The I^{2}R's system for IWSLT 2015 English ASR evaluation

Tran Huy Dat, **Jonathan Dennis, Ng Wen Zheng Terence**

Institute for Infocomm Research, A*STAR, Singapore

Agency for Science, Technology and Research
SINGAPORE

Outline

Introduction

- Task overview
- I²R's team
- I²R's system overview
- I²R's hardware system overview

System details

- SHR-based audio segmentation
- Feature extraction
- Auxiliary GMM-HMM
- DNN acoustic modelling
- Language modelling & rescoring
- Semi-supervised DNN adaptation

Results and Discussions

Conclusions

Task Overview

- To build ASR system(s) to transcribe TED/TEDx audio
- The speech in English TED talks are lectures related to Technology, Entertainment and Design (TED) in spontaneous speaking style, without segmentations
- Despite that the speech in the TED talks is in general planned, well articulated, and recorded in high quality, the task is challenging due to
  - the large variability of topics,
  - the presence of non-speech events,
  - the ascents of non-native speakers,
  - and the informal speaking style.

I²R Team

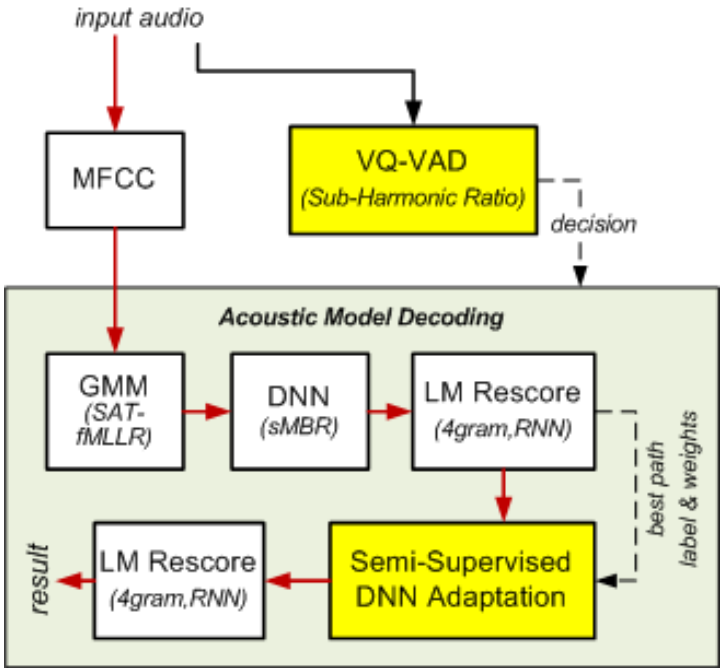
Tran Huy Dat
Dr., Senior Scientist,
HLT, I²R

Jonathan Dennis
Dr., Scientist I, HLT, I²R

Ng Zheng Terence
MSc., Engineer, HLT, I²R

I²R System Overview

- Sub-harmonic ratio –based audio segmentation
- MFCC feature extraction
- Progressive GMM-HMM training (LDA-SAT-fMLLR)
- DNN acoustic modelling with sequence minimum Bayesian risk discriminative training
- Semi-supervised DNN adaptations
- N-gram language modelling and Recurrent Neural Network (RNN) language model rescoring

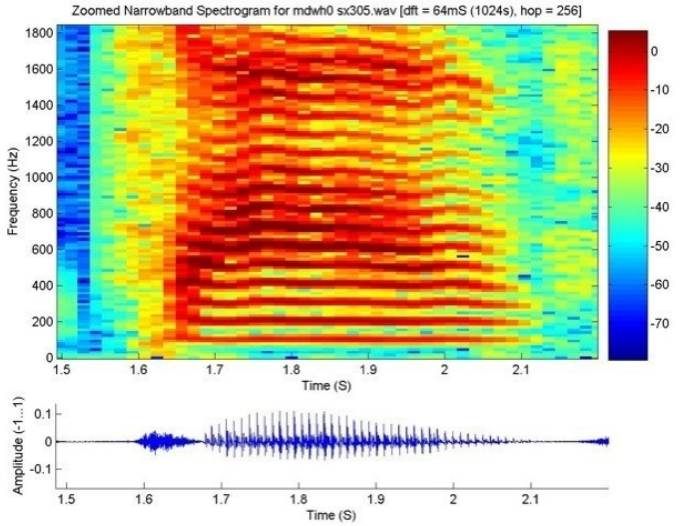


I²R Processing Hardware

- Single server available for training with the following configuration:
  - OS – CentOS release 6.6, 64-bit
  - Total number of CPUs – 4 (on one server machine)
  - Descriptions of CPUs – Intel Xeon E5-2680 @ 2.7 GHz, 8 cores per CPU
  - Total number of GPUs – 4 available, 1 used concurrently
  - Descriptions of GPUs – Tesla K20m, 2496 cores, 5 GB RAM
  - Total available RAM – 128 GB
  - Used Disk Storage – approx: 200 GB for training wave file storage, 100 GB files generated during training, 15 GB for decoding (including approximately 10 GB for LMs)

Sub-Harmonic Ratio

- The harmonic to sub-harmonic ratio (SHR) is used as a feature for voiced speech detection.
- Fundamental physic: voiced segments of speech has strong peaks in the amplitude spectrum occurring at harmonics of the fundamental frequency.
- It translates into maximum value of the ratio between total energy at harmonics and sub-harmonic reference
- To implement it, a simple search is conducted by maximizing the ratio calculated at each frequency f for each frame index t



SHR-based audio segmentation

- The whole audio segment is clustered into 3 groups of background noise, speech and clapping using physical knowledges of events
  - Voiced speech - high SHR and high energy (top highest 10%)
  - Background Noise - for the TED: low SHR and low energy (top lowest 10%)
  - Clapping - impulsive noise has a high energy but a low SHR (top highest and lowest 2%)
- VQ clustering is applied independently for each TED talk audio
  - MFCC vectors were used for VQ
  - 16 VQs for each class of speech, background noise and clapping
- VQ distances are calculated from each frame MFCC to the models' VQs
  - Speech index is determined by

$$\min(D{\downarrow}noise, D{\downarrow}clapping) - D{\downarrow}speech > 0$$

- Frame indexes are smoothed by applying joining and hangover of 500ms

Feature Extraction

- 13-dimensional MFCCs, without energy, which are mean normalised over the speech segments
- Each conversation is assumed to have a single speaker
- Spliced by 3 frames adjacent to the central frame and projected down to 40 dimensions using Linear Discriminant Analysis (LDA).
- Same base feature for GMM-HMM and DNN.

Auxiliary GMM-HMM Training

- All the experiments are carried out using Kaldi toolkit
- Mono-phone MFCC-GMM-HMM system is first trained using 20k shortest utterances from TEDLIUM corpus to provide the initial alignment
- Next triphone and LDA-GMM-HMM systems are trained with 2500 and 4000 tied states, respectively
- Then the whole training data is used to train the SAT-GMM-HMM with 6353 tied states and 150k Gaussian components

Wall Street Journal - this comprises of 81.1 hours of read speech, available from the Linguistic Data Consortium (LDC), from LDC93S6B and LDC94S13B.

HUB4 English Broadcast news - unlike [4] we use the full 201 hours of broadcast news data from LDC97S44 and LDC98S71.

TEDLIUM version 2 - this corpus contains 204 hours of lecture-style TED speech [5] consisting of 1481 talks after the removal of non-permissible talks.

Table1. Training data for acoustic modelling

DNN Acoustic Model Training

- The DNN acoustic model with 5 hidden layers is trained on top of SAT features.
- Each hidden layer has 2k sigmoid neurons, and a 7557 dimensional softmax output layer.
- The hidden layer weights are initialised using layer-wise restricted Boltzmann machine (RBM) pre-training, using 100 hours of randomly selected utterances.
- After pre-training, 2 stages of fine-tuning was performed using the 100 hour subset used for pre-training and a full training set for the second stage.
- Finally, the DNN is re-trained by sequence-discriminative training to optimise the state minimum Bayes risk (sMBR) objective, using the full training data.

Language Modelling and Rescoring

- Two LMs, trained by Kaldi, have been used: a 3-gram is used in first stage of decoding and a 4-gram is used in rescoring the output lattice
- Another RNN LM is used to rescoring the output lattice from 4-gram LM
- RNNLM package with 30k words in the vocabulary, 480 hidden units, 300 classes, 2000M direct connections, BPTT with truncated time order 5, perplexity 60, interpolation weight 0.3
- 2M sentences (i.e. 14%) were randomly selected from LM corpora for RNN training

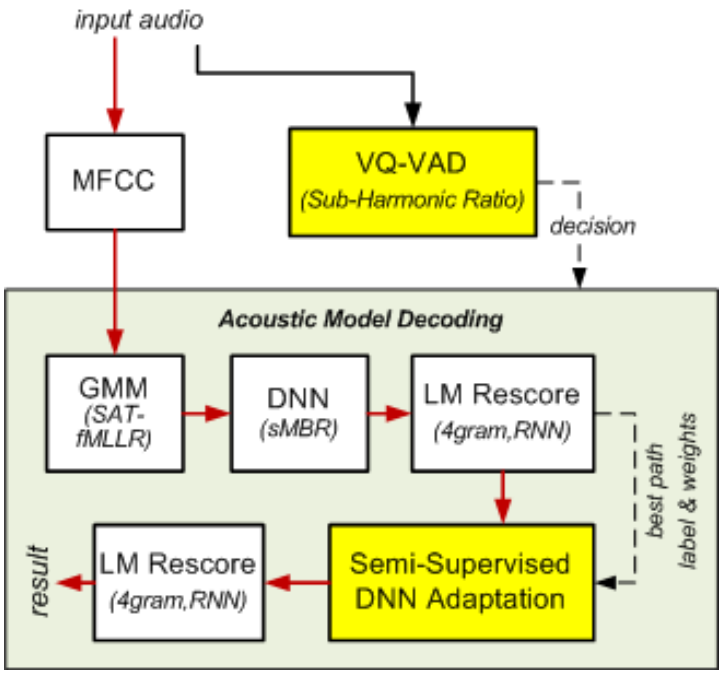
Category	Corpus	Sentences selected	Pct% of Original
In-domain	TED Talks	92k	-
Out-of-domain	CommonCrawl	770k	9%
	Europarl	140k	6%
	Gigaword FR-EN	0.9M	4%
	NewsCommentary	47k	19%
	News	12.3M	18%
	Yandex	310k	31%

Table2. Training data for language modelling

- The training data is the enhanced TEDLIUM corpus, where the selection of additional data is based on the XenC tool.
- Text from each corpus is concatenated together to form a single large set that is used for training each of the subsequent language models.

Semi-supervised DNN Adaptation

- Aims to adapt the acoustic model to the actual recording file (each TED (x)talk audio)
- First-pass decoding output is used to estimate weighted frame labels for the spoken text
- Further iterations of cross-entropy training are performed with weighted error back-propagation stopping is based on closed-set cross-entropy measure
- Adapted DNN then used to re-decode the same data



Results and Discussions

- Table 3 reports the WER of the system from each stage of processing
- Most important components are: segmentation, DNN acoustic modelling with discriminative learning (sMBR), semi-supervised DNN adaptations, and n-gram/ RNN language re-scoring
- It can be seen that the proposed segmentation even outperformed the human labeling ground truth at the final results

Processing Step	WER tst2013	
	Ground Truth	Segmentation
SAT GMM	21.3%	22.4%
DNN sMBR	12.3%	11.6%
+ LM Rescore	10.8%	10.1%
+ DNN Adapt 1	9.5%	8.7%
+ DNN Adapt 2	9.4%	8.5%
+ DNN Adapt 3	9.1%	8.4%

Table3. Detailed results from each stage of processing on development set tst2013

Results and Discussions

- The DNNsMBR system with proposed segmentation yields reasonable result at 11.6% and is very fast in decoding and hence recommended for online implementations
- Semi-supervised adaptations are useful as it could reduce the mismatch between training and actual testing records. After 3-round the improvement became minimum
- LM rescoring is some kind of adaptations in LM and has shown to be useful for the taks

Processing Step	WER tst2013	
	Ground Truth	Segmentation
SAT GMM	21.3%	22.4%
DNN sMBR	12.3%	11.6%
+ LM Rescore	10.8%	10.1%
+ DNN Adapt 1	9.5%	8.7%
+ DNN Adapt 2	9.4%	8.5%
+ DNN Adapt 3	9.1%	8.4%

Table3. Detailed results from each stage of processing on development set tst2013

WER Analysis

- DNN sMBR gives the most significant improvement in performance over the baseline SATGMM (9%-11%)
- In addition, the DNN decoding strategy gives a total of around 2-3% improvement, with the biggest contribution coming from the semi-supervised DNN speaker adaptation, combined with a consistent improvement achieved through language model rescoring.

Processing Step	WER tst2013	
	Ground Truth	Segmentation
SAT GMM	21.3%	22.4%
DNN sMBR	12.3%	11.6%
+ LM Rescore	10.8%	10.1%
+ DNN Adapt 1	9.5%	8.7%
+ DNN Adapt 2	9.4%	8.5%
+ DNN Adapt 3	9.1%	8.4%

Processing Step	WER Gain (tst2013)
DNN sMBR	9%
+ LM Rescoring	1.5%
+ Semi-supervised DNN	1.7%

Table 4. Approximate WER improvements given by the key components

WER Analysis

- The semi-supervised DNN adaptation is suitable for TED and TEDx talks since involving single speakers and long enough to be effective.
- However, a big jump of performance is normally seen in the first round of adaptation while it is very time consuming. Hence, in practical situations, using one round of adaptation is recommended.

Processing Step	WER tst2013	
	Ground Truth	Segmentation
SAT GMM	21.3%	22.4%
DNN sMBR	12.3%	11.6%
+ LM Rescore	10.8%	10.1%
+ DNN Adapt 1	9.5%	8.7%
+ DNN Adapt 2	9.4%	8.5%
+ DNN Adapt 3	9.1%	8.4%

Processing Step	WER Gain (tst2013)
DNN sMBR	9%
+ LM Rescoring	1.5%
+ Semi-supervised DNN	1.7%

Table 4. Approximate WER improvements given by the key components

Our final results

Development set (tst2013)	Test set (tst2015)
8.4%	7.7%

Results were obtained using single systems without any kind of fusion (ROVER)

Conclusions

- We participated in English ASR task within the IWSLT 2015 evaluation campaign:
  - Transcribing of TED talks
  - Non-segmented audio of spontaneous speeches
  - Wide selection of topics
  - Possible non-native speakers
  - Presence of non-speech events
  - Some variations in recording conditions
- Summary of I²R system
  - SHR-based audio segmentation
  - MFCC feature extraction
  - Progressive training steps combining GMM-HMM and DNN-HMM acoustic modelling:
    - GMM-HMM→LDA-GMM-HMM→SAT-GMM-HMM→DNN→DNNsBMR
  - Semi-supervised DNN adaptation and LM rescoring at decoding
    - 3-gram LM decoding→4-gram LM rescoring→RNN rescoring→SS-DNN
- The most significant contributions are from
  - SHR-based audio segmentation
  - Discriminative training DNN acoustic modelling
  - Semi-supervised DNN adaptations
  - RNN language modelling rescoring

Thank you!