3-1-7

The NAIST ASR for IWSLT: A Multi-architecture DNN System Combination Approach [*]

☆ Michael Heck, Quoc Truong Do, Sakriani Sakti,
Graham Neubig, Satoshi Nakamura (NAIST)

1 Introduction

The International Workshop for Spoken Language Translation (IWSLT) [1] is a host for evaluation campaigns for core tasks in spoken language translation. In 2015, all tasks were performed and evaluated on TED and TEDx conference talks (`http://www.ted.com`). We present a multi-architecture deep neural net (DNN) based automatic speech recognition (ASR) system for English that competed in the track for automatic transcription of unsegmented talks. Our system [2] uses several front-ends, DNN acoustic models and n-gram and recurrent neural network (RNN) language models [3], and recognizer output voting error reduction (ROVER) system combination.

2 Overall system

2.1 Acoustic features

As standard features, we extract Mel-frequency cepstral coefficients (MFCC), perceptual linear prediction (PLP) and log Mel-filter bank (FBANK) feature vectors for every 10 ms and a window length of 25 ms. Additionally, we also adopt i-vector features to capture speaker and channel information.

2.2 Acoustic model training

We trained the following models using the Kaldi speech recognition toolkit [4]:

GMM-HMM Speaker adaptively trained GMM-HMM models serve the purpose of producing fMLLR transforms as basis for the sigmoid DNN.

Sigmoid DNN The sigmoid activation function is applied in each hidden layer, and the softmax function is applied in the output layer. The input features are generated by linear discriminant analysis (LDA) + maximum likelihood linear transform (MLLT) + feature-space maximum likelihood linear regression (fMLLR) on top of spliced standard features and another splicing prior to DNN input. We performed pre-training with a restricted Boltzmann machine (RBM) deep belief network. The DNN was trained with the frame cross-entropy (CE) criterion.

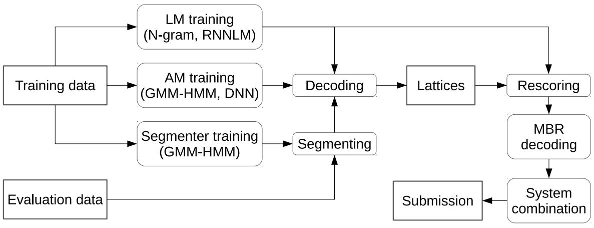

Fig. 1 The multi-architecture framework.

Rectified linear unit DNN The Rectified linear unit (ReLU) activation function is applied in each hidden layer. The input features are a standard feature vector and an i-vector stacked on top. For pre-training we train for a fixed number of epochs and average model parameters over the last few epochs of training. The parameters are optimized according to the frame CE criterion.

p-norm DNN The p-norm [5] is a “dimension-reducing” non-linearity that is inspired by maxout. The input features are the same as for the ReLU DNN. We trained two models, one by using frame CE, the other according to the state-level minimum Bayes risk (sMBR) criterion [6].

2.3 Decoding strategy

We run the decoding with a pruned 4-gram language model. Subsequent lattice rescorings make use of a full 5-gram language model and an RNNLM language model. Given the lattices, we apply minimum Bayes risk (MBR) [7] decoding for all systems to minimize the expected word error rate (WER). After rescoring, we perform system combination using ROVER.

3 Data resources

3.1 Acoustic model training data

The data we used for training acoustic models is selected from various resources including TED-LIUM corpus release 2 [8], Broadcast News, WSJ, and TIMIT. The data sums up to 439.08 h of labelled audio.

3.2 Language model training data

The data for training language models comes from different sources including WSJ, EUROPARL, GIGA, NC, and TED. The data is cleaned by removing all punctuation, and removing case sensitivity by uppercasing all characters. The training set contains 620.35 M tokens.

[*] IWSLT2015 における NAIST 音声認識システム：複数のアーキテクチャを持つ DNN の統合

3.3 Evaluation data

An official development set was provided by the organizers. Automatic segmentation of the raw audio data prior to decoding is a mandatory sub-task. We picked a GMM-based segmentation approach that uses a Viterbi decoder and GMM-HMM models to classify consecutively observed feature vectors into speech and non-speech.

4 ASR evaluation

4.1 Single system performance

We observed that PLP features generally led to the best performance, followed by MFCC features, and that there is a fairly large gap between the MFCC and FBANK features. The results also show that DNNs that use the p-norm activation function exceed the other nets' classification performance. Ultimately, the sMBR training criterion led to better accuracy than the commonly used cross-entropy criterion. We assume that the main cause for the comparably low performance of the sigmoid DNN is the usage of standard features only, without the i-vectors stacked on top. This matches our observations in [9], where we used the same layout for all DNNs but varying input features and observed a significant difference in system performance.

4.2 System combination performance

We ensured that the system combination covered all three front-ends by choosing the systems accordingly. Experiments confirmed that failing to do so leads to sub-optimal combinations that do not even manage to beat the single best system. The best ROVER combination made use of three systems and achieved a WER of 9.5% on the ground truth segmentation and 10.1% on the automatic segmentation. The design of this final setup is depicted in Fig. 2.

5 Conclusion

This paper described the structure and development of our multi-architecture system combination approach for the English ASR track of the IWSLT 2015 evaluation campaign. We incorporated various DNN acoustic models for the input feature types MFCC, PLP, FBANK and i-vector. Our results show that a p-norm DNN trained on combined MFCC + i-vector feature vectors using the sMBR training criterion achieves the best single system performance. System combination with ROVER, where the outputs of the best systems for each front-end were combined, further increases recognition performance and yields a WER of 10.1% given the automatic segmentation.

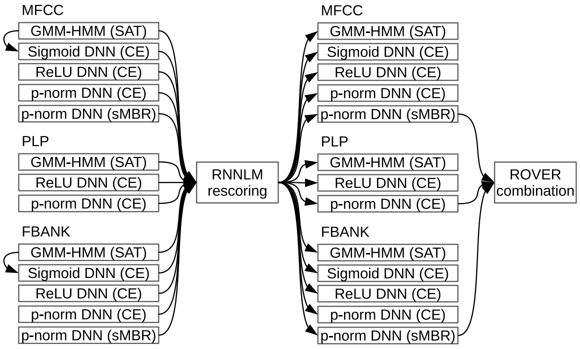

Fig. 2 Decoding pipeline and model selection for ROVER.

Models		Features		
Function	Criterion	MFCC	PLP	FBANK
Sigmoid	CE	14.4%	-	-
ReLU	CE	11.2%	10.9%	12.7%
p-norm	CE	10.8%	10.5%	12.6%
p-norm	sMBR	**9.8%**	-	11.2%

Table 1 Single system performances in WER after RNNLM rescoring.

6 Acknowledgements

Part of this research was supported by JSPS KAKENHI Grant Number 24240032 and 26870371, and the Commissioned Research of National Institute of Information and Communications Technology (NICT), Japan.

References

[1] M. Cettolo, J. Niehues, S. Stüker, L. Bentivogli, R. Cattoni, and M. Federico. The IWSLT 2015 evaluation campaign. In *Proceedings of IWSLT*, 2015.

[2] M. Heck, Q. T. Do, S. Sakti, G. Neubig, and S. Nakamura. The NAIST english speech recognition system for IWSLT 2015. In *Proceedings of IWSLT*, 2015.

[3] T. Mikolov, S. Kombrink, L. Burget, J. Cernocky, and S. Khudanpur. Extensions of recurrent neural network language model. In *Proceedings of ICASSP*, pages 5528–5531, 2011.

[4] D. Povey, A. Ghoshal, G. Boulianne, N. Goel, M. Hannemann, Y. Qian, P. Schwarz, and G. Stemmer. The Kaldi speech recognition toolkit. In *Proceedings of IEEE*, 2011.

[5] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur. Improving deep neural network acoustic models using generalized maxout networks. In *Proceedings of ICASSP*, pages 215–219, May 2014.

[6] D. Povey and B. Kingsbury. Evaluation of proposed modifications to MPE for large scale discriminative training. In *Proceedings of ICASSP*, volume 4, pages IV–321–IV–324, April 2007.

[7] 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*, 25(4):802 – 828, 2011.

[8] A. Rousseau, P. Delglise, and Y. Estve. Enhancing the TED-LIUM corpus with selected data for language modeling and more TED talks. In *Proceedings of LREC*, 2014.

[9] Q. T. Do, M. Heck, S. Sakti, G. Neubig, T. Toda, and S. Nakamura. The NAIST ASR system for the 2015 multi-genre broadcast challenge: On combination of deep learning systems using a rank-score function. In *Proceedings of ASRU*, 2015.