DNN Online Adaptation for Automatic Speech Recognition

Xinwei Li, Yue Pan, Daniel Willett, Puming Zhan
Nuance Communications, Inc.

Deep Neural Networks (DNN) techniques are widely used for acoustic modeling in the state-of-the-art ASR systems nowadays [1,2]. Although DNN-HMM based ASR systems can provide better accuracy than that of GMM-HMM based ASR systems in general, they are still pretty sensitive to speaker, channel, and background changes. This is reflected by large variance in accuracy across different speakers, channels, and environments that we have observed. Online adaptation is a very effective way for making an ASR system more robust under various environments and speaker characteristics. It has been widely used in HMM-GMM based ASR systems. However, it is very challenging for doing DNN online adaptation effectively, because a DNN acoustic model usually contains millions of parameters and adapting them w/ very small amount of data (i.e. single sentence) in a traditional way is usually not effective for improving accuracy. On the other hand, online adaptation is subject to strict constraint on computing cost because of latency requirement. In this paper, we propose two methods, namely i-vector and Linear Hidden Network (LHN), for doing DNN online adaptation and demonstrate that they each can provide significant accuracy improvement. We will also show that some combinations of the two methods can provide extra accuracy improvement.

I-vector is widely used for speaker verification and speaker recognition [4]. The work we present in this paper follows the similar way described in [5] for using i-vector to do speaker adaptation for DNN acoustic models. However, we focus on doing online adaptation, instead of offline batch-mode adaptation as in [5]. Our i-vector based online adaptation is carried out in a per sentence basis. That is to calculate i-vector with the data from the current sentence and use it in recognizing the next sentence. In order to make such online adaptation effective and affordable, we conducted research in a). i-vector initialization; b). i-vector and sufficient statistics carryover from one sentence to the next; c). i-vector normalization and quantization.

DNN adaptation via linear transformation has been proposed long time ago [6,7]. The general idea is to insert a linear layer into a trained DNN. Such linear layer can be placed either at the input of the DNN [6] or right before the output layer of the DNN [7]. Adaptation is done by training this newly added layer with the adaptation data via back-propagation algorithm. Such method only works well for offline batch-mode adaptation, because the adaptation process is too expensive and the amount of data is usually too small in online situation. To overcome the impact of data sparsity, [8] proposed a regularization method via KL-Divergence. Our main focus is on how to make LHN effective and affordable for online adaptation. Along this direction, we propose two methods, frame skipping and DNN internal layer feature caching for reducing the computational cost in this paper.

We conducted adaptation experiments based on a voice search data set which contains over three thousand hours of audio data. We trained large DNN acoustic models on this data set and evaluated performance on various testing scenarios. We observed over 3% relative word error reduction from the i-vector online adaptation and the LHN online adaptation individually. A further improvement of over 2% was observed from combining i-vector with LHN adaptation.

[1] F. Seide, G. Li, X. Chen, and D. Yu, “Feature engineering in context-dependent deep neural networks for conversational speech transcription,” in Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on, Dec 2011, pp. 24– 29.

[2] G.E.Dahl,D.Yu,L.Deng,andA.Acero, “Context-dependent pre-trained deep neural networks for large vocabulary speech recognition,” IEEE Trans. on Audio, Speech, and Language Processing, vol. 20, no. 1, pp. 30–42, 2012

[4] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Frontend factor analysis for speaker verification,” IEEE Trans. Audio, Speech and Language Processing, vol. 19, no. 4, May 2011

[5] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, “Speaker adaptation of neural network acoustic models using i-vectors,” in Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on, Dec 2013, pp. 55–59.

[6] J.Netoetal, “Speaker-adaptation for hybrid HMM-ANN continuous speech recognition system,” in EUROSPEECH, 1995.

[7] R. Gemello, F. Mana, S. Scanzio, P. Laface, and R. D. Mori, “Adaptation of hybrid ANN/HMM models using linear hidden transformations and conservative training,” in ICASSP, 2006, pp. 1189–1192.

[8]K. Yao, D. Yu, F. Seide, H. Su, L. Deng, and Y. Gong, “Adaptation of context-dependent deep neural networks for automatic speech recognition,” in Proc. SLT, 2012.

[9] D. Yu, K. Yao, H. Su, G. Li, and F. Seide, “KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition,” in ICASSP, May 2013, pp. 7893–7897.