Deep neural network context embeddings for model selection in rich-context HMM synthesis

Thomas Merritt, Junichi Yamagishi, Zhizheng Wu, Oliver Watts, Simon King
T.Merritt@ed.ac.uk
Centre for Speech Technology Research, University of Edinburgh
National Institute of Informatics, Japan

Abstract

Background:
• Across-linguistic context averaging is extremely detrimental to quality. Within-linguistic context averaging is much more preferable [1].
• Conventional rich-context synthesis system [2] - modelling within-linguistic contexts only.

Contribution:
• Bottleneck features extracted using [3] are used to identify closest rich-context models where out-of-training contexts are encountered.

Rich-context models

\[
\begin{align*}
\mu & \quad \sigma^2 \\
\mu' & \quad \sigma'^2
\end{align*}
\]

• Conventional trained decision tree
• Untie leaf nodes
• Update means (keep tied variances)

At training time
• Frame-wise bottleneck features (BN) generated using [3]
• HMM state alignments used
• Distributions of training context features calculated

At synthesis time
• Frame-wise bottleneck features generated
• HMM state alignments used
• Closest seen rich-context (RC) model selected based on distance in ‘bottleneck space’
• For each phoneme distances across all states summed together to guide selection

System comparisons

<table>
<thead>
<tr>
<th>System comparisons</th>
<th>Standard HMM system</th>
<th>Conventional rich-context system</th>
<th>Proposed system</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Select tied cluster from decision tree. Calculated using across-context averaging - reduces quality [1].</td>
<td>Pre-selection of rich contexts to use based on matching triphone.</td>
<td>No linguistic constraints placed - this is learnt by DNN.</td>
</tr>
<tr>
<td></td>
<td>Kullback-Leibler divergence (KLD) between standard tied cluster &amp; each rich context.</td>
<td>Kullback-Leibler divergence (KLD) between standard tied cluster &amp; each rich context.</td>
<td>Kullback-Leibler divergence (KLD) between standard tied cluster &amp; each rich context.</td>
</tr>
<tr>
<td></td>
<td>Smallest divergence selected.</td>
<td>Smallest divergence selected.</td>
<td>Smallest divergence selected.</td>
</tr>
</tbody>
</table>

Results

Figure: Boxplot of rank order of conditions from MUSHRA test

- \( N \) - natural
- \( V \) - vocoded
- \( D \) - Stacked bottleneck DNN [3]
- \( H \) - HTS demo
- \( F \) - Fully untied tree (MDL=0)
- \( CT \) - [2] w/ triphone pre-selection
- \( CB \) - [2] w/ triphone pre-selection
- \( ETS \) - proposed - Euclidean distance
- \( KL \) - proposed - KLD
- \( KLTS \) - proposed - KLD w/ tied source

Conclusions & future work

• Proposed system provides significantly improved selection of rich-context models.
• Pre-selection in [2] inadvertently hiding that target distribution is not optimal.
• DNN system no longer requires speech parameters as output - perceptually more relevant features can be used.

References