Improving DNN-Based Automatic Recognition of Non-native Children's Speech with Adult Speech

Yao Qian Xinhao Wang Keelan Evanini David Suendermann-Oeft

Educational Testing Service Research, USA

{yqian, xwang002, kevanini, suendermann-oeft}@ets.org

Abstract

Acoustic models for state-of-the-art DNN-based speech recognition systems are typically trained using at least several hundred hours of task-specific training data. However, this amount of training data is not always available for some applications. In this paper, we investigate how to use an adult speech corpus to improve DNN-based automatic speech recognition for non-native children's speech. Although there are many acoustic and linguistic mismatches between the speech of adults and children, adult speech can still be used to boost the performance of a speech recognizer for children using acoustic modeling techniques based on the DNN framework. The experimental results show that the best recognition performance can be achieved by combining children's training data with adult training data of approximately the same size and initializing the DNN with the weights obtained by pre-training using the full training set of the adult corpus. This system can outperform the baseline system trained on only children's speech with an overall relative WER reduction of 11.9%. Among the three speaking tasks studied, the picture narration task shows the largest gain with a WER reduction from 24.6 % to 20.1%.

Index Terms: speech recognition, non-native child spontaneous speech, DNN and i-vector

1. Introduction

Globalization has been accompanied by a significantly growing demand for English language proficiency and learning English as a foreign language in primary school is mandatory in many European countries. In many Asian countries, children also begin learning English in the primary grades. This large and growing population of young English learners has increased the demand for automated speech processing technology to facilitate the scoring of their spoken responses. Employing automated speech scoring for this population would both reduce the burden on the teachers and also provide a mechanism for real-time feedback to young learners when teachers are not available.

Automated assessment of several aspects of spoken language proficiency, including vocabulary, grammar, content appropriateness, and discourse coherence, depends heavily on how accurately the input speech can be recognized. While state-of-the-art acoustic models based on deep neural networks have significantly improved recognition performance of native speech, accurate recognition results are still challenging to obtain when the input is non-native spontaneous speech. This is due, in large part, to the fact that non-native spoken responses tend to contain substantially higher amounts of pronunciation errors, disfluencies, ungrammatical phrases, etc. [1].

Recognizing non-native children's speech is even more challenging. Conventional speech recognition systems that are modelled on adult data typically do not perform well on children's speech [2]. Many acoustic and linguistic variations are observed between adult speech and children's speech. Acoustically, shorter vocal tracts and smaller vocal folds in children lead to higher fundamental frequencies and formant frequencies than for adults [3-5]. In addition, children's speaking rates tend to be slower and more variable overall due to the fact that their articulators have not fully developed yet [6]. Linguistically, children's choices of vocabulary and syntax tend to differ from adult patterns, and children may use greater amounts of imaginative words and ungrammatical phrases [7]. The differences between non-native speech produced by adults and children lies more in the acoustic aspects than the linguistic aspects since similar linguistic errors tend to exist in both non-native adult speech and children's.

In this paper, we explore how to build an automatic speech recognizer for non-native children's speech using a corpus of children's speech of moderate size assisted by a large corpus of adult speech.

2. Previous Research

Many approaches have been tried to improve automatic speech recognition for children's speech. Most approaches have addressed the acoustic differences between adult and children's speech by vocal tract length normalization (VTLN), which aims to compensate for the fact that speakers have vocal tracts of different sizes. It is generally implemented by warping the frequency axis of one spectrum by expanding or compressing in different regions to minimize the distance to a canonical spectrum, typically towards a global average vocal tract length [8]. This warping is usually employed before both training and recognition. The warping functions can be linear, piecewise linear, bilinear, or nonlinear, and the corresponding warping factor, generally a single scalar parameter for each speaker, is usually estimated by a small amount of data. VTLN is one of the easiest ways of doing fast speaker adaptation and its effectiveness for improving recognition performance of children's speech has been reported in [7, 9,10]. However, when a DNN-HMM framework is used for acoustic modeling instead of a GMM-HMM, the gains obtained by VTLN have been much less or even nonexistent for deeper neural networks. Studied have demonstrated that the features automatically extracted by DNNs are far superior to those produced by feature-engineering techniques generally used in GMM-based acoustic modeling, such as VTLN [11].

Other commonly used approaches for speaker adaption apply linear transformations [12], either by transforming the parameters of a trained model towards the test speakers, e.g., maximum likelihood linear regression (MLLR), or by

transforming the features of test speakers towards the trained models, e.g., feature-space MLLR (fMLLR). Some studies have documented that this approach is very effective for adapting an existing adult GMM-based LVCSR system to produce a child-specific system [5,7]. However, the idea of GMM adaptation (MLLR) is not applicable for a DNN in the sense of nonlinear transformation for layer connections and discriminative training with back-propagation (BP) rather than maximum likelihood (ML) training with expectation-maximization (EM). It has also been demonstrated that the improvement from using CMLLR (constrained MLLR) adapted features is minimal when DNN acoustic models are used instead of GMMs [13].

Recently, with the significantly improved performance obtained by using deep learning to train acoustic models for recognizing adult speech, several studies have also show the power of DNNs for acoustic modeling of children's speech [14-16]. In [14], a comparison is made between GMM-HMM and DNN-HMM systems using various amounts of training data for recognizing non-native children's speech for spoken language assessment applications. This study found that the DNN models outperform the GMM models when enough training data was available to train the DNN parameters reliably, and it demonstrated that the improvement in recognition performance of the DNN models compared to the GMM models increased monotonically with more training data. In [16], the best results for children's speech recognition were obtained by training on a large amount of data, which better matched children's speech, aided by a neural network classifier. Many acoustic modeling techniques for improving children's speech recognition shown in the literature, e.g., spectral smooth, VTLN and using pitch features, are found not to be effective, given their voice search task with trained acoustic model by 1.9 million utterances [16]. .

3. Data and Task

Two corpora of non-native spontaneous English drawn from the domain of spoken English proficiency assessment are used in this study. The first corpus contains non-native children's speech drawn from a pilot version of the TOEFL Junior Comprehensive assessment administered in late-2011 [17]. The TOEFL Junior Comprehensive is a computer-based test containing four sections: Reading Comprehension, Listening Comprehension, Speaking, and Writing. It is intended for middle school students around the ages of 11 - 15, and is designed to assess a student's English communication skills through a variety of tasks. This study focuses on the Speaking section, which contains the following three task types eliciting spoken responses:

- Read Aloud (RA): the test taker reads a paragraph (containing approximately 90 - 100 words) presented on the screen out loud
- Picture Narration (PN): the test taker is shown six images that depict a sequence of events and is asked to narrate the story in the pictures
- Listen Speak (LS): the test taker listens to an audio stimulus (approximately 2 minutes in duration) containing information about a non-academic topic (for example, a homework assignment) or an academic topic (for example, the life cycle of frogs) and provides a spoken response containing information about specific facts in the stimulus

The responses to each of three task types are approximately 60 seconds in duration, and they are scored on a scale of 1 - 4 by expert human raters. Each speaker provided 5 responses: one RA, one PN, and three LS. This corpus is hereafter referred to as the children's corpus and mainly used to build a speech recognizer for children.

The second corpus is drawn from a large-scale standardized spoken language proficiency test, TOEFL iBT, which measures a non-native speaker's ability to use and understand English at the university level. The speaking tasks in this test elicit monologs of 45 or 60 seconds in duration; example tasks include expressing an opinion on a familiar topic or summarizing information presented in a lecture. Each speaker provides 6 responses. Human experts were recruited to rate the responses using holistic rubrics on a 1-4 scale that cover the following three aspects of speaking proficiency: delivery, language usage and topic development. This corpus is hereafter referred to as the adult corpus and is used to improve the performance of speech recognizer for children.

The children's corpus includes responses from 1,685 test takers from over 10 native language backgrounds. It is divided into the following three sets (with no speaker overlap) for the current study: ASR training (AsrTrain), ASR development (AsrDev) and ASR evaluation (AsrEval). The corresponding number of speakers, number of responses, and hours of speech are presented in Table 1.

Table 1. *Number of speakers, number of responses, and duration of speech for each data partition in the children's corpus.*

	AsrTrain	AsrDev	AsrEval
#Speakers	1,625	30	30
#Responses	8,125	150	150
Duration (hours)	137.2	2.5	2.5

The adult corpus contains over 800 hours of non-native spontaneous speech covering over 100 native languages across 8,900 speakers. Table 2 presents the number of speakers, number of responses, and duration of speech for the three partitions, AsrTrain, AsrDev and AsrEval in the adult corpus; there is no speaker overlap across the three partitions.

Table 2. *Number of speakers, number of responses and duration of speech for each data partition in the adult corpus.*

	AsrTrain	AsrDev	AsrEval
# Speakers	8,700	100	100
# Responses	52,200	600	600
Duration (hours)	819	9.4	9.4

4. Speech Recognizer for Children

As discussed in the Section 2, DNN models in combination with large amounts of training data can significantly improve the performance of a speech recognition system. However, large corpora are not always available for deployed applications. In this study, we explore how to build an automatic recognizer for non-native children's speech with a corpus of moderate size containing approximately 100 hours of speech.

4.1. DNN-based Speech Recognizer with I-vectors

The spoken responses contain many non-scorable cases, e.g., non-English responses or response with large amounts of background noise. These responses were excluded from the AsrTrain set for training. After these exclusions, a total of 7,594 responses from the TOEFL Junior Comprehensive assessment were used to train a baseline speech recognizer for non-native English produced by children.

A GMM-HMM is first trained to obtain senones (tied tri-phone states) and the corresponding aligned frames for DNN training. The input feature vectors used to train the GMM-HMM contain 13-dimensional MFCCs and their first and second derivatives. Contextual dependent phones, tri-phones, are modeled by 3-state HMMs and the pdf of each state is represented by a mixture of 8 Gaussian components. The splices of 9 frames (4 on each side of the current frame) are projected down to 40-dimensional vectors by linear discriminant analysis (LDA), together with maximum likelihood linear transform (MLLT), and then used to train the GMM-HMM using ML. To alleviate the mismatch between the training criterion and performance metrics, the parameters of the GMM-HMM are then refined by discriminative maximum mutual information (MMI) training.

An i-vector is a popular auxiliary feature for improving DNN-based ASR. It is a compact representation of a speech utterance that encapsulates speaker characteristics in a low-dimensional subspace [18, 19] and has become the state-of-the-art approach in the field of speaker recognition. Using i-vectors is also a promising approach to speaker adaptation for speech recognition and appending the i-vector to frame-level acoustic features has been reported to improve the performance of ASR based on DNN acoustic modeling [20-22]. The AsrTrain partition of the children's corpus is also used to train the following hyper-parameters: GMM-UBM and T-matrix for i-vector extraction.

The features used to train the DNN are concatenated MFCC features and i-vector features. The MFCC features have the same dimensions as those used in GMM-HMM, while the i-vector features have 100 dimensions extracted from each response and are appended to the frame-level MFCC features. The input features stacked over a 15 frame window (7 frames to either side of the center frame for which the prediction is made) are used as the input layer of DNN. The output layer of the DNN has 3,957 nodes, the senones of the HMM obtained by decision-tree based clustering. The input and output feature pairs are obtained by frame alignment for senones with the GMM-HMM. The DNN has 7 hidden layers, and each layer contains 1024 nodes. The sigmoid activation function is used for all hidden layers. All the parameters of the DNN are first initialized by "layer-wise BP" pre-training [11], then trained by optimizing the cross-entropy function through back-propagation (BP), and finally refined by sequence-discriminative training, state-level minimum Bayes risk (sMBR).

The CMU pronunciation dictionary [23] is used to build a grapheme-to-phoneme (G2P) converter by data-driven joint-sequence models [24]. After text normalization for the transcriptions, we use G2P to automatically generate pronunciations for the words not contained in the CMU dictionary in the transcription and combine them with the CMU dictionary to create a new pronunciation dictionary. A trigram LM is trained from the transcriptions of the AsrTrain set using the IRSTLM toolkit [25].

4.2. Improving Speech Recognition for Children using Adult Data

Despite the large number of differences between children's speech and adult speech, there are still many acoustical and linguistic commonalities between the two varieties of speech. In addition, a DNN can represent high-level abstractions of complex data sets through multiple non-linear transformation [26]. In this study, we investigate whether using an adult corpus can improve the performance of DNN-based speech recognition for children. Our motivation is that the features extracted or transformed by the DNN can share the speech commonalities between children and adults and may potentially compensate for (or normalize) the mismatch between the two varieties. The following three approaches are tested to improve speech recognition for children using adult data.

A. Speaker adaption of DNN with i-vectors

We trained a DNN-based ASR system using a fairly large corpus of non-native English adult speech, as described in Section 4.1, and made it self-adaptive to a test speaker in the children's corpus by i-vector-based speaker adaptation. This approach has been shown to be very effective for cross-task adaptation, from monologic speech to dialogic speech [27].

To compensate for the mismatch in content between the children's and adult corpora, linear interpolation is used to combine the two LMs, which are trigram LMs trained from the transcriptions of the AsrTrain sets of the adult and children's corpora separately. The interpolation weight was optimized by minimizing the WER on the AsrDev set in the child corpus. The interpolated (combined) LM is finally represented as a finite state transducer (FST) for decoding using weighted FSTs (WFSTs).

B. Combining children's and adult training data

We added utterances from the AsrTrain partition of the adult corpus into the training set of children's corpus for the DNN training procedure described in Section 4.1. The DNN topology and i-vector dimension were both kept the same. The LM is also retrained using the merged transcriptions from the training sets of the adult and children's corpora.

This approach is more computationally expensive comparing to approach A. In addition, it is also unknown what amount of data should be added to optimize the performance of the combined speech recognizer for children's speech. Adding the adult corpus in its entirety would result in the adult corpus dominating the estimation of DNN weights, since the child corpus contributes only a relatively small faction of the training data.

C. DNN pre-training with adult data

Pre-training has been shown to be crucial for training deep structured models for speech recognition tasks [28, 29]; furthermore, it has been demonstrated that pre-training can initialize the DNN weights to a better starting point than random initialization prior to BP that allows the BP to facilitate a rapid global learning. Thus, DNNs have outperformed traditional shallow networks [30,31].

In this approach, the DNN weights obtained by pre-training with the full training set from the adult corpus is first employed to initialize the weights of the DNN for children's speech acoustic modeling, then the DNN weights are updated using the BP procedure with the combined children's and adult training data. Compared to approach B, this approach saves time for pre-training, which is usually very time-consuming.

We also experimented with retraining only the final phoneme-dependent soft-max layer instead of all layers, or using deep bottleneck features generated from the adult DNN to build a tandem-based children ASR, since these approaches have been suggested to be beneficial in low-resource speech recognition [32-34]. However, preliminary results demonstrated that these approaches were not able to outperform the baseline system built in Section 4.1. We think that the high-

level feature representation learned by the adult DNN still has a mismatch to the children's speech and that just updating the final layer transform is not enough to model the variation exhibited in a moderately sized children's speech corpus.

5. Experimental Results and Analysis

ASR systems were constructed using the Kaldi toolkit [35] based on the approaches described in Section 4.These systems are as follows:

- **Baseline:** DNN-based speech recognizer with i-vectors described in Section 4.1. The training data used for this system is the AsrTrain partition of the child corpus.
- **Adp-Adult**: Speaker adaption of DNN with i-vectors described in Section 4.2 (A). The training data used for this system is the AsrTrain partition of the adult corpus. None of the data from the children's corpus is included.
- **Adp-Comb:** Combining children's and adult training data as described in Section 4.2 (B). The AsrTrain partition of the children's corpus and 10,000 responses (~150 hours) randomly selected from the AsrTrain partition of the adult corpus are used as training data for this system. We also tried adding more adult speech into the training set, but this addition of more data resulted in worse performance.
- **Adp-Pretr:** DNN pre-training with Adult data as described in Section 4.2 (C). The same data as in the Adp-Comb system is used for acoustic modeling except that the weights of the DNN are initialized by pre-training with the full AsrTrain set from the adult corpus.

The performance in terms of word error rate (WER) is reported on the AsrEval set of the children's corpus.

Table 3 shows the WER of the Baseline and Adp-Comb systems using the following acoustic models: GMM-HMM and DNN-HMM, with/without discriminative training and i-vectors. In the Baseline system, the DNN-HMM in combination with i-vectors can improve the recognition performance on children's speech over the GMM-HMM: WER is reduced from 23.9% to 21.9% (8.4% relative WER reduction). This improvement is much smaller compared to that reported in the literature for larger corpora, which has been shown to be over 20%. This indicates that the DNN could likely not demonstrate its strong learning ability for acoustic modeling when the training corpus is only moderately sized. Combining the adult data with the children's data does not improve the performance of children's speech recognition using the GMM-HMM, but it can significantly improve the performance of acoustic modeling with the DNN-HMM (a relative WER reduction of 10.1%). This further demonstrates that the features learned by the DNN are more invariant and selective.

Table 3. *Word error rates (%) for the Baseline and Adp-Comb systems using different acoustic models.*

Model	Baseline	Adp-Comb
GMM	27.3	27.8
GMM + MMI	23.9	23.5
DNN + i-vector	22.4	20.3
DNN+ sMBR+ i-vector	21.9	19.7

The recognition performance using the Adpt-Adult system with/without LM interpolation is provided in Table 4, which indicates that LM interpolation is very effective when using an adult speech recognizer for children's speech. The interpolation weight 0.9 is the optimal weight according to our previous experience on LM interpolation for cross-task speech recognition [27]. However, the WER of 24.8% is higher than that obtained by the best GMM-HMM system, in which acoustic models are refined by MMI training. This is likely caused by the large acoustic mismatch between children's and adult speech.

Table 4. *Word error rates (%) for the Adp-Adult system with or without LM interpolation with the children's corpus.*

Adp-Adult	LM weight=0	LM weight=0.9
DNN+ sMBR+ i-vector	45.9	24.8

The WERs obtained using the Baseline, Adp-Comb and Adp-Pretr systems with DNN-based acoustic modeling are shown in Table 5, where Adp-Pretr achieves the best performance among these three systems: a relative WER reduction of 11.9% is obtained in comparison to the baseline system. DNN pre-training with the full training set of adult speech is also effective in improving the performance of speech recognition for children.

Table 5. *Word error rates (%) for the Baseline, Adp-Comb and Adp-Pretr systems with DNN based acoustic modeling.*

Model	Baseline	Adp-Comb	Adp-Pretr
DNN+ sMBR+ i-vector	21.9	19.7	19.3

Non-native speech contains many disfluency, including filled pauses, e.g., *um, uh*, and partial words,. These are often filtered out for automated assessment of language proficiency based on spoken language understanding. Table 6 shows the WERs obtained by the four systems (Baseline, Adp-adult, Adp-Comb and Adp-pretr.) across the three tasks with filled pauses and partial words removed. The recognition performance for all tasks is improved in the Adp-Comb and Adpt-Pretr systems compared to the baseline. The largest gain is achieved for the PN task, in which WER is reduced from 24.6 % to 20.1%, i.e., a relative WER reduction of 18.3%.

Table 6. *Word error rates (%) of the best performance obtaind by the four systems across all 3 with filled pauses and partial words removed.*

	RA	PN	AS	Overall
Baseline	7.5	24.6	27.5	22.3
Adp-Adult	9.9	26.8	29.5	24.4
Adp-Comb	6.8	22.7	25	20.3
Adp-Pretr	7.1	20.1	24.5	19.6

6. Conclusions

This paper has described three approaches to improving DNN-based automatic recognition of non-native children speech with the help of an out-of-domain corpus of adult speech. The experimental results show that the approach of combining children's training data with adult data is effective in improving the performance of speech recognition for children. This can reduce the WER of the DNN-based system from 21.9% to 19.7%. In addition, DNN pre-training with the full training set of a large adult corpus can further improve the performance of speech recognizer for children's speech. Future work will explore the benefits of improved recognition performance for the automated assessment of spoken language proficiency.

7. References

[1] K. Evanini, S. Singh, A. Loukina, X. Wang and C. M. Lee, "Content-based automated assessment of non-native spoken language proficiency in a simulated conversation", in NIPS Workshop on Machine Learning for Spoken Language Understanding and Interaction, 2015.

[2] D. Elenius and M. Blomberg, "Comparing speech recognition for adults and children", in Proc. of FONETIK, 2004.

[3] S. Das, D. Nix and M. Picheny, "Improvements in children's speech recognition performance", in Proc. of ICASSP, 1998.

[4] L. Måhl, "Speech recognition and adaptation experiments on children's speech", Master of Science thesis at the Department of Speech, Music and Hearing, KTH (The Royal Institute of Technology).

[5] P. G. Shivakumar, A. Potamianos, S. Lee, and S. Narayanan, "Improving speech recognition for children using acoustic adaptation and pronunciation modeling," in Proc. of Workshop on Child, Computer and Interaction (WOCCI), 2014.

[6] A. Potamianos, S. Narayanan, and S. Lee, "Automatic speech recognition for children," in Proc. of Eurospeech, 1997.

[7] S. S. Gray, D. Willett, J. Lu, J. Pinto, P. Maergner, and N. Bodenstab, "Child automatic speech recognition for US English: Child interaction with living-room-electronic-devices," in Proc. of Workshop on Child, Computer and Interaction (WOCCI), 2014.

[8] M. Blcmberg and K. Elenius, "Nonlinear frequency warping for speech recognition," in Proc. of ICASSP, pp. 2631-2634, 1986.

[9] R. Serizel and D. Giuliani, "Vocal tract length normalisation approaches to dnn-based children's and adults' speech recognition," in Proc. of the IEEE Workshop on Spoken Language Technology (SLT), 2014.

[10] S. Umesh, R. Sinha, and D. R. Sanand, "Using vocal-tract length normalization in recognition of children speech," in Proc. of National Conference on Communications, Kanpur, 2007.

[11] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription, " in Proc. of IEEE ASRU, 2011.

[12] M. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," Computer Speech and Language, vol. 12, 1998.

[13] T. N. Sainath, A. R. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," in Proc. of ICASSP, pp. 8614–8618, 2013.

[14] J. Cheng, X. Chen, and A. Metallinou, "Deep neural network acoustic models for spoken assessment applications", Speech Communication, vol. 73, pp. 14–27, 2015.

[15] A. Metallinou and J. Cheng, "Using Deep Neural Networks to improve proficiency assessment for children English language learners, " in Proc. of Interspeech, pp. 1468–1472, 2014.

[16] H. Liao, G. Pundak, O. Siohan, M. K. Carroll, N. Coccaro, Q.-M. Jiang, T. N. Sainath, A. Senior, F. Beaufays, and M. Bacchiani, "Large vocabulary automatic speech recognition for children", in Proc. of Interspeech, pp. 1611-1615, 2015.

[17] K. Evanini and X. Wang, "Automated Speech Scoring for Non-native Middle School Students with Multiple Task Types", in Proc. of InterSpeech, pp. 2435-2439, 2013.

[18] N. Dehak, P. Kenny, R. Dehak, P. Ouellet, and P. Dumouchel, "Front end factor analysis for speaker verification, " IEEE Trans. Acoust., Speech, Signal Processing, vol. 19, no. 4, pp. 788–798, 2011.

[19] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of inter-speaker variability in speaker verification, " IEEE Transactions on Audio, Speech and Language Processing, vol. 16, no. 5, pp.980 – 988, 2008.

[20] G. Saon, H. Soltau, D. Nahamoo and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors", in Proc. of ASRU, 2013.

[21] V. Gupta, P. Kenny, P. Ouellet and T. Stafylakis, "I-vector-based speaker adaptation of deep neural networks for French broadcast audio transcription", in Proc. of ICASSP, pp. 6334-6338, 2014.

[22] Y. Miao, H. Zhang and F. Metze, "Speaker adaptive training of deep neural network acoustic models using i-vectors", IEEE/ACM Transactions on Audio, Speech and Language Processing, vo. 23, no. 11, 2015.

[23] http://svn.code.sf.net/p/cmusphinx/code/trunk/cmudict/

[24] M. Bisani and H. Ney. "Joint-Sequence Models for Grapheme-to-Phoneme Conversion". Speech Communication, Vol. 50, Issue 5, pp. 434-451, 2008.

[25] A. Stolcke, SRILM - An Extensible Language Modeling Toolkit, in Proc. of Intl. Conf. Spoken Language Processing, Denver, Colorado, 2002.

[26] Y. Bengio. "Learning deep architectures for AI," Foundations and Trends in Machine Learning, Vol.2:No.1, pp.1-127, 2009.

[27] Y. Qian, X. Wang, K. Evanini and D. Suendermann-Oeft, "Self-Adaptive DNN for Improving Spoken Language Proficiency Assessment", in Proc. of Interspeech, 2016.

[28] D. Erhan, Y. Bengio, A. Courville, P.A. Manzagol, P. Vincent, and S. Bengio, "Why does unsupervised pre-training help deep learning?," JMLR, 2010.

[29] D. Yu, L. Deng, and G. Dahl, "Roles of pre-training and fine-tuning in context-dependent DBN-HMMs for real-world speech recognition," in Proc. of NIPS Workshop on Deep Learning and Unsupervised Feature Learning, 2010.

[30] G. E. Dahl, D. Yu, L. Deng, and A. 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.

[31] T.N. Sainath, B. Kingsbury, B. Ramabhadran, P. Fousek, P. Novak, and A.R. Mohamed, "Making deep belief networks effective for large vocabulary continuous speech recognition," in Proc. of IEEE ASRU, 2011.

[32] S. Thomas, S. Ganapathy, and H. Hermansky, "Multilingual MLP features for low-resource LVCSR systems," in Proc. of ICASSP, 2012.

[33] K. Vesel´y, M. Karafi´at, and F. Gr´ezl, "Convolutive bottleneck network features for LVCSR," in Proc. of ASRU, pp. 42–47, 2011.

[34] Y. Zhang, E. Chuangsuwanich, and J. Glass, "Extracting deep neural network bottleneck features using lowrank matrix factorization," in Proc. of ICASSP, pp. 185–189, 2014

[35] 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. Vesel, "The kaldi speech recognition toolkit," in Proc. of ASRU, 2011.