Speech recognition in SUMMA

Peter Bell, Sibo Tong, Andrea Carmantini, Simon Vandieken,
Ondřej Klejch, Roberts Darģis, Phil Garner, Steve Renals

Speech recognition challenges in SUMMA

- Our systems need to have wide language coverage:

Well resourced	English, Arabic, Spanish, German
Some resources	Portuguese, Russian, Farsi
Poorly resourced	Ukranian, Latvian

- Systems need to be trained on broadcast TV material – models trained on standard corpora of read speech will not perform well on TV speech
- TV audio data may be captioned or scripted, but these are often not usually a verbatim transcription of the words spoken, and time markers may be unreliable

Our systems

- Our standard systems use feedforward deep neural networks (DNNs) trained using cross-entropy followed by sequence training
- Models are able to process live streams of speech in a continous, online manner using the **CloudASR** platform which is optimised for fast decoding, allows rapid scalability, and is compatible with all neural network frameworks contained in the Kaldi toolkit
- Single-language baseline models are generally trained on the GlobalPhone corpus, which comprises small quantities of read speech in many different languages
- Systems adapted to TV data are trained on data from the BBC, Deutsche Welle, and Aljazeera, amongst others, but suitable resources are not available for all langauges

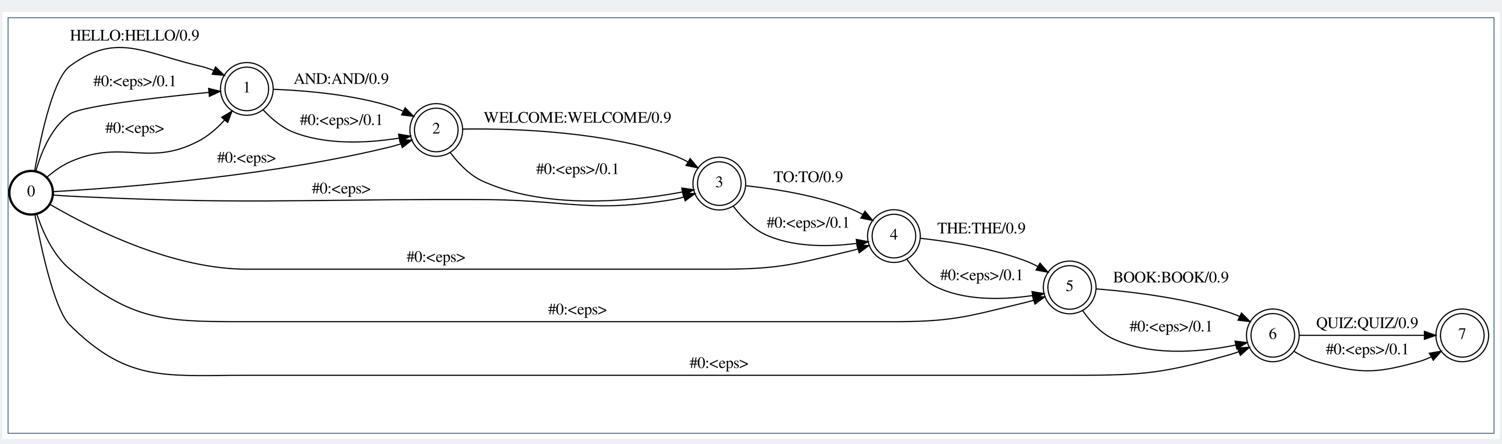
Lightly supervised alignment

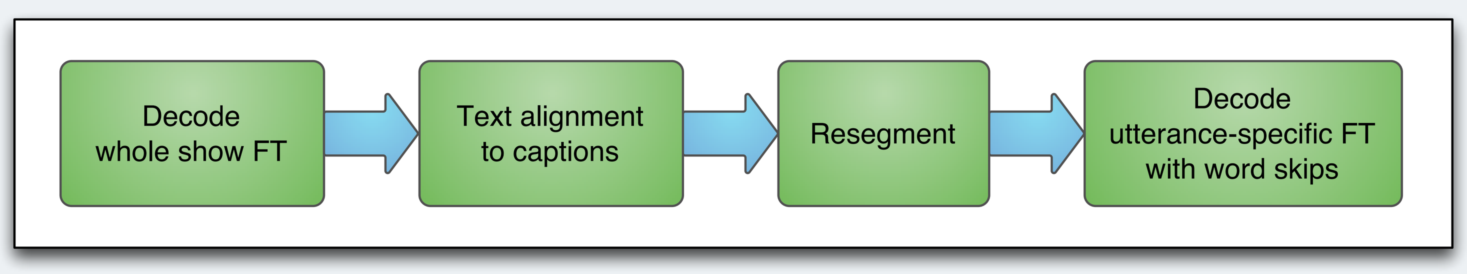
To train models on data with some matching TV captions or script material, we need a method for aligning caption text with the audio data in a way which is robust to mismatches between the speech and the captions...

```
he loves your ******** ** PICTURE he thinks ****** YOU'LL do ********** well in milan
```

```
he loves your PICTURES SO MUCH    he thinks YOU'RE GONNA  do INCREDIBLY well in milan
```

- We apply a **two-pass factor transducer approach**.
- In the first pass, a single grammar transducer, G, is generated for each show.
- In the second pass, WFSTs are generated dynamically per utterance by selecting surrounding text, and word skips are allowed giving robustness to deletions
- Important to set appropriate penalty for word skips to avoid excessive word removal

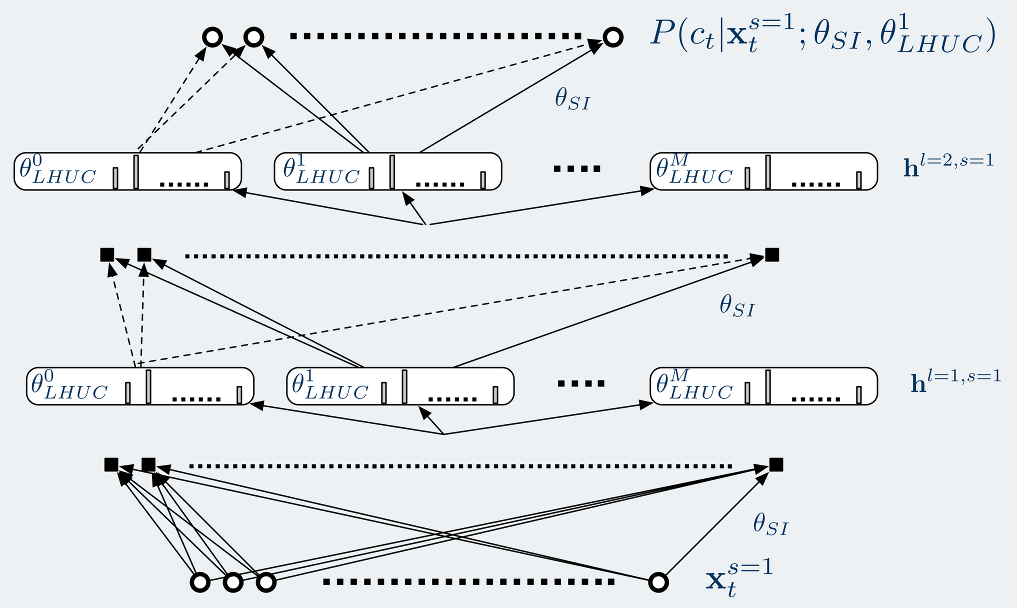




Adapting to low-resource languages

Many techniques investigated:

- Multi-task adaptive networks with multi-lingual bottneck features
- Learning hidden unit contributions (LHUC) and cluster adaptive training (CAT) to create models shared across all languages
- Dirichlet output distributions
- End to end systems with connectionist temporal classification
- Backpropagation with early stopping



LHUC adaptation

Example: adapting German models to TV domain

- 9.6 hours of TV data supplied by Deutsche Welle with accompanying text material → 8.1 hours of audio remain after automatically selecting only the segments containing speech
- A large proportion of the text material does not directly relate to the speech, eg. metadata, scene descriptions, speaker labels, translations into other languages
- Out of 124k words, 54k are successfully aligned to the audio using the lightly supervised method → represents a rate of 100 successfully aligned words per minute of speech – a reasonable averaging speaking rate
- We ran a similar exercise on a 15,000 short clips of web-scraped audio from Euronews, harvesting 380 hours of speech data.
- Due to the small quantity of data, DW models must be adapted from the our initial baseline model, trained on mismatched data from TUM.

System	Word Error Rate (%)
155 hr baseline TUM model	39.6
8hr DW model	39.8
TUM model, adapted to DW	33.9
+ sMBR training, 4gram rescore	29.2
380 hr Euronews model	31.6
+ LF-MMI training, 4gram rescore	28.7

Current results on SUMMA test sets

Language	Word Error Rate† (%)
Arabic (MGB Challenge)	14.7
English (MGB Challenge)	26.1
German	28.7
Latvian	32.7
Russian	34.5
Ukranian	44.2
Farsi	53.8
Portuguese	68.4

†Results may differ when systems are used in an online mode within the SUMMA platform

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 688139 (SUMMA).