Asynchronous Factorisation of Speaker and Background in Speech Recognition

Oscar Saz

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

The University Of Sheffield.

NST User Group Meeting

May 23rd, 2013

Overview

- Introduction: NST and canonical models
- Adaptation in speech recognition
  - Factorisation
  - Asynchronous factorisation
- Experiments
  - Simulated data: WSJCAM0
  - Diverse data: BBC
- Conclusions

NST and canonical models

Natural Speech Technology

- Canonical models aim to achieve the goals stated in the "Learning and Adaptation" section of the NST proposal
  - *"Models (...) can compactly represent and adapt to new scenarios (...) and seamlessly adapt to new situations and contexts almost instantaneously"*

- This requires automatic speech recognition to
  - Explicitly or implicitly track the background
  - Have a correct representation of speech for each background
  - Perform this preferably in an on-line unsupervised fashion

NST and canonical models

Natural Speech Technology

- Adaptation and normalisation techniques are good at dealing with synchronous changes of speaker and scenario
  - But in real life the context often changes asynchronously with speech
- This is, for example, prevalent in media data
  - Crowds shout and cheer as a sport broadcaster comments the game
  - Laughs are inserted in sit-coms while a character delivers the punchline
  - A programme tune runs in the background while the anchor speaks

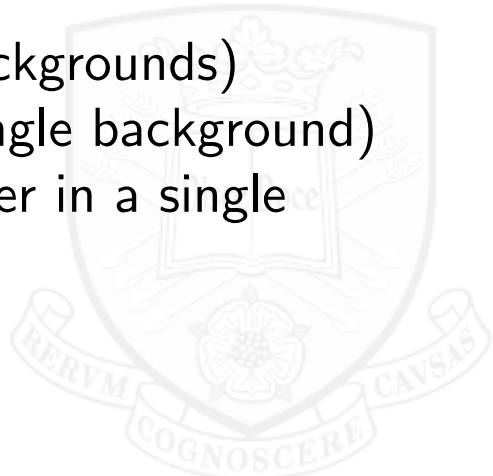
Adaptation: Feature transforms

Natural Speech Technology

- Constrained MLLR transforms reduce the mismatch between the speech and the acoustic models
  - x: Input speech feature
  - $W = \{A, b\}$: Transformation matrix and bias vector
  - y: Transformed speech feature, matched to the model

$$y = Ax + b$$

- They can model all kinds of influence factors:
  - Speaker transforms (single speaker in multiple backgrounds)
  - Background transforms (multiple speakers in a single background)
  - Speaker and background transforms (single speaker in a single background)



Factorisation: Feature transforms

Natural Speech Technology

- Factorisation is used in many tasks to separate factors that influence speech, for instance, background and speaker
- Can be applied with feature transforms [Seltzer and Acero, 2011]
  - Train background transforms $W_{bck} = \{A_{bck}, b_{bck}\}$
  - Train speaker transforms after applying background transforms $W_{spk} = \{A_{spk}, b_{spk}\}$
- And then use them for each pair of background and speaker

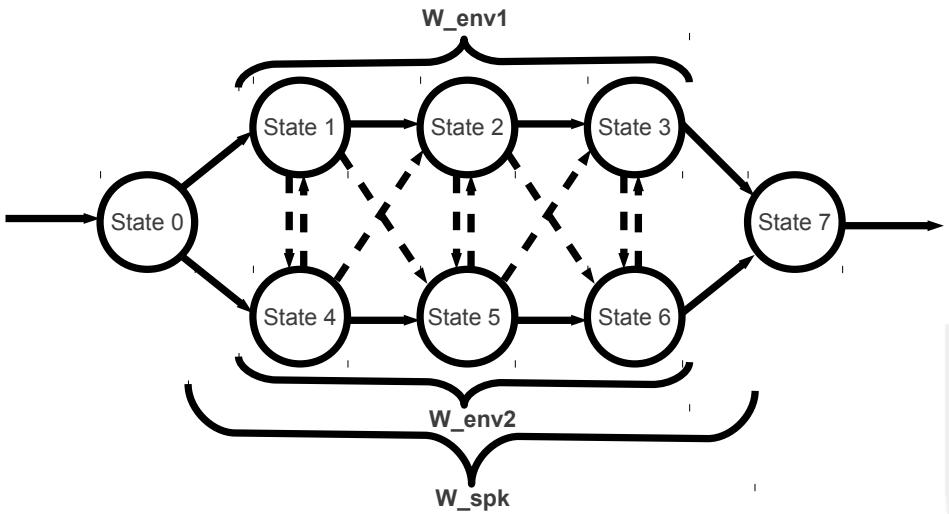
$$y = A_{spk}(A_{bck}x + b_{bck}) + b_{spk}$$

Asynchronous factorisation: feature transforms

Natural Speech Technology

- We propose sets of transforms which can be switched on-line during ASR to reflect changes in the background

$$y(t) = A_{spk}(A_{bck}(t)x(t) + b_{bck}(t)) + b_{spk}$$



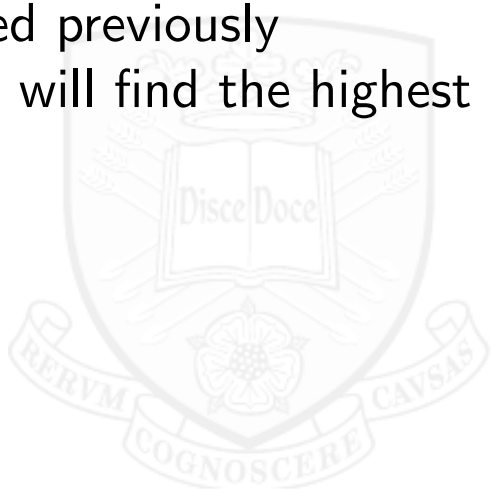
- The transform changes can be done
  - Following a priori information
  - Automatically, following the ML criterion

Asynchronous factorisation: feature transforms

Natural Speech Technology

- Training process
  - Generate synchronous CMLLR transforms based on seed backgrounds
  - Align data asynchronously to these transforms, retrain them
  - Apply background asynchronous transforms to the data and train factorised speaker transforms

- Decoding process
  - Provide speaker and background transforms trained previously
  - Decode with ML criterion, asynchronous topology will find the highest likelihood path



Experiments on WSJCAM0

- WSJCAM0: British English read speech
- We created an artificial dataset adding bursts of music to the background
  - 50% of frames contaminated by music
- Baseline with speaker indepedent HMM-GMM models and standard WSJ language models

Train	Test	si_dt5a	si_dt5b	si_dt20a	si_dt20b	Global WER
Clean	Clean	6.3%	6.4%	13.8%	14.0%	**10.1%**
Clean	Music	19.4%	22.6%	33.4%	30.1%	**26.4%**
Music	Music	10.9%	11.8%	22.4%	21.1%	**16.6%**

WSJCAM0: Asynchronous factorisation

Natural Speech Technology

- Applied adaptation, factorisation and asynchronous factorisation with CMLLR transforms
  - Two background transforms: One initialised for silent background and another initialised for music background

Adaptation	Global WER
None	**26.4%**
CMLLR	**18.5%**
Factorised CMLLR	**18.8%**
Asynchronous factorised CMLLR	**17.7%**

WSJCAM0: Music and applause

Natural Speech Technology

- We replaced the music in half our data and added bursts of applause instead
  - Three background transforms: Add a new transform initialised for applause background

Adaptation	Global WER
None	**29.0%**
CMLLR	**22.5%**
Factorised CMLLR	**22.8%**
Asynchronous factorised CMLLR	**20.9%**

WSJCAM0: Background classification

Natural Speech Technology

- How well is the asynchronous topology tracking the background?
  - 2-class scenario

	Speech	Speech & Music
Speech	**83.6%**	16.4%
Speech & Music	18.5%	**80.5%**

  - 3-class scenario

	Speech	Speech & Music	Speech & Applause
Speech	**62.4%**	17.5%	20.1%
Speech & Music	14.2%	**71.7%**	14.1%
Speech & Applause	9.1%	16.4%	**74.5%**

Experiments on BBC data

Natural Speech Technology

- We used the following data:
  - 1 whole day of Radio4 programming
  - 14 episodes of a BBC drama series

	Training		Testing	
	Programmes	Speech time	Programmes	Speech time
Radio4	30	15.80 hours	6	1.56 hours
Drama	12	4.98 hours	2	0.80 hours
Total	42	20.78 hours	8	2.36 hours

- Besides programme name and description, no a priori information on existing backgrounds

Experiments on BBC data

- Baseline system:
  - Speaker independent HMM-GMM model retrained on BBC training data
  - Language model trained on several sources and interpolated with BBC training data
  - Dictionary with pronunciation probabilities
- It performs well in Radio4, falls short in the drama data

Adaptation	Drama	Radio4	Global
None	61.0%	20.6%	**34.5%**

BBC data: Asynchronous factorisation

Natural Speech Technology

- Applied adaptation, factorisation and asynchronous factorisation
- Pre-defined 3 backgrounds for initialisation
  - Clean background
  - Studio/Radio4 background
  - Effects-noise/drama background

Adaptation	Drama	Radio4	Global
None	61.0%	20.6%	**34.5%**
CMLLR	59.6%	19.1%	**33.1%**
Factorised CMLLR	59.2%	18.8%	**32.8%**
Asynchronous CMLLR	58.8%	18.8%	**32.6%**

BBC data: Background classification

Natural Speech Technology

- No possible objective evaluation of background tracking
  - We lack ground truth of acoustic events so far
- We can see the use of transforms by programmes depending on genre

Programme	Xform 1	Xform 2	Xform 3
Radio4 - Studio	**38%**	**34%**	28%
Radio4 - Location	27%	26%	**47%**
Drama	21%	23%	**56%**

Conclusions

- Our technique can provide improvement in the presence of rapidly changing backgrounds
  - No ground truth is required to either train or apply the CMLLR transforms
- Improvement is also provided in diverse media data
  - The variety of backgrounds is more challenging
- The use of metadata will be relevant in this context
  - Basic a priori knowledge of a programme content can help us make an educated guess when initialising transforms
  - More complex information must surely help, for instance, indication of acoustic events in subtitling files

Demo

Natural Speech Technology

- Come see the demo this afternoon

BBC Radio4 transcription demo with background factorisation

This demo presents the output of the automatic transcription of BBC Radio 4 content
The Average Word Error Rate for this task is around 19%
The automatic speech recognition system also factorises the background according to three possible backgrounds

Green: High quality studio background

Blue: Studio background with minor disturbances

Red: Location background presenting noise or reverberation

The background factorisation is on-line and unsupervised and it is performed frame-by-frame.
For clarity, the demo presents each sentence with the colour corresponding to the most selected background in the sentence

HUNDREDS OF MY CONSTITUENTS HAVE BEEN DEPRIVED AND ARE BEING DEPRIVED FORWARD OPPONENTS TO WHICH THEY SHOULD BE ENTITLED PARTICULARLY IN UPLAND AREAS BECAUSE OF THE WAY IN WHICH THE TEMPERATURE IS BEING MADE FOR EXAMPLE THE TEMPERATURE IN A MORE