The IOIT English ASR system for IWSLT 2015

Van Huy Nguyen[1], *Quoc Bao Nguyen*[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, nqbao@ictu.edu.vn, {vtthang,lcmai}@ioit.ac.vn

Abstract

This paper describes the speech recognition system of IOIT for IWSLT 2015. This year, we focus on improving acoustic and language models by applying some new training techniques based on deep neural networks compared to the last year system. There are two subsystems which are combined by using lattice minimum Bayes-Risk decoding. On the 2013 evaluations set, provided as a test set, we are able to reduce the word error rate of our transcription system from 22.7% of the last year system to 17.6%.

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 Talks, which are a collection of public lectures on a variety of topics, ranging from Technology, Entertainment 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 talks, 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 TED ASR track of the IWSLT 2015 evaluation campaign. The system is a further development of our last year's evaluation system [1]. There are two hybrid acoustic models in our system. The first one is built by applying a convolutional deep neural network with the input feature of log Mel filter bank feature (FBANK). The second one is applied a feed-forward deep neural network. Its input feature is a speaker-dependent feature that is extracted by applying a feature space maximum likelihood linear regression (fMLLR) in the speaker adaptive training (SAT) stage of the baseline system. These models and an interpolated language model are used to produce decoding latices which are then used to generate the N-best lists for re-scoring.

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 [2] 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 [3]. It contains 360 hours of speech sampled at 16 kHz, and is available for training and evaluating speech recognition system.

Table 1: Traning 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, two kinds of acoustic feature are used for developing the acoustic models. The first one is a Mel-frequency Cepstral Coefficients (MFCC). A Hamming window of 25ms, which is shifted at the interval of 10ms, is applied. Each MFCC vector consists of 39 coefficients which are 13 MFCCs, the first and the second order derivatives. The second kind is a combination of a log Mel filter bank feature and a pitch feature (FBANK+P). FBANK+P consists of 43 coefficients including 40 FBANK coefficients, 1 the pitch value, the first derivative of the pitch value, and the probability of voice for the current frame. Both MFCC and FBANK+P are extracted by using the Kaldi toolkit [4][5].

4. Acoustic Model

4.1. Baseline Acoustic Model

The baseline acoustic model was built by using the Kaldi toolkit [4] 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

The hybrid Deep Neural Network and Hidden Markov Model (DNN-HMM) acoustic model were built in which the HMM models were the baseline model's HMM, and their deep neural networks were built in different architectures. Fig. 1 describes the process for training these models. The first hybrid model was applied 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 second one was applied a convolution neural network (CNN-DNN) which has 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 DNN and CNN-DNN, 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) [6] was adopted for the second stage training. The input feature for the DNN was a fMLLR-based feature that was calculated 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 is finally applied a feature space maximum likelihood linear regression (fMLLR) in the speaker adaptive training (SAT) stage. The LDA, MLLT and fMLLR transforms are estimated during the training of the baseline model. The concatenation of 11 neighbor vectors of FBANK+P, the first and the second order derivatives was used as input feature of CNN-DNN.

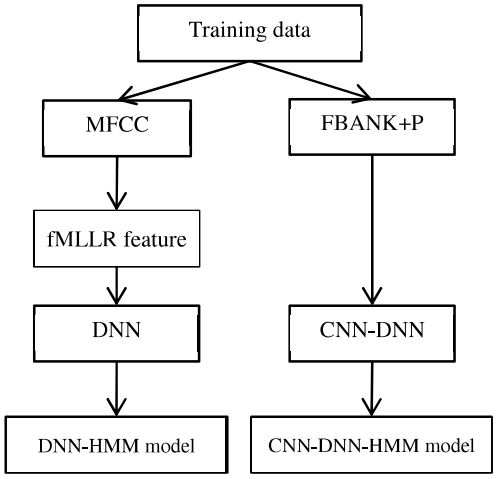

Figure 1: Training process of hybrid acoustic models

5. Language Model and Dictionary

Two categories of textual corpora was used for estimating the language model (LM) (as shown in Table 2). The first one is the transcript of Libri360 data set that was used for training the acoustic models. The second one is the subtitles of all TED talks published before June-2015 (TED2015) which is provided by Fondazione Bruno Kessler (FBK) (https://wit3.fbk.eu). TED2015 was used for training the language model after rejecting all disallowed TED talks according to the suggestion of IWSLT-2015 committee.

Table 2: Traning data for language model

Corpus	Utts
Libri360	104014
TED2015	517098

For training the language model, a vocabulary set is firstly extracted from textual sets. This vocabulary set has 73491 words and is then used to build the language model by using the SRILM toolkit [7]. The perplexity (PPL) score of the trained language model is 184 on the tst2013 test set. In order to improve the performance, it is then combined in weight of 0.65 with a 3-gram Gigaword Language model that

is available on [8] by using the linear interpolation method. We implemented combinations with difference weights from 0.1 to 0.9 (step is 0.5). The weight of 0.65 is the weight that gave a minimum PPL of 151 on tst2013.

The vocabulary set, obtained in the training stage of the language model, is used to make the dictionary. The lexicon is 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 is based on the ARPAbet symbol set.

6. Decoding Procedure and Results

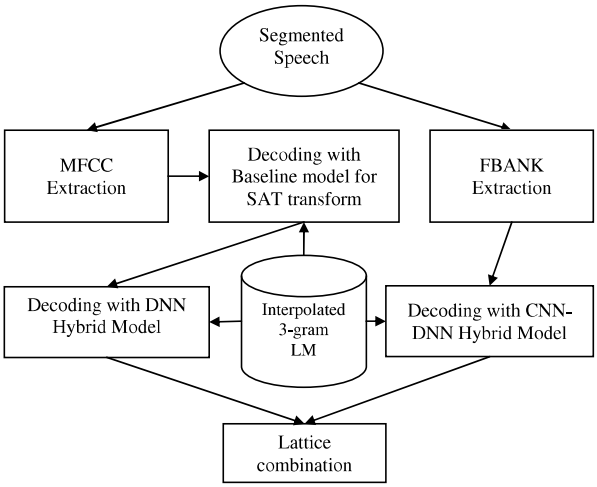

Figure 2: The full decoder architecture

During development, we evaluated our system on the tst2013 test 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 two 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 from the this pass are combined by using Lattice Minimum Bayes-Risk (MBR) decoding as described in [10].

Table 3 lists the performance of our system in terms of the word error rate (WER). Both tst2013 and dev2012 sets were segmented manually. Regarding the performance of the baseline system, the WER is 18.53% on dev2012 and 22.86% on tst2013. The first row is the number of the best system from last year [1] on the same test set. As we can see on the Table, all of our hybrid models give better WERs which are approximately 3% absolute compared to the baseline model. The last row on the table shows the final combination results of the hybrid models that give a further 1% absolute WER reduction as compared to the best single system. For this year's test set which was segmented automatically like last year system [1], we obtained 14.4% WER (about 2 % loss compared to manual segmentation).

Table 3: Experiment results

Denoted	**Model**	**WER%**	
		dev2012	**tst2013**
Last year	Combination	18.7	22.7
Baseline	MMI-SAT/HMM-GMM	18.53	22.86
S1	DNN-HMM	15.19	18.85
S2	CNN-DNN-HMM	15.81	19.30
S1+S2	Combination	14.5	17.6

7. Conclusions

In this paper, we presented our English LVCSR system, with which we participated in the 2015 IWSLT evaluation. The acoustic model was improved by using deep neural networks for this year evaluation. On the 2012 development set for the IWSLT lecture task our system achieved a WER of 14.5%, and a WER of 17.6% on the 2013 test set. The final combination model gives about 5% absolute WER reduction on tst2013 compared to the last year system.

In the future, we intend to improve language model using deep neural network as in [11] as well as will apply a hybrid DNN on top of deep bottleneck features [12] to improve acoustic model for the systems.

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] Q. B. Nguyen, T. T. Vu, and C. M. Luong, "The speech recognition systems of ioit for iwslt 2014," in *Proceedings of the 11th International Workshop for Spoken Language Translation (IWSLT)*, Lake Tahoe, USA, Dec-2014 2014.

[2] 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.

[3] 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.

[4] 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.

[5] P. Ghahremani and D. . R. K. . T. J. . K. S. BabaAli, B. ; Povey, "A pitch extraction algorithm tuned for automatic speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 2494 – 2498.

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

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

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

[9] S. Meignier and T. Merlin, "Lium spkdiarization: an open source toolkit for diarization," in *in CMU SPUD Workshop*, 2010.

[10] H. Xu, D. Povey, L. Mangu, and J. Zhu, "Minimum bayes risk decoding and system combination based on a recursion for edit distance," *Computer Speech & Language*, vol. 25, no. 4, pp. 802–828, 2011.

[11] 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.

[12] 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.