An extension of Kaldi at The University of Edinburgh

Liang Lu, Pawel Swietojanski, Peter Bell and Steve Renals

The University of Edinburgh

Introduction

We introduce the following recipes and extension to the Kaldi speech recognition toolkits

- A recipe for the AMI corpus
- A recipe for Multi-Genre Broadcast (MGB) challenge
- A linkage between Kaldi and CNTK

Kaldi AMI recipe

Data -The AMI Meeting Corpus is a multi-modal data set consisting of 100 hours of meeting recordings. Which gives the split for automatic speech recognition experiments of about 80 hours for training and 9 hours for development and evaluation sets.

Recipe - The recipe currently supports the following acoustic scenarios:

- Individual Headset Microphones (IHM) - each of participants in the meeting has a head mounted microphone capturing speech at close distance
- Multiple Distant Microphones (MDM) - channels captures by distant microphones are combined using BeamformIt toolkit [1] and acoustic models are then trained on the enhanced channel
- Single Distant Microphone (SDM) - as the acoustic resource one uses one (1st by default) of MDM from microphone array.

Baseline results - We show the standard baseline results using the Kaldi toolkits.

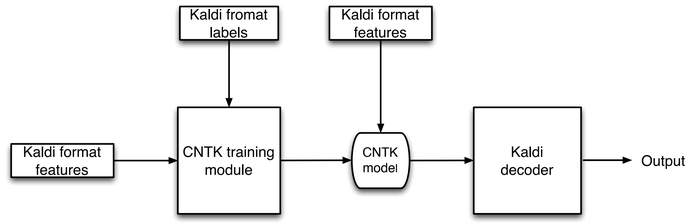
System	SDM	IHM	MDM
GMM	71.8	43.7	64.4
+ LDA_MLLT	69.6	40.8	61.9
+fMLLR	68.0	34.8	-
+bMMI	65.9	31.7	-
DNN CE	58.0	27.0	-
+ realign	56.6	26.5	-
CNTK_DNN CE	56.0	26.8	-

Linkage between Kaldi and CNTK

1. Why link Kaldi with CNTK?

- The two toolkits are complimentary to each other in that Kaldi is a state-of-the-art speech recognition toolkits, while CNTK is more flexible in terms of training neural network models and it supports the automatic gradient computation.

2. The current integration - CNTK can read Kaldi format features and labels and use Kaldi decoder for decoding. The current work is on sequence training of CNTK using Kaldi lattices.



3. Example

```
/* general configuration */
NdlDir=
precision=float
...
/* CNTK model training configuration */
TrainDNN=[
   action=
   modelPath=

/* CNTK neural network configuration */
   NDLNetworkBuilder=[
      ndlMacros=
      networkDescription= ]

/* SGD configuration */
   SGD=[
      epochSize=
      minibatchSize=
      ... ]
/* read Kaldi features and labels */
  reader = [
     readerType=Kaldi2Reader
   ...]
]
```

Kaldi recipe for MGB challenge

- The ***MGB Challenge*** is a new competition in speech recognition and speaker diarization organised by the three sites of the NST project in collaboration with the BBC.
- The are four tasks, all centered on multi-genre TV output from the BBC.
- We have released a Kaldi recipe for building baseline speech recognition acoustic and language models on up to 1,000 hours of TV data, 600M words of subtitle text, and the Comblix British English dictionary, resources which are provided to all participants. Our aim is that this recipe will allow participants to focus on innovative ASR research in their Challenge systems, rather than spending a long time preparing the data and tuning basic parameters.
- One of our aims in designing the challenge has been to ensure that each participant has access to the same data resources. Giving participants a common starting point for system-building will further help make systems more comparable.
- The MGB Challenge data is highly diverse, noisy, fast speech, presenting a difficult problem for ASR systems – the baseline recipe GMM system scores more than 50% word error rate. In future, we plan to include more modern DNN methods to give a much stronger baseline.

References

[1] Acoustic beamforming for speaker diarization of meetings", Xavier Anguera, Chuck Wooters and Javier Hernando, IEEE Transactions on Audio, Speech and Language Processing, September 2007, volume 15, number 7, pp.2011-2023

Acknowledgement There are numerous people who have contributed on the interface between Kaldi and CNTK, in particular, Yu Zhang from MIT. The research was supported by EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology)