COMPARISON OF UNSUPERVISED SEQUENCE ADAPTATIONS FOR DEEP NEURAL NETWORKS

Akio Kobayashi

NHK Engineering System. Inc.
Tokyo, Japan

Kazuo Onoe, Manon Ichiki, Shoei Sato

NHK (Japan Broadcasting Coporation)
Science and Technology Research Laboratories
Tokyo, Japan

ABSTRACT

This paper compares unsupervised sequence training techniques for deep neural networks (DNN) for broadcast transcriptions. Recent progress in digital archiving of broadcast content has made it easier to access large amounts of speech data. Such archived data will be helpful for acoustic/language modeling in live-broadcast captioning based on automatic speech recognition (ASR). In Japanese broadcasts, however, archived programs, e.g., sports news, do not always have closed-captions used typically as references. Thus, unsupervised adaptation techniques are needed for performance improvements even when a DNN is used as an acoustic model. In this paper, we compared three unsupervised sequence adaptation techniques: maximum a posteriori (MAP), entropy minimization, and Bayes risk minimization. Experimental results for transcribing sports news programs showed that the best ASR performance is brought about by Bayes risk minimization which reflects information as to expected errors, while comparable results are obtained with MAP, the simplest way of unsupervised sequence adaptation.

Index Terms— acoustic modeling, deep neural network, unsupervised adaptation, Bayes risk minimization, deep denoising autoencoder

1. INTRODUCTION

NHK (Japan Broadcasting Corp.) has studied closed-captioning for automatic-speech-recognition (ASR) to resolve accessibility issues for hard-of-hearing persons and launched a closed-captioning system for live news shows in 2012 [1, 2]. While the system achieves high ASR performance when decoding read and clean speech, its accuracy drastically deteriorates under low-SNR/conversational speech conditions. Thus, a large amount of training data is required for performance improvements to be made.

On the other hand, many broadcasters have been making available digital archives. Naturally, these archives offer useful sets of training data for ASR as well as collections of information for public use. However, there are difficulties in utilizing the information in the archives. First, in Japanese broadcasts, programs do not always have closed-captions that reflect their content. Since there is no information as to the content, i.e., speech, except for the program name and genre, transcriptions must be obtained without clues. In addition, background sounds such as noise and music make it more difficult to acquire accurate transcriptions. Thus, denoising may be required to get useful information from program audio streams. As an example of such conditions, we can point to sports news programs that do not have any closed-captions and where background noise exists behind the announcers' speech.

As for the challenges accompanying acoustic modeling, recent progress in the study of deep neural networks (DNN) has shown that DNN-based acoustic models outperform conventional Gaussian mixture models (GMM). Moreover, the deep denoising autoencoder (DDA), which is represented as a DNN regression model, performs well under low SNR conditions [3, 4]. Neural-network-based approaches would be expected to achieve a considerable increase in ASR performance, and ones such as semi-supervised training for DNN would also boost performance [5]. In our case, however, the model adaptation should be conducted in an unsupervised manner due to the lack of closed-captions for the semi-supervised training. In the literature, regular adaptation approaches were proposed in [6, 7] for DNN acoustic modeling. In a speaker adaptation manner, DNN adaptation techniques have also been used [8, 9, 10, 11]. The models are adapted according to speaker information estimated from a small amount of speech data in a fashion similar to cMLLR/fMLLR.

Of the variety of approaches to unsupervised DNN adaptation, we decided to focus on unsupervised adaptation techniques from the perspective of sequence training for DNN [12, 13]. We compared three methods using training data on the scale of tens of hours. a) maximum a posteriori (MAP), b) entropy minimization, and c) Bayes risk minimization. MAP is the simplest way of sequence adaptation and maximizes the posteriors of 1-best sentence hypotheses. Entropy minimization defines competition among the hypotheses as an entropy and minimizes it. Bayes risk minimization is the most

complex technique and minimizes the expected risk, which is derived from the posteriors and edit distances. We explored the efficacy of these techniques through experiments on transcribing broadcast programs under noisy conditions.

2. UNSUPERVISED DNN ADAPTATION

This section describes the three ways of unsupervised sequence adaptation for DNN. Note that all the adaptation techniques are assumed to use n-best lists as training data.

2.1. MAP Adaptation

Supposing that 1-best sentence hypotheses can be regarded as reference labels, one of the simplest sequence adaptation methods is MAP, which maximizes the posteriors or conditional log-likelihoods. The training objective in batch mode is defined as

$$L_{\mathrm{post}}(\Lambda) = \frac{1}{M}\sum_{m}\log p(\boldsymbol{w}_0^m|\boldsymbol{x}^m;\Lambda), \tag{1}$$

where $p(\boldsymbol{w}_0^m|\boldsymbol{x}^m;\Lambda)$ is a posterior of the 1-best sentence hypothesis, $\boldsymbol{w}_0^m$, for the m-th acoustic feature vector, $\boldsymbol{x}_m$. Λ is the set of parameters to be estimated in an unsupervised manner.

The posterior is given by

$$p(\boldsymbol{w}|\boldsymbol{x}) = \frac{1}{Z(\Lambda)}\exp\left\{\lambda_{\mathrm{lm}} f_{\mathrm{lm}}(\boldsymbol{w}) + \lambda_{\mathrm{am}}\sum_{t} f_{\mathrm{ac}}(x_t|s_t;\Lambda)\right\}, \tag{2}$$

where $f_{\mathrm{lm}}(\boldsymbol{w})$ is the logarithmic language score for the sentence hypothesis, $\boldsymbol{w}$, and $f_{\mathrm{am}}(x_t|s_t;\Lambda)$ is the logarithmic acoustic score for the t-th feature of x_t. $Z(\Lambda)$, a normalization factor, i.e., the sum over the n-best hypotheses, is given by

$$Z(\Lambda) = \sum_{\boldsymbol{w}}\exp\left\{\lambda_{\mathrm{lm}} f_{\mathrm{lm}}(\boldsymbol{w}) + \lambda_{\mathrm{am}}\sum_{t} f_{\mathrm{ac}}(x_t|s_t;\Lambda)\right\}.$$

Note that $Z(\Lambda)$ reflects the competition among the hypotheses, which is an essential factor for discriminative training.

When the DNN is utilized as an acoustic model, the acoustic score is decomposed as

$$f_{\mathrm{ac}}(x|s;\Lambda) = f_{\mathrm{dnn}}(s|x;\Lambda) - f_{\mathrm{prior}}(s), \tag{3}$$

where $f_{\mathrm{prior}}(s)$ is the logarithmic prior of the senone, s, which is computed from the training data. $f_{\mathrm{dnn}}(s|x;\Lambda)$ is the output from the DNN, or the logarithmic posterior of the senone, and it is typically activated by the softmax function. However, some decoders, e.g., Kaldi decoder, can use outputs not through the softmax as acoustic scores for convenience [14]. Then, we compute the gradients required for the model parameter updates in a simple manner by removing the softmax layer.

Abbreviating the posterior as $p_n^m \equiv p(\boldsymbol{w}_n^m|\boldsymbol{x}^m;\Lambda)$ for readability, the gradient w.r.t. the element, $\lambda_{ij}^{(\ell)} \in \Lambda$, of the matrix connecting between the j-th and i-th units of the top two layers is defined as

$$\Delta_{ij} = p_n^m\left\{\sum_t\left(\delta_{n,t}^m(j) - \sum_k p_k^m\delta_{k,t}^m(j)\right)\right\}y_i, \tag{4}$$

where y_i is the i-th input to the top layer and $\delta_{n,t}^m(j)$ is a delta function defined as

$$\delta_{n,t}^m(j) = \begin{cases}1 & \text{if } s_{n,t}^m \text{ mathces } j\text{-th senone,}\\ 0 & \text{otherwise.}\end{cases} \tag{5}$$

According to Eq. (4), the propagation error for the j-th output unit of the m-th training data can be derived as

$$\varepsilon_j^m = \sum_t\left(\delta_{0,t}^m(j) - \sum_k p_k^m\delta_{k,t}^m(j)\right). \tag{6}$$

The propagation error leads to the conventional backpropagation via the stochastic gradient decent algorithm for the parameter update.

2.2. Entropy Minimization

In unsupervised training, a conditional entropy function is often defined as a regularizer to be minimized on the training data [15, 16]. The training objective is defined as

$$L_{\mathrm{ent}}(\Lambda) = -\frac{1}{M}\sum_m\sum_n p_n^m\log p_n^m. \tag{7}$$

The entropy regularizer is designed under the assumption that the uncertainty associated with hypothesis discrimination for the unlabeled data should be reduced by the estimated model. As the regularizer can be viewed as an expectation for the scores (log-probabilities) of sentence hypotheses, minimizing the objective is equivalent to making the scores of possibly correct hypotheses larger and the scores of unpromising ones smaller. Unlike the MAP approach, the competition among the hypotheses appears explicitly in entropy minimization.

As with derivation of Eq.(4), the propagation error can be formulated as

$$\begin{aligned}\varepsilon_j^m &= -\sum_n\frac{\partial p_n^m}{\partial\lambda_{ij}^{(\ell)}}\left(\log p_n^m + 1\right)\\ &= -\sum_n p_n^m\left(\log p_n^m + 1\right)\\ &\times \left\{\sum_t\left(\delta_{n,t}^m(j) - \sum_k p_k^m\delta_{k,t}^m(j)\right)\right\}.\end{aligned} \tag{8}$$

2.3. Bayes Risk Minimization

The Bayes risk minimization approach is performed on n-best lists in order to obtain hypotheses with minimum error probabilities [17, 18]. Regarding it as inductive learning, we can introduce a training objective, which reflects error information, as follows:

$$L_{\mathrm{risk}}(\Lambda) = \frac{1}{M} \sum_m \sum_n p_n^m \sum_k r_{n,k} p_k^m, \tag{9}$$

where $r_{n,k}$ is a cost defined as the distance (conventionally, the Levenshtein distance) between the n-th sentence hypothesis and the k-th one.

In a similar fashion, the propagation error can be derived as

$$\begin{aligned} \varepsilon_j^m &= \sum_n \left(\frac{\partial p_n^m}{\partial \lambda_{ij}^{(\ell)}} \sum_k r_{n,k} p_k^m + p_n^m \sum_k r_{n,k} \frac{\partial p_k^m}{\partial \lambda_{ij}^{(\ell)}} \right) \\ &= \sum_n p_n^m \sum_k \left(\phi_n + \phi_k \right) r_{n,k} p_k^m, \end{aligned} \tag{10}$$

where, as a shorthand, ϕ_n is defined as

$$\phi_n \equiv \sum_t \left(\delta_{n,t}^m(j) - \sum_\nu p_\nu^m \delta_{\nu,t}^m(j) \right). \tag{11}$$

From the point of view of calculation amount, the latter two approaches, Bayes risk minimization and entropy minimization, are more complex than the MAP approach because the round-robin competition among the sentence hypotheses are required.

3. EXPERIMENTS

3.1. Setup

3.1.1. Evaluation

As listed in Table 1, we used sports news shows for testing the ASR performance. The shows are characterized by their topics, including sumo tournaments, professional baseball games, and domestic league soccer games. Moreover, the speech is often against a background of music and noise, such as opening/closing themes, jingles, and cheers in the stadium. For the reference labels, the perplexity (PP) and out of vocabulary (OOV) rate were measured using a baseline trigram LM. The LM was trained from Japanese closed-captions (239M) with a 200k vocabulary. For the evaluation, we used a deep denoising autoencoder (DDA) for front-end feature processing followed by a DNN-HMM hybrid decoder based on the Kaldi toolkit [14].

3.1.2. Training and Adaptation

Fig. 1 shows our adaptation scheme. First, the training data were decoded by using a baseline trigram language model (LM) and a DNN-based acoustic model (AM) to obtain the transcriptions. Then, the baseline LM was interpolated with the model estimated from the transcriptions for the adapted LM. The n-best lists were obtained through decoding with the adapted LM. Finally, unsupervised sequence adaptation was conducted on the n-best lists.

Table 1. *Evaluation data*

#utts	#words	PP	PP(adpt)	OOV(%)
1374	12.2k	120.7	80.7	0.6

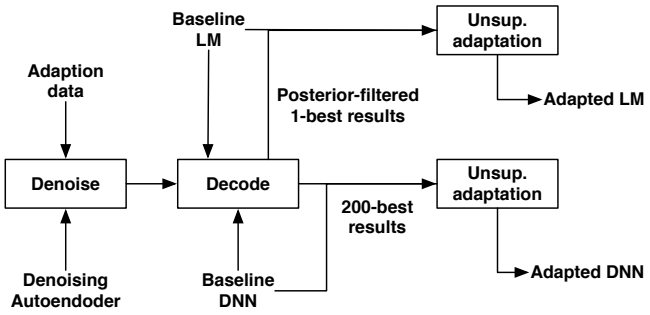

Fig. 1. Diagram of unsupervised sequence adaptation

The DDA was trained as a front-end from 480 hours of broadcast programs mixed with 116.2 hours of noise and music data while changing the SNR from -5 dB to 20 dB. We utilized 40-dimensional log-mel-filterbank with log-energy as an acoustic feature and configured 15-frame splicing features as inputs to the DDA. The DDA network consisted of three hidden layers with 1024 units each and was constructed in a conventional manner, specifically, pre-training followed by fine-tuning.

We used an eight-hidden-layer deep neural network (DNN) for the acoustic modeling. 11-frame splicing features were fed as inputs into the network, and the posteriors of 7000 units were output. Each hidden layer had 2048 units activated by sigmoid functions. The DNN was trained from 1000 hours of speech from broadcast programs in the way of pre-training and fine-tuning and utilized as baseline for adaptation.

For unsupervised sequence adaptation, we utilized 73.4 hours of speech data from similar sports news programs as the adaptation data (Table 2). The n-best lists for adaptation were obtained by decoding with the adapted trigram LM and the baseline DNN. Then, according to Eqs. 6,8 and 10, all the DNN parameters including affine transforms and biases were fully updated and adapted by the conventional backpropagation. In addition, for the detailed analysis, we prepared the subsets of training data by thresholding on the basis of 1-best sentence posteriors.

Table 2. *Training data for unsupervised sequence adaptation*

threshold	hours	#segments	#words
0.0(all)	73.4	75.9k	78.7k
0.1	47.5	53.5k	50.9k
0.3	29.7	35.6k	31.5k
0.5	19.8	24.6k	21.0k
0.7	12.9	16.6k	13.6k
0.9	7.0	9.6k	7.3k

Table 3. *Overall results (%)*

		WER(%)	ERR(%)
Baseline		49.1	-
+DDA		31.8	35.2
+Adpt. LM		29.3	7.9
+Adpt. AM	MAP	28.5	2.7
	MinEnt	28.5	2.7
	MBR	**28.4**	3.1

Table 4. *Detailed results (subsets, %)*

threshold	MAP	MinEnt	MBR
0.0(all)	28.5	28.5	**28.4**
0.1	28.5	28.6	**28.4**
0.3	28.6	28.5	28.5
0.5	28.8	28.5	28.6
0.7	28.8	28.7	28.8
0.9	28.9	28.8	28.9

Table 5. *Detailed results (speakers, %)*

		Anchors		Others	
		WER	ERR	WER	ERR
Baseline		42.3	-	87.3	-
+DDA		24.5	42.1	72.9	16.5
+Adpt. LM		21.8	11.0	71.5	1.9
+Adpt. AM	MAP	21.0	3.8	70.8	1.0
	MinEnt	21.0	3.8	70.8	1.0
	MBR	20.9	4.1	70.8	1.0

3.2. Experimental Results

Table 4 shows the word error rates (WERs) and relative error reduction rates (ERRs) for the evaluation data. The **Baseline** result was obtained from the baseline trigram LM and the DNN without using the DDA front-end processing. **MAP**, **MinEnt** and **MBR** denote the results from the DNNs estimated by using MAP, entropy minimization and Bayes risk minimization, respectively. As shown in the table, the effect of DDA appeared to be extremely large, and it produced a relative reduction of 35.2 % compared with the **Baseline** result. The results from the adapted LM also reduced WER by 7.9 % against the DDA result. Clearly, this is because the transcriptions of the training data matched the evaluation data in terms of topic.

Compared with these improvements, the gains from unsupervised sequence adaptation methods remained modest, with error reductions of around 3 % absolute to the adapted LM result. Among the three unsupervised approaches, **MBR** achieved the best WER of 28.4 % and reduced WER by 42.2 % compared with the **Baseline**. Although matched-pair testing [19] showed that there are no significant differences between **MBR** and the others, Bayes risk minimization would probably lead to the best performance when it is trained from a larger amount of data as we reported in discriminative language modeling [20].

We further explored the efficacy of the unsupervised sequence adaptation methods from the perspective of data amount. Table 5 shows the results from the DNNs estimated while changing the subsets of training data. As the amount of data increased by reducing the threshold posteriors, the WERs of the training methods steadily improved. Regardless of the adaptation method, the DNNs trained from the same amount of data achieved similar WERs. The adaptation methods did not cause any differences in performance when the DNNs were adapted by matched-conditioned training data.

Finally, Table 6 shows the results for the different kinds of speaker. The evaluation data can be classified into two speaker types: **Anchors** (10.8k words) and **Others** (1.8k words). The **Anchors** subset includes commentary in the studio and field, while **Others** consists of utterances by players in stadiums, on baseball fields, and soccer pitches. As is clear from the table, fewer word errors were reduced in the results for **Others** than in those for **Anchors**, even when DDA and the adapted LM were used. Moreover, the unsupervised DNN adaptation yielded only small gains. Such poor performance could have been caused by the degraded speech in **Others** that was recorded with much reverberation in addition to background noise. Naturally, as the quality of transcriptions obtained from such difficult conditions is low, using DNN adaptation on them would probably be inefficient and insufficient.

4. CONCLUSION

We explored the sequence model adaptation approaches of DNNs in an unsupervised manner. The experimental results showed that the Bayes risk minimization method performed the best, while the MAP approach achieved comparable results. However, a further experimental study would be required for comparison among the adaptation methods because the results were preliminarily obtained by a small amount of training data and small n-best lists. Moreover, we will explore lattice-based approaches to achieve further WER reductions.

5. REFERENCES

[1] T. Imai, S. Homma, A. Kobayashi, T. Oku, and S. Sato, "Speech recognition with a seamlessly updated language model for real-time closed-captioning," in *Proc. Interspeech*, 2010, pp. 262–265.

[2] A. Kobayashi, Y. Fujita, T. Oku, S. Sato, S. Homma, T. Arai, and T. Imai, "Live closed-captioning system using hybrid automatic speech recognition for broadcast news," in *Proc. NAB Broadcast Engineering Conference*, 2013, pp. 277–283.

[3] X. Feng, Y. Zhang, and J. Glass, "Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition," in *Proc. ICASSP*, 2014, pp. 1759–1763.

[4] P. Vincent, H. Larochele, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion," *The Journal of Machine Learning Research*, vol. 11, pp. 2271–3408, 2010.

[5] H. Liao, E. McDermott, and A. Senior, "Large scale deep neural network acoustic modeling with semi-supervised training data for YouTube video transcription," in *Automatic Speech Recognition and Understanding (ASRU)*, 2013, pp. 368–373.

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

[7] K. Vaselỳ, M. Hannemann, and L. Burget, "Semi-supervised training of deep neural networks," in *Proc. ASRU*, 2013, pp. 267–272.

[8] T. Yoshioka, A. Ragni, and M.J.F. Gales, "Investigation of unsupervised adaptation of DNN acoustic models with filter bank input," in *Proc. ICASSP*, 2014, pp. 6344–6348.

[9] M. Mimura and T Kawahara, "Unsupervised speaker adaptation of dnn-hmm by selecting similar speakers for lecture transcription," in *Proc. APSIPA*, 2014, pp. 1–4.

[10] P. Swietojanski and S. Renals, "Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models," in *Spoken Language Technology Workshop*, 2014, pp. 171–176.

[11] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Automatic Speech Recognition and Understanding (ASRU)*, 2013, pp. 24–29.

[12] K. Veselý, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Proc. Interspeech*, 2013, pp. 2345–2349.

[13] H. Su, G. Li, D. Yu, and F. Seide, "Error back propagation for sequence training of context-dependent deep networks for conversational speech transcription," in *Proc. ICASSP*, 2013, pp. 6664–6668.

[14] 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. Vesely, "The Kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011.

[15] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," *Advances in neural information processing systems*, pp. 529–536, 2005.

[16] A. Kobayashi, T. Oku, T. Imai, and S. Nakagawa, "Multi-objective optimization for semi-supervised discriminative language modeling," in *Proc. ICASSP*, 2012, pp. 4997–5000.

[17] V. Goel and W. Byrne, "Minimum Bayes-risk automatic speech recognition," *Computer Speech and Language*, vol. 14, pp. 115–135, 2000.

[18] R. Schlueter, M. Nussbaum-Thom, and H. Ney, "On the relation of Bayes risk, word error, and word posteriors in ASR," in *Proc. Interspeech*, 2010, pp. 230–233.

[19] L. Gillick and S.J. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proc. ICASSP*, 1989, pp. 532–535.

[20] A. Kobayashi, T. Oku, S. Homma, T. Imai, and S. Nakagawa, "Lattice-based risk minimization training for unsupervised language model adaptation," in *Proc. Interspeech*, 2011, pp. 1453–1456.