The 2017 KIT IWSLT Speech-to-Text Systems for English and German

Thai-Son Nguyen, Markus Mueller, Matthias Sperber, Thomas Zenkel, Sebastian Stueker and Alex Waibel
Outline

- IWSLT 2017 ASR Tasks
- System Overview
- Setups
  - Feature Extraction
  - 4-gram and FFNN LM
  - GMM & DNN Systems
  - Speaker Adaption Models
- Results and Discussions
- Conclusion
IWSLT 2017 ASR

- English and German Lecture task
- TED talks and lecture talks.
- Various topics, spontaneous speaking style
- Not segmented
System Overview

Feature Extraction

Speaker Adaptive Features

Regular Features

GMM/DNN

DNN1

GMM1

DNN2

GMM2

Lattices Combination

Model Adaption

Vote

DNN1'

GMM1'

DNN2'

GMM2'
Setups

- **Feature Extraction**
  - Bottleneck features
  - Speaker adaptive feature (SAF)

- **Speaker Adaption Models**
  - GMMs and DNNs using SAF

- **Language Models**
  - 4-gram LM
  - Feed-forward LM
Feature Extraction

Pipeline for extracting Speaker Adaptive Feature (SAF)
Input Features for GMMs

We used FBank and MFCC features to build two GMMs

Feature Extraction for GMMs
Input Features for DNNs

- Also FBank and MFCC features for DNNs

Feature Extraction for DNNs
DNN & GMM Models

- **FFNNs**
  - 8k states of CD-Phone for English systems, 18k states for German systems
  - SAF-IMEL and SAF-MFCC

- **GMMs**
  - The same number of CD-Phone states
  - The same front-ends
Model Adaption

- Use transcriptions from the CNC system
- Align and eliminate the frames with confidence score less than 0.7
- GMMs
  - MLLR
- DNNs
  - An adapted DNN per speaker
  - Training one more epoch on the adaption data with a small learning rate
Training Data

About **480 hours** and **360 hours** for acoustic modeling of English and German systems

<table>
<thead>
<tr>
<th>Source</th>
<th># Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quaero from 2010 to 2012</td>
<td>200 hours</td>
</tr>
<tr>
<td>Broadcast news [8]</td>
<td>80 hours</td>
</tr>
<tr>
<td>TED-LIUM v2 [9]</td>
<td>203 hours</td>
</tr>
<tr>
<td>excluding disallowed talks</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>483 hours</strong></td>
</tr>
</tbody>
</table>

**English acoustic modeling data**

<table>
<thead>
<tr>
<th>Source</th>
<th># Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quaero from 2009 to 2012</td>
<td>180 hours</td>
</tr>
<tr>
<td>Broadcast news</td>
<td>24 hours</td>
</tr>
<tr>
<td>Baden-Württemberg parliament</td>
<td>160 hours</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>364 hours</strong></td>
</tr>
</tbody>
</table>

**German acoustic modeling data**
System Training

Deep bottleneck network and FFNN network

- Input layer of 11-15 stacked frames
- 5-6 hidden layers with 2000 units per layer
- Bottleneck layer of 42 units
- Fine-tuning with cross-entropy loss function
- Newbob training schedule
Language Models

- **4-gram LM** from 7B words for English (150k vocab) and 2B words for German (300k vocab)

- **Feed-forward Neural Network LM**
  - 4 sigmoid layers of 600 units
  - 200-dimensional word embedding for the vocabulary size of 20k
  - To be used directly while decoding
### English Lecture Task

<table>
<thead>
<tr>
<th>System</th>
<th>tst2015</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNN(IMEL)</td>
<td>12.9</td>
</tr>
<tr>
<td>GMM(SAF-MFCC)</td>
<td>11.6</td>
</tr>
<tr>
<td>DNN(SAF-IMEL)</td>
<td>10.2</td>
</tr>
<tr>
<td>DNN(SAF-MFCC)</td>
<td>11.2</td>
</tr>
<tr>
<td>CNC</td>
<td>9.4</td>
</tr>
<tr>
<td>GMM(SAF-MFCC) adapted</td>
<td>9.3</td>
</tr>
<tr>
<td>DNN(SAF-IMEL) adapted</td>
<td>8.8</td>
</tr>
<tr>
<td>DNN(SAF-MFCC) adapted</td>
<td>9.3</td>
</tr>
<tr>
<td>Kaldi 4-gram LM rescored</td>
<td>9.3</td>
</tr>
<tr>
<td>ROVER</td>
<td>8.3</td>
</tr>
</tbody>
</table>

*Results for English lecture task on tst2015 testset*
German Lecture Task

<table>
<thead>
<tr>
<th>System</th>
<th>dev2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>18k DNN(BSV BN-1MEL+T) NNLM</td>
<td>26.7</td>
</tr>
<tr>
<td>18k DNN(Mod-M2+1MEL+T)</td>
<td>27.1</td>
</tr>
<tr>
<td>10k DNN(SAF-BN-M2+T) NNLM</td>
<td>25.2</td>
</tr>
<tr>
<td>10k DNN(SAF-BN-1MEL+T) NNLM</td>
<td>25.7</td>
</tr>
<tr>
<td>CNC</td>
<td>24.5</td>
</tr>
</tbody>
</table>

*Results for German lecture task on dev2017*
Conclusion

- Used techniques
  - Speaker Adaptive Feature
  - Model Adaption
  - System Combinations

- WER results:
  - 8.3% on English tst2015
  - 24.5% on German dev2017