The IOIT English ASR system for IWSLT 2016

Van Huy Nguyen[1], *Trung-Nghia Phung*[2], *Tat Thang Vu*[3], *Chi Mai Luong*[3]

[1]Thai Nguyen University of Technology, Vietnam
[2]University of Information and Communication Technology, Thai Nguyen University, Vietnam
[3]Institute of Information and Technology (IOIT),
Vietnamese Academy of Science and Technology, Vietnam

huynguyen@tnut.edu.vn, ptnghia@ictu.edu.vn, {vtthang,lcmai}@ioit.ac.vn

Abstract

This paper describes the speech recognition system of IOIT for IWSLT 2016. Four single DNN-based systems were developed to produce the 1st-pass lattices for the test sets using a baseline language model. The 2nd-pass lattices were further obtained by applying N-best list rescoring on topic adapted language models which were constructed from closed topic sentences by applying a text selection method. The final transcriptions of test sets were finally produced by combining the rescored results. On the 2013 evaluation set, we are able to reduce the word error rate of 1.62% absolute. On the 2014, provided as a development set, the word error rate of our transcription is 11.3%.

1. Introduction

The International Workshop on Spoken Language Translation(IWSLT) is a yearly scientific workshop, associated with an open evaluation campaign on spoken language translation. One part of the campaign focuses on the translation of TED, QED Talks, and the conversations conducted via Skype. TED and QED talks are a collection of public lectures on a variety of topics, ranging from Technology, Entertainment and Education to design. As in the previous years, the evaluation offers specific tracks for all the core technologies involved in spoken language translation, namely automatic speech recognition (ASR), machine translation (MT), and spoken language translation (SLT).

The goal of the ASR track is the transcription of audio coming from unsegmented TED, QED talks, and Microsoft Speech Language Translation (MSLT) Corpus that was drawn from Skype conversations [1], in order to interface with the machine translation components in the speech-translation track. The quality of the resulting transcriptions is measured in word error rate (WER).

In this paper, we describe our speech recognition system which participated in the ASR track of the IWSLT 2016 evaluation campaign. The system is a further development of our last year's evaluation system [2]. There are four single hybrid acoustic models in our system. These models and an interpolated language model were used to produce lattices which were further applied N-best list rescoring with a topic adapted language model. The final transcriptions of the test sets were combinations of the rescored results.

The organization of the paper is as follows. Section 2 describes the data that our system is trained on. This is followed by Section 3 which provides a description of the way to extract acoustic features. An overview of the techniques, used to build our acoustic models, is given in Section 4. Language model and dictionary are presented in Section 5. We describe the decoding procedure and results in Section 6 and conclude the paper in Section 7.

2. Training Corpus

For training acoustic models, we used two types of corpus as described in Table 1. The first corpus is TED talk lectures (http://www.ted.com/talks). Approximately 220 hours of audio, distributed among 920 talks, were crawled with their subtitles, which are deliberately used for making transcripts. However, the provided subtitles do not contain the correct time stamps corresponding with each phrase as well as the exact pronunciation for the spoken words, which lead to the necessity for long-speech alignment. Segmenting the TED data into sentence-like units, used for building a training set, is performed with the help of SailAlign tool [3] which helps us to not only acquire the transcript with exact timing, but also to filter non-spoken sounds such as music or applause. A part of these noises are kept for training noise models while most of them are abolished. After that, the remained audio used for training consists of around 160 hours of speech. The second corpus is Libri360 which is the Train-clean-360 subset of the LibriSpeech corpus [4]. It contains 360 hours of speech sampled at 16 kHz, and is available for training and evaluating speech recognition system.

Table 1: Training data for acoustic models

Corpus	Type	Hours	Speakers	Utts
Ted	Lecture	160	718	107405
Libri360	Audiobook	360	921	104014

3. Feature Extraction

In this work, four kinds of combination features were used to build the acoustic models. These features were obtained by directly concatenating raw frames which were MFCC, FBank, Pitch (P) and i-vector (I) features using Kaldi recipes [5][6]. A Hamming window of 25ms, which was shifted at the interval of 10ms, was applied to calculate MFCC and FBank. MFCC consists of 39 coefficients which are 13 MFCCs, the first and the second order derivatives. FBank consists of 40 log-scale filterBank coefficients. Pitch consist of 3 coefficients including 1 the pitch value, the first derivative of the pitch value, and the probability of voice for the current frame. i-vectors were 100-dimensional vectors that were generated from i-vector extractors that were trained over MFCC using alignments from a baseline system. The combined features are denoted as MFCC, FBank+P, MFCC+P+I, FBank+P+I according to their components.

4. Acoustic Model

4.1. Baseline Acoustic Model

The baseline acoustic model was built by using the Kaldi toolkit [5] with MFCC feature. First, this model was trained as a basic context dependent tri-phone model, followed by a speaker adaptive training (SAT) with a feature space maximum likelihood linear regression (fMLLR). A discriminative training based on the maximum mutual information (MMI) was applied at the end. This model (MMI-SAT/HMM-GMM) had 6496 tri-phone tied states with 160180 Gaussian components, and it was then used to produce a forced alignment in order to get the labeled data for training deep neural networks.

4.2. Hybrid Acoustic Model

This year, we reapplied two hybrid models from last year [2] which are denoted as fMLLR-DNN and FBank-CNN for our transcription system. The fMLLR-DNN model was built by applying a feedforward deep neural network (DNN) congured as 440-1024*5-6496 (input layer with 440 neurons, 5 hidden layers with 1024 neurons for each, output layer with 6496 neurons). The input feature for this model was a fMLLR-based feature that was calculated over MFCC as follow: The MFCC was adjusted by concatenating 11 neighbor vectors (5 ones for each left and right side of the current MFCC vector) to make the context dependent feature, afterward the dimension of the concatenated vector was reduced to 40 by applying a linear discriminate analysis (LDA) and decorrelated with a maximum likelihood linear transformation (MLLT). It was finally applied a feature space maximum likelihood linear regression (fMLLR) in the speaker adaptive training (SAT) stage. The LDA, MLLT and fMLLR transforms were estimated during the training of the baseline model. The FBank-CNN model using FBank+P was applied a convolution neural network (CNN-DNN) which had one convolutional layer with convolution and polling operations. The conguration of the convolutional layer was as follows: 128 filters with filter size and shift as 9 and 1 for each. The pooling width and shift is set to 2 and 2, respectively. The output from the pooling layer was further processed with feedforward DNN with 5 hidden layers (1024 neurons each), and output layer with 6496 neurons. For training MFCC-DNN and FBank-CNN, a frame-based cross-entropy criterion was first applied in the first stage, then a sequential discriminative training based on a state level minimum Bayesian risk criterion (sMBR) [7] was adopted for the second stage training.

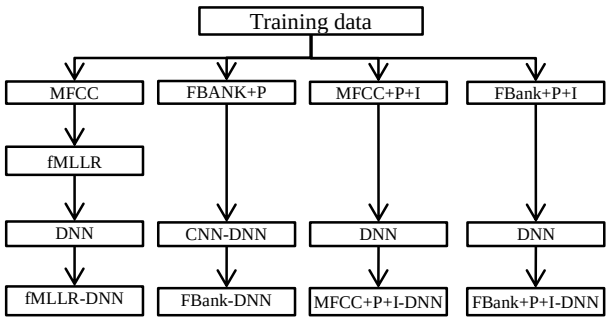

Figure 1: Training process of hybrid acoustic models

Two more models, MFCC+P+I-DNN and FBank+P+I-DNN, were built using the same architecture and training process as the FBank-DNN model, but the input features were MFCC+P+I and FBank+P+I. The i-vectors were combined to improve speaker information for the features. The processes to train the models are represented in Fig.1.

5. Language Model and Dictionary

5.1. Baseline language model

A 3-gram, so called a far-topic language model (FLM), was firstly built. This model was used to generate lattices using the acoustic models. Three categories of textual corpora were used for estimating the model (as shown in Table 3). The first one was the transcript of Libri360 data set that was used for training the acoustic models. The second one was the subtitles of all TED talks published before April-2016 (TED2016) which is provided by Fondazione Bruno Kessler (FBK) (https://wit3.fbk.eu). The third one was QED corpus, version 1.4, provided by Qatar Computing Research Institute (http://alt.qcri.org/resources/qedcorpus/). TED2016 and QED corpora were used for training the language model after rejecting all disallowed talks according to the suggestion of IWSLT-2016 committee.

For training FLM, a vocabulary set was firstly extracted from textual sets. This vocabulary set has 73491 words and was then used to build the language model by using the SRILM toolkit [8]. The perplexity (PPL) score of the trained language model was 184 on the tst2013 test set. In order to

Table 2: Experimental results

System	Model	WER%			
		tst2013		tst2014	
		FLM	Adapted LM	FLM	Adapted LM
S1	fMLLR-DNN	18.85	17.23	14.59	12.64
S2	FBank-CNN	-	-	14.19	12.11
S3	MFCC+P+I-DNN	-	-	14.78	12.96
S4	FBank+P+I-DNN	-	-	15.05	12.91
S1+S2+S3+S4	Combination	-	-	-	11.3

Table 3: Text corpora for training language models

Corpus	Utts
Libri360	100k
TED2016	250k
QEDv1.4	1460k

improve the performance, it was then combined in weight of 0.65 with a 3-gram Gigaword Language model that is available on [9] by using the linear interpolation method. We implemented combinations with difference weights from 0.1 to 0.9 (altering was 0.5). The weight of 0.65 was the weight that gave a minimum PPL of 151 on tst2013.

The vocabulary set, obtained in the training stage of the FLM, was used to make the dictionary. The lexicon was built based on the Carnegie Mellon University (CMU) Pronouncing Dictionary v0.7a. The phoneme set contains 39 phonemes. This phoneme (or more accurately, phone) set was based on the ARPAbet symbol set.

5.2. Topic-adapted language model

FLM was firstly used to generate the first pass lattices (1st-lattice) using the acoustic models, and they were further combined to produce the first pass transcript for the tst2013 and tst2014 sets. This transcript was considered as an closed topic reference to select closed topic sentences from the our text corpora. The topic adapted language model, a 5-gram model, was constructed by using the only selectted sentences based on a cross-entropy difference metric [10] that was biased towards sentences that were both similar to the in-topic data and unlike the average of the out-topic data using XenC toolkit [11]. The sentence cross entropy was measured between two n-gram LMs, one was built by using the first pass transcript, another was built by using our corpora. The final closed topic corpus was the top 100k sentences from the scored sentences of text corpora. Three rounds of this process were performed to construct the final topic adapted model which has PPL score on the tst2014 was 86, and on the tst2013 was 113.

6. Decoding and Results

During development, we evaluated our system on the tst2013 and tst2014 set that released by the IWSLT organizers. Fig. 2 shows our complete decoding process. After feature extraction step, followed by decoding with the baseline system to estimate the transforms LDA, MLLT, and fMLLR, we operated four parallel decoding sequences for the hybrid acoustic models. For each model, the complete process consists of a decoding with the 3-gram LM applying Kaldi decoder. Lattice outputs were applied N-best list rescoring and combined to produce the first pass transcriptions which were further used as the closed topic reference for selecting closed topic sentences from our whole text corpora. The selected sentence were used for training the 5-gram topic adapted language model, and this language model was used for decoding and combination in the second pass with the same way as the first pass decoding.

Table 2 lists the performance of our system in terms of the word error rate (WER). Both tst2013 and tst2014 sets were segmented manually. As we can see on the Table, the topic adapted language model absolutely reduced a significant WER of about 2% of WER. The last row of the table shows the final combination results of the hybrid models that was 11.3% of WER.

7. Conclusions

In this paper, we presented our English LVCSR system, with which we participated in the 2016 IWSLT evaluation. The transcription was improved by improving the combination system with two more DNN based systems compared to the last year system. By applying the text selection, we got a significant improvement. This result shown that it is possible to adapt a ASR system to a new domain or topic by adapting its language model on a closed topic corpus that can be drawn from the training text corpus based on the first pass decoding results. On the tst2013, the WER of the best single system, built in last year, was reduced from 18.85% to 17.23%. On the tst2014 development set we got the best WER of 11.3% which was obtained from the combination system.

In the future, we intend to improve language model using deep neural network as in [12] as well as will apply a hybrid DNN on top of deep bottleneck features [13] to improve acoustic model for the systems.

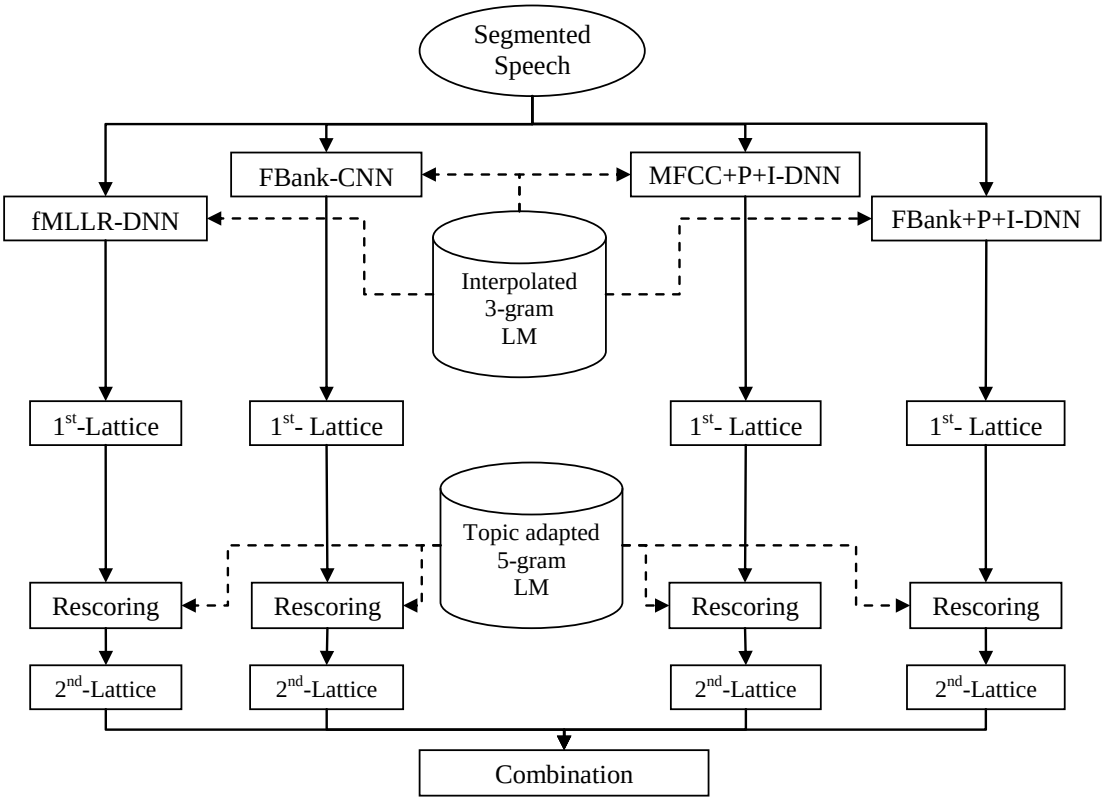

Figure 2: The decoding architecture

8. Acknowledgements

This work is partially supported by Project: "Development of spoken electronics newspaper system based on Vietnamese text-to-speech and web-based technology", VAST01.02/14-15

9. References

[1] W. D. L. Christian Federmann, "Microsoft speech language translation (mslt) corpus: The iwslt 2016 release for english, french and german," in *IWSLT*, USA, 2016.

[2] V. H. Nguyen, Q. B. Nguyen, T. T. Vu, and C. M. Luong, "The speech recognition systems of ioit for iwslt 2015," in *Proceedings of the 12th International Workshop for Spoken Language Translation (IWSLT)*, Da Nang, Vietnam, Dec-2015 2015.

[3] A. Katsamanis, M. Black, P. G. Georgiou, L. Goldstein, and S. S. Narayanan, "Sailalign: Robust long speech-text alignment," in *Proc. of Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*, jan 2011.

[4] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Acoustics, Speech and Signal Processing (ICASSP)*. South Brisbane: IEEE, 2015, pp. 5206 – 5210.

[5] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[6] Y. Miao, L. Jiang, H. Zhang, and F. Metze, "Improvements to speaker adaptive training of deep neural networks." in *IEEE Spoken Language Technology Workshop*, California and Nevada, Dec 2014.

[7] K. Vesely, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Interspeech*, Lyon, 2013.

[8] A. Stolcke, "Srilm - an extensible language modeling toolkit," in *International Symposium on Chinese Spoken Language Processing (ISCSLP)*, Hong Kong, 2012.

[9] K. Vertanen, *English Gigaword language model training recipe*, Std. [Online]. Available: https://www.keithv.com/software/giga/

[10] R. C. Moore and W. Lewis, "Intelligent selection of language model training dat," in *Association for Computational Linguistics (ACL)*, Uppsala, Sweden, July 2010, p. 220224.

[11] R. Anthony, "Xenc: An open-source tool for data selection in natural language processing," *The Prague Bulletin of Mathematical Linguistics*, no. 100, pp. 73–82, 2013.

[12] N. Q. Pham, H. S. Le, T. T. Vu, , and C. M. Luong, "The speech recognition and machine translation system of ioit for iwslt 2013," in *Proceedings of the International Workshop for Spoken Language Translation (IWSLT)*, 2013.

[13] Q. B. Nguyen, J. Gehring, K. Kilgour, and A. Waibel, "Optimizing deep bottleneck feature extraction," in *Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), 2013 IEEE RIVF International Conference on*, Nov 2013, pp. 152–156.