Introduction

• The aim of this research to improve confidence scores in the context of speech processing.
• Automatic speech recognition aims to generate a transcription for a given speech recording.
• A good confidence score is able to predict errors in the generated transcription.
• This information is useful for applications such as speaker adaptation [1] and error detection [2].

Alignment

<table>
<thead>
<tr>
<th>Word</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>quick</td>
<td>0.81</td>
</tr>
<tr>
<td>brown</td>
<td>0.75</td>
</tr>
<tr>
<td>fox</td>
<td>0.68</td>
</tr>
</tbody>
</table>

Fig. 1: Example audio recording of the phrase “quick brown fox” with word and confidence score predictions.

Confidence scores for 1-best sequences

• Traditionally confidence scores use 1-best hypotheses.
• Propagate information through the sequence in one direction to generate confidence scores.

| 0.81 | quick | 0.75 | brown | 0.68 | fox |

Fig. 2: A 1-best hypothesis with the predicted word, confidence score, and hidden state $h_i$.

Lattice representation

• Represents N-best lists in an efficient and compact form.
• Additional information from each confusion is considered.

Fig. 3: A simple word-marked lattice

Lattice enrichment

$y_i, c_i = f(h_i, x_i)$

Fig. 4: Edge $e_i$ with enriched features

LatticeRNN

• Bidirectional recurrent architecture which considers the forward and backward probabilities distinctly.
• The existing arc combination procedure can be improved by applying attention over arc neighbourhoods $\mathcal{N}$ and $\mathcal{N}'$ rather than just the incoming and outgoing arcs.

References