Semi-supervised Training for Bottle-neck Feature based DNN-HMM Hybrid Systems

Haihua Xu[1], *Hang Su*[2], *Eng-Siong Chng*[1], *Haizhou Li*[3]

[1]Nanyang Technological University, Singapore
[2] International Computer Science Institute, Berkeley, California, US
[3] Institute for Infocomm Research, Singapore

Abstract

In this paper, we investigate semi-supervised training (SST) method in various state-of-the-art acoustic modeling techniques, using bottle-neck and corresponding tandem features. These techniques include subspace GMM, tanh-neuron deep neural network (DNN), and a generalized soft-maxout (p-norm) DNN. We demonstrate that SST may lead up to 2% Word Error Rate (WER) reduction using all these techniques in each case, and the best one comes from tandem feature based p-norm DNN system. In addition to recognition performance, effectiveness of the SST on keyword search performance is also investigated. Results on Actual Term Weighted Value (ATWV) are reported, with an analysis on lattice density. It is shown that SST may not necessarily increase ATWV due to the shrink of lattices size.

Index Terms: Semi-supervised training, Bottle-neck features, Maxout Networks, keyword search, Speech recognition

1. Introduction

Semi-supervised training (SST) method is an important topic for speech recognition, especially when transcription is limited while unlabeled audio data is available [1–8].

Research work on SST was very successful for conventional GMM-HMM method [1–3,5]. Using a stronger language model and a large volume of unlabeled data, one may obtain significant recognition performance improvements. Meanwhile, data selection strategy on unsupervised data is also a primary concern [4, 5]. Once remarkable recognition performance improvement is achieved by SST, iteration also plays a role [3].

Recently, Neural network (NN) based SST is surging [5–8]. In [6], Deep Neural Network (DNN) is employed as a front-end feature extractor and a GMM-HMM acoustic model is trained using the obtained Bottle-Neck Feature (BNF). In [7] stacked bottle-neck (SBN) NNs are also shown as an effective front-end feature extractor for SST. Alternatively, [8] shows that combination of frame-weighted data selection for SST and sequence-discriminative DNN training [9] yields remarkable recognition performance improvement.

In this paper we make a further step to investigate BNF based SST approach using various acoustic modeling techniques such as Subspace GMM (SGMM) [10], tanh-neuron based DNN, and a recently proposed generalized maxout DNN, called p-norm DNN [11]. That is, SST training is conducted on stacked BN NNs [7, 12] to get robust front-end BNFs and acoustic models are still trained with supervised data. We investigate cases using BNF, and corresponding tandem features by combining BNF with PLP together. As keyword search (KWS) performance [13–15] is also a primary concern, we investigate the effectiveness of the SST on KWS performance.

Section 2 reviews the prior related work and our contributions. Section 3 introduces our experimental setup. Section 4 and 5 report supervised and semi-supervised recognition results respectively. Section 6 presents SST based keyword search results and analysis. Section 7 concludes.

2. Prior related work

The idea of using BNF to address the SST problem has been popular [6, 7] recently, but one used to be concerned with using BNF or corresponding tandem features to train conventional GMM-HMM system. In this paper we extend the previous work to train SGMM, tanh-neuron DNN, and p-norm DNN acoustic models using semi-supervised (SS) learned BNF and corresponding tandem features. Our work is also distinct from the direct use of sigmoid-neuron DNN approach to SST [8] in that we compare different feature and classifiers respectively.

In addition to these, the keyword search performance concerned with the SST is also studied. Recognition accuracy and keyword search performance are two closely related metrics, but we may still run into cases when lower WER does not give better ATWV [16]. It is shown in [17] that ATWV is improved by SST, while details such as lattice oracle error rate or density are not given.

3. Experimental setup

All the experiments in this paper are conducted on `babel107b-v0.7` Vietnamese telephony data set, using KALDI toolkit [18]. The original data is released for `OpenKWS13` [1] by IARPA Babel program. It contains three packs: Full Language Pack (FullLP), Limited Language Pack (LimitedLP) and scripted pack. In this paper, only FullLP and LimitedLLP are used for training. Data is preprocessed by removing silence-dominated and noise-contaminated utterances. We also restrict endpoint silence length for each utterance. Table 1 shows the size of data after preprocessing. In Table 1, LimitedLP is used for supervised training and FullLP for semi-supervised training. *Dev* data is for evaluation of recognition

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] http://www.nist.gov/itl/iad/mig/upload/OpenKWS13-EvalPlan.pdf

Table 1: Size of data before/after preprocessing (hours)

	FullLP	LimitedLP	*Dev*
Original release	158.7	20.1	20.0
Preprocessed data	42.3	5.4	9.9

and KWS performance. Dictionary and trigram Language Models (LM) are built using LimitedLP resource.

The work flow of SST proceeds as below: 1) baseline system is trained from supervised training data (LimitedLP); 2) decoding for unlabeled data (from FullLP) is conducted using baseline system; 3) SBN NN is trained with the mixture of supervised and unsupervised training data; 4) so-called semi-supervised BNF for supervised training data is generated; 5) acoustic models are retrained using the semi-supervised BNF or semi-supervised BNF+PLP feature.

3.1. Stacked bottle-neck NN training

Details of training SBN NNs refer to [7, 12]. Original feature for training is 18-dim: 15 Mel filter-bank log energies plus 3 KALDI pitch features [19]. Hamming window and DCT transformation are applied on 11 contextual features. The first BN NN is configured as 108-1500-1500-80(BN)-1500-884 and the second 400-1500-1500-30(BN)-1500-884. All BN layers employ linear-neurons while the other hidden-layers are sigmoid-neurons. Networks are trained on GPU using gradient descent cross-entropy criterion without pre-training.

Note that we used monophone state labels as SBN output. To obtain better alignment triphone system was used to conduct forced-alignment instead, and then phone conversion is performed.

3.2. GMM-HMM system

Three baseline GMM-HMM systems are trained with different features: PLP plus KALDI pitch, BNF, and BNF+PLP[2]. Features are transformed using LDA+MLLT within three-frame context before SAT training for the former two systems, while no LDA+MLLT transformation is used for the BNF+PLP tandem feature system. All the three systems have 2500 states with 10 Gaussian mixtures for each state on average. For decoding unsupervised fMLLR method is employed.

3.3. Subspace GMM system

In this paper, SGMM systems are built on top of corresponding GMM-HMM systems in previous section. The UBM has 400 Gaussians. The state number is 2500 with around 4 sub-states for each. We followed the standard procedure in KALDI to train SGMM system, and speaker vector training is also enabled [10].

3.4. Tanh-neuron DNN system

Recently, KALDI is upgraded to support new DNN training recipes and parallel training in KALDI can be conducted across external and internal machines with or without GPU support. As a baseline, we trained various tanh-neuron DNNs with three sets of features as stated in 3.2. All DNNs have 7 hidden layers with 1000 neurons in each layer, and neurons in output layers correspond to tied triphone states. Details of this DNN training method are stated in [11].

3.5. P-norm DNN system

Inspired by the idea of maxout network [20] with y=$\max_i x_i$ as activation function, [11] proposed a p-norm nonlinear activation function written as y=$(\sum_i |x_i|^p)^{\frac{1}{p}}$ and showed that $p=2$ and input group size being 10 can yield the best performance. In this paper we adopt this novel DNN framework to show its effectiveness under SST circumstance. We keep $p=2$, but change the input group size to 6 in this paper.

4. Supervised experiments

This section investiges various front-end feature and acoustic modeling methods using limited supervised training data (see Table 1).

4.1. PLP plus pitch features

Table 2 reports WER from various acoustic modeling methods using PLP plus KALDI pitch features.

Table 2: WER with various acoustic modeling methods using PLP+Kaldi pitch features

acoustic modeling methods	WER(%)
PLP+GMM-HMM+fMLLR	77.3
PLP+pitch+GMM-HMM+fMLLR	71.8
PLP+pitch+fMLLR+SGMM	69.8
PLP+pitch+fMLLR+tanh-neuron DNN-HMM	68.5
PLP+pitch+fMLLR+p-norm DNN-HMM	66.6

Table 2 shows that: (1) KALDI pitch feature is very effective in improving recognition performance (5.5% absolute WER reduction), (2) stat-of-the-art acoustic modeling recipes are consistently better than conventional GMM-HMM, (3) p-norm DNN is better than tanh-neuron DNN, giving a 1.9% absolute WER reduction.

4.2. Bottle-neck features

Bottle-neck features (BNFs) are generated after training stacked bottle-neck neural networks. Table 3 shows WER in various acoustic models. Compared with Table 2, results in Table 3 clearly demonstrate the effectiveness of using BNFs for acoustic modeling. Also, we notice that GMM-HMM system gets the largest improvement(71.8%-67.0%). Taking BNF GMM-HMM as baseline, p-norm DNN-HMM gives an additional 1.2% WER reduction, while tanh-neuron DNN-HMM yields no benefits. This again demonstrates the effectiveness of p-norm DNN modeling technique.

Table 3: WER with various acoustic modeling methods using BNFs

acoustic modeling methods	WER (%)
BNF+GMM-HMM+fMLLR	67.0
BNF+fMLLR+SGMM	66.4
BNF+fMLLR+tanh-neuron DNN-HMM	67.0
BNF+fMLLR+ p-norm DNN-HMM	65.8

4.3. Tandem features

In this section, WERs using BNF+PLP tandem features are reported, Table 4 indicates our results. As is shown in Table 4, a marginal WER reduction could be obtained by combing BNF

[2] The PLP feature here means PLP+ΔPLP+$\Delta\Delta$PLP in all tandem scenarios.

Table 4: WER with various acoustic modeling methods using BNF+PLP (69 *dim*) tandem features

acoustic modeling methods	WER (%)
tandem+GMM-HMM+fMLLR	66.8
tandem+fMLLR +SGMM	66.3
tandem+fMLLR +tanh-neuron DNN-HMM	66.5
tandem+fMLLR + p-norm DNN-HMM	65.7

with PLP features. We tried various tandem methods, however, no obvious advantage from one over the other is observed.

5. Semi-supervised experiments

Based on experimental results in Section 4, we conduct SST recognition experiments with two feature sets: BNF and BNF+PLP tandem features.

5.1. Overall unsupervised data selection

Experiments in [7,8] already show that SST can improve speech recognition performance even when no confidence measure is employed, so we start off using all unsupervised data. Table 5 shows results of various modeling methods using semi-supervised trained BNFs. WERs shown in Table 5 are about 2 point lower then those in Table 3 and 4 respectively, indicating the effectiveness of SST.

Table 5: WER with various acoustic modeling methods using semi-supervised trained BNFs

acoustic modeling methods	WER (%)
tandem+GMM-HMM+fMLLR	64.7
tandem+fMLLR +SGMM	64.3
tandem+fMLLR +tanh-neuron DNN-HMM	64.6
tandem+fMLLR + p-norm DNN-HMM	63.6

Table 6 further presents results using semi-supervised trained tandem features (only the BNF part is related to SST). In general, comparison between Table 6 and Table 5 is consistent with those corresponding experiments without SST. Simply put, DNN-HMM based recipes still give a slight WER improvement using tandem features, while GMM-HMM and SGMM systems don't. Particularly, p-norm DNN-HMM shows better WER improvement (0.6% absolute). These might be due to that DNN is more insensitive to feature diversity, in addition to taking advantage of higher dimensional features.

Table 6: WER with various acoustic modeling methods using semi-supervised trained tandem features

acoustic modeling methods	WER (%)
tandem+GMM-HMM+fMLLR	64.8
tandem+fMLLR +SGMM	64.3
tandem+fMLLR +tanh-neuron DNN-HMM	64.5
tandem+fMLLR + p-norm DNN-HMM	63.0

5.2. Confidence based data selection

In this section we investigate how much benefit can be obtained by using utterance based confidence score to select machine transcribed data. To this end we define $C_w = \max_{t=start_w:w_i=w}^{end_w} C_{w_i}$ as word confidence in lattice, and $T_w = end_w - start_w$ as corresponding duration, then utterance confidence score is computed as $C_u = \frac{1}{T}\sum_{w\in lat} T_w C_w$. These are similar to what is advocated in [7,21].

Since the unsupervised data has ground-truth transcripts in our experiments, we can use these transcripts as reference to verify if our confidence estimate method is working. Figure 1 plots absolute WER reduction versus confidence on unsupervised data. From Figure 1, we can observe that the confidence score is strictly correlated with the WER reduction, demonstrating the method is working. Note that the best WER on the unsupervised data from our supervised tandem+fMLLR+p-norm DNN-HMM is 69.1%.

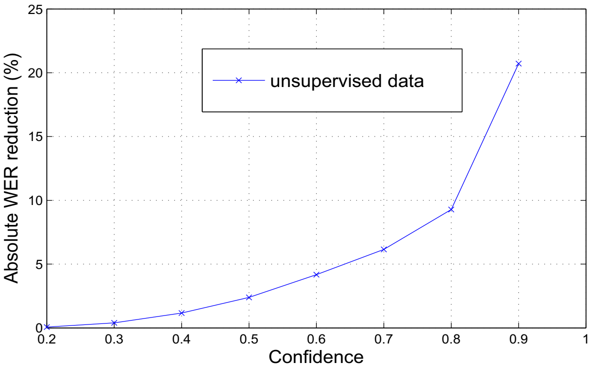

Figure 1: Confidence score versus Word Error Rate reduction on unsupervised data using tandem+fMLLR+p-norm DNN-HMM recipe in supervised case

Once the confidence score for each utterance is estimated, we vary threshold to select machine transcribed data to merge with the supervised data, and SST is started on SBN NNs accordingly. To visualize how much WER will be affected by confidence based data selection method, Figure 2 illustrates the details. First of all Figure 2 shows confidence score in BNF based SST is not crucial, probably due to too many errors within the machine transcribed data (about 70% WER). The best absolute WER reductions are only around 0.2 with confidence score being about 0.4, compared with no confidence score used at all. Secondly, It also shows the performance of the SST is getting worse once the confidence threshold is larger (over 0.4) because of less data being utilized. Actually, what we observed from Figure 2 is basically consistent with ones in [7].

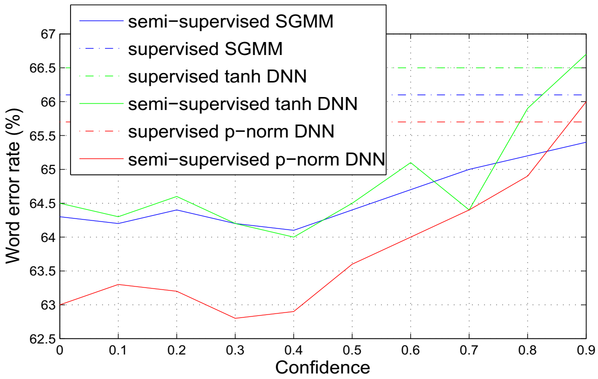

Figure 2: Confidence score versus various acoustic model performance with semi-supervised trained BNF based tandem method

6. Keyword search experiments

One of the primary goals in Babel program is to conduct keyword search (KWS) using corresponding SR system. In this section we are investigating several aspects regarding to lattice density, lattice oracle error rate, and Actual Term Weighted Value (ATWV) change before and after SST, within the three acoustic modeling methods. This is worthwhile since WER and ATWV are two disjoint performance metric, and better WER does not always mean better ATWV as shown in [16]. We also note though it is clearly shown SST can improve recognition and KWS performance with similar data in [17], no other details are given there.

6.1. Task description

We use OpenKWS13 *eval* keyword list to conduct KWS experiments on the *dev* data in Table 1. The list contains 4065 KW instances, 2180 of which have OOV words to the decoding dictionary, resulting 53.6% OOV rate in terms of detection. However such a higher OOV issue is irrelevant with our concern.

6.2. Lattice density versus decoding beam

Figure 3 illustrates lattice density versus decoding beam. From Figure 3 we can infer feature discrimination ability is improved and thinner lattices are generated with corresponding models after SST. This also implies models after SST generally generate lattice faster.

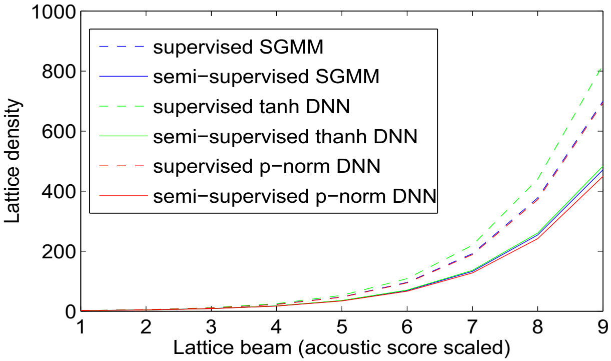

Figure 3: Lattice decoding beam versus density in various acoustic modeling methods

6.3. Lattice oracle error versus density

Denser lattice normally means lower oracle error. However besides this common sense it is also meaningful to know how much oracle error reduction can be achieved by SST. Figure 4 plots lattice oracle error rate versus lattice density for the three systems. From Figure 4 we observe that there is no obvious difference in lattice oracle error rate for SGMM and tanh-DNN systems before and after SST. This might be due to there is only limited WER reduction with SST. However the p-norm DNN clearly demonstrates lattice oracle error rate reduction after SST.

6.4. ATWV versus lattice density

Figure 5 plots ATWV KWS performance versus lattice density. From Figure 5 SGMM and tanh-DNN systems have no ATWV improvement with SST. On the contrary the ATWV of the SGMM system is degraded after SST. However the p-norm DNN system can make ATWV improvement with SST. We guess that ATWV is more closely correlated with lattice oracle error rate than one-best WER from Figure 5. Surprisingly SGMM system yields much worse ATWV than thanh-DNN overall. This is beyond our expectation since it has better oracle error rate than thanh-DNN system from Figure 4. We are suspecting SGMM has much different recognition pattern style compared with DNN system and they should be well complementary.

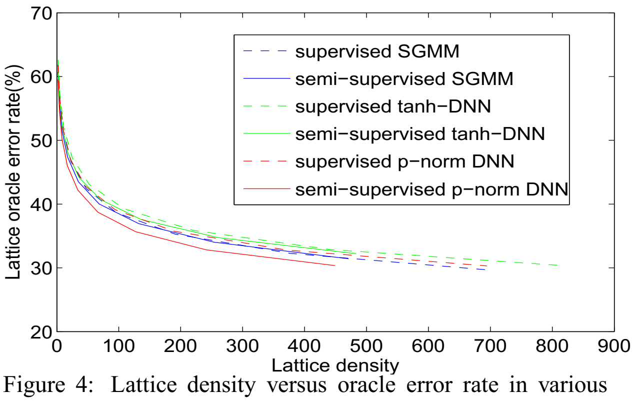

Figure 4: Lattice density versus oracle error rate in various acoustic modeling methods

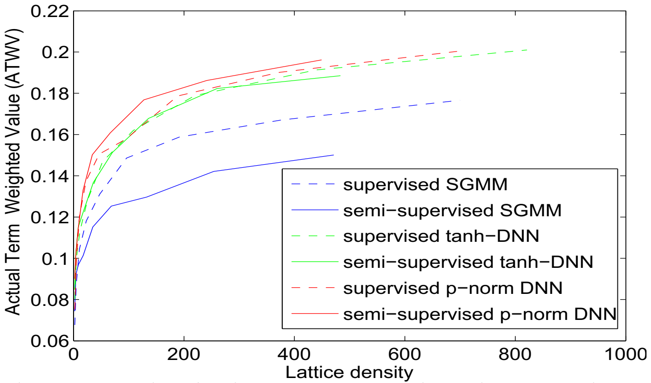

Figure 5: Lattice density versus ATWV in various acoustic modeling methods

7. Conclusions and future work

In this paper, we investigated bottle-neck feature based semi-supervised training using SGMM, tanh-neuron DNN and a generalized maxout p-norm DNN acoustic modeling methods. WER improvements are achieved using SST in all scenarios. We also found p-norm DNN consistently yields the best performance in terms of both SR and KWS tasks with or without semi-supervised training. Future work will be aimed at model based SST training, and hopefully it is complementary to the feature based SST.

8. References

[1] F. Wessel and H. Ney, "Unsupervised training of acoustic models for large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 23–31, 2005.

[2] K. Yu, M. J. F. Gales, L. Wang, and P. C. Woodland, "Unsupervised training and directed manual transcription for lvcsr," *Speech Communications*, vol. 52, no. 7–8, pp. 652–663, 2010.

[3] S. Novotney, R. M. Schwartz, and J. Z. Ma, "Unsupervised acoustic and language model training with small amounts of labelled data," in *Proc. of IEEE ICASSP*. IEEE, 2009, pp. 4297–4300.

[4] 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.

[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. of INTERSPEECH*, 2013.

[6] 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.

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

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

[9] K. Veselý, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Proc. of INTERSPEECH*, 2013.

[10] D. Povey, L. Burget, and et al., "The subspace gaussian mixture model-a structured model for speech recognition," *Computer Speech & Language*, vol. 10, no. 2, pp. 249–264, April 2011.

[11] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, "Improving deep neural network acoustic models using generalized maxout networks," in *Proc. of ICASSP*, 2014.

[12] M. Karafiát, F.Grézl, and M. Hannermann, "But babel system for spontaneous cantonese," in *Interspeech*. ISCA, 2013.

[13] J.G.Fiscus, J. Ajot, J. S. Garofolo, and G. Doddingtion, "Results of the 2006 spoken term detection evaluation," in *SIGIR*. ACM, 2007.

[14] D. R. Miller, M. Kleber, C. Kao, O. Kimball, T. Colthurst, S. A. Lowe, R. M. Schwartz, and H. Gish, "Rapid and accurate spoken term detection," in *Interspeech*. ISCA, 2007.

[15] S. Wegmann, A. Faria, A. Janin, K. Riedhammer, and N. Morgan, "The tao of atwv: probing the mysteries of keyword search performance," in *ASRU*, 2013.

[16] K. M. Knill, M. J. F. Gales, S. P. Rath, P. C. Woodland, C. Zhang, and S. X. Zhang, "Investigation of multilingual deep neural networks for spoken term detection," in *ASRU*, 2013.

[17] 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*, 2013.

[18] 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.

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

[20] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, "Maxout networks," in *arXiv preprint arXiv:1302.4389*, 2013.

[21] 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.