Natural Speech Technology

Steve Renals
University of Edinburgh
Natural Speech Technology

5-year UK programme in core speech technology research 2011–2016

Focus on
Speech Recognition
Speech Synthesis
Learning & Adaptation
Motivations

• Weakly-factored models
  • Factor the underlying causes of observed variability in speech
• Domain fragility
  • Rapid transfer to new domains, with minimal supervision
• Synthesis and recognition developed independently
• Lack of reaction to the environment or context
  • Respond and adapt to changes in the acoustic or linguistic environment
• Relatively little speech knowledge incorporated
• Cannot rely on gold standard transcription
  • Work somewhere on the supervised-unsupervised spectrum
Natural Speech Technology
Natural Speech Technology

Exemplar applications

Deep generative models

Domain transfer

Multi-task learning

Applications

Technology

Theory

Adaptation and canonical models

Adaptation and canonical models

Distant speech recognition

Multi-genre transcription MGB Challenge
Adaptation

multiple speakers
acoustic environment
different channels
Adapting NN acoustic models

- Neural network adaptation is challenging
  - models with large numbers of parameters potentially need a lot of adaptation data
  - relatively little structure in the weights
  - unsupervised adaptation is preferable to supervised
  - compact adaptation is preferable
  - joint optimisation of core acoustic model parameters and adaptation parameters

- Baseline
  - feature-space MLLR using a CD-GMM-HMM system to adapt input features for NN acoustic model
i-vectors:
- low-dimension speaker representation
- can be estimated from small amounts of data
- for ASR: Karafiat et al (ASRU-2011); Saon et al (ASRU-2013)

Append (and optimise) additional speaker-based features to the input.
Factorised i-vectors

- Extract two sets of i-vectors
  - speaker information
  - acoustic environment information
- Estimate the i-vectors as weights for a cluster adaptive training GMM system
- Orthogonal factor representations allow adaptation to account for wide range of speaker/environment conditions
- On WSJ with added noise, factorised i-vectors result in 5–10% relative reduction in WER

Karanasou et al, Interspeech-2014
i-Vector priors

- With limited data (1 utterance) use a prior to improve the robustness of the i-vector estimate
- Default: Gaussian prior – sensitive to amount of data/speaker, and to mismatches between training and test duration
- Count-smoothing prior interpolates between prior and observed statistics (cf MAP)
  - speaker-independent prior statistics (estimate over all speakers)
  - gender-dependent prior statistics (two clusters)
- YouTube data, WER improves ~1–3% relative without prior, 3–5% relative with SI prior

Karanasou et al, Interspeech-2015
Unsupervised domain discovery

Discovery of hidden acoustic domains using LDA

Experiments on highly diverse data
- radio
- television
- conversational telephone speech
- meetings
- read speech
- lectures
LDA-DNN

8% relative reduction in WER on MGB Challenge, compared with speaker adapted DNN

Doulaty et al, ASRU-2015
Model-based adaptation

- **Speaker codes** (Bridle & Cox 1990; Abdel-Hamid & Jiang 2013) – model-based adaptation using auxiliary features
- Adaptation of different **weight subsets** (Liao, ICASSP-2013)
  - 5% relative decrease in WER when all 60M weights adapted
- Automatically adapt specific parameter subsets – **output biases** (Yao et al, SLT-2012), slope and bias of hidden units (Siniscalchi et al, TASLP-2013)
- Adaptation cost based on **KL divergence** between SI and SA output distributions (Yu et al, ICASSP-2013)
  - 3% relative decrease in WER on Switchboard
- Increase compactness by **SVD factorisation** of weight matrix (Xue et al, ICASSP-2014)
Learning Hidden Unit Contributions

Key idea: add a learnable speaker-dependent amplitude to each hidden unit

\[ h^l_m = a(r^l_m) \circ \sigma^l(W^{l\top}h^{l-1}_m) \]

Speaker dependent parameter \( r \)

SI Model: set amplitudes to 1
SD Model: learn amplitudes from data, per speaker
LHUC – Adaptation data

 amount of adaptation data [seconds]

 TED
 IWSLT, tst2010

 WER (%)
LHUC – Improvement per speaker

Combined results from TED, AMI, Switchboard
Multi-basis adaptive NN

C Wu & Gales, Interspeech-2015

2–4% WER relative reduction (YouTube)
Adaptation by speaker selection for dysarthric speech

- Dysarthric speech is highly talker dependent
- UA-Speech: SD 45% WER, SI+MAP 49% WER
- Select SI speaker pool based on WER
- Pooled SI model + MAP – 40% WER

Christensen et al, SLT-2014
Multiple average voice model

Personalised speech synthesis for people with speech disorders

Adaptation by interpolating into a speaker eigenspace spanned by mean vectors of speaker-adapted AVMs

Improvements in intelligibility and naturalness over ‘tailored’ synthetic voice

Lanchantin et al, Interspeech-2014
Adaptation in DNN speech synthesis

- 259D inputs
  - 60 melcep + Δ + Δ Δ
  - 25 BAP + Δ + Δ Δ
  - F0 + Δ + Δ Δ
  - Voicing
- 6 tanh hidden layers (1536 units), linear output layer
- SD normalisation of vocoder parameters

Z Wu et al, Interspeech-2015
Naturalness evaluation

MUSHRA test, 30 listeners
Similarity evaluation

MUSHRA test, 30 listeners
DNN vs HMM

Preference test, 30 listeners

DNN adapted using i-vector+LHUC+fMLLR
HMM adapted using CSMAPLR
Multi-task learning
Multi-task DNNs in speech synthesis

Main task: vocoder parameters
Secondary task: Glimpse-based perceptual measure (STEP)

Z Wu et al, ICASSP-2015
Multi-task learning for ASR

Bell & Renals, ICASSP-2015

3–5% WER relative reduction (TED)
Deep generative modelling
Trajectory RNADE

Uria et al, ICASSP-2015
RNADE synthesis

<table>
<thead>
<tr>
<th>MDN Sample</th>
<th>MDN Mean</th>
<th>Traj. RNADE Sample</th>
<th>Traj. RNADE Mean</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>19.4%</td>
<td>-</td>
<td>80.6%</td>
<td>-</td>
<td>900</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>16.0%</td>
<td>84.0%</td>
<td>865</td>
</tr>
<tr>
<td>- 33.6%</td>
<td>-</td>
<td>- 66.4%</td>
<td>-</td>
<td>794</td>
</tr>
</tbody>
</table>
Exemplar Applications
Voice banking and personalised TTS

Veaux et al
Transcription of Youtube clips using Automatic Speech Recognition

MH370 - Experts investigate mystery plane wreckage - BBC News
Inside Out Trailer 2 UK - Official Disney Pixar _ HD
Asian Stocks Fall To Three-Year Low

The History of Pork Pies in Britain - The Great British Bake Off
007 Spectre Official Trailer #2 (2015) Daniel Craig James Bond Movie HD
Branson - Volkswagen scandal is a wake-up call
### Browsing Oral Histories

#### Green et al.

<table>
<thead>
<tr>
<th>Time</th>
<th>Transcription</th>
</tr>
</thead>
<tbody>
<tr>
<td>00:07:2</td>
<td>&quot;travelled to doncaster&quot;</td>
</tr>
<tr>
<td>00:07:12</td>
<td>&quot;and he was taking some dispatches to doncaster&quot;</td>
</tr>
<tr>
<td>00:01:20</td>
<td>&quot;we didn't go back to school in doncaster for&quot;</td>
</tr>
<tr>
<td>00:13:18</td>
<td>&quot;and she had been a teacher at doncaster highschool actually&quot;</td>
</tr>
</tbody>
</table>
GlobalVox

And for those you weren't there the I a c is the largest scientific experiment ever attempted twenty-seven kilometers.....

Contacted peacefully in nineteen fifty eight in nineteen fifty seven five missionaries attempted contact made a critical mistake they drop.....

There is one corner by the way that i'm not going to tell anybody about where you actually where the.....

Now for almost twenty years when we sequence the human genome was going from the analogue world the biology into.....

But this understates the seriousness of this particular problem because it doesn't show the thickness of the ice the arctic.....

That's why we have a refrigerators air conditioning can make a modern materials and do so many things so we're.....

I'd i out see also hurt useless and we and to them made healthy strong capable i was reading this.....

Lay people thinking about their own happiness and the price of scholars thinking about happiness because it turns out with.....

One particular cause of death say accidents right away i see there's a different pattern emerges this is because in.....

And now the envelope push back and i was told by ah the folks at my company that we were.....
Concluding remarks

- Some recent advances from the NST project
- Other things include
  - distant speech recognition
  - disordered speech recognition
  - end-to-end RNN speech recognition
  - disfluent speech synthesis
  - speaker verification spoofing challenge
  - multilingual / cross-lingual recognition & synthesis
  - software: HTK v3.5, NN LM estimation, …