Transcription of multi-genre media archives using out-of-domain data

PJ Bell¹, MJF Gales², P Lanchantin², X Liu², Y Long², S Renals¹, P Swietojanski¹, PC Woodland²

(1) Centre for Speech Technology Research, University of Edinburgh, UK;
(2) Cambridge University Engineering Department, Cambridge UK

This work

- We describe a speech recognition system for multi-genre media archives, a highly challenging ASR task.
- To allow systems to benefit from out-of-domain data, we present multi-level adaptive networks (MLAN), incorporating information from out-of-domain posterior features using deep neural networks.

Motivation: unlocking media archives

- Large quantities of broadcast content exist in media archives, but automatic transcription, metadata extraction, and indexing is underdeveloped.
- Media organisations – such as the British Broadcasting Corporation (BBC) in the UK – wish to “unlock” historic content, sometimes dating back many decades.
- Research on transcription of broadcast data has focused strongly on news and related content. Transcription of arbitrary multi-genre content is much harder, due to the high diversity of data, which may include – “on location” broadcasts in diverse environments, including sports and documentary features – drama with highly-emotional speech and overlaid sound effects.
- There may be limited amounts of training data from each genre – can we improve performance by using mismatched out-of-domain (OOD) data?

Acoustic modelling with tandem DNNs

- Deep neural networks (DNNs) were trained to model model frame posterior probabilities over monophones.
- DNNs were pre-trained in an unsupervised manner using restricted Boltzmann machines, and fine-tuned with stochastic gradient descent.
- DNN structure was fixed following frame-error analysis on held-out training data, finally using four hidden layers, nine frames of acoustic context and 1024 units in each hidden layer.
- Monophone log-posterior probabilities were decorrelated with a PCA transform and projected to 30 dimensions.
- GMMs were trained on 69-dimensional tandem features.

Domain adaptation

Prior work

Posterior features derived from neural networks have been successfully used in cross-domain and cross-lingual applications in tandem systems. Domain adaptation may be performed in a number of ways:
- Retraining or adapting tandem GMMs in OOD features.
- Performing additional training of OOD nets on in-domain data.
- Combining outputs from nets trained on different domains using a merger MLP.

MLAN: Multi-level adaptive networks

We propose the MLAN architecture to most effectively incorporate information from multiple OOD features:
- Generate tandem features for the in-domain data by combining standard acoustic features with posterior features generated using OOD nets.
- Train a DNN on the OOD tandem features to generate MLAN posterior features.
- Train final MLAN GMMs on these features.

Experiments

- We performed experiments on a corpus of TV and radio broadcasts in UK-English from the BBC, including studio speech, programmes made on location, and a TV drama series.
- 20.8 hours of in-domain BBC data was available for training.
- For out-of-domain data we used 276 hours of US-English conversational telephone speech (CTS) from Switchboard, and 127 hours of multi-party meetings from the AMI corpus. Both were highly mismatched to the BBC task.
- We built two systems: a development system, with ML-trained models and a one-pass decoder architecture; and a final evaluation system, using MPE-trained models and speaker-adaptation.

<table>
<thead>
<tr>
<th>Feature set</th>
<th>Studio Location Drama</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLP</td>
<td>17.0 32.5 67.3</td>
<td>39.4</td>
</tr>
<tr>
<td>BBC tandem</td>
<td>14.4 27.5 59.2</td>
<td>34.1</td>
</tr>
<tr>
<td>AMI tandem</td>
<td>14.3 26.9 59.2</td>
<td>33.8</td>
</tr>
<tr>
<td>CTS tandem</td>
<td>14.3 25.6 62.3</td>
<td>35.5</td>
</tr>
<tr>
<td>AMI+CTS MLAN</td>
<td>13.5 25.0 56.1</td>
<td>32.0</td>
</tr>
<tr>
<td>CTS MLAN</td>
<td>12.5 25.5 56.6</td>
<td>31.9</td>
</tr>
<tr>
<td>AMI+CTS MLAN</td>
<td>12.5 24.3 54.9</td>
<td>31.0</td>
</tr>
</tbody>
</table>

Development system results (WER%) without posterior features (PLP) and using in-domain (BBC) and OOD (AMI, CTS) tandem features

<table>
<thead>
<tr>
<th>Feature set</th>
<th>1-pass (unadapted)</th>
<th>2-pass (adapted)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLP</td>
<td>12.0 25.9 58.8</td>
<td>11.5 23.6 58.9</td>
</tr>
<tr>
<td>BBC tandem</td>
<td>11.7 23.3 54.9</td>
<td>11.3 22.3 54.4</td>
</tr>
<tr>
<td>AMI tandem</td>
<td>11.3 22.6 55.0</td>
<td>11.1 21.5 54.2</td>
</tr>
<tr>
<td>AMI+CTS MLAN</td>
<td>10.2 20.9 50.5</td>
<td>9.8 20.0 50.2</td>
</tr>
</tbody>
</table>

Final system results (WER%) using selected PLP, tandem, and MLAN features

Conclusions and future work

- MLAN provides substantial reductions in word error rate on the BBC task: 15% relative over a PLP baseline, 9% over in-domain tandem features and 8% over the best out-of-domain tandem features.
- In future we will investigate the use of the MLAN technique to hybrid DNN systems, and add speaker-adaptative training.