Multi-softmax Deep Neural Network for Semi-supervised Training

Hang Su[1 2], *Haihua Xu*[3]

[1] International Computer Science Institute, Berkeley, California, US
[2] Dept. of Electrical Engineering & Computer Science, University of California, Berkeley, CA, USA
[3]Nanyang Technological University, Singapore

suhang3240@gmail.com hhx502@gmail.com

Abstract

In this paper we propose a Shared Hidden Layer Multi-softmax Deep Neural Network (SHL-MDNN) approach for semi-supervised training (SST). This approach aims to boost low-resource speech recognition where limited training data is available. Supervised data and unsupervised data share the same hidden layers but are fed into different softmax layers so that erroneous automatic speech recognition (ASR) transcriptions of the unsupervised data have less effect on shared hidden layers. Experimental results on Babel data indicate that this approach always outperform naive SST on DNN, and it can yield 1.3% word error rate (WER) reduction compared with supervised DNN hybrid system. In addition, if softmax layer is retrained with supervised data, it can lead up to another 0.8% WER reduction. Confidence based data selection is also studied in this setup. Experiments show that this method is not sensitive to ASR transcription errors.

Index Terms: Semi-supervised training, Low resources, Deep Neural Networks.

1. Introduction

Semi-supervised training (SST) in low-resource condition is an important topic. When labeled data is limited, the performance of speech recognizer training may benefit a lot from automatically transcribed data. However, SST could be hard when there is no enough supervised data for baseline seed system training. In this case, transcriptions for unlabeled data obtained from ASR may contain so many errors that they even hurt acoustic model training.

DNN has been shown to be effective for SST in low-resource condition. In a hybrid DNN approach [1] where DNN are directly used for decoding after SST [2], it is shown that a 2.2% absolute WER reduction could be achieved by setting a threshold for dropping frames from unsupervised data. In a bottle-neck NN approach [3] where semi-trained DNN are used as feature extractor for other acoustic models [4], more than 2% absolute WER reduction is achieved for several different acoustic models. Research into confidence measure [5–7], data selection [8,9], SST sequence-training [10] and bootstrapping [11] in the framework of SST have shown to be successful as well, and other research on combining multilingual training with semi-supervised training [12, 13] also show their effectiveness in different settings.

Comparing SST on bottle-neck and hybrid approach, we find that the bottleneck approach tends to produce better results when labeled data and unlabeled data are blent together and no confidence threshold is set for data selection. We suppose this is because alignment errors tend to influence top most layers of neural networks. Deep neural network could be seen as a concatenation of feature extractor and log-linear classifier where lower layers are responsible for feature extraction and a softmax layer for classification. In this sense, unlabeled data will help more in lower layers training while it may hurt classification layers due to its error in alignments. SST on bottle-neck NN thus utilizes this property by removing top layers in neural network and preserve lower layers for feature extraction.

However, the disadvantage of bottle-neck approach is that it involves more parameters, and makes the training pipeline longer. Acoustic models such as Gaussian Mixture Model (GMM) or DNN need to be retrained after bottle-neck features are extracted. So, it would be better if we could combine the benefits of these two approaches to mitigate the alignment issue in hybrid approach.

Shared Hidden Layer Multilingual DNN (SHL-MDNN) proposed in [14] is shown to be effective in utilizing training data from different languages. In this setup, multilingual training data share the same hidden layers but use different softmax layers to do senone classification. Similar structures are also used in [15–18] for multilingual speech acoustic modeling. Multilingual stacked bottle-neck neural network is used in [19–21] for feature extraction. This approach solves the problem of training DNN using data with different alignments models. In this paper, we introduce the same model to SST task to mitigate the alignment issue for unsupervised data. [1]

The rest of this paper is organized as follows: Section 2 introduces SHL-MDNN based semi-supervised training. Section 3 describes the experimental setup. Section 4 analyze baseline DNN SST results. Section 5 reports experiments on semi-supervised training, together with results on several different languages. Section 6 concludes our work.

2. SHL-MDNN based semi-supervised training

Figure 1 shows the structure of the SHL-MDNN. In this approach, supervised data and unsupervised data share the same hidden layers, while the softmax layers are not shared.

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense US Army Research Laboratory contract number W911NF-12-C-0014. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.

[1] The meaning of ”M” in SHL-MDNN is changed from ”multilingual” to ”multi-softmax” to fit the topic.

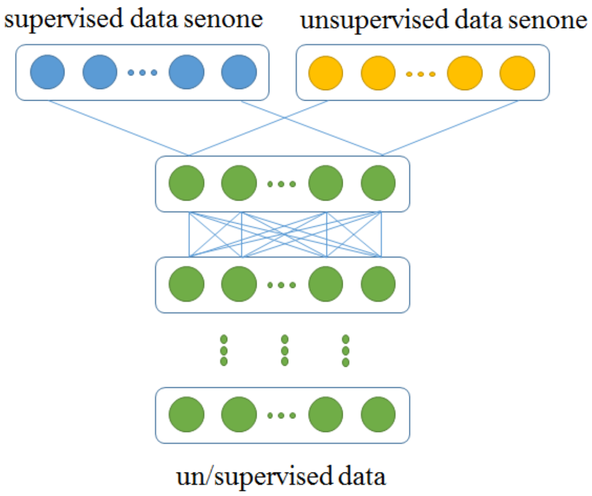

Figure 1: SHL-MDNN for semi-supervised training

Input layers covers a long contextual window of acoustic feature. In this paper, we use PLP feature transformed by linear discriminative analysis (LDA), maximum likelihood linear transformation (MLLT) and speaker adaptive training (SAD). Since supervised and unsupervised data use the same models for transformation estimation, it is reasonable to mix those transformed feature together and send them into shared hidden layers [2].

We train the SHL-MDNN using supervised data and unsupervised data simultaneously, using mini-batch stochastic gradient descent (SGD). Each mini-batch may contain different number of supervised and unsupervised frames. This can be effectively done by frame randomization.

The SHL-MDNN are pre-trained layer-wisely, using Contrastive Divergence algorithm for Restricted Boltzmann Machines (RBM) [22, 23]. Both supervised and unsupervised data are used for pretraining. The second stage of model training is carried out by backpropagation (BP) algorithm. In each pass, the gradient of supervised frames and unsupervised frames are used to update softmax layers respectively. For shared hidden layers, however, the gradients are combined for weight and bias update.

After training, the softmax layer for unsupervised data is thrown away and only softmax for supervised data is preserved. Since the gradient for updating softmax layer and shared hidden layers are different, it helps to fine-tune the neural network using supervised data only for a few iterations.

3. Experimental Setup

In this paper, we report detailed experimental results on the Vietnamese dataset provided by the IARPA Babel program [3], together with final results on several OP1 languages including Tamil, Assamese, Bengali and Zulu.

BABEL data in Table 1 are used for experiments. Each language pack provides lexicons and a limited amount of training data which has been transcribed at the word level. The data is divided into subsets called the full language pack (FLP) and the limited language pack (LLP) which have approximately 65 hours and 10 hours of training data respectively. Development data sets (10 hours) are provided for performance testing and parameter tuning. In this work, we used the LLP as supervised training data, and FLP data that does not appear in LLP as unsupervised data. Dev set is used for evaluating the performance of SST. Language pack statistics are summarized in Table 2.

	version
Vietnamese	IARPA-babel107b-v0.7
Assamese	IARPA-babel102b-v0.5a
Bengali	IARPA-babel103b-v0.4b
Zulu	IARPA-babel206b-v0.1e
Tamil	IARPA-babel204b-v1.1b

Table 1: Babel data for different languages

	vocab	FLP(h)	LLP(h)
Vietnamese	3205	87.7	11.0
Assamese	7661	60.8	10.0
Bengali	7933	61.7	10.3
Zulu	13674	62.1	10.4
Tamil	14265	69.4	11.7

Table 2: statistics for Babel data

The Kaldi toolkit [24] is used for speech recognition framework. Standard 13-dim PLP feature, together with 3-dim Kaldi pitch feature [25], is extracted and used for maximum likelihood GMM model training. Features are then transformed using LDA+MLLT before SAT training. After GMM training is done, a tanh-neuron DNN-HMM hybrid system is trained using the the 40-dimension transformed fMLLR (also known as CMLLR [26]) feature as input and GMM-aligned senones as targets. This DNN serves as a baseline system in this paper, and is used for decoding / ASR of unsupervised data to do semi-supervised training. fMLLR is estimated in an EM fashion for both training data, unsupervised data and test data. Despite the high WER in Babel setup, we still find it helpful to use fMLLR feature on top of LDA and MLLT transformations.

Details of DNN training follows Section 2.2 in [2]. In this paper, we use 6 hidden layers, where each hidden layer has 2048 neurons with sigmoids. Input layer is 440 dimension (i.e. the context of 11 fMLLR frames), and output layer for Vietnanmese is 1921 dimension. Mini-batch SGD is used for backpropagation. The training starts with an initial learning rate of 0.008 and halves the rate when the improvement in training objective (cross-entropy) on a cross-validation set between two successive epochs falls below 0.01%. The optimization terminates when the objective improves by less than 0.0001%. Cross validation is done on 10% of supervised data only.

4. Analysis on supervised experiments

4.1. Supervised Experiments

In order to get an idea how much semi supervised data may help in DNN acoustic modeling, we train GMM models using LLP and FLP data separately. Both models are then used to align LLP and FLP data for DNN model training. WERs of these DNN models on dev set are reported in Table 3.

As is shown in the table, 64.4% is the baseline supervised experiment where only LLP data is used. When LLP is used

[2] This is different from the multi-lingual task where feature sent into shared hidden layers are not transformed

[3] This is consistent with related works that have been published

DNN \ alignment model	LLP	FLP
LLP	64.4	59.5
FLP	57.0	51.5

Table 3: WER for supervised experiments

for GMM training and FLP is aligned and used for DNN training, the WER for DNN system is 57.0%, which we would like to consider as upper bound of semi-supervised training for DNN. The WER 51.5% is the upper bound in theory for semi-supervised training on all the models (including GMM flat start).

4.2. The problem of ASR transcription

It is reasonable to expect a GMM model trained on FLP data helps DNN training (compared with a GMM trained on LLP), because it can provide better alignments. As is shown in Table 3, the difference between these GMMs result in a 5% WER reduction (from 64.4% to 59.5%). On the other hand, semi-supervised training uses the same GMM for alignment generation for supervised and unsupervised data, but the transcriptions for these data are different – supervised data are transcribed manually while unsupervised data are transcriped by ASR. Thus, it is also necessary to check the difference between these alignments.

We compare the alignments for unsupervised data produced by ground truth transcription and ASR. Because unsupervised channels need to be segmented into utterances before decoding, the timing information may not match the ground truth segmentation. We summarized statistics using both the ground truth segmentation and VAD in Table 4. Here, 'total' denotes the total number of frames of the alignments and 'matched' denotes the number of frames that have the same alignment senone from decoded text and manual transcription.

segments	# total (10^6)	# matched (10^6)
ground truth	27.5	10.1 (36.9%)
VAD	20.0	5.0 (24.9%)

Table 4: Alignment analysis for unsupervised data

As is shown in Table 4, the percentage of aligned frames between decoded best-path alignments and ground truth text alignments is 24.9%. Even with ground truth segmentation, the percentage of matched frames is just around 1/3, of which many of them are silence. Admittedly, these statistics does not necessarily show the decoded text alignments are bad for SST, because alignments from ground truth text may not be correct anyway, and different alignment may still give similar decoding result because words may have different pronunciations. But these numbers show that alignments from human transcription and ASR usually do not agree with each other, which may make the acoustic model confused during semi-supervised training.

5. Semi-supervised Training

Our baseline bottleneck based SST follows the work in [4], where stacked bottle-neck neural network is trained using both supervised and unsupervised data with alignments generated by the seed GMM. With a semi-trained bottleneck feature extractor, we start the whole acoustic training pipeline (with GMM flat start) again using bottleneck features extracted for supervised data only. The final acoustic model for decoding is a bottleneck feature based deep neural network. This approach has been proven to be effective for SST under low-resource condition [4, 27, 28], but the disadvantage of this approach is that the training pipeline is longer and the model contains more parameter.

Our SST pipeline for DNN generally follows the work in [2]. After the seed GMM is trained, we align supervised data against ground truth, and decode unsupervised data to get alignments from the lattice. Then supervised data and unsupervised data are mixed together and send to train DNN. We did not use DNN for alignment, because we found it influences semi-supervised training results only a little, but it makes the training pipeline longer. Using GMM as seed system makes semi-supervised training pipeline shorter, but also more difficult. For naive semi-supervised DNN training, we blend supervised and unsupervised data together, and perform standard DNN training pipeline on the mixed data. We do not perform any data selection or frame weighting. As is shown in Table 7, the naive semi-supervised training degrades the WER of the recognizer.

To mitigate the issue of poor alignment for unsupervised data, we throw away the top softmax layer after the naive SST DNN is trained, and add a new randomly initialized layer. Then, we train the whole network using supervised data only, using the naive SST DNN as an initialization. This approach is similar to the fine-tuning step in [28]. Our experiments show that this generally give better performance than directly using the semi-trained DNN – 1.5 WER improvement.

5.1. SHL-MDNN for SST

As is stated in Section 2, SHL-MDNN is trained using supervised and unsupervised data collectively. After the SHL-MDNN is trained, we remove the semi-supervised branch and remain the supervised one. This method gives a WER of 63.1, i.e. 1.3 WER improvement.

Furthermore, we throw away the softmax layers from SHL-MDNN, add a new randomly initialized layer and train the whole network using supervised data only, this gives us an additional 0.8 WER improvement, making it comparable to the bottle-neck approach. Table 7 shows the experimental results for different settings.

Systems	WER (%)
Baseline DNN	64.4
Baseline SST bottle-neck	62.5
Naive SST DNN	66.2
SST DNN soft-retrain	62.9
SST SHL-MDNN	63.1
SST SHL-MDNN soft-retrain	62.3

Table 5: WER of Vietnamese DNN system

5.2. Confidence based data selection

Confidence based data selection has been widely used in semi-supervised training [2, 4, 10] for better acoustic modeling. In this section, we try to explore how data selection affect SHL-MDNN based SST.

We follow the sentence level confidence measure proposed in [2], i.e.

$$c_{sent} = \frac{1}{N} \sum_{i=1}^{N} c_{w_i} \tag{1}$$

where c_{w_i} is the posterior probability of word w_i obtained by Minimum Bayes Risk decoding [29]. The confidence evaluation produced by NIST scoring tool SCLite is shown in Figure 2.

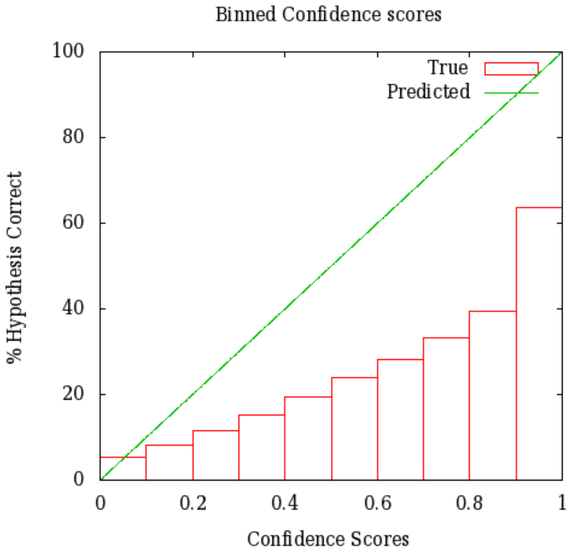

Figure 2: Decode confidence score v.s. WER

As is shown in the figure, the correlation between decoding word confidence and WER are consistent.

To evaluate how confidence-based data selection affects the performance of SST, we sort the utterance by the sentence confidence scores, and select top/bottom 20% of them to mix with supervised data for STT [4]. The number '20%' is set to make sure supervised data and unsupervised data are of similar size. Table 6 shows the WER for different settings.

Systems \ conf. score	top 20%	bottom 20%
Baseline DNN	64.4	
Baseline SST bottle-neck	62.5	
Naive SST DNN	64.7	66.2
SST DNN soft-retrain	64.1	64.6
SST SHL-MDNN	64.3	65.3
SST SHL-MDNN soft-retrain	63.2	63.3

Table 6: WER of Vietnamese using data selection

Comparing columns in row 'Naive SST DNN' and 'SST SHL-MDNN', we can see that unsupervised sentences with different confidence scores do make a difference in SST. But if we apply softmax layer retraining on top of SST, the difference is largely covered. This shows that this method makes SST less sensitive to alignment errors. Comparison between Table 6 and Table 7 also show that the benefit of data selection are counterweighed by the benefits of more data.

5.3. Results over languages

To show the effectiveness of our approach, we apply the method to 4 other languages, Assamese(AS), Bengali(BE), Zulu(ZU), Tamil(TA), and the WER results are shown in Table 7. As we can see in the figure, SHL-MDNN based SST with softmax retraining always have better results compared with naive DNN SST. It gives comparable results against bottleneck based SST, but with less model parameter and a shorter training pipeline.

Languages	AS	BE	ZU	TA
Base DNN	65.3	68.3	70.8	77.9
Base SST BN	64.6	67.3	69.8	75.4
Naive SST DNN	67.5	69.4	71.0	78.4
SST DNN retrain	66.0	68.5	70.6	76.7
SST SHL-MDNN	66.2	69.0	70.8	77.8
SST SHL-MDNN retrain	64.5	67.1	69.7	76.1

Table 7: WER of DNN systems on 4 other Babel languages

5.4. Other refinement experiments

We tried to combine SHL-MDNN SST with bottle-neck approach, using bottle neck feature with/out STT to do SHL-MDNN based SST. But both experiments show little improvement in terms of WER. We also tried to bootstrap the performance by realignment, but no significant improvement is observed.

6. Conclusions and future work

In this paper, we apply SHL-MDNN to semi-supervised training to treat alignments for un/supervised data differently. Together with softmax layer retraining, we achieve an 2% absolute WER improvement on Babel Vietnamese LLP task. Recognition results on 4 other languages are provided, and they show SHL-MDNN based SST consistently outperforms naive DNN SST. Confidence based data selection experiments show that transcribed utterances with higher confidence score tend to yield better SST results, but it becomes unnoticeable when we perform softmax layer retraining. It would be interesting to see if this SHL-MDNN based SST also works on languages with thousands hours of untranscribed training data.

7. Acknowledgements

We would like to thank Nelson Morgan and Karel Vesely for their help on DNN based SST. We also thank Daniel Povey for his advice on semi-supervised training in Kaldi.

8. References

[1] H. A. Bourlard and N. Morgan, *Connectionist speech recognition: a hybrid approach.* Springer Science & Business Media, 1994, vol. 247.

[2] K. Veselýl, M. Hannemann, and L. Burget, "Semi-supervised training of deep neural networks," in *ASRU*. IEEE, 2013.

[3] F. Grézl and M. Karafiát, "Hierarchical neural net architectures for feature extraction in asr." in *Proc. Interspeech*, 2010.

[4] H. Xu, H. Su, E.-S. Chng, and H. Li, "Semi-supervised training for bottle-neck feature based dnn-hmm hybrid systems," in *Proc. Interspeech*, 2014.

[5] Y. Huang, D. Yu, Y. Gong, and C. Liu, "Semi-supervised gmm and dnn acoustic model training with multi-system combination and confidence re-calibration," in *Proc. Interspeech*, 2013.

[4] To ensure the amount of data are of similar size, the 20% here refers to number of frames.

[6] F.Wessel, R. Schlüter, K. Macherey, and H. Ney, "Confidence measures for large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, pp. 288–298, 2001.

[7] P. Zhang, Y. Liu, and T. Hain, "Semi-supervised dnn training in meeting recognition," in *SLT*. IEEE, 2014.

[8] D. Yu, B. Varadarajan, L. Deng, and A. Acero, "Active learning and semi-supervised learning for speech recognition: a unified framework using the global entropy reduction maximization criterion," *Computer Speech and Language*, 2009.

[9] K. Wei, Y. Liu, K. Kirchhoff, C. Bartels, and J. Bilmes, "Submodular subset selection for large-scale speech training data," in *ICASSP*. IEEE, 2014.

[10] R. Hsiao, T. Ng, F. Grézl, D. Karakos, S. Ksakalidis, L. Nguyen, and R. Schwartz, "Discriminative semi-supervised training for keyword search in low resource languages," in *ASRU*. IEEE, 2013.

[11] F. Greźl and M. Karafiát, "Semi-supversied bootstrapping approach for neural network feature extractor training," in *ASRU*. IEEE, 2013.

[12] S. Thomas, M. Seltzer, K. Church, and H. Hermansky, "Deep neural network features and semi-supervised training for low resource speech recognition," in *ICASSP*. IEEE, 2013.

[13] F. Grézl, , E. Egorova, and M. Karafiát, "further investigation into multilingual training and adaptation of stacked bottleneck neural network structure," in *SLT*. IEEE, 2014.

[14] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *ICASSP*. IEEE, 2013.

[15] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual acoustic models using distributed deep neural networks," in *ICASSP*. IEEE, 2013.

[16] Y. Miao, H. Zhang, and F. Metze, "Distributed learning of multilingual dnn feature extractors using gpus," in *Proc. Interspeech*, 2014.

[17] N. T. Vu, D. Imseng, D. Povey, P. Motlicek, T. Schultz, and H. Bourlard, "Multilingual deep neural network based acoustic modeling for rapid language adaptation," in *ICASSP*. IEEE, 2014.

[18] A. Ghoshal, P. Swietojanski, and S. Renals, "Multilingual training of deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013.

[19] Y. Zhang, E. Chuangsuwanich, and J. Glass, "Language id-based training of multilingual stacked bottleneck features," in *Proc. Interspeech*, 2014.

[20] F. Grézl, M. Karafiát, and K. Vesely, "Adaptation of multilingual stacked bottle-neck neural network structure for new language," in *ICASSP*. IEEE, 2014.

[21] N. T. Vu, J. Weiner, and T. Schultz, "Investigating the learning effect of multilingual bottle-neck features for asr," in *Proc. Interspeech*, 2014.

[22] G. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[23] A.-r. Mohamed, T. N. Sainath, G. Dahl, B. Ramabhadran, G. E. Hinton, and M. A. Picheny, "Deep belief networks using discriminative features for phone recognition," in *ICASSP*. IEEE, 2011.

[24] D. Povey, A. Ghoshal, G.Boulianne, L. Burget, O.Glembek, N. Goel, M. Hannermann, P. Motlíček, Y. Qian, P. Schwartz, J. Silovský, G. Stemmer, and K. Veselý, "The kaldi speech recognition toolkit," in *ASRU*. IEEE, 2011.

[25] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *ICASSP*. IEEE, 2014.

[26] M. J. Gales, *The generation and use of regression class trees for MLLR adaptation*. University of Cambridge, Department of Engineering, 1996.

[27] S. Thomas, M. L. Seltzer, K. Church, and H. Hermansky, "Deep neural network features and semi-supervised training for low resource speech recognition," in *ICASSP*. IEEE, 2013.

[28] F. Grézl and M. Karafiát, "Combination of multilingual and semi-supervised training for under-resourced languages," in *Proc. Interspeech*, 2014.

[29] H. Xu, D. Povey, L. Mangu, and J. Zhu, "An improved consensus-like method for minimum bayes risk decoding and lattice combination," in *ICASSP*. IEEE, 2010.