The Development of the Cambridge University RT-04 Diarisation System

S.E. Tranter, M.J.F. Gales, R. Sinha, S. Umesh, P.C. Woodland

9th November 2004

Cambridge University

RT-04 workshop: November 2004
Overview

• The Diarisation Task and Data

• The CU Diarisation System

• Development Results

• Results on RT-04f Evaluation Data

• Summary and Future Plans
The Diarisation Task

• Task
  – Label ‘who spoke when’ from audio data.
  – Essentially a speaker segmentation and clustering task.

• Data
  – Development sets: each consisted of 6 shows of approx. 30 minutes
    didev03: RT-03s dev data, epoch Oct-Dec 2000.
    sttdev04: manually marked at CU, epoch Jan 2001.
    devall: represents sum of all dev sets.

The CU Diarisation System - Overview

3 stage process: Segmentation → Gender Labelling → Clustering

- **Coding**
  - mfcc, plp, plp-nb data

- **Segmentation**
  - Music + silence removed
  - Segments with bandwidth and putative speaker labels

- **Gender-Labelling**
  - Silence removed
  - Added gender labels

- **Clustering**
  - Speaker segmentation
The CU Diarisation System - Segmenter

- MFCC/PLP coding
- Speech/Music/Narrowband GMMs
- Bandwidth labels
- Phone recognition
- Divergence-based change detection
- Silence removed
- Inter-silence segments
- Over-segmented data

- Build model for each segment
- Viterbi Segmentation and re-estimation
- Threshold-based clustering
- no change or max iterations

- Over-segmented data is combined using LIMSI-style iterative scheme.
The CU Diarisation System - Segmenter (2)

- Single Gaussian model is built for each segment.
- Segments having loss likelihood less than a threshold if merged are combined.
- Viterbi decoding using new models then resegments the data.
- First few iterations used diagonal covariance to model segments as there were many short segments.
- In subsequent iterations a full covariance model is used.
- RT-04 segmenter also produces speaker labels unlike RT-03s segmenter.
The CU Diarisation System - Gender Labelling

- The first-pass of CU BN STT system is run to transcribe the segmenter output.

- Segments with no transcription are discarded from segmenter output.

- A forced alignment with GD models then determines the most likely gender of each segment.
The CU Diarisation System - Clusterer

• Clustering is done bandwidth and gender dependently.

• Clusterer uses only the start/end times of segments and ignores the segmenter speaker labels.

• Segments are sorted by mid-time based segment cluster-id before clustering.

• Clustering is done top-down using AHS distance metric and BIC-based stopping criterion.

• Single full correlation matrix of static PLP features is used to model segments.
### Development Results - Segmenter Improvement

<table>
<thead>
<tr>
<th>Segmentation</th>
<th>Dataset</th>
<th>Segment-Impurity †</th>
<th>Seg DER</th>
<th>+Clust DER</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MS/FA/SPE/SI @ NumSeg</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RT-03s</td>
<td>didev03</td>
<td>0.1/3.0/1.9/5.07 @ 875</td>
<td>-</td>
<td>18.8</td>
</tr>
<tr>
<td></td>
<td>eval03</td>
<td>0.3/1.9/1.7/3.92 @ 869</td>
<td>-</td>
<td>19.8</td>
</tr>
<tr>
<td></td>
<td>sttdev04</td>
<td>1.0/0.9/2.1/4.01 @ 913</td>
<td>-</td>
<td>22.9</td>
</tr>
<tr>
<td></td>
<td>dev04f2</td>
<td>1.3/4.1/1.0/6.33 @ 1077</td>
<td>-</td>
<td>32.7</td>
</tr>
<tr>
<td></td>
<td>devall</td>
<td><strong>0.69/2.34/1.70/4.74 @ 3734</strong></td>
<td>-</td>
<td><strong>23.2</strong></td>
</tr>
<tr>
<td>RT-04</td>
<td>didev03</td>
<td>0.6/1.6/1.0/3.16 @ 790</td>
<td>27.9</td>
<td>18.0</td>
</tr>
<tr>
<td></td>
<td>eval03</td>
<td>0.6/0.7/0.9/2.17 @ 706</td>
<td>31.2</td>
<td>15.9</td>
</tr>
<tr>
<td></td>
<td>sttdev04</td>
<td>2.2/0.3/0.9/3.36 @ 786</td>
<td>30.1</td>
<td>21.2</td>
</tr>
<tr>
<td></td>
<td>dev04f2</td>
<td>1.5/1.8/0.6/3.93 @ 632</td>
<td>39.9</td>
<td>26.9</td>
</tr>
<tr>
<td></td>
<td>devall</td>
<td><strong>1.26/1.03/0.85/3.14 @ 2914</strong></td>
<td><strong>29.7</strong></td>
<td><strong>20.3</strong></td>
</tr>
</tbody>
</table>

† sometimes called oracle clustering

- 34% relative drop in SI along with 22% reduction in # of segments on devall.

- DER using RT-03s clusterer improved by 12% relative.
Development Results - Segmenter Tuning

- Best Segmenter DER = 19.3% (X), Best Clusterer DER = 17.1% (Y)
Development Results - Silence Removal

- Silence stripping after phone recogniser stage in segmenter

<table>
<thead>
<tr>
<th>Silence Threshold</th>
<th>Segment Impurity (MS/FA/SPE/SI) @ NumSeg</th>
<th>Segmenter DER (on devall)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5s</td>
<td>3.62/0.39/0.85/4.86 @ 5190</td>
<td>36.1</td>
</tr>
<tr>
<td>1.0s</td>
<td>1.22/1.08/0.85/3.15 @ 3005</td>
<td>32.0</td>
</tr>
<tr>
<td>2.0s</td>
<td>0.77/2.12/0.94/3.83 @ 3045</td>
<td>32.9</td>
</tr>
</tbody>
</table>

*Segment impurity as well as DER lowest for 1s value.*

- Empty segments removal in gender-labelling stage

<table>
<thead>
<tr>
<th>Stage</th>
<th>Segment Impurity (MS/FA/SPE/SI) @ NumSeg</th>
</tr>
</thead>
<tbody>
<tr>
<td>before P1</td>
<td>1.22/1.08/0.85/3.15 @ 3005</td>
</tr>
<tr>
<td>after P1</td>
<td>1.26/1.03/0.85/3.14 @ 2914</td>
</tr>
</tbody>
</table>

*Number of segments dropped by 3% with no to loss segment purity.*
Development Results - Clusterer Initialisation

Aim: sorting the segments before clustering to help initialisation

<table>
<thead>
<tr>
<th>sorting</th>
<th>didev03</th>
<th>eval03</th>
<th>sttdev04</th>
<th>dev04f2</th>
<th>devall</th>
</tr>
</thead>
<tbody>
<tr>
<td>none</td>
<td>18.0</td>
<td>15.9</td>
<td>21.2</td>
<td>26.9</td>
<td>20.3</td>
</tr>
<tr>
<td>time</td>
<td>17.5</td>
<td>16.7</td>
<td>21.5</td>
<td>25.7</td>
<td>20.2</td>
</tr>
<tr>
<td>spkr-start</td>
<td>17.5</td>
<td>17.9</td>
<td>22.6</td>
<td>17.5</td>
<td>19.0</td>
</tr>
<tr>
<td>spkr-mid</td>
<td>14.0</td>
<td>15.2</td>
<td>22.2</td>
<td>23.5</td>
<td><strong>18.7</strong></td>
</tr>
</tbody>
</table>

- Relative improvement of 8% in DER on devall.
- Clusterer highly sensitive to initialisation.
**Development Results - Changing Feature in Clustering**

**Motivation**: DER dropped by 5% absolute on dev04f2 by using PLP without c0

- Clustering uses c0 if and only if mean c0 value is above a threshold.
- c0-switching reduced DER by 1% absolute on devall.
Results on RT-04f Evaluation Data - Progress

<table>
<thead>
<tr>
<th>Coding</th>
<th>Segmentation</th>
<th>Clustering</th>
<th>DER main</th>
<th>DER c0switch</th>
</tr>
</thead>
<tbody>
<tr>
<td>RT-03s</td>
<td>RT-03s</td>
<td>RT-03s</td>
<td>36.33</td>
<td>-</td>
</tr>
<tr>
<td>RT-03s</td>
<td>RT-03s</td>
<td>RT-04f</td>
<td>27.90</td>
<td>24.45</td>
</tr>
<tr>
<td>RT-03s</td>
<td>RT-04f</td>
<td>RT-04f</td>
<td>22.48</td>
<td>22.35</td>
</tr>
<tr>
<td>†RT-04f</td>
<td>RT-04f</td>
<td>RT-04f</td>
<td>23.86</td>
<td>24.12</td>
</tr>
</tbody>
</table>

† Official evaluation submission

- New segmenter and clusterer resulted in 14% absolute drop in primary DER.

- Slight degradation in evaluation submission performance due to compiler switch affecting only *coding*. 
Results on RT-04f Eval Data - Impact of Different Strategies

Different strategies to pick segmenter and clusterer on dev data:

(a) use segmenter with best segmenter DER on devall
(b) use clusterer with best DER on devall
(c) use clusterer with best DER on dev04f2
(e) RT-04 evaluation system

<table>
<thead>
<tr>
<th>Likelihood Threshold</th>
<th>Segmenter DER</th>
<th>RT-03s Clusterer DER</th>
<th>RT-04 Clusterer DER</th>
</tr>
</thead>
<tbody>
<tr>
<td>3000</td>
<td>35.15</td>
<td>22.03</td>
<td>22.48(e)</td>
</tr>
<tr>
<td>11000</td>
<td>18.72(a)</td>
<td>22.90</td>
<td>21.02</td>
</tr>
<tr>
<td>16000</td>
<td>21.17</td>
<td>20.50(c)</td>
<td>22.18</td>
</tr>
<tr>
<td>17000</td>
<td>22.05</td>
<td>22.06(b)</td>
<td>21.44</td>
</tr>
</tbody>
</table>

• NB using segmenter output only, eval04f DER could be reduced to 18.7%.

• using static only coefficients in final segmentation stage reduced this to 18.1%.
Summary and Future Plans

• On RT-04f evaluation data the final system gave DER of 23.9%.

• Modifications since RT-03s resulted in 34% relative improvement in DER.

• DER of 18.1% is possible using segmenter output directly.

• Clustering stage rather sensitive to segmentation.

• Possible future work includes :
  – exploiting speaker labels from segmenter in clusterer.
  – cluster voting of segmenter and clusterer outputs.
  – investigating the use of proxy speaker models (a la MIT).