Exploring CTC-Network Derived Features with Conventional Hybrid System

Thai-Son Nguyen, Sebastian Stüker, Alex Waibel

ASR with CTC Model
- Using LSTMs, the training with CTC criterion can efficiently model the dependencies between a small number of units (e.g., phonemes or characters) and speech frames.
- The CTC criterion automatically handles possible alignments between a label sequence and the speech frames.
- Eliminating complex steps in the conventional hybrid system, e.g., HMM topology definition, CD phonemes, and frame-wise alignment.
- Good performance on many thousands hours of speech data but severely overfit with less training data [1].
-Bad performance without incorporating language model during inference.
-Optimizing for CTC decoding is hard, e.g. with or without incorporating the priors.

CTC Alignment
- The CTC posteriors have peaky behaviors in which blank has the highest probability in almost all frames, except for short periods where sparse labels dominate.
- Do the phone probabilities assigned by the CTC model still correlate to the fixed labels of a traditional Viterbi alignment?

Our Approach
- We train CTC model with phone labels and use CTC posterior probabilities as input features (so-called C-Phone) in hybrid i-vector/HMM-ANN system.
- To benefit from the strengths of the CTC network at label discrimination on the one side and the highly optimized decoding stack of conventional hybrid systems on other side.
- Taking advantages of combining different features e.g., i-vectors, bottleneck features for further improving phonemes classification performance.

Related Works
- The posterior output of MLP was originally proposed as input features to Tandem GMM models [2].
- When the multiple HMM states per phone and CD states were features to Tandem GMM models [2].
- The posterior output of MLP was originally proposed as input.
- The posterior output of MLP was originally proposed as input.
- The probability of the blank does not carry useful information.

Experimental Setups
- Training data includes 300 hours of the Switchboard-1 Release 2 (LDC97S62).
- CTC modeling with Bi-directional LSTM with 5 layers of 320 units on input features of 40 filter-bank co-efficients with 45 English phonemes as labels.
- FFNN architecture of 7 layers of 1600 units for all hybrid HMM/ANN models.
- Evaluated on Hub500 evaluation data (LDC2002S09).
- Used 4-gram language model from Fisher corpus and the training data.

Features Combination
- +Features
- Window Hub5 e (dB)
- FBank 1/1 23.0 (17.7)
- 3/3 18.9 (13.8)
- 5/5 19.1 (13.7)
- BNF 1/1 18.5 (13.3)
- 3/3 18.2 (13.7)
- 5/5 18.2 (13.7)
- IMLR-BNF 1/1 18.4 (13.2)
- 3/3 18.2 (13.7)
- 5/5 18.2 (13.7)

CTC-Phone Performance
<table>
<thead>
<tr>
<th>Model</th>
<th>Features</th>
<th>Window</th>
<th>Hub5 e (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FFNN</td>
<td>+Features</td>
<td>1/5</td>
<td>23.0 (17.7)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3/3</td>
<td>18.9 (13.8)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5/5</td>
<td>19.1 (13.7)</td>
</tr>
<tr>
<td>C-Phone-L</td>
<td></td>
<td>1/1</td>
<td>18.5 (13.3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3/3</td>
<td>18.2 (13.7)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5/5</td>
<td>18.2 (13.7)</td>
</tr>
<tr>
<td>IMLR-BNF</td>
<td></td>
<td>1/1</td>
<td>18.4 (13.2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3/3</td>
<td>18.2 (13.7)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5/5</td>
<td>18.2 (13.7)</td>
</tr>
</tbody>
</table>

Conclusions
- A feed-forward network system using our proposed CTC-network derived features with cross-entropy training outperforms a strong CTC baseline by a margin of 5% rel.
- With the same model, we achieved further improvements of 9% rel. when combining them with bottleneck features.
- We are examining the gain when performing sequence training as well as the performance of the presented systems on different training data sets.

References