Multi-Accent Deep Neural Network Acoustic Model with Accent-Specific Top Layer Using the KLD-Regularized Model Adaptation

Yan Huang, Dong Yu, Chaojun Liu, and Yifan Gong

Microsoft Corporation, One Microsoft Way, Redmond, WA 98052

{yanhuang; dongyu; chaojunl; ygong}@microsoft.com

Abstract

We propose a multi-accent deep neural network acoustic model with an accent-specific top layer and shared bottom hidden layers. The accent-specific top layer is used to model the distinct accent specific patterns. The shared bottom hidden layers allow maximum knowledge sharing between the native and the accent models. This design is particularly attractive when considering deploying such a system to a live speech service due to its computational efficiency. We applied the KL-divergence (KLD) regularized model adaptation to train the accent-specific top layer. On the mobile short message dictation task (SMD), with 1K, 10K, and 100K British or Indian accent adaptation utterances, the proposed approach achieves 18.1%, 26.0%, and 28.5% or 16.1%, 25.4%, and 30.6% word error rate reduction (WERR) for the British and the Indian accent respectively against a baseline cross entropy (CE) model trained from 400 hour data. On the 100K utterance accent adaptation setup, comparable performance gain can be obtained against a baseline CE model trained with 2000 hour data. We observe smaller yet significant WER reduction on a baseline model trained using the MMI sequence-level criterion.

Index Terms: Accent speech recognition, model adaptation, KL-divergence regularization

1. Introduction

Speech with foreign accent can largely degrade the intelligibility and result in poor automatic speech recognition (ASR) performance [2, 3]. The deep learning acoustic model technology [4, 5, 6] can help improve the foreign-accented-speech ASR performance due to its layer-by-layer invariant and selective feature extraction [7, 8]. Nevertheless, the performance gap between the native and the foreign-accented-speech remains large in the deep neural network (DNN) acoustic model.

The accented-speech is usually perceived as an interpolation of the native and the target language. The degree of the accentedness depends on many factors, such as the language competency, the education background, the articulation habit, among others.

Previously, much work has been conducted in the area of the accented-speech ASR, which can be roughly categorized into the model adaptation approach [9, 10, 11] and the lexicon adaptation approach [12]. The model adaptation approach is typically found to be more effective than the lexicon adaptation [2]. In this paper, we mainly focus on the model adaptation solution.

We propose an efficient and effective multi-accent deep neural network with an accent-specific top layer and shared bottom hidden layers. The accent-specific top layer is used to model the distinct accent specific patterns distilled from small amount of accent speech. We adopted the model adaptation technique and applied the KL-divergence (KLD) regularized deep neural network model adaptation methodology [1] to train the accent-specific top layer.

We found that with limited amount of accent adaptation data, conducting the model adaptation on the full neural net does not necessarily always result in the best performance. Constraining the adaptation to the top layers, as one way of regularization, can yield competitive adaptation performance. More importantly, this design is appealing when considering deploying such a multi-accent model in a live speech service due to its computational efficiency.

On a mobile short message dictation task (SMD), with 1K, 10K, and 100K British or Indian accent adaptation utterances, 18.1%, 26.0%, and 28.5% or 16.1%, 25.4%, and 30.6% word error rate reduction (WERR) for the British or the Indian accent respectively against a 400 hour cross entropy (CE) model. Comparable performance gain can be obtained against a baseline CE model trained from 2000 hour data. We observe smaller performance gain (19.4% and 16.8% WERRs for the British or the Indian accent respectively with the 100K utterance adaptation setup) on a baseline model trained using the MMI sequence-level criterion.

The remainder of this paper is organized as follows: Section 2 compares the speech recognition performance of the native and the foreign-accented speech on different acoustic models; Section 3 presents our proposed multi-accent model framework; Section 4 reviews the KLD-regularized model adaptation methodology; Section 5 presents the experiments and results; Section 6 concludes this paper.

2. ASR Performance of the Native and the Foreign-accented Speech

We conducted a comparative study on the ASR performance between the native and the foreign-accented-speech using the mobile SMD task. In particular, we would like to find out how the generally practiced model improvement methodologies, such as applying better modeling techniques, increasing the training data size, and enlarging the model capacity, can help improve the native and the accented-speech ASR performance.

A set of context-dependent Gaussian mixture hidden Markov model (GMM) and deep neural network hidden Markov models (DNN) were trained for this study. The GMM is a discriminative model trained with the feature-space minimum phone error rate (fMPE) [15] and the boosted MMI (bMMI) [14]. The DNNs are the deep neural network models trained using the cross entropy (CE) [4] or the sequence-level

MMI criterion (SE) [16]. All three models (GMM, DNN.CE, DNN.SE) share the 400 hour training data and use the same decision tree with 6000 tied senone states. The last model (DNN.CE.2K) is a cross entropy DNN with 9000 tied senone states trained from 2000 hour data including 1600 hour data with automatically inferred transcription.

The testing material consists of a native test set (en-US), a British-accent test set (en-BR), and an Indian-accent test set (en-IN), all collected from our mobile SMD application. We evaluated and compared the ASR performance of the native and the foreign-accented-speech against GMM, DNN.CE, DNN.SE, and DNN.SE.2K with results summarized in Table 1:

Table 1: *Performance comparison of the native en-US, the British accent, and the Indian accent against an fMPE+bMMI model (GMM) and three DNNs (DNN.CE, DNN.SE, and DNN.CE.2K).*

Model	en-US (%)	en-BR (%)	en-IN (%)
GMM	21.4	42.7	52.1
DNN.CE	16.2	34.4	48.4
DNN.SE	13.8	29.3	40.6
DNN.CE.2K	13.8	30.3	40.0

- Comparing GMM to DNN.CE or DNN.CE to DNN.SE, we can see that the improved modeling techniques yield significant WER reduction on both the native and the accented-speech. The native and the foreign-accented-speech share common error patterns which could be corrected by the improved modeling techniques; Moreover, the deep learning technique may extract better invariant features and thus be more robust to accent variation.
- Comparing DNN.CE.2K to DNN.CE, we can see that the increased training data size and the enlarged model capacity also results in accuracy improvement both for the native and the non-native speech. It is interesting to observe that the performance gain due to the better sequence-level MMI criterion roughly equals to the gain obtained from adding four times more semi-supervised data in combination with the enlarged model capacity.

In summary, the model learning process can be shared among the native and the accented-speech modeling. The generally practiced model improvement methodologies can improve both the native and the accented-speech ASR. Nevertheless, as the model performance is further improved, one needs to introduce the accent-specific modeling strategies to continue improve the accented-speech ASR performance.

The more than doubled WER on the accented-speech as comparing to the native speech shown in Table 1 indicates that it is imperative to improve the non-native speech ASR accuracy for better overall user experience in the mobile SMD tasks.

3. Multi-accent Deep Neural Network with Accent-Specific Top Layers

The main challenges of solving the accented-speech ASR in the context of a practical speech service are the potentially significant increase in the run-time cost and the usually limited amount of available accent training data. Our proposed multi-accent deep neural network directly addresses these two challenges. It consists of an accent-specific top layer and shared accent-independent bottom hidden layers as depicted in Figure 1.

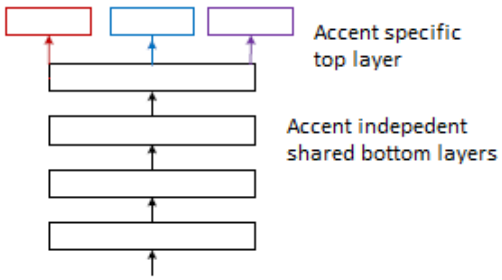

Figure 1: *Multi-accent deep neural network framework with an accent-specific top layer and shared bottom layers.*

The shared accent-independent bottom layers allow maximal data sharing and knowledge transfer between the accented-speech and the native speech. This is especially important when only small amount of accented-speech data is available. The shared hidden layers can be viewed as a type of regularization.

This design is also particularly appealing on the run-time cost for a practical deployment system. The hidden layer computation can be shared across different accent origins during the decoding. In the multi-model one-pass decoding design, we only need to evaluate the top layer separately for each accent origin with the bottom layer calculation shared; in the two-pass decoding design, the bottom layer evaluation can also be reused during the second pass when the accent-specific model is activated following the accent identification. Here we suggest treating the proposed multi-accent model as one single model with split top layers instead of multiple models.

This approach is related to the previous multi-lingual deep neural network work [13], but with different assumptions. In this work, only small amount of data (e.g. a few thousand utterances) is available for learning the accent-specific information, while the previous multi-lingual work [13] assumes the availability of much larger amount of multi-lingual data for each language. Therefore, we propose to use the KL-divergence regularized model adaptation methodology [1] to train the accent-specific top layer to avoid overfitting which will be discussed in the next section.

4. Review of the KLD-Regularized Model Adaptation

The KLD-regularized model adaptation was first proposed in [1]. In this methodology, an additional term that measures the KL-divergence of the base model $p^{SI}(y|x_t)$ and the adapted model $p(y|x_t)$ is added to the standard cross entropy objective function $\bar{D}$ to regularize the adaptation model.

This formulation prevents the model from drifting too far away from the base model due to overfitting as depicted in Eq.(1). Here x_t denotes the t-th input sample, y denotes the output label, N is the total number of samples, S is the total number of senones, and ρ is the regularization weight.

$$\hat{D} = (1-\rho)\bar{D} + \rho\frac{1}{N}\sum_{t=1}^{N}\sum_{y=1}^{S} p^{SI}(y|x_t)\log p(y|x_t). \quad (1)$$

An excellent property of this formulation is that it is equivalent to replace the original 0/1 target with the soft target $\hat{p}(X|Y_t)$ calculated as a linear interpolation of the posterior

$p^{SI}(y|x_t)$ estimated from the base model and the standard 0/1 target $\tilde{p}(y|x_t)$ determined by the senone state alignment:

$$\hat{p}(X|Y_t) = (1-\rho)\tilde{p}(y|x_t) + \rho p^{SI}(y|x_t). \quad (2)$$

The model training can simply proceed as the standard CE training with no need to change the learning machine. This formulation can be applied to adapt the full neural net as in [1] or to adapt the specified layers. In this paper, we only adapt the top layer while keeping the hidden layers fixed. The analysis on the regularization weight and model adaptation capacity will be discussed in Section 5.1.

5. Experiments and Results

In this section, we present our experimental results on the multi-accent DNN with an accent-specific top layer using the KLD-regularized model adaptation.

We first study the model adaptation overfitting behavior with respect to the regularization weight and the model adaptation capacity; then present the accent model adaptation performance with 1K, 10K, or 100K British accent or Indian accent data; last, we investigate the accent model adaptation performance with respect to different quality baseline models. All experiments were conducted using the mobile SMD task with similar experimental setup as in Section 2.

5.1. Regularization and Adaptation Capacity

Starting from the baseline DNN.CE, we conducted the KLD-regularized accent model adaptation using 10K British accent utterances. The adaptation was configured with the top 1, 2, 4, or 6 (the full net) layers to be adapted. The regularization weight was set to 0.5 or 0.3. Fig. 2 depicts the frame accuracy change as the training progresses.

Fig. 3 presents the corresponding model performance evaluated on the British accent test set (only for the regularization weight set to 0.3). The native speech test results are also included for comparison. We make the following observation by studying Fig. 2 and Fig. 3 together:

- Higher frame accuracy on the training set can be achieved when allowing more layers to be adapted or setting the regularization term to a smaller value.
- The higher frame accuracy achieved by allowing more layers to be adapted does not always suggest a better performed model. For example, conducting the model adaptation on the full net does not yield better performance gain comparing to adapt only the top layer. When the full net is adapted, the model converges faster and quickly starts exhibiting overfitting.
- With 10K British accent adaptation utterances, the best performance is achieved when adapting the top two layers. Comparable performance can be achieved by adapting only the top one layer.
- A contrastive performance pattern can be observed between the native and the accented speech in Fig. 3. An adaptation setting resulting in better accent adaptation performance yields a model with worse performance on the native speech. As the learning proceeds, the accent adapted model performs increasingly better on the accented speech while progressively worse on the native speech with roughly the same *tempo*.

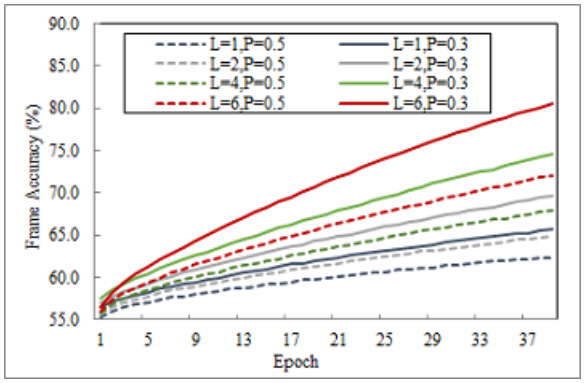

Figure 2: *The frame accuracy change with respect to the training epochs in the 10K utterance British accent adaptation. L: Number of layers adapted; P: Regularization weight.*

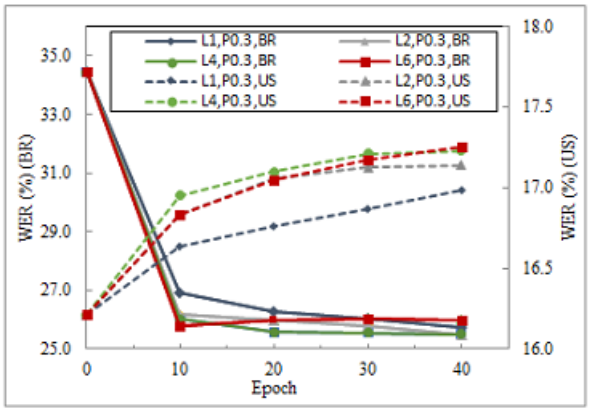

Figure 3: *Model adaptation performance with respect to the training epochs in the 10K utterance British accent adaptation. L: Number of layers adapted; P: Regularization weight; BR: British-Accent; US: Native US-English.*

This set of experiments explored the interrelationship between the model adaptation capacity, regularization, and overfitting. Allowing more parameters to be adapted or setting a smaller regularization weight can potentially result in overfitting. With limited amount of data, *regularizing* the model by restricting the adaptation only on the top layer is both effective and efficient. For the rest of this section, all accent model adaption experiments were conducted on the top layer unless otherwise specified.

Lastly, the contrastive performance pattern on the native and the accented-speech further verified that the model adaptation procedure can extract the distinct accent specific information and move the model towards the target accent.

5.2. Amount of Accent Adaptation Data

This series of experiments studies how the accent model adaptation performance is affected by different amount of adaptation data. Starting with the baseline DNN.CE as before, we conducted the accent model adaptation using 1K, 10K, and 100K British or Indian accent utterances. The regularization weight was set to 0.3 for the 1K and 10K, 0.1 for the 100K utterance adaptation. The resulting models were evaluated using the British accent or the Indian accent test sets with results summarized in Table 2:

- With 1K British or Indian accent adaptation utterances, 18.1% or 16.1% WERRs were achieved for the British or the Indian accent respectively. Distinct accent specific information can be distilled and effectively modeled using the accent-specific top layer with only 1K utterances.
- As the adaptation data increases to 10K utterances, the WERRs increase to 26.0% or 25.4%. More adaptation data resulted in significantly larger performance gain.
- Further increasing the adaptation data to 100K utterances, the resulting models yield 28.5% or 30.6% WERRs for the British or the Indian accent. The additional performance gain is smaller as the adaptation data increases from 10K to 100K.

We also experimented with the full net model adaptation on the 100K utterance setup and found only small extra performance gain can be achieved comparing to only adapt the top layer. This suggests that the top layer effectively models the accent variance with sufficient accent adaptation capacity.

In summary, increasing the model adaptation data can achieve better model adaptation performance. The gain becomes smaller with sufficient amount of adaptation data.

Table 2: *The WERs and WERRs (in parentheses) of the KLD-regularized top-layer adapted accent model adaptation with 1K, 10K, or 100K British or Indian accent adaptation utterances. The baseline model is DNN.CE.*

Model (DNN.)	EN-US (%)	EN-BR(%)	EN-IN(%)
CE (Baseline)	16.2	34.4	48.4
CE.Adapt (1K)		28.2 (18.1)	40.6 (16.1)
CE.Adapt (10K)		25.5 (26.0)	36.1 (25.4)
CE.Adapt (100K)		24.6 (28.5)	33.6 (30.6)

5.3. Baseline Model Quality and Training Criterion

This series of experiments investigate how the quality of the baseline models, e.g., those trained with the MMI sequence-discriminative criterion or with significantly more training data and the enlarged model capacity, affect the accent adaptation performance.

We conducted accent model adaptation starting from the three baseline models CE.DNN, SE.DNN, and CE.DNN.2K with increasingly better accuracy using 100K British or Indian accent adaptation utterances with results summarized in Table 3.

DNN.CE and DNN.CE.2K were both trained using the cross-entropy criterion and the latter has better accuracy due to the increased training data amount and the enlarged model size. When conducting accent adaptation from these two models, we observe 28.5% and 31.7% WERRs on the British accent or 30.6% and 23.0% WERRs on the Indian accent respectively. All these WERRs are quite significant albeit the baseline DNN models were very well trained.

DNN.SE was trained using the MMI sequence-discriminative training criterion with comparable performance with DNN.CE.2K. When performing accent adaptation from DNN.SE, we observe notably smaller WER reduction as comparing to adapting from DNN.CE.2K, e.g. 19.4% versus 31.7% for the British accent or 16.8% versus 31.7% WERRs for the Indian accent. As a result, we obtain lower WER when conducting model adaptation from DNN.CE.2K even though DNN.CE.2k and DNN.SE have almost identical performance.

The gap in the accent model adaptation WER reduction is even larger when comparing to adapting from DNN.CE, which differs from DNN.SE only in the model training criterion. As a result, we obtain almost identical WER even though the baseline DNN.SE is a significantly better performed model comparing to DNN.CE.

We think the primary reason is that SE-DNN was trained using the sequence-level criterion while our current KLD-regularized model adaptation uses the cross entropy criterion. Applying the cross entropy based model adaptation on top of a model trained with the sequence-discriminative criterion could be less effective. We are currently working on extending the KLD-regularized model adaptation to the MMI sequence-discriminative training to further improve the model adaptation performance.

Table 3: *The performance of the top-layer adapted accent models adapted from three different baseline models DNN.CE, DNN.SE, and DNN.CE.2K using the KLD-regularized adaptation measured by WER and WERRs (in parentheses). 100K British accented or Indian accented adaptation utterances were used in this series of experiments.*

Model (DNN.)	EN-US (%)	EN-BR(%)	EN-IN(%)
CE (Baseline)	16.2	34.4	48.4
CE.Adapt		24.6 (28.5)	33.6 (30.6)
CE.2K (Baseline)	13.8	30.3	40.0
CE.2K.Adapt		20.7 (31.7)	30.8 (23.0)
SE (Baseline)	13.8	29.3	40.6
SE.Adapt		23.6 (19.4)	33.8 (16.8)

6. Conclusion

In summary, we presented a multi-accent deep neural network with an accent-specific top layer and shared bottom hidden layers using the KL-divergence regularized model adaptation. The accent-specific top layer is used to model the distinct accent specific features, while the shared bottom layers allow maximum share and knowledge transfer with the native speech. This approach is also practically appealing due to its computational efficiency when considering deploying such a multi-accent model to live speech services. We applied the KLD-regularized model adaptation to train the accent-specific top layer.

On the mobile short message dictation task (SMD), with 1K, 10K, or 100K British or Indian accented adaptation utterances, the proposed approach achieves 18.1%, 26.0%, and 28.5% or 16.1%, 25.4%, and 30.6% WERRs for the British or the Indian accent respectively over a 400 hour baseline cross entropy (CE) model. Comparable performance gain can be obtained from a baseline CE model trained with 2000 hour data. We observe smaller performance gain on the baseline model trained using the MMI sequence-discriminative criterion.

Our ongoing research topics include the neuron selective accent model adaptation based on the neuron firing footprint, the regularization weight adjustable model adaptation, and the sequential accent model adaptation with regularization. Furthermore, we are investigating an alternative accent invariant deep learning methodology to normalize the accent distinct aspects of the speech using the additional neural network input which encode the accent.

7. References

[1] Yu, D., Yao, K., Su, H., Li, G., and Seide, F., "KL-Divergence Regularized Deep Neural Network Adaptation for Improved Large Vocabulary Speech Recognition", in the Proceedings of the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013.

[2] Huang, C., Chen, T., and Chang, E., "Accent Issues in Large Vocabulary Speech Recognition", in the International Journal of Speech Technology , vol. 7, no. 2, pp. 141-153, 2004.

[3] Wang, Z., Schultz, T., and Waibel, A.,"Comparison of Acoustic Model Adaptation Techniques on Non-native Speech", in the Proceedings of the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013.

[4] Dahl, G.E., Yu, D., Deng, L., and Acero, A., "Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition", IEEE Transactions on Audio, Speech, and Language Processing (TASLP) - Special Issue on Deep Learning for Spech and Language Processing, Volume: 1, No. 1, Page(s): 33-42, Jan 2012.

[5] Seide, F., Li, G., and Yu, D., "Conversational Speech Transcription Using Context-Dependent Deep Neural Networks", in the Proceedings of Interspeech 2012.

[6] Kingsbury, B., Sainath, N. T., and Soltau, H., "Scalable Minimum Bayes Risk Training of Deep Neural Network Acoustic Models Using Distributed Hessian-free Optimization", in the Proceedings of Interspeech 2012.

[7] Yu, D., Seltzer, M., Li, J. ,Huang, J., Seide, F., "Feature Learning in Deep Neural Networks - Studies on Speech Recognition Tasks", in the Proceedings of 2013 International Confernece on Learning Representation, 2013.

[8] Goodfellow, I. J., Le, Q. V., Saxe, A. M., Lee, H., and Ng, A. Y., "Measuring Invariances in Deep Networks", Advances in Neural Information Processing Systems (NIPS) 22, 2009.

[9] Vergyri, D., Lamel, L.,and Gauvain, L., "Automatic Speech Recognition of Multiple Accented English Data", in the Proceedings of the 2010 Interspeech conference, 2010.

[10] Yanli Zheng, Y., Sproat, R., Gu L., Shafran, I., Zhou, H., Su, Y., Jurafsky, D., Starr R., and Yoon, S., "Accent Detection and Speech Recognition for Shanghai-accented Mandarin", in the Proceedings of the 2005 Interspeech conference, 2005.

[11] Nallasamy, U., Metze, F., and Schultz, T., "Enhanced Polyphone Decision Tree Adaptation for accented-speech Recognition, in the Proceedings of the 2012 Interspeech conference, 2012.

[12] Arslan, L. M. and Hansen, J. L., "A Study of the Temporal Features and Frequency Characteristics in American English Foreign Accent, Journal of the Acoustic Society, America, December, 1996.

[13] Huang, J., Li, J., Yu, D., Deng, L., and Gong, Y., "Cross-Language Knowledge Transfer Using Multilingual Deep Neural Network With Shared Hidden Layers", in the Proceedings of the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013.

[14] Povey, D, Kingsbury, B., Ramabhadran, B., Saon, G., Soltau H., and Visweswariah, K., "Boosted MMI for model and feature-space discriminative training", in the Proceedings of the 2008 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2008.

[15] Povey, D, Kingsbury, B., Mangu, L., Saon, G., Soltau, H., and Zweig, G., "fMPE: Discriminatively Trained Features for Speech Recognition", in the Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2005.

[16] Su, H., Li, G., Yu, D., and Seide, F., "Error Back Propagation For Sequence Training Of Context-Dependent Deep Networks For Conversational Speech Transcription", in the Proceedings of the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013.