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
- DNN1'
- GMM1'
- DNN2'
- GMM2'

Vote

IWSLT 2017 ASR

System Overview

Evaluation Setups

Results

Conclusion

Thai-Son Nguyen - 2017 KIT IWSLT Speech-to-Text

Interactive Systems Lab - Institute for Anthropomatics and Robotics
Setups

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

- Speaker Adaptation 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.
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

### English acoustic modeling data

| Source                                     | # Amount |
|--------------------------------------------|----------|
| Quaero from 2010 to 2012                   | 200 hours|
| Broadcast news [8]                         | 80 hours |
| TED-LIUM v2 [9]                            |          |
| excluding disallowed talks                 | 203 hours|
| **Total**                                  | **483 hours**|

### German acoustic modeling data

| Source                                                | # Amount |
|-------------------------------------------------------|----------|
| Quaero from 2009 to 2012                              | 180 hours|
| Broadcast news                                        | 24 hours |
| Baden-Württemberg parliament                          | 160 hours|
| **Total**                                              | **364 hours**|
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

| System                          | tst2015 |
|---------------------------------|---------|
| DNN(IMEL)                       | 12.9    |
| GMM(SAF-MFCC)                   | 11.6    |
| DNN(SAF-IMEL)                   | 10.2    |
| DNN(SAF-MFCC)                   | 11.2    |
| CNC                             | 9.4     |
| GMM(SAF-MFCC) adapted           | 9.3     |
| DNN(SAF-IMEL) adapted           | 8.8     |
| DNN(SAF-MFCC) adapted           | 9.3     |
| Kaldi 4-gram LM rescored        | 9.3     |
| ROVER                           | 8.3     |

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

| System | dev2017 |
|--------|---------|
| 18k DNN(BSV BN-1MEL+T) NNLM | 26.7 |
| 18k DNN(Mod-M2+1MEL+T)  | 27.1 |
| 10k DNN(SAF-BN-M2+T) NNLM | 25.2 |
| 10k DNN(SAF-BN-1MEL+T) NNLM | 25.7 |
| CNC | 24.5 |

*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