webASR 2

Improved cloud based speech technology

Thomas Hain and the Machine Intelligence for Natural Interfaces group

Speech and Hearing Group, Department of Computer Science, University of Sheffield, UK

T.Hain@sheffield.ac.uk

The University Of Sheffield.

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

Background

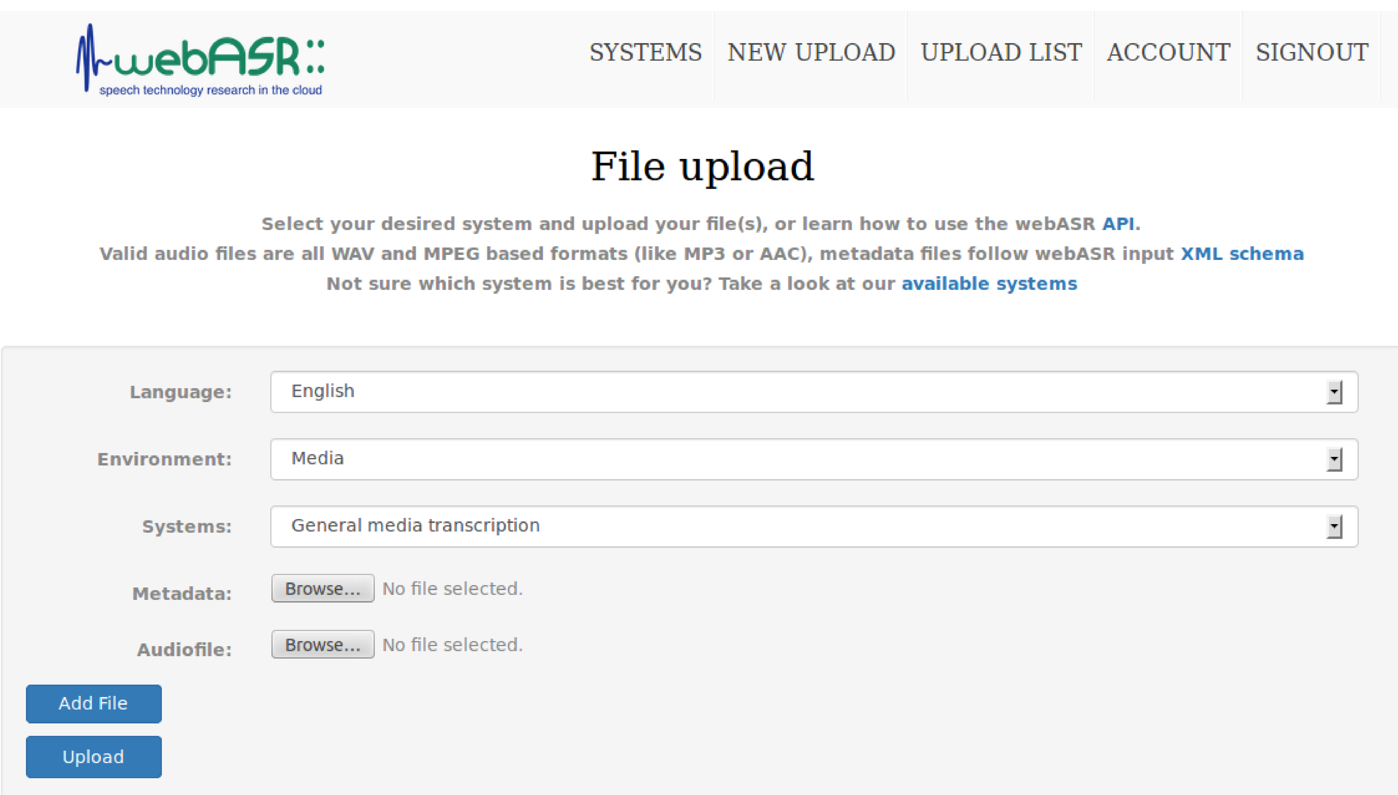
This poster[1] presents webASR 2, a redevelopment of the original webASR web service.

- In 2008, www.webasr.org became the world's first cloud–based speech recognition engine
  - 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

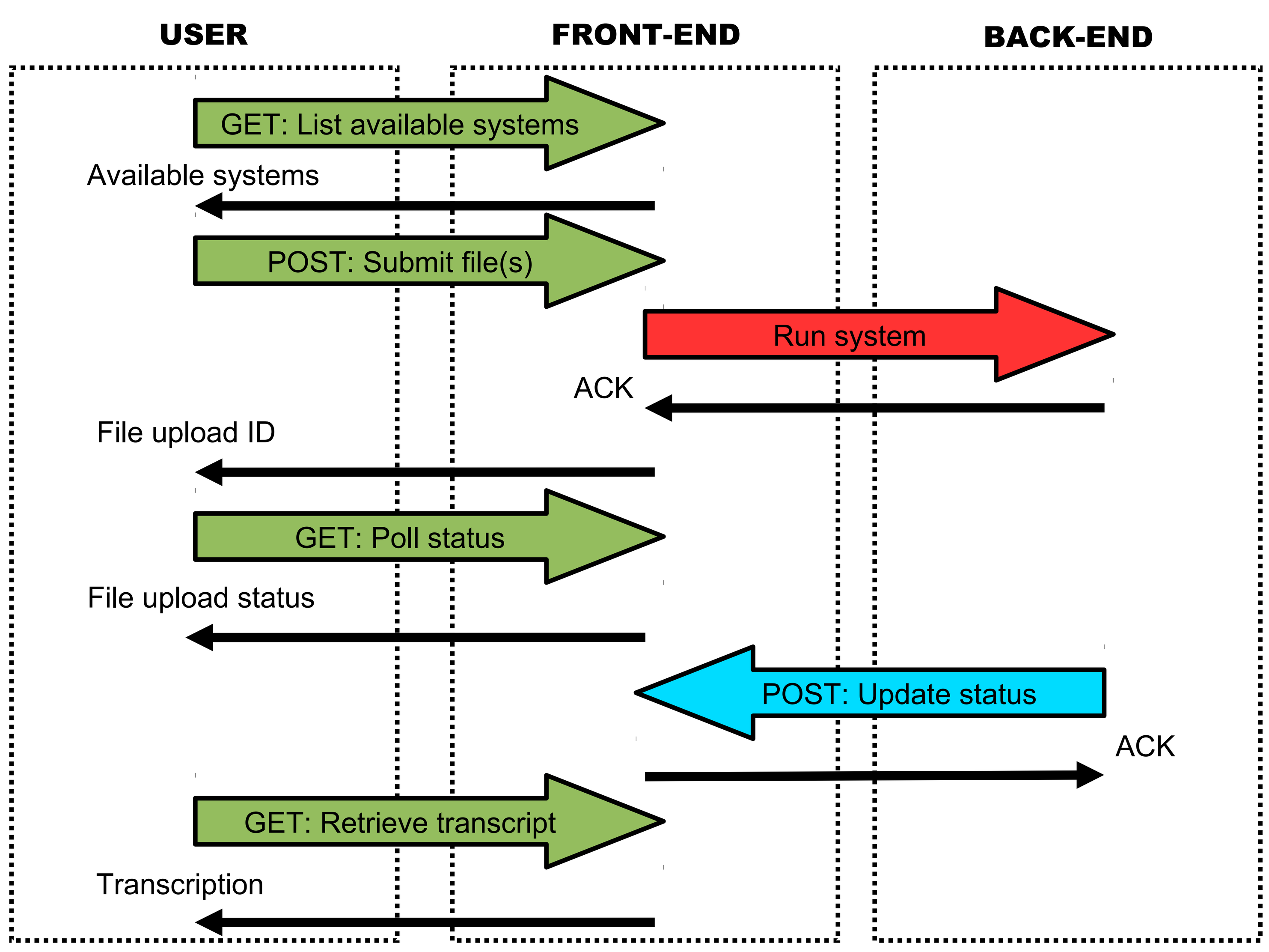
Extending the scope – Version 2

webASR 2 is a complete redevelopment 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



- Users can now submit metadata files (XML schema).
  - Provide a manual segmentation to be used instead of automatic segmentation, and/or
  - Provide a rough transcript/summary to be used for language model adaptation

```xml
<?xml version="1.0" encoding="UTF-8"?>
<body>
<segments>
        &lt;segment start="1.64" end="5.47" speaker="Thomas Hain"/&gt;
        &lt;segment start="6.22" end="10.30" speaker="Thomas Hain"/>
</segments>
<transcript>
        <p>This is webASR,</p>
        <p>the world first cloud-based speech recognition engine</p>
</transcript>
</body>
```

- Expansion to support multiple tasks:
  - Currently transcription, diarisation, alignment, translation

[1] *Supported by the EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology)*

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.

System	Benchmark	Substitutions	Deletions	Insertions	WER
Multichannel Meeting Transcription	RT'09	18.4%	6.8%	3.3%	**28.5%**
Lecture Transcription	IWSLT'12	8.0%	2.3%	2.6%	**12.9%**
General Media Transcription	MGB'15	14.1%	10.7%	3.2%	**28.0%**

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

System	Benchmark	Missed speech	False alarm	Speaker error	SER/DER
Meeting Segmentation	RT'07	11.8%	10.7%	-	**22.5%**
General Media Diarisation	MGB'15	1.9%	6.4%	41.1%	**49.3%**

- Lightly supervised alignment benchmark based on MGB'15.

System	Benchmark	Precision	Recall	F–measure
General Media Alignment	MGB'15	0.8818	0.8689	**0.8753**

- Translation benchmark based on IWSLT'12, English to French (true–cased and no–punctuated)

System	Benchmark	WER(English)	BLEU(French)
Lecture translation	IWSLT'12	12.5%	**31.28**

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:
  - Transcription of YouTube videos (http://mini-vm20.dcs.shef.ac.uk/youtube/)
  - Translation of TED Talks (http://mini-vm20.dcs.shef.ac.uk/ted/)