EVALUATION OF MULTI-LEVEL CONTEXT-DEPENDENT ACOUSTIC MODEL FOR LARGE VOCABULARY SPEAKER ADAPTATION TASKS

Hung-An Chang and James Glass

MIT Computer Science and Artificial Intelligence Laboratory
32 Vassar Street, Cambridge, MA, 02139, USA
Email: {hung_an,glass}@csail.mit.edu

ABSTRACT

In this paper, we investigate the ability of a recently proposed discriminatively trained, multi-level context-dependent acoustic model to adapt to a new speaker in both supervised and unsupervised adaptation scenarios. Speaker adaptive speech recognition experiments performed on a large-vocabulary spoken lecture task show that the multi-level model reduces word error rates by more than 10% in both cases as compared to the conventional clustering-based decision-tree context-dependent acoustic model approach.

Index Terms— Multi-level acoustic model, context-dependent model, speaker adaptation, discriminative training, LVCSR

1. INTRODUCTION

Speaker adaptation methods have been studied extensively for automatic speech recognition (ASR). One of the simplest and most effective methods is maximum a posteriori (MAP) adaptation that updates Gaussian mixture model (GMM) parameters of a speaker independent (SI) model to maximize the posterior probability of the adaptation data with respect to the updated parameters [1]. Maximum likelihood linear regression (MLLR) groups Gaussian components and estimates a linear transform of the SI GMM means to the corresponding speaker dependent (SD) distribution in order to maximize the likelihood of the adaptation data [2]. Eigenvoice analysis and reference speaker weighting use multiple reference speakers to represent a speaker vector that is a concatenation of Gaussian means. The adapted speaker vector is determined using a maximum likelihood (ML) criterion to derive a linear combination of the reference speaker vectors [3, 4]. Note that while these methods are all effective with limited adaptation data, MAP-based adaptation typically provides the largest improvement in ASR word error rate (WER) when there is a significant quantity of adaptation data [3].

While the aforementioned methods use ML-based criterion, discriminative methods have also been successfully used for speaker adaptation. Instead of seeking to model parameters that maximize the likelihood of adaptation data, discriminative training methods seek parameters that can minimize the amount of confusion reflected in a computable objective function. Several types of objective functions have been applied to construct discriminative speaker adaptation frameworks. For example, maximum mutual information (MMI) training statistics have been used to formulate a conditional MLLR adaptation framework [5]. Minimum phone error (MPE) training has also been shown to be effective in estimating the regression transformation matrix [6]. If enough adaptation data are available, the entire set of GMM parameters can be adapted via a discriminative criterion with the discriminative MAP method [7]. In addition, training criterion such as minimum classification error (MCE) have also been shown to be effective for speaker adaption [8].

Recently, we introduced a novel method for context-dependent acoustic modeling that is based on a discriminatively trained, multi-level framework for integrating acoustic models with differing degrees of contextual granularity. In our initial speaker-independent experiments we were able to reduce ASR WERs on a large vocabulary spoken lecture transcription task when compared to a similar model that was based on the conventional clustering-based decision tree approach for determining contextual equivalence classes [9]. In this paper, we extend this research by examining the ability of the multi-level model to adapt to a new speaker within a discriminatively trained framework. In the following sections, we describe the multi-level modeling, discriminative training, and MCE-based adaptation we implemented for this model. We then describe a series of speaker adaptation experiments for a large vocabulary spoken lecture processing task.

2. MULTI-LEVEL ACOUSTIC MODELS

Context-dependent (CD) acoustic modeling has become a standard modeling procedure for most large vocabulary speech recognizers as a mechanism to model coarticulatory variation that occurs during speech production. Typically, the local phoneme context is used as a means to define CD units. Since the number of possible units grows exponentially with the length of the local context, many units do not have enough training examples to produce a robust model.

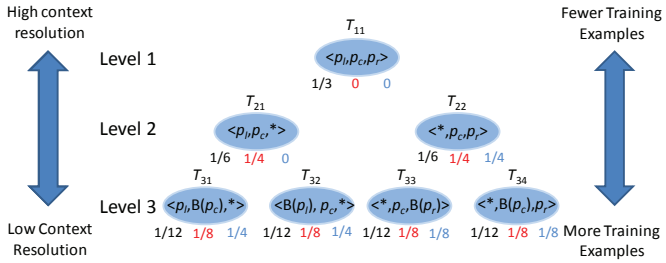

Fig. 1. Classifiers and combination weights for the triphone "p_l-p_c+p_r". The "*" symbol means to ignore certain contexts. The function $B(\cdot)$ reduces a phonetic unit into its corresponding broad-class, e.g. B("n") ="Nasal". Under this notation, T_{33}("k-oy+n") =<*,"oy",B("n")> which represents a classifier that identifies an "oy" with right context "Nasal". The numbers of the same color represent a possible setting of combination weights. Different weight settings can be used to avoid data sparsity effect.

For many years, this data sparsity issue has been addressed by a clustering-based approach that learns a decision tree to group contexts into clusters that each have enough training examples to create a robust model [10]. While clustering addresses the sparsity issue, it also inherently quantizes the contexts; that is, different CD units within a cluster will always have the same acoustic likelihood, making the units acoustically indistinguishable from each other. This quantization effect is not completely negligible. Typically, the number of clustered states in a conventional triphone-based large vocabulary ASR system is on the order of $10^3 \sim 10^4$, which is one or two orders of magnitude smaller than the potential number of triphones. This difference can hinder the discriminative ability of the resulting CD acoustic model.

In order to address the sparsity and quantization issues, we have recently explored a multi-level CD acoustic modeling framework [9]. The basic idea of the multi-level model is to associate each CD unit with a set of GMM classifiers that identify contexts at multiple levels of resolution, linearly combining the classifier outputs for scoring. By appropriately choosing the classifiers, every pair of CD units will have at least one differing classifier, making them mutually distinguishable to the speech recognizer.

Figure 1 illustrates the multi-level concept as implemented in [9] to setup classifiers and combination weights for a triphone "p_l-p_c+p_r". An important aspect of the multi-level model is that while each classifier at a lower level ignores certain contextual details, if each classifier at the same level contributes a likelihood for observation $\mathbf{x}$, then the full context of $\mathbf{x}$ can be identified. Another result of this criterion is that at least one classifier will differ between any pair of CD units, which will generally produce differing acoustic scores for all CD units.

Having selected the classifiers, the acoustic score of $\mathbf{x}$ with respect to a CD label s can be computed by

$$a_{\boldsymbol{\lambda}}(\mathbf{x}, s) = \sum_{i=1}^{3} \sum_{j=1}^{J_i} w_{ij}^s l_{\boldsymbol{\lambda}}(\mathbf{x}, T_{ij}(s)), \quad (1)$$

where $l_{\boldsymbol{\lambda}}(\mathbf{x}, T_{ij}(s))$ denotes the log-likelihood of $\mathbf{x}$ with respect to the GMM classifier $T_{ij}(s)$, J_i denotes the number of classifiers at level i, and w_{ij}^s is a non-negative combination weight satisfying the convex constraint $\sum_{i=1}^{3}\sum_{j=1}^{J_i} w_{ij}^s = 1$. To address the data sparsity issue, a classifier combination weight is zeroed when it does not have enough training examples. This prevents it from contributing to the overall acoustic score, as shown by the red/blue colored weights in Figure 1.

2.1. Discriminative Training of Multi-Level Model

As is the case for the conventional cluster-based CD model, discriminative training is performed on the multi-level CD model by taking the gradient of the objective function with respect to the GMM parameters. This can be computed by first taking partial derivatives with respect to each acoustic score and then summing up the contribution of the gradient with respect to each acoustic score

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\lambda}} = \sum_{n=1}^{N} \sum_{\mathbf{x} \in \mathbf{X}_n} \sum_{s} \frac{\partial \mathcal{L}}{\partial a_{\boldsymbol{\lambda}}(\mathbf{x}, s)} \frac{\partial a_{\boldsymbol{\lambda}}(\mathbf{x}, s)}{\partial \boldsymbol{\lambda}}. \quad (2)$$

Since the acoustic score can be decomposed into a linear combination of the log-likelihood of GMMs as in Eq. (1), the gradient of the acoustic score can be further computed by

$$\frac{\partial a_{\boldsymbol{\lambda}}(\mathbf{x}, s)}{\partial \boldsymbol{\lambda}} = \sum_{i=1}^{3} \sum_{j=1}^{J_i} w_{ij}^s \frac{\partial l_{\boldsymbol{\lambda}}(\mathbf{x}, T_{ij}(s))}{\partial \boldsymbol{\lambda}}. \quad (3)$$

In this way, computing the gradient of the multi-level CD model can be done by first computing the partial derivative with respect to each acoustic score as in conventional discriminative training, and then distributing the contribution of each GMM with respect to the combination weights.

While the prefixed weights in Figure 1 work reasonably well, the combination weights can also be automatically learned by performing a constrained optimization on the objective function used for discriminative training [9]. Let $\mathbf{W}$ be the combination weights, and $\overline{w}_{ij}^s$ be the prefixed weight of the j^{th} classifier at level i for label s. The optimization for the weights can be expressed as follows

$$\min_{\mathbf{W}} \mathcal{L}(\mathbf{W}) + \alpha \sum_{s,i,j} ||w_{ij}^s - \overline{w}_{ij}^s||^2 \quad \text{s.t.}$$
$$\sum_{i,j} w_{ij}^s = 1 \ \forall s, w_{ij}^s \geq 0 \ \forall s, i, j, w_{ij}^s = 0 \ \forall \overline{w}_{ij}^s = 0, \quad (4)$$

where the first term is the objective function used for discriminative training, and the second term is a regularization term. The first two sets of constraints ensure convexity; the last set ensures no unreliable classifier contributes to the scores.

2.2. Multi-Level Model Adaptation

To adapt model parameters in a discriminative way, we applied the MCE criterion to perform speaker adaptation. Instead of interpolating the speaker-independent model with MCE updated speaker-dependent model as in [8], we directly apply MCE training to update the SI model parameters over the adaptation data. The MCE loss function used in the experiment can be expressed as

$$\mathcal{L} = \sum_{n=1}^{N} \ell(-L_{\boldsymbol{\lambda}}(\mathbf{X}_n, \mathbf{Y_n}) + \log([\tfrac{1}{K} \sum_{\mathbf{S} \in \boldsymbol{\mathcal{S}}_n^K} \exp(\eta L_{\boldsymbol{\lambda}}(\mathbf{X}_n, \mathbf{S}))]^{\frac{1}{\eta}})), \quad (5)$$

where N is the number of adaptation utterances, $L_{\boldsymbol{\lambda}}(\mathbf{X}_n, \mathbf{Y_n})$ denotes the recognition score of the reference path $\mathbf{Y_n}$, $L_{\boldsymbol{\lambda}}(\mathbf{X}_n, \mathbf{S})$ denotes the score for the hypothesis $\mathbf{S}$, $\boldsymbol{\mathcal{S}}_n^K$ is the best K incorrect hypotheses of the n^{th} utterance, η is a parameter that determines the relative importance of the hypotheses, and $\ell(\cdot)$ is a sigmoid function that maps the score difference into a value between 0 and 1. The quickprop algorithm in [11] was used for parameter optimization.

3. EXPERIMENTS

3.1. Task and ASR Configuration

Experiments were performed on the MIT lecture corpus consisting of approximately 119 hours of audio recordings and transcriptions from a variety of talkers and topics[12]. A preliminary set of SI acoustic models were trained from these data. Feature vectors were extracted at a 10ms frame rate. Each vector consisted of average values of 14 MFCCs in 8 telescoping regions spanning 150ms. The 112 dimensional vectors were reduced to 50 dimensions by a composition of neighborhood component analysis and principal component analysis as in [13].

For the baseline clustering-based (CL) model, a decision tree was used to cluster the triphone states [10]. The stopping criteria for number of clusters and model size were tuned on two held out lectures. For the multi-level (Multi) model, 9 broad-classes were used to construct low-level classifiers. For both types of models, parameters were initialized by ML criterion and were refined by MCE training [9]. For notation purposes, we use SI-ML-CL and SI-ML-Multi to refer to the SI ML-trained models that use the two CD methods, while SI-MCE-CL and SI-MCE-Multi correspond to the SI MCE-trained CD models.

A standard vocabulary of 37K words was used for this task; a trigram language model was trained on training lecture texts, Switchboard conversations, and the Michigan Corpus of Academic Spoken English via the SRILM toolkit [14]. The trigram language model was converted to a finite-state transducer (FST) representation by the MIT FST toolkit [15], and was composed with other lexicon-level FSTs to form the search module [9].

3.2. Supervised Adaptation

A series of lectures on introductory mechanical physics taught by a Dutch-accented lecturer were used for the speaker adaptation experiments. The audio and transcripts of first 30 lectures of the series were used as potential adaptation data, while the last 3 lectures were used as test data. To obtain speaker adaptive (SA) models, an additional MCE training run was applied to the SI-MCE-CL and SI-MCE-Multi models on the adaptation lectures, resulting in SA-MCE-CL and SA-MCE-Multi models, respectively. To compare ML-based adaptation with discriminative-based adaptation, MAP adaptation was also performed on the SI-ML-CL models to produce a SA-ML-CL model. A clustering-based speaker-dependent (SD) model was also trained on the available adaptation data. MCE training was applied to produce a SD-MCE-CL model.

As shown in solid lines in Figure 2, the WER results of the four models were measured using different amounts of supervised adaptation data (i.e., transcripts known). As shown in the figure, both MAP and MCE adaptation achieved significant WER reductions over the SI models (i.e. 0 adaptation lectures). When less than 10 lectures of adaptation data were used, the SA-MCE-CL models performed better than SD-MCE-CL, although as more adaptation became available the opposite was observed. In contrast, the SA-MCE-Multi model consistently outperformed all CL models over all adaptation amounts, demonstrating better model adaptability.

3.3. Unsupervised Adaptation

Since adaptation transcripts are not always available, we also performed unsupervised adaptation (UA) experiments using these data to compare the clustering-based and multi-level CD models. For these experiments we used the baseline SI acoustic model to decode the available adaptation lectures. We then used the recognition hypothesis as a reference and performed MCE adaptation with a unigram language model. The performance of the resulting UA-MCE-CL and UA-MCE-Multi models are shown in dashed lines in Figure 2. Although the gain over the SI model was much smaller than observed for the supervised scenario, the multi-level model still provided about 10% relative improvement over the clustering-based model over these adaptation conditions.

3.4. Discussion

Although the multi-level CD models performed well under both supervised and unsupervised adaptation conditions, there are several areas of ongoing investigation. In particular, we have not achieved significant gains by optimizing the classifier weights as described in Eq. (4). As compared to the default weight settings shown in Figure 1, we found slight WER improvements if the weights were optimized prior to MCE GMM parameter adaptation, but these gains vanished post MCE GMM adaptation. Moreover, if weight optimiza-

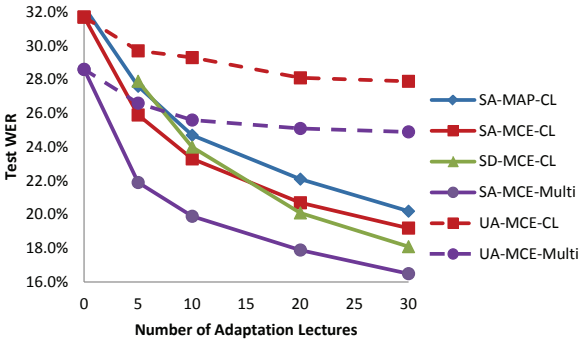

Fig. 2. Results for supervised and unsupervised adaptation.

tion was performed after MCE GMM parameter adaptation, the test WER actually increased slightly. By observing the optimized classifier weights, it appears that many top-level classifiers (i.e., most context-dependent) had too large of a weight which suggests over-training on the adaptation data, and reducing the model generalizability to test data.

We also continue to investigate unsupervised adaptation approaches. Instead of always using the SI models to produce transcript hypotheses, we have also explore an incremental approach whereby we adapt in "chunks" of five lectures . In this scenario, the SI models are used to produce transcripts for the first five lectures, upon which an initial SA model is produced. This new model is used to produce transcripts for the next five lectures, before a new SA model is produced on all ten lectures. This process can be iterated through all 30 adaptation lectures. Our initial tests of the incremental procedure did not produce a monotonically decreasing WER however. Preliminary analyses indicated that the hypotheses generated by the incremental models had a higher insertion rate of unnecessary fillers and short function words. Our future research will attempt to address this issue, as well as explore alternative adaptation mechanisms including the use of confidence scoring for unsupervised adaptation. Currently however, using the SI models to generate transcript hypotheses seems to be the most stable way to adapt to a new speaker.

4. CONCLUSION

In this paper, we compared the performances of a newly proposed multi-level context-dependent acoustic model with a conventional clustering-based model on a large vocabulary speaker adaptive ASR task. Based on a series of experiments, the multi-level model had more than a 10% WER relative improvement over the baseline model for both supervised and unsupervised adaptation scenarios. Future work includes further investigation of classifier weight optimization, alternative methods for unsupervised adaptation, as well as experiments with "mostly correct" transcripts that have been generated via crowdsourcing-based methods [16].

Acknowledgements
This work is supported by the T-Party Project, a joint research program between MIT and Quanta Computer Inc., Taiwan.

5. REFERENCES

[1] J. Gauvain and C.-H. Lee, "Maximum a posterior estimation for multivariate gaussian mixture observations of Markov chains," *IEEE Trans. on Speech and Audio Processing*, vol. 2, pp. 291–298, Apr. 1994.

[2] C. J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[3] R. Kuhn, J. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Trans. on Speech and Audio Processing*, vol. 8, pp. 695–707, Nov. 2000.

[4] T. J. Hazen and J. R. Glass, "A comparison of novel techniques for instantaneous speaker adaptation," *Eurospeech*, pp. 2047–2050, 1997.

[5] A. Gunawardana and W. Byrne, "Discriminative speaker adaptation with conditional maximum likelihood linear regression," *Eurospeech*, pp. 1203–1206, 2001.

[6] L. Wang and P. C. Woodland, "MPE-based discriminative linear transform for speaker adaptation," *Proc. ICASSP*, pp. 321–324, 2004.

[7] D. Povey, P. C. Woodland, and M. J. F. Gales, "Discriminative MAP for acoustic model adaptaion," *Proc. ICASSP*, pp. 312–315, 2003.

[8] T. J. Hazen and E. McDermott, "Discriminative MCE-based speaker adaptation of acoustic models for a spoken lecture processing task," *Proc. Interspeech*, pp. 1577–1580, 2007.

[9] H.-A. Chang and J. Glass, "Multi-level context-dependent acoustic modeling for automatic speech recognition," *ASRU*, 2011.

[10] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modeling," *Proc. Human Language Technology*, pp. 307–312, 1994.

[11] E. McDermott, T. J. Hazen, J. L. Roux, A. Nakamura, and S. Katagiri, "Discriminative training for large-vocabulary speech recognition using minimum classification error," *IEEE Trans. Audio, Speech, and Language Processing*, pp. 203–223, Jan. 2007.

[12] A. Park, T. J. Hazen, and J. R. Glass, "Automatic processing of audio lectures for information retrieval: vocabulary selection and lagnage modeling," *Proc. ICASSP*, pp. 497–500, 2005.

[13] N. Singh-Miller, *Neighborhood analysis methods in acoustic modeling for automatic speech recognition*, Ph.D. thesis, Massachusetts Institute of Technology, Massachusetts, 2010.

[14] A. Stokle, "Srilm: an extensible language modeling toolkit," *Proc. ICSLP*, pp. 901–904, 2002.

[15] I. L. Hetherington, "MIT finite-state transducer toolkit for speech and language processing," *Proc. ICSLP*, pp. 2609–2612, 2004.

[16] C.-Y. Lee and James Glass, "A transcription task for crowdsourcing with automatic quality control," *Proc. Interpseech*, pp. 3041–3044, 2011.