Equivalence of Segmental and Neural Transducer Modeling: A Proof of Concept


1 Human Language Technology and Pattern Recognition
Lehrstuhl Informatik 6
Fakultät für Mathematik, Informatik und Naturwissenschaften
RWTH Aachen University
Introduction

Sequence Classification

Monotonic tasks:
- automatic speech recognition
- text image recognition

Most general case:
- input sequence:
  \[ x^T_1 := x_1 \ldots x_t \ldots x_T \]
- output sequence (of unknown length \( N \)):
  \[ a^N_1 := a_1 \ldots a_n \ldots a_N \]

Acoustic/joint modeling:
- posterior distribution \( pr(a^N_1|x^T_1) \)
- modeling approaches:
  \- hybrid HMM
  \- transducer (RNN-T/RNA)
  \- segmental model/direct HMM
  \- attention-based encoder-decoder

Goal: Compare direct models involving explicit alignment
Alignment in Direct Posterior Models

Alignment: **Transducer Model** (here: RNN-T)

Examplary alignment of label sequence “A  B  C  D”

- length modeling via frame-wise blank label ε emissions: horizontal transitions
- actual labels: vertical transitions
- mapping from transducer labels $\overline{a}_{1}^{T+N}$ incl. blank to actual labels $a_{1}^{N} = G(\overline{a}_{1}^{T+N})$
  e.g. here: $G(\varepsilon\varepsilon\text{A}\varepsilon\varepsilon\text{B}\varepsilon\text{CD}\varepsilon\varepsilon) = \text{A} \ \text{B} \ \text{C} \ \text{D}$
Alignment: **Segmental** Model (also: direct HMM)

Examplary alignment of same label sequence “A B C D”

- single label $a_n$ per segment $n$, with segment end time $t_n$
- requires explicit segment length model
Alignment in Direct Posterior Models

Alignment: **Segmental** vs. **Transducer** Model

Compare examplary alignment of label sequence “A B C D”

- comparison: similar structure of alignment on transducer and segmental model level
- each alignment has unique correspondence, both ways
- open questions: relation of **modeling potential?**
  - relation of **search & pruning behaviour?**
Direct ("end-to-end") modeling of posterior involving explicit alignment:

**Transducer Model**

\[
p(a_1^N \mid x_T^1) = \sum_{\bar{a}_1^{T+N} : G(\bar{a}_1^{T+N}) = a_1^N} p(\bar{a}_1^{T+N} \mid x_T^1) = \sum_{\bar{a}_1^{T+N} : G(\bar{a}_1^{T+N}) = a_1^N} \prod_{u=1}^{T+N} p(\bar{a}_u \mid \bar{a}_1^{u-1}, x_T^1)
\]

**Segmental Model**

\[
\hat{p}(a_1^N \mid x_T^1) = \sum_{t_1^N} \hat{p}(a_1^N, t_1^N \mid x_T^1) = \sum_{t_1^N} \prod_{n=1}^{N+1} \hat{p}(t_n \mid a_1^{n-1}, t_0^{n-1}, x_T^1) \cdot \hat{p}(a_n \mid a_1^{n-1}, t_0^n, x_T^1)
\]

**Research question:** How are transducer model and segmental model related?
Proof of Equivalence of Transducer and Segmental Model

Transformation of \textit{given transducer} into segmental model:

\[
\hat{p}(t_n \mid a_{1}^{n-1}, t_{1}^{n-1}, x_{1}^{T}) = \prod_{t=t_{n-1}}^{t_n-1} \left\{ p(\epsilon \mid \overline{a}_{1}^{t+n-2}, x_{1}^{T}) \cdot \left[ 1 - p(\epsilon \mid \overline{a}_{1}^{t+n-2}, x_{1}^{T}) \right] \right\}
\]

and

\[
\hat{p}(a_n \mid a_{1}^{n-1}, t_{1}^{n}, x_{1}^{T}) = \frac{p(\overline{a}_{t+n-1} \mid \overline{a}_{1}^{t+n-2}, x_{1}^{T})}{1 - p(\epsilon \mid \overline{a}_{1}^{t+n-2}, x_{1}^{T})}
\]

for \( n = 1, \ldots, N + 1 \) and with transducer label sequence \( \overline{a}_{1}^{T+N} \) defined as follows:

\[
\overline{a}_u = \begin{cases} 
    a_n & \text{iff } \exists \ n : u = t_n + n - 1 < T + N \\
    \epsilon & \text{otherwise}
\end{cases}
\]

\[
\Rightarrow \quad \hat{p}(a_{1}^{N}, t_{1}^{N} \mid x_{1}^{T}) = p(\overline{a}_{1}^{T+N} \mid x_{1}^{T}) \quad \Rightarrow \quad \hat{p}(a_{1}^{N} \mid x_{1}^{T}) = p(a_{1}^{N} \mid x_{1}^{T})
\]

equality on alignment level

equality on sequence level
Proof of Equivalence of Transducer and Segmental Model

Transformation of \textbf{given segmental} into transducer model:

\[
p(a_u | \bar{a}_1^{u-1}, h_1^T) = \begin{cases} 
1 - \sum_{\tau=t_{n-1}}^{t} \hat{p}(\tau | t_1^{n-1}, a_1^{n-1}, h_1^T) & \text{iff } a_u = \epsilon, \\
1 - \sum_{\tau=t_{n-1}}^{t-1} \hat{p}(\tau | t_1^{n-1}, a_1^{n-1}, h_1^T) \\
\hat{p}(a_n | a_1^{n-1}, t_1^{n-1}, t, h_1^T) \cdot (1 - p(\epsilon | \bar{a}_1^{u-1}, h_1^T)) & \text{otherwise},
\end{cases}
\]

with the number of segments \( n \), labels \( a_1^n \) and boundaries \( t_1^n \) defined as:

\[
\begin{align*}
    n &= \left| \{ u' \in \{1, \ldots, u-1 \} : \bar{a}_{u'} \in V \} \right| + 1, \\
t &= u - n + 1, \\
t_{n'} &= \min \{ \tau \in \{ t_{n'-1}, \ldots, t \} : \bar{a}_{\tau+n'-1} \in V \}, \\
a_{n'} &= \bar{a}_{t_{n'}+n'-1} \quad \forall \; n' = 1, \ldots, n-1.
\end{align*}
\]

\[
\Rightarrow \quad p(a_1^{T+N} | x_1^T) = \hat{p}(a_1^N, t_1^N | x_1^T) \quad \Rightarrow \quad p(a_1^N | x_1^T) = \hat{p}(a_1^N | x_1^T)
\]

equality on alignment level

equality on sequence level
Equivalence of Transducer and Segmental Model.

**Analytic Result:** transducer and segmental model can be rewritten into each other with equal sequence posterior.

→ In principle, transducer and segmental model have same modeling potential.

- Blank probability \( p(\epsilon|\ldots) \) can be interpreted as label continuation probability.
- Required final blank at transducer step \( U = T + N \) can be interpreted as sequence end label.
- (Strictly) monotonic transducer also is covered (zero length probability to stay in time frame).

Still to be investigated:

- Not answered by equivalence proof: what is the best underlying model?
  - e.g. inclusion of inherently segmental features/model.

Search:

- Transducer model: inherently time-synchronous.
- Segmental model: inherently label-synchronous.
- Equivalence allows for direct comparison of different search variants.
Experiments comparing time- and label-synchronous search:

- strictly monotonic transducer model
- phoneme labels in first order context: feed-forward decoder model
- external 4-gram word-level language model
- dynamic programming beam search with standard look-ahead and pruning methods
- ASR task: TED-LIUM 2

Results:

<table>
<thead>
<tr>
<th>$Q_{\text{prune}}$</th>
<th>WER [%]</th>
<th>dev set</th>
<th>Utterances [%] with same:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>time-sync.</td>
<td>label-sync.</td>
<td>transcription</td>
</tr>
<tr>
<td>4</td>
<td>8.6</td>
<td>26.4</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>7.5</td>
<td>20.9</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>7.2</td>
<td>17.5</td>
<td>-</td>
</tr>
<tr>
<td>10</td>
<td>7.1</td>
<td>15.7</td>
<td>30.2</td>
</tr>
<tr>
<td>12</td>
<td>14.4</td>
<td>36.7</td>
<td>33.1</td>
</tr>
<tr>
<td>14</td>
<td>13.5</td>
<td>42.4</td>
<td>37.5</td>
</tr>
<tr>
<td>20</td>
<td>12.4</td>
<td>55.8</td>
<td>51.9</td>
</tr>
</tbody>
</table>
Experimental Results

Experiments comparing time- and label-synchronous search:
- strictly monotonic transducer model
- character-based byte-pair encoding (BPE) labels with full context: LSTM-RNN decoder model
- fixed-size beam search w/o (external) language model
- label-synchronous search: optionally, keep target time hypotheses separate

Results:

<table>
<thead>
<tr>
<th>search approach</th>
<th>beam size: ( B_t ) times per hypothesis, ( B ) hypotheses, ( B )</th>
<th>Hub5’00 WER [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>time-sync.</td>
<td>-</td>
<td>14.0</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>label-sync.</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>12</td>
</tr>
</tbody>
</table>
Conclusions

Analytic proof of equivalence of transducer and segmental model.
- Models can be rewritten into each other with exact same label sequence posterior.
- Also applies on the level of latent variable (alignment/blank configuration).

In principle, transducer and segmental model provide same modeling potential.
- Direct consequence of equivalence.

Insight into search & pruning strategies for direct sequence posterior models:
- Allows for direct comparison of time- and label-synchronous search.
- Time-synchronous search shows advantage.

Thank you for your interest!