"Funny Moth" - BUT ATC system

Martin Karafiát, František Grézl, Igor Szoke, Ján Švec, Bhargav Pulugundla
Brno University of Technology, Speech@FIT group, Czech Republic
e-mail: {karafiat,grezl,szoke,ipulugundla}@fit.vutbr.cz, jan.svec@phonexia.com

1 Introduction
ASR system was based on fusion of two complementary systems:
• BLSTM - pre-trained on external noised data.
• F-TDNN [1] - trained on provided data only.

1.1 Data
Provided ATC data were split on:
• Train - 37h
• Dev - 1h

External training data:
• English Fisher1+2, Switchboard 1 Release2, Call Home English, AMI and ICSI-meetings - 2240h.

1.1.1 External data enhancement
• Reverberation using about 2000 RIRs generated by ISM method (Habec's tool).
• Noising (additive) with radio static noises (www.freesound.org and self collected), passengers cabin (self collected), motor and FAN "white" noise. With the SNR range 45 - -5.
• Noised and reverberated data were passed through various codecs including GSM, MP3 and TETRA.
• Random selection of SNR level/reverberation/codec for each speaker so size of the data stays intact.

2 BLSTM ASR System
• Architecture: 6 bi-directional LSTM layers, 512 memory units, 300 dimensional projection layer.
• Features: 24 log Mel-filter bank energies + F0.
• Pre-trained on external noised data with cloned GMM alignment from clean data.
• Fine-tuning to ATC provided data.
• No speaker based adaptation.

2.1 TDNN-F ASR System
• Architecture and training followed standard kaldi AMI TDNN recipe.
• Trained on ATC train data only.
• i-vector based adaptation.

3 Call Sign detection
• A simple approach based on ASR output was chosen.
• Three dictionaries were created.
  1. the company names as used in call signs (alaska, german wings,...) – c
  2. the spelling alphabet (Alpha, Bravo, ...) – a
  3. the numbers – n
• Call sign composed of two words as gulfstar was decomposed, and both sub-words added to call sign dictionary.
• The sentence was converted into string of letters, so that each word is replaced by letter denoting the dictionary the word is coming from. If the word does not belong to any of the above dictionaries, the letter o (as other) is used.
• A regular expression search was done on resulting string to find the call sign. Several patterns are created to capture differently shortened call signs.

3.1 Results
• Dictionary: CMU based
• Language Model: 3gram Knesser-Ney trained on ATC train data
• ROVER fusion of BLSTM and TDNN-F system

<table>
<thead>
<tr>
<th>System</th>
<th>dev WER [%]</th>
<th>eval WER [%]</th>
<th>F1 callsign [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLSTM</td>
<td>31.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLSTM+FineTune</td>
<td>12.2</td>
<td>11.04</td>
<td>73.2</td>
</tr>
<tr>
<td>F-TDNN</td>
<td>12.0</td>
<td>10.43</td>
<td>73.97</td>
</tr>
<tr>
<td>Rover BLSTM(Fine Tune)+F-TDNN</td>
<td><strong>10.0</strong></td>
<td>18.4</td>
<td><strong>75.07</strong></td>
</tr>
</tbody>
</table>

4 Conclusion
• Importance of highly complementary systems.
• Unfortunate bug in final fusion.

References