Background

This poster presents webASR 2, a re-development of the original webASR web service.
  - It provided a web interface where users could freely sign up and submit their audio files for transcription with one of the available systems developed at the University of Sheffield.
  - Upon registration, users could upload files via a Java Applet and retrieve the transcriptions of those files.
- Highly flexible and scalable speech processing back-end that is hosted by the University of Sheffield. This back-end uses the Resource Optimisation Toolkit (ROTK) workflow engine.
- By 2015, several weaknesses of the web service implementation were identified:
  - The use of a servlet and a Java Applet in the front-end was not user-friendly.
  - There was no integration with an API, the web was the main and only interface.

Extending the scope – Version 2

webASR 2 is a complete re-development of the web service, aimed to overcome the weaknesses of the original implementation, while also providing extended functionality for the speech technology systems provided.
- Web service now follows a Representation State Transfer (REST) architecture using the Django web framework.
  - Easier implementation of new functionalities of the web service, fully HTML5 compliant.
- Due to the REST implementation, the user can use the same calls used in the web service from an application.
  - Developers can integrate webASR in their applications with only 4 HTTP calls.

Implemented systems

Transcription

- Multichannel meeting transcription
  - Cross-talk speech segmentation or single channel BIC speech detection
  - 3-pass speech recognition
    - First pass: Speaker independent MPE–trained GMM–HMM system
    - Second pass: VTLN normalised MPE–trained DNN–GMM–HMM system
    - Third pass: Identical to previous pass, but MLLR and CMLLR adapted
- Faster decoding using WFSTs

General media transcription
- DNN-based speech segmentation
- Combination of 3 independent systems:
  - Speaker and background adapted DNN–GMM–HMM system
  - Speaker adapted DNN–HMM system
  - Speaker normalised DNN–HMM system
- N-best rescoring using RNNLms

Lecture transcription
- DNN-based speech segmentation
  - Speaker adapted DNN–HMM system

Segmentation and diarisation

Meeting segmentation
- DNN-based speech segmentation

General media diarisation
- DNN-based speech segmentation with DNN fine-tuning
- Agglomerative speaker clustering with BIC
- Re-clustering using a fine-tuned speaker separation DNN

Lightly supervised alignment

General media alignment
- Lightly supervised decoding using LM adaptation on the General media transcription system
- Alignment of decoding output to input subtitles using dynamic programming
- Removal of insertions using regression techniques

Machine translation

Lecture translation (French)
- Decoding using the Lecture transcription system
- Translation using Moses

Benchmark results

- Transcription benchmarks based on RT’09, IWSLT’12 and MGB’15.

<table>
<thead>
<tr>
<th>System</th>
<th>Benchmark</th>
<th>Substitutions</th>
<th>Deletions</th>
<th>Insertions</th>
<th>WER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multichannel Meeting Transcription</td>
<td>RT’09</td>
<td>18.4%</td>
<td>6.8%</td>
<td>2.3%</td>
<td>12.9%</td>
</tr>
<tr>
<td>Lecture Transcription</td>
<td>IWSLT’12</td>
<td>8.0%</td>
<td>10.7%</td>
<td>10.7%</td>
<td>28.0%</td>
</tr>
<tr>
<td>General Media Transcription</td>
<td>MGB’15</td>
<td>14.1%</td>
<td>10.7%</td>
<td>3.2%</td>
<td>28.0%</td>
</tr>
</tbody>
</table>

- Segmentation benchmarks based on RT’07 and diarisation benchmarks based on MGB’15

<table>
<thead>
<tr>
<th>System</th>
<th>Benchmark</th>
<th>Missed speech</th>
<th>False alarm</th>
<th>Speaker error</th>
<th>WER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meeting Segmentation</td>
<td>RT’07</td>
<td>11.8%</td>
<td>10.7%</td>
<td>-</td>
<td>22.5%</td>
</tr>
<tr>
<td>General Media Diarisation</td>
<td>MGB’15</td>
<td>1.9%</td>
<td>6.4%</td>
<td>41.1%</td>
<td>49.3%</td>
</tr>
</tbody>
</table>

- Lightly supervised alignment benchmark based on MGB’15.

<table>
<thead>
<tr>
<th>System</th>
<th>Benchmark</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Media Alignment</td>
<td>MGB’15</td>
<td>0.8818</td>
<td>0.8689</td>
<td>0.8753</td>
</tr>
<tr>
<td>Lecture Transcription</td>
<td>IWSLT’12</td>
<td>12.5%</td>
<td>31.26</td>
<td></td>
</tr>
</tbody>
</table>

Conclusions

- Improved webASR freely available for the research community and the general public.
- New and improved systems with state-of-the-art results across several benchmarks.
- Easy integration for developers with the RESTful API.
- Demo examples available:
  - Translation of YouTube videos (http://mini-vm20.dcs.shef.ac.uk/youtube/)
  - Translation of TED Talks (http://mini-vm20.dcs.shef.ac.uk/ted/)