Structured Output Layer with Auxiliary Targets for Context-Dependent Acoustic Modelling

Pawel Swietojanski, Peter Bell, Steve Renals

The Centre for Speech Technology Research, University of Edinburgh, United Kingdom

{p.swietojanski, peter.bell, s.renals}@ed.ac.uk

Abstract

In previous work we have introduced a multi-task training technique for neural network acoustic modelling, in which context-dependent and context-independent targets are jointly learned. In this paper, we extend the approach by structuring the output layer such that the context-dependent outputs are dependent on the context-independent outputs, thus using the context-independent predictions at run-time. We have also investigated the applicability of this idea to unsupervised speaker adaptation as an approach to overcome the data sparsity issues that comes to the fore when estimating systems with a large number of context-dependent states, when data is limited. We have experimented with various amounts of training material (from 10 to 300 hours) and find the proposed techniques are particularly well suited to data-constrained conditions allowing to better utilise large context-dependent state-clustered trees. Experimental results are reported for large vocabulary speech recognition using the Switchboard and TED corpora.

Index Terms: multitask learning, structured output layer, adaptation, deep neural networks

1. Introduction

Modelling context-dependent (CD) phones using tied-state clustered trees, initially proposed by Young and Woodland [1], has been a cornerstone of acoustic modelling for more than two decades, providing a flexible data-driven framework for managing the trade-off between the amount of training material and the final size of the model. Combining this technique with deep neural network (DNN) / hidden Markov model (HMM) hybrid models [2, 3] was one of the major factors in the recent success of DNNs for acoustic modelling [4, 5, 6]. The use of CD states as targets for a single DNN allows for a greater level of parameter sharing, in comparison with Gaussian mixture models (GMMs), where distinct clusters are modelled by a different mixtures of Gaussian components, as well as with earlier approaches to modelling context-dependency in hybrid models, where an ensemble of distinct networks was trained to estimate a set of conditional probabilities necessary to derive a CD likelihood score [7, 8, 9].

Despite its widespread and successful use, the optimal clustering for GMM-based systems is often suboptimal for DNNs [10, 11]. Under data-constrained conditions some additional initialisation techniques [5, 12, 13, 14, 15] need to be applied to fully utilise large CD trees. We propose a structured output layer – an approach that allows the optimisation and prediction of CD and context-independent (CI) targets jointly, with an explicit dependence of CD targets on CI targets. This makes it possible to use CI predictions at test time as well as learning a more difficult task in combination with an easier one.

2. Structured Output Layer

We build our model based on a multi-task learning approach [16] and its applications to robust [17] and cross-lingual [18, 19, 20] acoustic modelling, where the hidden representation is shared and jointly optimised across tasks. In this paper we are concerned with multi-task training within a single language. The choice of an auxiliary task was inspired by the work of Zhang and Woodland [13] who found the use of CI targets for layer-wise discriminative pre-training followed by CD fine-tuning leads to the models that better generalize, and, due to low dimensionality of the CI task, are also faster to pre-train. The idea of layer-wise pre-training itself was proposed by Bengio et al. [21] and was further explored in acoustic modelling for speech recognition by Seide et al. [12]. However, in [12], contrary to [13], pre-training and fine-tuning relied on the same context-dependent task. More recently we extended the CI-based initialisation technique to multi-task training [22] where both context-independent and context-dependent targets are jointly trained. All these methods implicitly implement a form of curriculum learning [23] where a lower entropy task (with respect to the complexity of classification task or the number of the optimised weights used for intermediate predictions) is employed to iteratively place some relevant prior on the parameters: for example, by forcing the model to predict simpler (but related) concepts first, or using initially fewer parameters which are then expanded as the training progresses.

In this work we further extend [13, 22] by using the CI layer not only at the (pre-)training stage but also to compute CD outputs at run-time – the structured output layer (SOL). The SOL estimates the CI outputs $m_t$ as an auxiliary task – (1) and (2). In the original multitask formulation, the CD outputs $s_t$ would be estimated independent of the CI outputs at runtime – (3) and (4) – whereas using the SOL, the CD outputs are given by (5) and (6). If $a_m$ represents the CI layer activations, and $a_s$ and $a_{sm}$ represent the CD layer activations with and without dependency on the the CI layer, then we have:

\[
\begin{align*}
a_m &= (x, M + m) \\
P(m_t | o_t) &= \text{softmax}(a_m) \\
a_s &= (x, S + b) \\
P(s_t | o_t) &= \text{softmax}(a_s) \\
a_{sm} &= (x, S + \psi(a_m)C + b) \\
P(s_{sm} | o_t) &= \text{softmax}(a_{sm})
\end{align*}
\]

where $o_t$ is the acoustic input. The SOL layer, depicted in Fig 1, is then composed of parameters $\theta_{SOL} = \{S, M, C, b, m\}$, where $S \in \mathbb{R}^{X \times S}$ and $b \in \mathbb{R}^{S}$ represent hidden to CD weight matrix and bias, respectively. $M \in \mathbb{R}^{X \times M}$ and $m \in \mathbb{R}^{M}$ are...
Figure 1: The model with the structured output layer (SOL). $P(s|x)$ can be computed with or without a dependency on the monophone layer to compute either $a_m$ (5) or $a_s$ (3).

for hidden to CI targets while C $\in \mathbb{R}^{M \times S}$ are the CI to CD connection weights, allowing us to use the easier monophone prediction task when deciding on (harder) context-dependent tied-state both in training and also at run-time. $\psi$ is the non-linearity used for the activations of the CI layer in the SOL. The remaining part of the model follows the usual structure with L hidden layers $\{h^1, \ldots, h^L\}$. In the remainder of this paper, we will be focused mostly on the SOL layer itself, rather than the model as a whole. As such, we introduce an auxiliary variable $x_t \in \mathbb{R}^N$ which denotes the vector of top hidden layer activations at a time $t$, or when considering a mini-batch of examples, $x_t$ becomes to be $x \in \mathbb{R}^{B \times N}$, where $B$ is the mini-batch size.

We learn the model by optimising a global cost, the weighted sum of each of the separate costs:

$$F = (1 - \alpha)F_s + \alpha F_m,$$  

where both CD ($F_s = -\sum \log P(s_t|o_t; \theta_s)$) and CI ($F_m = -\sum \log P(m_t|o_t; \theta_m)$) components are expressed as a gradient descent on a negative log likelihood over $T$ training examples. Note that we obtain both predictions in parallel in a single forward-pass which is different from our earlier work [32], but we set $\alpha = 0$ and scale the error signals are:

$$\frac{\partial F_s}{\partial x_t} = \frac{\partial \log P(s_t|o_t; \theta_s)}{\partial x_t} \frac{\partial s_t}{\partial a_s} + \frac{\partial \log P(s_t|o_t; \theta_s)}{\partial a_{sm}} \frac{\partial a_{sm}}{\partial x_t}$$

3. Experiments

We work with the TED talks corpus [29] following the IWSLT (www.iwslt.org) evaluations and the Switchboard corpus of conversational telephone speech [30].

For TED talks we primarily work on 143 hours of data obtained with light supervision [31], following the recipe described in detail in [32]. For the purpose of this work, we additionally sub-sample random subsets of 10 and 30 hours of training material to simulate data constrained conditions. We do most experiments on the 30 hours split, reporting the most promising configurations on 10 hours and the full 143 hours. This work, compared to [32], benefits from better language models developed for 2014 IWSLT evaluation campaign [33], in
Table 1: WER(%) results on tst2010 set. Models trained on 30 hour data-split with $\alpha = 0.3$

<table>
<thead>
<tr>
<th>Model</th>
<th>tst2010</th>
<th>+4gm</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1 CD-NN (1k)</td>
<td>23.1 ± 0.1</td>
<td>19.7</td>
</tr>
<tr>
<td>S2 SOL-NN const $\partial F_s/\partial {M, m}$</td>
<td>22.8</td>
<td>19.5</td>
</tr>
<tr>
<td>S3 SOL-NN $\partial F_s/\partial \mu$</td>
<td>21.9 ± 0.2</td>
<td>18.7</td>
</tr>
<tr>
<td>S4 SOL-NN + PL Monophones</td>
<td>22.7</td>
<td>19.4</td>
</tr>
<tr>
<td>S5 SOL-NN + Retrained CD</td>
<td>22.5</td>
<td>19.2</td>
</tr>
<tr>
<td>S6 CD-NN (2k)</td>
<td>22.6</td>
<td>19.2</td>
</tr>
<tr>
<td>S7 SOL-NN (2k) $\partial F_s/\partial \mu$</td>
<td>21.6</td>
<td>18.5</td>
</tr>
</tbody>
</table>

Table 2: WER(%) on tst2010 set for different $M$ to $C$ $\psi$ activations. The base model is SOL-NN $\partial F_s/\partial \mu$, $\alpha = 0.3$

<table>
<thead>
<tr>
<th>$\psi$-activation</th>
<th>linear</th>
<th>softmax</th>
<th>sigmoid</th>
<th>relu</th>
<th>tanh</th>
</tr>
</thead>
<tbody>
<tr>
<td>WER(%)</td>
<td>21.9</td>
<td>22.2</td>
<td>22.5</td>
<td>22.4</td>
<td>23.1</td>
</tr>
</tbody>
</table>

Figure 2: WER and FER as a function of $\alpha$.

Figure 3: Top) Convergence plots for SOL-NN and baseline CD/CI models on dev2010. Bottom) Distribution of CD states obtained from ground-truth labels and the estimates of CD-NN and SOL-NN models.

We found that both truncation of $C$ (scenario 1) and optimising $C$ as a post-processing step (scenario 4) resulted in very high frame error rates in comparison with the baseline. Row (S2) gives word error rates (WERs) for the case where the CD cost is not used to update $M$ and $m$; row (S3) shows the opposite scenario indicating that updating the CI-dependent parameters using the $F_s$ cost yields the lowest WER, 21.9%, with a 6% relative improvement over the baseline. Row (S4) is a model trained on 45 position-independent phones (compared to the other models utilising 186 position-dependent phones). Model (S5) is built from the hidden representation of (S3) with a new CD regression layer showing that both the SOL layer and multitask training are important. Finally, rows (S6) and (S7) present WERs for larger models showing over 1% absolute (or 0.7% for re-scored lattices) gain for SOL-NN structure.

We found that both truncation of $C$ (scenario 1) and optimising $C$ as a post-processing step (scenario 4) resulted in very high frame error rates in comparison with the baseline. Row (S2) gives word error rates (WERs) for the case where the CD cost is not used to update $M$ and $m$; row (S3) shows the opposite scenario indicating that updating the CI-dependent parameters using the $F_s$ cost yields the lowest WER, 21.9%, with a 6% relative improvement over the baseline. Row (S4) is a model trained on 45 position-independent phones (compared to the other models utilising 186 position-dependent phones). Model (S5) is built from the hidden representation of (S3) with a new CD regression layer showing that both the SOL layer and multitask training are important. Finally, rows (S6) and (S7) present WERs for larger models showing over 1% absolute (or 0.7% for re-scored lattices) gain for SOL-NN structure.

We did some sanity checks, and the corresponding models (S1) and (S3) with retrained top layers converged to their base model accuracies where all layers were jointly optimised.
3.2. Multi-task adaptation

In this section we investigate the feasibility of using the CI targets to perform unsupervised two-pass adaptation. Our motivation is that CI modelling is usually characterised by a lower frame error rate, and at the same time there is less sparsity in the distribution of CI targets, given the same amount of adaptation data, hence potentially obtaining better adaptation results compared to a CD-only objective. A similar approach, but using hierarchy of output layers and CI-only adaptation targets was proposed by Price et al. in [39].

We adapt our speaker-independent models with a technique which learns hidden unit contributions (LHUC) [40, 41] given unsupervised adaptation data. We report the adaptation results for two scenarios, using both a limited amount of 10 seconds of speech per speaker as well as full two pass adaptation. For the 10s scenario we repeated the experiments 5 times, for randomly selected utterances, and report the average WERs.

The results on tst2010 are reported in Table 3 showing that around 0.2-0.3% absolute gain was obtained on top of CD-only adaptation for both scenarios. Interestingly, when using all adaptation data available for a given speaker, interpolated adaptation does not bring a WER reduction with a pruned trigram LM, but re-scoring with a 4-gram brings up to 0.3% absolute gain on top of CD targets only. The cost objectives for adaptation with different $\alpha$ are plotted in Fig 4.

We observe similar gains when adapting models on other test sets. With $\alpha = 0.5$, WER on dev2010 is 0.3% abs. lower for the 10s adaptation scenario. Similarly to tst2010, interpolated ($\alpha = 0.5$) adaptation on de2010 with the whole speaker’s data reduced the WER by 0.2% abs., when compared to CD-only adaptation. On tst2011 adapting with 10s gave smaller reductions for both methods (0.2% abs.) regardless of $\alpha$; this could be due to the fact tst2011 is better matched to training conditions and benefits less from adaptation.

3.3. Full TED and Switchboard

Finally, we report summary results on three scenarios for TED (10, 30 and 143 hours) in Table 4, as expected, we observe similar gains from the proposed method on even more constrained scenario (10 hours) and also, as expected, the adaptation brings larger gains there since the SI models see less speaker variability during training. For larger models and more training data the gains diminish, with a small improvement of 0.3% on tst2011, but not on dev2010 and tst2010 for Full TED. On Switchboard task (Table 5) the SOL-NN model reduced WERs on Switchboard (SWB) part at the same time increasing the metric on CallHome (CHE) test data, making the model falling back 0.2% WER in average behind the baseline.

### Table 3: Detailed results on tst2010 and unsupervised adaptation with LHUC using auxiliary targets on 30 hours models.

<table>
<thead>
<tr>
<th>System</th>
<th>$\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>21.9</td>
</tr>
<tr>
<td>+4gm</td>
<td>18.6</td>
</tr>
<tr>
<td>Adaptation with 10 seconds per speaker</td>
<td></td>
</tr>
<tr>
<td>+LHUC</td>
<td>21.3</td>
</tr>
<tr>
<td>+4gm</td>
<td>18.0</td>
</tr>
<tr>
<td>Adaptation with all speaker’s data</td>
<td></td>
</tr>
<tr>
<td>+LHUC</td>
<td>18.3</td>
</tr>
<tr>
<td>+4gm</td>
<td>16.0</td>
</tr>
</tbody>
</table>

### Figure 4: Context-dependent frame error rates as a function of adapting iterations for different values of $\alpha$ on tst2010.

### Table 4: Summary results on the remaining TED test-sets and different amounts of training material and LHUC adaptation with pruned (4gm) language models.

<table>
<thead>
<tr>
<th>System</th>
<th>dev2010</th>
<th>tst2010</th>
<th>tst2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 hour</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CD-NN</td>
<td>27.0</td>
<td>28.1</td>
<td>22.7</td>
</tr>
<tr>
<td>SOL-NN +LHUC (All)</td>
<td>25.6</td>
<td>26.6</td>
<td>21.3</td>
</tr>
<tr>
<td>30 hour</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CD-NN</td>
<td>22.9</td>
<td>23.2</td>
<td>19.2</td>
</tr>
<tr>
<td>SOL-NN +LHUC (All)</td>
<td>22.0</td>
<td>21.9</td>
<td>17.8</td>
</tr>
<tr>
<td>143 hour</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CD-NN</td>
<td>18.3</td>
<td>17.9</td>
<td>14.6</td>
</tr>
<tr>
<td>SOL-NN +LHUC (All)</td>
<td>18.3</td>
<td>17.9</td>
<td>14.4</td>
</tr>
</tbody>
</table>

### Table 5: WER(%) on Switchboard Hub00

<table>
<thead>
<tr>
<th>Model</th>
<th>Hub5’00</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD-NN</td>
<td>15.8</td>
<td>28.4</td>
</tr>
<tr>
<td>SOL-NN</td>
<td>15.6</td>
<td>28.9</td>
</tr>
</tbody>
</table>

4. Conclusions

We have proposed a structured output layer, an approach in which an auxiliary (context-independent) task is used as a regularizer during training but also as an auxiliary predictor in deriving context-dependent tied states for decoding. We have investigated various training strategies for this technique, and have shown that this approach is an effective way of addressing unsupervised adaptation with sparse data. For future work we are interested in evaluating with respect to other techniques proposed for low-resource speech recognition as well as extending to sequence discriminative training.

Acknowledgement: This research was supported by EPSRC Programme Grant EP/I03022/1, Natural Speech Technology (NST). The NST research data collection may be accessed at http://datashare.is.ed.ac.uk/handle/10283/786. Experiments were carried using Kaldi Speech Recognition Toolkit [38] while the acoustic modelling part was developed in Pylearn2 [42].
5. References


