Automatic speech recognition for media monitoring in SUMMA

Peter Bell, Sibo Tong, Ondřej Klejch, 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 – this makes it hard to train acoustic models

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

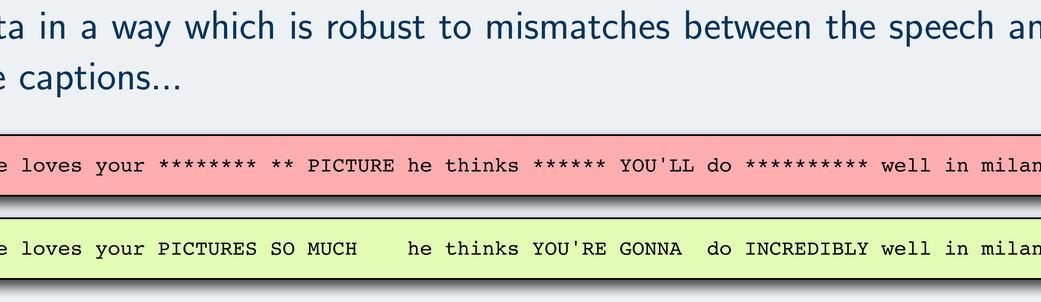
Lightly supervised alignment

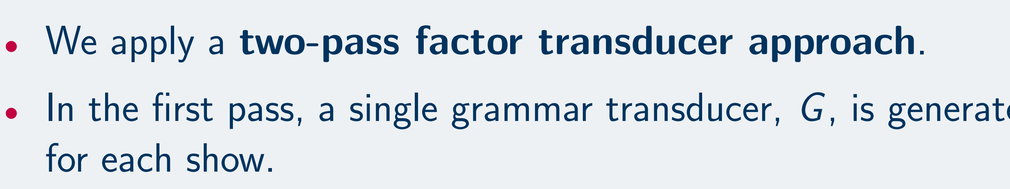
To train models on data with some matching TV captions or script material, we need a method for aligning caption text with the audio data in a way which is robust to mismatches between the speech and the captions...

```
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

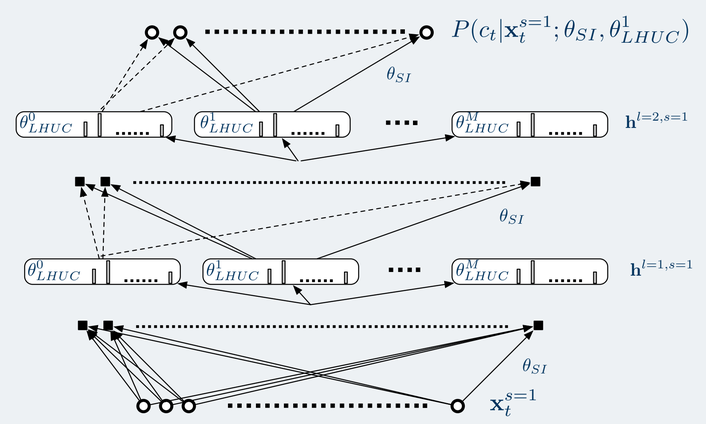




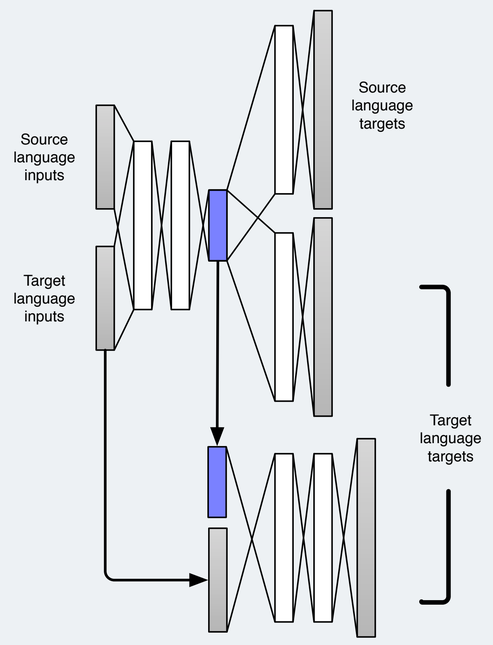
Adapting to low-resource languages

We are investigating a range of techniques to train robust DNN models on small amounts of TV data per language:

- Multi-task adaptive networks
- 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
- Multi-lingual training using shared IPA symbols



LHUC adaptation



Multitask adaptive networks

Text normalisation

- Text normalisation presents an issue for the SUMMA stream processing pipeline comprising speech recognition, punctuation insertion and machine translation.
- In speech recognition, phrases are normally output exactly as they are spoken, for example: "*one hundred dollars*"
- This creates a mismatch with machine translation systems trained purely on text, where the same token would appear as "$100" $\rightarrow$ may lead to poor quality translations
- Currently looking at methods to map between the verbal and written forms of these non-lexical tokens (which are common in TV broadcasts) that are not language dependent

Current results on SUMMA test sets

Language	Word Error Rate† (%)
English (MGB Challenge)	26.1
Arabic (MGB Challenge)	14.7
German	34.6
Russian	40.0

Work is ongoing to transcribe test sets for the remaining languages

†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).