0 values in right window</td>
</tr>
<tr>
<td>Avg F0 L/R</td>
<td>the mean of the good F0 values in left window divided by right window</td>
</tr>
<tr>
<td>Eng L/R</td>
<td>the RMS energy in left window divided by right window</td>
</tr>
<tr>
<td>Cnt F0 L</td>
<td>the number of good F0s in left window</td>
</tr>
<tr>
<td>Cnt F0 R</td>
<td>the number of good F0s in right window</td>
</tr>
<tr>
<td>Eng Ratio</td>
<td>the RMS energy in right window divided by left window</td>
</tr>
<tr>
<td>50Hz &lt; Good F0 values &lt; 400Hz</td>
<td></td>
</tr>
</tbody>
</table>
Insert the required SU token after every word in the training data:

```
< s OKAY SU s >

> s OKAY SU s ARE WE READY SU Q I THINK WE SHOULD GIVE
```

Various SULMs built using standard LM tools:

| N-gram SULMs (i.e., $t_g = \text{4gram}$ and $f_g = \text{4gram}$). |
| Class-based SULMs (i.e., $c_l = \text{40 class}$). |
| Interpolated SULMs (i.e., $t_g \times c_l = \text{interpolated}$ $t_g$ and $c_l$). |

Perplexities (PPs) calculated using the dev03f test data.

Interpolation Weights (IWs) calculated using the dev03f test data.

### The Prosodic Feature Model

<table>
<thead>
<tr>
<th>Num PFM Vvecs</th>
<th>Num Tree Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>397 (183 terminal)</td>
<td>380 (153 terminal)</td>
</tr>
<tr>
<td>752.938</td>
<td>737</td>
</tr>
<tr>
<td>336 (170 terminal)</td>
<td>600.201</td>
</tr>
<tr>
<td>152</td>
<td>27.25</td>
</tr>
<tr>
<td>232.07</td>
<td>94.76</td>
</tr>
<tr>
<td>80.67</td>
<td>12.24</td>
</tr>
<tr>
<td>693.24</td>
<td>12.24</td>
</tr>
<tr>
<td>397</td>
<td>12.24</td>
</tr>
<tr>
<td>397</td>
<td>12.24</td>
</tr>
</tbody>
</table>

All training data

SR1+ meteor-mapped V5 data

LDG train-3rd-third data

LDG train-2nd-third data

LDG train-1st-third data

LDG train-batch-meteor40 data

LDG train-dynun

LDG train-simple-pilot
The Slash Unit Language Models

Two different types of stream information for SULM interpolation:

- ST: obtain stream info for all tokens in training data.
- STS: obtain stream info only for SU tokens in training data.

ST and STS give different PPs and IWs.

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

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.

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

Interpolating a TG, a CG40-TG and a CG40-RG:

ST-T and ST-S give different PPs and IWs.

ST-S: obtain stream info only for SU tokens in training data.

ST-T: obtain stream info for all tokens in training data.

Two different types of stream information for SULM interpolation:
The Decoder

Comparing two decoding strategies:

- **VITERBI-1-BEST**
  - Expand initial lattices using SULM.
  - Select hypotheses with highest likelihood.
  - Generate confusion network.
  - Sum the posteriors of the SU subtypes.
  - Estimate word-level posterior probs.
  - Expand initial lattices using SULM.

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

### The Decoder

**Figure 2:** Initial SU Decoder Lattice

The SU Decoder: lattice-based combination of the PFM and SULM scores.
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 meteor-mapped V5 data (c.40 hrs)
- Test data: dev03f test set
- Scores obtained using su-eval-v15.pl with the `-w -W -t 1.00` settings.

Decoding Method

<table>
<thead>
<tr>
<th>Decoding Method</th>
<th>% Ins</th>
<th>% Del</th>
<th>% Err</th>
</tr>
</thead>
<tbody>
<tr>
<td>VITERBI-1-BEST</td>
<td>31.36</td>
<td>15.09</td>
<td>46.45</td>
</tr>
<tr>
<td>POSTDEC-1-BEST</td>
<td>31.75</td>
<td>14.12</td>
<td>45.88</td>
</tr>
</tbody>
</table>

Key Results: Dec02-Oct03

Three CTS Su-detection Systems:

1. **RT-03f-Sys**:
   - SULM (6p).
   - PFM (397 nodes [183 terminal], 10 prosodic features).
   - Interpolated SULMs (tg, cl40-tg, cl40-fg).
   - IWs obtained from SU stream info.

2. **Post-RT-03s-Sys**:
   - TB3 data (c.90 hrs).
   - SULM (1456 nodes [729 terminal], 10 prosodic features).
   - PFM (1456 nodes [729 terminal], 10 prosodic features).

3. **Dec02-Sys**: Simple rule-based system used for Dec 2002 dryrun.

---

Scores obtained using su-eval-v15.pl with the `-w -t 1.00` settings.

Experimental Set-up:
- Insertion penalty = 0.0
- Acoustic scale factor = 2.0
- Grammar scale factor = 1.0
- Interpolated SULM (tg, cl40-tg, cl40-fg) and PFM

SU-Detection System: The Decoder
Tomalin et al.: SU Detection for RT-03f at Cambridge University

Key CTS Results: Dec02-Oct03

<table>
<thead>
<tr>
<th>System</th>
<th>%Err (Dev03f)</th>
<th>%Err (Eval03f)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RT-03f-Sys</td>
<td>34.62</td>
<td>34.96</td>
</tr>
<tr>
<td>RT-03f-Sys</td>
<td>30.29</td>
<td>49.52</td>
</tr>
</tbody>
</table>

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

The Ref condition task:

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

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

Tomalin et al.: SU Detection for RT-03f at Cambridge University
su-eval-v12.pl and rteval-v2.3.pl used for system development.

results obtained for the following systems:

<table>
<thead>
<tr>
<th>Code</th>
<th>System</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>pfm+(tg*cl40-fg)</td>
</tr>
<tr>
<td>b</td>
<td>pfm+(tg*cl40-tg)</td>
</tr>
<tr>
<td>c</td>
<td>pfm+fg</td>
</tr>
<tr>
<td>d</td>
<td>pfm+tg</td>
</tr>
<tr>
<td>e</td>
<td>pfm+(tg*cl40-fg)</td>
</tr>
<tr>
<td>f</td>
<td>pfm+(tg*cl40-tg)</td>
</tr>
<tr>
<td>g</td>
<td>pfm+(tg<em>cl40-tg</em>cl40-fg)</td>
</tr>
</tbody>
</table>

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

results obtained for the following systems:

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

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

Comparison of scoring tools for different systems:
SU %Err falls at a rate of c. 0.25 %/h (abs) per hour of training data.

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

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

- 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+ meter-mapped V5 data (c.9 hrs)

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

Training Data and SU %Err
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.

SU %Err for CTS task reduced from 62.59 to 45.88 since May 03.

Posterior decoding strategy reduces SU %Err (c.0.6% abs).

Calculating IWs using SU stream into reduces SU %Err (c.0.3% abs).

Interpolating SULMs reduces SU %Err (c.2.5% abs).

Future Plans

• Continue to provide feedback concerning tools, tasks definitions etc.
• Continue to provide feedback concerning tools, tasks definitions etc.

• Develop PFMs (i.e. experiment with other kinds of features).
• Develop BN system.
• Use syntactic parser as post-processing stage (work in progress).

• Explore system combination strategies.
• Explore system combination strategies.

• Consider impact of STT performance upon the SU detection task.
• Consider impact of STT performance upon the SU detection task.

• Develop BN system.
• Develop BN system.