ASR and alignment systems for multi-genre media data

Peter Bell

Automatic Speech Recognition— ASR Lecture 14
13 March 2017

This lecture

- The MGB Challenge
- Building ASR systems from captioned TV broadcasts
- Lightly supervised alignment
- Speech activity detection

What are we working on in CSTR?

Topics Wide domain coverage, understanding diverse data, cross-lingual recognition, environment and speaker modelling

Methods Deep learning, canonical models, adaptation, factorisation, generalisation

Applications Talks and lectures, TV broadcasts, multiparty meetings, spoken dialogue systems

What are we working on in CSTR?

Topics Wide domain coverage, understanding diverse data, cross-lingual recognition, environment and speaker modelling

Methods Deep learning, canonical models, adaptation, factorisation, generalisation

Applications Talks and lectures, **TV broadcasts**, multiparty meetings, spoken dialogue systems

Case study: multi-genre TV broadcasts

Automatic speech processing of TV broadcasts has an obvious commercial need, but is still very difficult for current systems

The MGB Challenge

- We proposed an open challenge to work on English **M**ulti-**G**enre **B**roadcast data at the 2015 ASRU workshop
- Our aim was to encourage researchers from around to world to work on this kind of data
- Create a standard experimental setup so that cutting edge research methods can be compared in a controlled setting
- Repeated on Arabic TV data for the 2016 SLT workshop

Data supplied to all participants

- 1,600 hours of TV, taken from 7 complete weeks of BBC output over four channels, with accompanying subtitle text
- 600M words of subtitle text from 1988 onwards
- XML metadata for all shows, generated in a standard format
- Data supplied freely for the purpose of participation in the challenge

Why is this task difficult

- Many different background noise conditions
- Diverse range of accents and speaking styles – including fast dramatic speech, and natural, spontaneous speech
- Speaker identities are usually not known
- Although lots of training data is available, the captions available are not very accurate.

Example...

Two contrasting programmes...

The tasks

- **Transcription of multi-genre TV shows**
  - we supplied around 16 TV shows to be completely transcribed
  - show names and genre labels are provided
  - some shows are from series appearing in the training data; some are not
- **Subtitle alignment**
  - for the same shows as Task 1, the subtitle text as originally broadcast were provided
  - these differ from the verbatim audio content for a range of reasons
  - participants must produce time stamps for all words in the subtitles

The tasks

- **Longitudinal transcription**
  - aim to evaluate ASR in a realistic longitudinal setting
  - participants transcribed complete TV series, where the output from shows broadcast earlier could be used to adapt and enhance the performance of later shows
- **Longitudinal diarization and speaker linking**
  - aim to label speakers uniquely across a complete series
  - realistic longitudinal setting again: participants must process shows sequentially in date order

Our ASR system

Some features of our best system:

- Models trained on 640 hours of broadcasts

Our ASR system

Some features of our best system:

- Models trained on 640 hours of broadcasts
- DNNs with 6 hidden layers, an input window of 9 frames and 28k output states used in combination with CNNs with a similar structure

Our ASR system

Some features of our best system:

- Models trained on 640 hours of broadcasts
- DNNs with 6 hidden layers, an input window of 9 frames and 28k output states used in combination with CNNs with a similar structure
- Networks trained with cross-entropy criterion, followed by minimum Bayes risk full-sequence training

Our ASR system

Some features of our best system:

- Models trained on 640 hours of broadcasts
- DNNs with 6 hidden layers, an input window of 9 frames and 28k output states used in combination with CNNs with a similar structure
- Networks trained with cross-entropy criterion, followed by minimum Bayes risk full-sequence training
- Training using a complex recipe of multiple iterations, with all training data re-aligned several times – the complete procedure takes several weeks, even on GPU machines!

Our ASR system

Some features of our best system:

- Models trained on 640 hours of broadcasts
- DNNs with 6 hidden layers, an input window of 9 frames and 28k output states used in combination with CNNs with a similar structure
- Networks trained with cross-entropy criterion, followed by minimum Bayes risk full-sequence training
- Training using a complex recipe of multiple iterations, with all training data re-aligned several times – the complete procedure takes several weeks, even on GPU machines!
- No speaker adaptation, but mean and variance normalisation used, based on speaker clusters

Some results on development data

System	3gram	4gram
210 hours training data		
GMM	53.1	-
DNN	40.9	37.4
+ sequence training	37.1	33.7
640 hours training data		
Final DNN	31.3	28.2
Final CNN	30.8	28.0
ROVER	30.1	**27.3**

Using broadcast captions for training

Problems with using closed captions as training data labels:

- Timings may not be accurate
- Not all words spoken are captioned
- Words may appear in the captions that were never actually spoken
- Limited speaker information is available (in the form of colour changes in the subtitles)

```
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
```

Training acoustic models on TV data

The basic recipe:

1. Using the captions and a previous ASR system, identify words and their timings within the audio
2. Select a set of utterances to use in training
3. Generate a pronunciation for every word from a base dictionary, and use this to create a phone alignment for each utterance
4. Train GMM and then DNN models using these phone alignments, frequently re-aligning the data

Lightly supervised training

- The problem of identifying words from the captions and using them to update the models is an example of *lightly supervised training*
- We don't have perfect labels for each training sample, but we do know something about them
- The main challenge is in identifying reliable labels and learning from them, without also learning from unreliable labels, or past mistakes

Lightly supervised training

A standard method [Braunschweiler et al]:

1. Train an *biased* language model on the captions, interpolated with a background LM

$$p(w_t|h_t) = \lambda p_{bias}(w_t|h_t) + (1-\lambda)p_{bg}(w_t|h_t)$$

Lightly supervised training

A standard method [Braunschweiler et al]:

1. Train an *biased* language model on the captions, interpolated with a background LM

$$p(w_t|h_t) = \lambda p_{bias}(w_t|h_t) + (1-\lambda)p_{bg}(w_t|h_t)$$

2. Decode the training data with a pre-existing acoustic model, and the biased LM

Lightly supervised training

A standard method [Braunschweiler et al]:

1. Train an *biased* language model on the captions, interpolated with a background LM

$$p(w_t|h_t) = \lambda p_{bias}(w_t|h_t) + (1-\lambda)p_{bg}(w_t|h_t)$$

2. Decode the training data with a pre-existing acoustic model, and the biased LM
3. Align the captions with the ASR output

Lightly supervised training

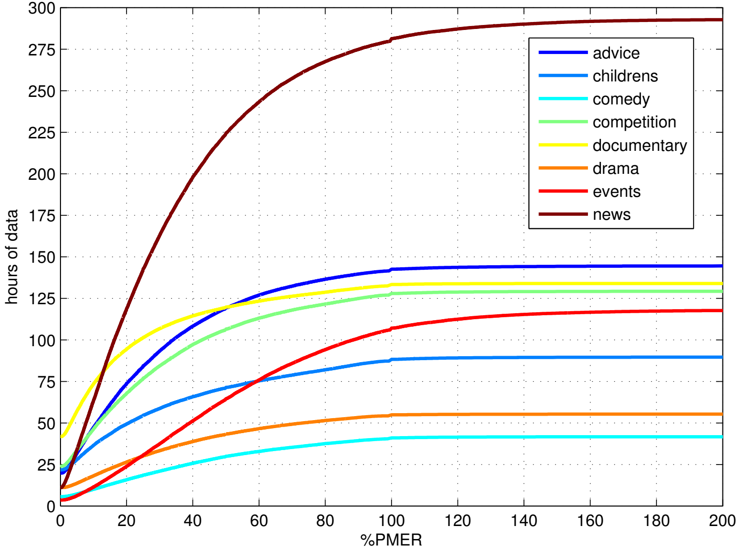
A standard method [Braunschweiler et al]:

1. Train an *biased* language model on the captions, interpolated with a background LM

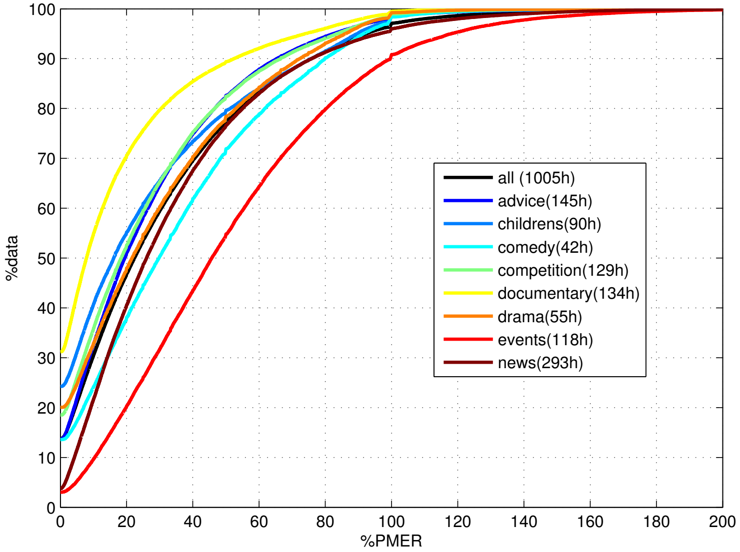
$$p(w_t|h_t) = \lambda p_{bias}(w_t|h_t) + (1-\lambda) p_{bg}(w_t|h_t)$$

2. Decode the training data with a pre-existing acoustic model, and the biased LM
3. Align the captions with the ASR output
4. Select utterances where there is a good match between the captions and the automatic output

Data selection by genre



Data selection



An alternative alignment method

- The biased LM approach is quite computationally costly, and can lead to bias towards data that we can already recognise well

An alternative alignment method

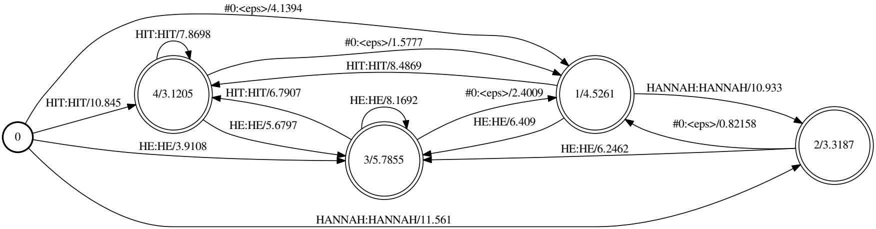
- The biased LM approach is quite computationally costly, and can lead to bias towards data that we can already recognise well
- We have used an alternative approach based on constructing weighted finite state transducers for each utterance

An alternative alignment method

- The biased LM approach is quite computationally costly, and can lead to bias towards data that we can already recognise well
- We have used an alternative approach based on constructing weighted finite state transducers for each utterance
- This allows us to use much stronger constraints – based on the captions – at decoding time

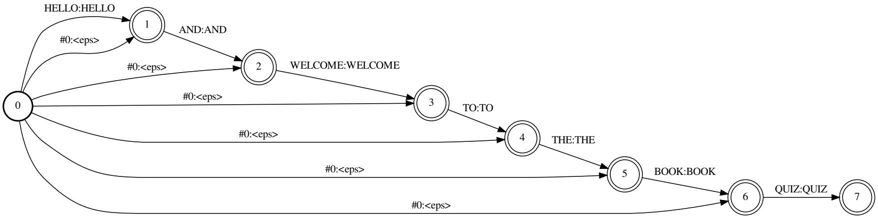
Recap: ASR with WFTs

- Most modern decoders use a transducer approach to combine the acoustic model, lexicon and language model in a unified framework
- Find the lowest-cost path through a composed transducer $H \circ C \circ L \circ G$



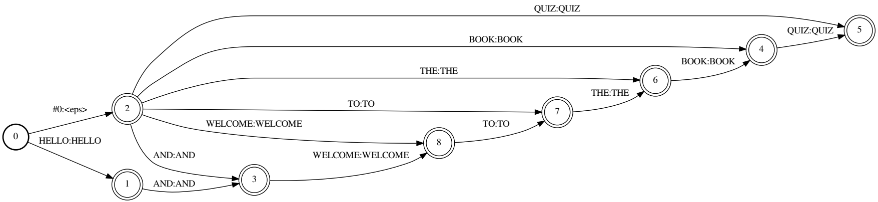
Alignment with WFSTs

A *G* transducer that allows any substring of the original captions – known as a *factor transducer*



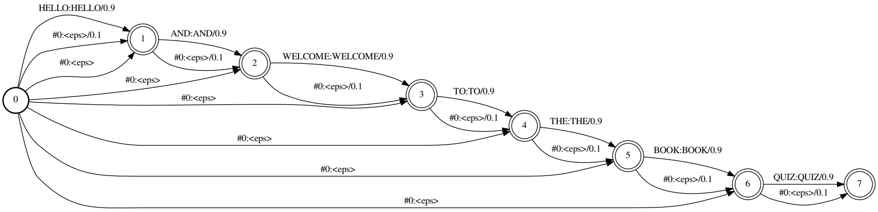
Alignment with WFSTs

A determinized version of the G transducer



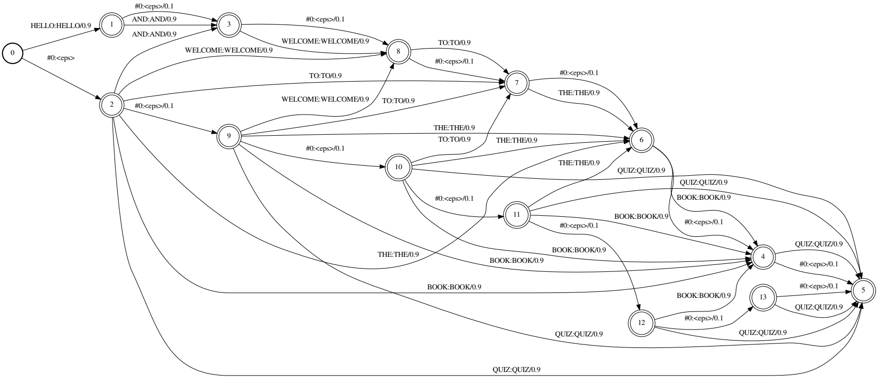
Alignment with WFSTs

What about when word appears in the captions that was not actually spoken? We need to alter the design to be robust to this by allowing deletions (at a cost)



Alignment with WFSTs

A determinized version



The complete alignment process

1. Decode with a factor-transducer for the each programme
2. Align the output to the original captions
3. Re-segment the data, to potentially include missed speech
4. Decode again with utterance-specific factor transducers, allowing word-skips

Another example

Spot how the automatically-aligned captions differ from the words actually spoken...

Scoring alignment for MGB

- Scoring with respect to a forced-alignment of human-generated verbatim transcription
- Words spoken but not in the captions are ignored
- For words in both, systems judged correct if supplied timings are correct within a 100ms window
- Evaluated in terms of f-score

$$P = \frac{N_{match}}{N_{hyp}}, R = \frac{N_{match}}{N_{ref}}, F = 2 \times \frac{P \times R}{P + R}$$

- Segments with overlapped speech are ignored

Our alignment results

System	Precision	Recall	F-score
Preliminary DNN AMs			
Pass 1 FT	0.8816	0.7629	0.8180
+ force align	0.8290	0.7855	0.8066
Pass 2 FT+skip	0.8679	0.8563	0.8620
Final DNN AMs			
Pass 1	0.9009	0.8128	0.8546
Pass 2 FT+skip	*0.8856*	*0.9013*	*0.8934*

Evaluation results

Participant	F-score
Cambridge	0.900
Edinburgh/Quorate	*0.877*
CRIM	0.863
Vocapia/LIMSI	0.846
Sheffield	0.834
NHK	0.797

Speech activity detection

- SAD is the task of deciding which portions of the audio contain speech

Speech activity detection

- SAD is the task of deciding which portions of the audio contain speech
- Aims to segment to audio into “reasonable length” utterances

Speech activity detection

- SAD is the task of deciding which portions of the audio contain speech
- Aims to segment to audio into "reasonable length" utterances
- It's surprisingly difficult! We need good models for non-speech as well as speech

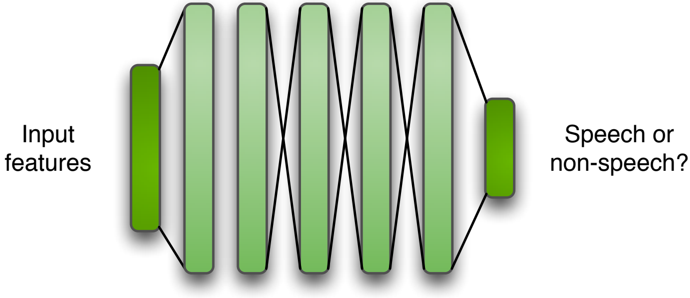
Speech activity detection

- SAD is the task of deciding which portions of the audio contain speech
- Aims to segment to audio into “reasonable length” utterances
- It’s surprisingly difficult! We need good models for non-speech as well as speech
- Training non-speech models on the TV data is effectively unsupervised learning, as we can’t be sure that uncaptioned portions of audio don’t actually contain speech

Speech activity detection

- SAD is the task of deciding which portions of the audio contain speech
- Aims to segment to audio into "reasonable length" utterances
- It's surprisingly difficult! We need good models for non-speech as well as speech
- Training non-speech models on the TV data is effectively unsupervised learning, as we can't be sure that uncaptioned portions of audio don't actually contain speech
- One solution is to train non-speech models only on the short pauses between known words

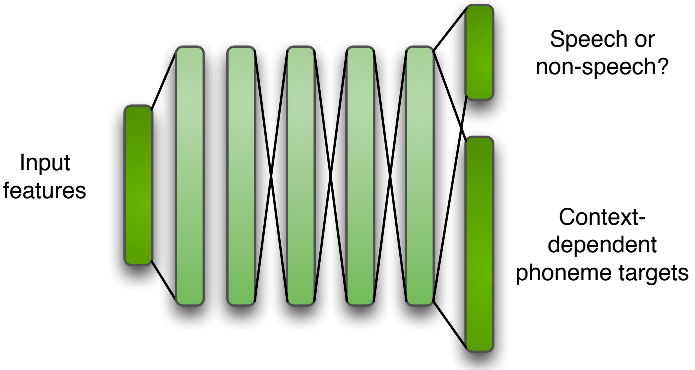
SAD with DNNs



Problems with this DNN

- Sensitive to the (frequent) errors in the frame labels
- It's hard to learn a good hidden representation of speech when modelling only two output classes
- We're not making use of the knowledge we have from the lightly supervised alignment of the captions

Alternative multi-task architecture



Reading

- P. Bell and S. Renals "A system for automatic alignment of broadcast media captions using weighted finite-state transducers," in *Proc. ASRU*, 2015.
- P.C. Woodland et al. "Cambridge University transcription systems for the Multi-Genre Broadcast Challenge," in *Proc. ASRU*, 2015.
- P. Moreno and C. Alberti, "A factor automaton approach for the forced alignment of long speech recordings," in *Proc. ICASSP*, 2009.
- N. Braunschweiler, M. Gales, and S. Buchholz, "Lightly supervised recognition for automatic alignment of large coherent speech recordings," in *Proc. Interspeech*, 2010.
- P. Bell et al. "The MGB Challenge: evaluating multi-genre broadcast media recognition" in *Proc. ASRU*, 2015.