Discriminative MCE-Based Speaker Adaptation of Acoustic Models for a Spoken Lecture Processing Task

Timothy J. Hazen[1] *and Erik McDermott*[2]

[1]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA 02139, USA
[2]NTT Communication Science Laboratories, NTT Corporation, Kyoto 619-0237, Japan

Abstract

This paper investigates the use of minimum classification error (MCE) training in conjunction with speaker adaptation for the large vocabulary speech recognition task of lecture transcription. Emphasis is placed on the case of supervised adaptation, though an examination of the unsupervised case is also conducted. This work builds upon our previous work using MCE training to construct speaker independent acoustic models. In this work we explore strategies for incorporating MCE training into a model interpolation adaptation scheme in the spirit of traditional maximum a posteriori probability (MAP) adaptation. Experiments show relative error rate reductions between 3% and 7% over a baseline system which uses standard ML estimation instead of MCE training during the adaptation phase.

Index Terms: acoustic modeling, MCE training, speaker adaptation.

1. Introduction

Discriminative training of speech recognizers, using either Minimum Classification Error (MCE) or Maximum Mutual Information (MMI), has been shown to improve performance over standard maximum likelihood (ML) training in a number of studies [1, 2, 3]. The benefit of discriminative training over ML-based training is that model parameters can be tuned specifically to improve recognition accuracy. In particular, the MCE framework uses a smooth estimate of the classification risk as the criterion function to be minimized. This approach allows for a more efficient use of model parameters, typically resulting in better performance and smaller model sizes.

After incorporating discriminative techniques into the training of acoustic models, it is natural to examine whether the use of MCE training can be extended to the task of supervised speaker adaptation. Along these lines, maximum likelihood linear regression (MLLR) adaptation has been recast in an MCE framework [4, 5, 6, 7]. Similarly, a method for revising MMI estimation to incorporate a prior distribution leads to a discriminative form of maximum *a posteriori* probability (MAP) adaptation [8]. In the case of unsupervised adaptation, it is not as clear if discriminative techniques such as MCE can be used effectively, because they require knowledge of the correct answer during training. In these cases, it is useful to examine if the generative ML and MAP estimation techniques can be used effectively in conjunction with baseline models trained using the MCE technique.

In this paper we examine the use of MCE training applied to the problem of speaker adaptation within a spoken lecture transcription task. This large vocabulary transcription task typically involves the processing of audio files which are 45 to 90 minutes in length and contain (almost entirely) speech from a single primary speaker. In addition, in cases such as the video lectures available on MIT's OpenCourseWare web site, data from a whole semester's worth of classes taught by a single lecturer may be available. The potential access to large amounts (i.e., many hours) of speech from individual speakers could allow discriminative speaker adaptation techniques to yield significant improvements over traditional ML or MAP based adaptation techniques. In this paper we apply the use of MCE training to a MAP-style model interpolation adaptation technique.

2. MCE Training Overview

The approach to MCE-based optimization in this paper follows that described in [3]. A loss function is defined in terms of a comparison between the log-likelihood for the correct word sequence with that for the best incorrect word sequence. The overall loss function, summed over the training set, is minimized using a simple batch-oriented second-order modified Newton's method, Quickprop [9].

The MCE formalism uses the speech recognition likelihood expression $\mathrm{p}(\mathbf{x}|\mathbf{\Lambda}, S_j)\mathrm{P}(S_j)$ for a hypothesized string S_j as its discriminant function, where $\mathbf{x}$ represents the acoustic feature vectors of the utterance and $\mathbf{\Lambda}$ represents a set of model parameters. A misclassification measure $d_k(\mathbf{x}, \mathbf{\Lambda})$ is used to compare the match between an utterance's correct string S_k and the best incorrect string as follows:

$$d_k(\mathbf{x}, \mathbf{\Lambda}) = \log \frac{\max_{i \neq k} \mathrm{p}(\mathbf{x}|\mathbf{\Lambda}, S_i)\mathrm{P}(S_i)}{\mathrm{p}(\mathbf{x}|\mathbf{\Lambda}, S_k)\mathrm{P}(S_k)} \qquad (1)$$

This expression is positive when the best incorrect string S_i has a larger score than that of the correct string S_k, and negative otherwise. A loss function then maps the misclassification measure to a 0-1 continuum. The loss function used in this work is a sigmoid,

$$\ell(d_k(\mathbf{x}, \mathbf{\Lambda})) = \frac{1}{1 + e^{-\alpha d_k(\mathbf{x}, \mathbf{\Lambda})}}, \qquad (2)$$

though many other loss functions are possible. When the derivative of the sigmoid loss function is taken, the sigmoid function assigns more weight to training utterances with a misclassification measure near zero, and less weight to utterances whose recognition is either strongly correct or strongly incorrect, i.e., those with large misclassification magnitudes.

The core component of the MCE training algorithm is the computation of the derivative of the loss function (Equation 2) with respect to each of the model parameters in $\mathbf{\Lambda}$ to be trained. The Quickprop method then uses those derivatives to update the model parameters.

3. Speaker Adaptation

3.1. MAP-Style Model Interpolation

Maximum *a posteriori* probability (MAP) estimation is a commonly used technique for speaker adaptation of Gaussian mixture models. MAP adaptation performs an interpolation process which slowly shifts the parameters of an *a priori* (or speaker independent) model towards the maximum likelihood (ML) estimated parameter settings for a particular speaker as more data from that speaker is observed. Full details of the algorithm can be found in [10]. Theoretically, MAP adaptation has the desirable property of converging to the ML estimated model asymptotically with increasing amounts of adaptation data. In practice, MAP adaptation requires experimental effort to effectively set its learning rate parameters.

An alternative approach is one we will refer to as MAP-style model interpolation (MI) adaptation. This approach interpolates the output scores of two density functions as follows:

$$\mathrm{p}_{sa}(\mathbf{x}) = \frac{c}{c+\tau}\mathrm{p}_{sd}(\mathbf{x}) + \frac{\tau}{c+\tau}\mathrm{p}_{si}(\mathbf{x}) \qquad (3)$$

Here, $\mathrm{p}_{sa}(\mathbf{x})$ is the speaker adapted (SA) model for scoring observation $\mathbf{x}$, $\mathrm{p}_{sd}(\mathbf{x})$ is the speaker dependent (SD) ML estimated model for the new speaker, $\mathrm{p}_{si}(\mathbf{x})$ is the speaker independent (SI) model, c is number of adaptation observations available for the model, and τ is a learning rate parameter (τ is set to 50 in our experiments). In this expression, rather than adjusting the parameters of an existing SI model, the system learns a new ML estimated SD density function for the speaker from scratch. This allows the SD estimated model to use a smaller number of Gaussian components than the full SI model, especially when the number of adaptation observations c is small. As c grows larger, the number of Gaussian components can be grown to improve the modeling power of the SD model. Eventually, as c grows very large, the contribution of the SI model becomes negligible and the full SD model can replace the interpolated SA version. In practice, we have observed that MI adaptation performs very similarly to traditional MAP adaptation on tasks we have investigated and is slightly easier to tune.

3.2. Extensions to MCE Training

To implement supervised MCE training within the model interpolation adaptation framework, we examine two possible approaches. The first approach, which we will refer to as MCE-ML adaptation, starts with an ML estimated speaker dependent model trained from the adaptation data. During each training iteration, the MCE procedure computes the loss function from a recognition pass over the adaptation data that uses only the ML estimated SD model, and then adjusts the SD model. When training is completed, the updated SD model is then combined with the SI model using model interpolation to create the SA model.

The second approach, which we will refer to as MCE-MI adaptation, works similarly to MCE-ML. However, during each MCE recognition pass the system instead computes the loss function using the MI combination of the SD model and the SI model. Then, during the MCE update phase, only the parameters of the original SD model are adjusted. The original SI model components in the model interpolation remain fixed at all times. In this approach, the MCE updates of the SD model are scaled by the interpolation weight of the SD model. This causes models with fewer adaptation observations to receive smaller MCE adjustments than models with more adaptation observations, thus limiting MCE over-training of the SD model.

3.3. Supervised vs. Unsupervised Adaptation

Speech recognition training algorithms generally use knowledge of the underlying words spoken in the training utterances. Unsupervised learning is required when the underlying word sequence is not known. For example, in the lecture processing task, it is often desirable to create speaker adapted models for a new speaker who is present in an audio file that the system is transcribing. Typically, this form of unsupervised adaptation is performed by assuming that the word sequence obtained from an initial recognition pass over the data can be used as a proxy for the true, but unknown, underlying word sequence.

Even when errors are present in the assumed transcription, improvements can be obtained from unsupervised adaptation. Multiple recognition passes over the adaptation data can also be conducted to iteratively refine the speech recognition models and hopefully the improve subsequent transcriptions.

4. Experimental Results

4.1. Task Overview

Our experiments are conducted on the task of spoken lecture transcription. A core set of 121 hours of lecture audio, primarily obtained from the MIT World collection of lectures, was available for the training of our SI speech recognizer. For our experiments, additional audio from a wide variety of academic courses from the MIT OpenCourseWare video lecture collection are also available. Details on the properties of the data in our corpus can be found in [11].

In developing a system to transcribe this data, we anticipate three different usage scenarios. In the first scenario, the system is requested to transcribe a lecture from a previously unseen lecturer. In this case, the system must rely on unsupervised adaptation performed on the same lecture.

In the second scenario, one to two lectures of data from a lecturer are manually transcribed. This could allow the system to produce a quality speaker adapted model for that lecturer, while requiring only a moderate expense for manual transcription. The speaker adapted model can then be used to automatically transcribe the remainder of lectures recorded for that speaker, where a typical academic course at MIT accumulates about 30 hours of recorded audio over the span of one semester.

In the third scenario, a full semester of lectures from a speaker could be manually transcribed to allow a highly accurate model for that speaker to be built and used on lectures from subsequent courses taught by that lecturer. For one lecturer in our collection, we have two full semesters worth of lectures manually transcribed, allowing us to examine long-term adaptation in this scenario.

4.2. Baseline System

For our experiments we are using the SUMMIT speech recognition system [12]. For language modeling, a standard word trigram with a core vocabulary of 37.4K words is used. A topic independent language model is trained from a combination the Switchboard corpus (containing 3.1M total words) and a collection of transcribed academic lectures obtained from three universities (containing 3.5M total words). For specific lectures, supplemental materials may be available, including companion textbooks, lecture slides, or relevant materials obtained from the web via a Google search. These materials are used to augment the recognizer's vocabulary and perform topic adaptation to the language model.

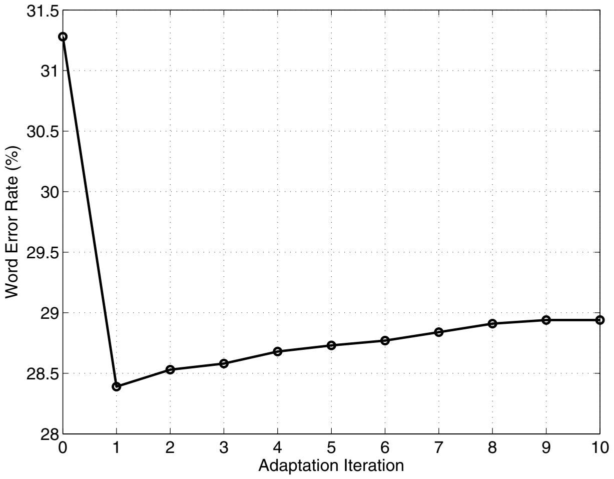

Figure 1: Word error rates vs. iterations of unsupervised MAP-style model interpolation speaker adaptation using MCE-trained speaker independent models interpolated with ML-estimated speaker dependent models.

The SUMMIT recognizer models the acoustic signal using the same diphone-based landmark modeling techniques used in the real-time recognition systems employed in our interactive dialogue systems [12]. Diagonal Gaussian mixture models are used to model each of the 1871 context-dependent acoustic landmark models in our system. This yielded a speaker independent set of acoustic models containing 31873 total Gaussian components (or around 17 Gaussians per model on average). The SI models were initialized with ML estimation, and then refined with 13 iterations of MCE training.

4.3. Unsupervised Speaker Adaptation

To examine unsupervised speaker adaptation, we conducted experiments on 5 held-out test lectures from 5 different speakers, containing an aggregate of 6 hours of audio. Three of the lectures were from an MIT course on automatic speech recognition and two of the lectures were one-time public seminars given at MIT by outside speakers. For the 3 speech recognition lectures, the language model and vocabulary were adapted based on the lecturer's presentation slides. For the 2 external lectures, the language model and vocabulary were adapted based on a small collection of documents about the lecturer and topic obtained via a Google search. Using our baseline speaker independent acoustic models, the topic-independent recognizer achieved a word error rate of 33.6% and the topic-adapted recognizer achieved a word error rate of 31.3%.

To perform unsupervised adaptation, the top-choice recognition result from the topic-adapted recognizer was used to guide the training of a new set of acoustic models for each lecture (from scratch) using ML estimation. In our experiments, no data filtering based on confidence scores was used. As a result, errors in the phonetic labels used during the training run are commonly present. In addition, no unsupervised speaker diarization is applied to the audio file. Thus, a small subset of utterances from speakers other than the lecturer (e.g., audience questions) are present in the audio and also not filtered out.

The resulting ML estimated SD models for each lecture are interpolated with the MCE trained SI models using MAP-style model interpolation. These speaker adapted models can then be reapplied to the test data to obtain new transcriptions and hence new speaker adapted models. This process can be iterated under the assumption that improved transcriptions will lead to improved speaker adapted models. In similar experiments using ML estimated SI models, we typically observe improvements in performance when iterating the adaptation process 3 to 5 times, with performance asymptoting with additional iterations. Figure 1 shows our adaptation results when using our SI-MCE models instead of SI-ML models. The figure shows a 9% relative reduction in error rate (from 31.3% to 28.4%) after one iteration of unsupervised speaker adaptation. After the first iteration, the figure shows an unusual decrease in accuracy with additional iterations, until the accuracy begins asymptoting at 29.0% after 10 iterations. This implies that, despite the improved fidelity of the recognition transcripts from the initial iteration of speaker adaptation, the learned ML estimated SD models may actually be harming some of the discriminative abilities provided by the baseline MCE trained SI models, after model interpolation is performed. It is unclear if this trend would continue to exist if data-filtering to remove poorly recognized words was applied before adaptation. We leave this experiment to future work.

4.4. Supervised Speaker Adaptation

For supervised adaptation we have a full semester of manually transcribed lectures available from two different academic courses taught by the same professor. The professor is a non-native (i.e. Dutch) speaker, who speaks in a clear, though animated, speaking style. We used 35 lectures totaling 29 hours from a physics course on electromagnetics as the adaptation material. We have randomly selected 4 lectures from a physics course on classical mechanics for the test material. By chance 2 hours of audio from other lectures given by this speaker are also contained in the training material for the original SI model. The language model for our experiments is adapted from physics materials, predominantly collected from two physics textbooks as well as the 35 electromagnetics lectures used for the adaptation data. The baseline speaker independent topic adapted recognizer achieves a performance of 30.7% on the test data.

We conducted 2 supervised adaptation experiments. In the first experiment we use only 2 lectures (or roughly 100 minutes of audio) for adaptation. In the second experiment we use the full 29 hours of available data. Both experiments begin with baseline speaker adapted models created using model interpolation where the SD models are derived with ML estimation.

Figure 2 shows the performance of the system using 100 minutes of adaptation data for model interpolation adaptation using the original ML SD models, as well as models from both the MCE-ML and MCE-MI training approaches. The initial speaker adapted model achieves a word error rate of 25.9%. Both MCE-ML and MCE-MI training improve the performance of the baseline speaker adapted model, with MCE-MI training outperforming MCE-ML. In particular, MCE-ML only improves performance during its first eight MCE iterations achieving a minimum word error rate of 25.5%. After eight iterations, performance begins to deteriorate, presumably because the MCE training is over-training the SD model. The MCE-MI approach exhibits better learning, achieving a minimum error rate of 25.1%, before also starting to deteriorate in performance. However, the deterioration of the MCE-MI models is slower, with a plateau in performance of around 25.2% word error rate from the 11^{th} iteration through the 24^{th} iteration. At it's best

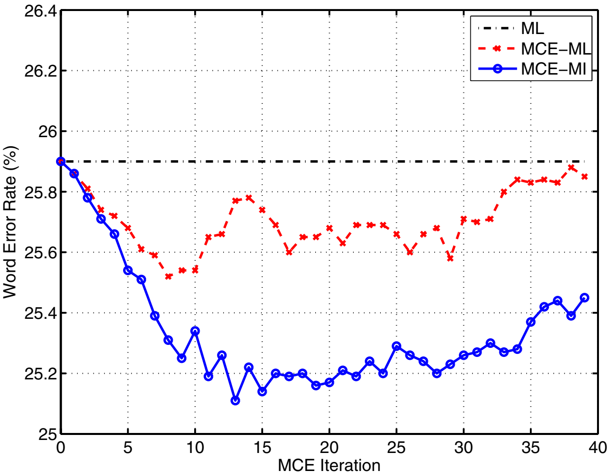

Figure 2: Word error rates for supervised model interpolation adaptation from 100 minutes of adaptation data when the SD model is created using ML estimation, using MCE-ML training, and using MCE-MI training. Results are shown for up to 40 iterations of MCE training during the adaptation process.

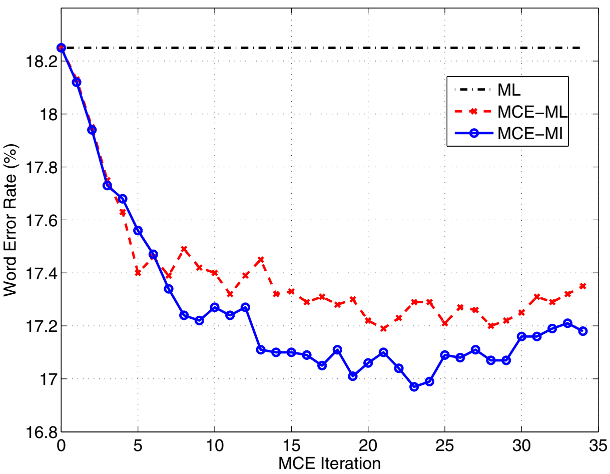

Figure 3: Word error rates for supervised model interpolation adaptation using 29 hours of adaptation data when the SD model is created using ML estimation, using MCE-ML training, and using MCE-MI training. Results are shown for up to 34 iterations of MCE training during the adaptation process.

point, the MCE-MI adaptation achieves an error rate which is a 3% relative improvement over the baseline MI adaptation technique using the ML estimated SD model. This also represents an 18% relative error rate reduction from the speaker independent performance of 30.7%.

Figure 3 shows the performance of the system under the same conditions but using 29 hours of adaptation data. The models show similar trends, with MCE-MI training outperforming MCE-ML training slightly but not as significantly as in the 100-minute adaptation case. Both models also begin to show a slight deterioration in performance as the number of iterations approaches 30, but this deterioration is not as severe as in the 100-minute case. The MCE-MI approach achieves a minimum word error rate of 17.0% at it best point, which is a relative reduction in error rate of 7% from the 18.2% error rate of the baseline MI adaptation approach using the ML estimated SD models. This also represents a 45% relative error rate reduction from the speaker independent performance of 30.7%.

5. Summary

In this paper we have explored strategies for incorporating MCE training into a model interpolation speaker adaptation scheme in the spirit of traditional maximum a posteriori probability (MAP) adaptation. Supervised adaptation experiments show relative error rate reductions between 3% and 7% over a baseline system which uses standard ML estimation instead of MCE training during the adaptation phase.

6. Acknowledgements

This research was supported in part by NTT Corporation.

7. References

[1] W. Chou, C.-H. Lee, and B.-H. Juang, “Minimum error rate training based on N-best string models,” in *ICASSP*, Minneapolis, April 1993.

[2] P. Woodland and D. Povey, “Large scale discriminative training of hidden Markov models for speech recognition,” *Comp. Speech and Lang.*, vol. 16, pp. 25–47, 2002.

[3] E. McDermott, *et al*, “Discriminative training for large vocabulary speech recognition using minimum classification error,” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 203-223, January 2007.

[4] R. Chengalvarayan, “Speaker adaptation using discriminative linear regression on time-varying mean-parameters in trended HMM,” *IEEE Signal Processing Letters*, vol. 5, no. 3, pp. 63-65, March 1998.

[5] F. Wallhoff, D. Willett, and G. Rigoll, “Frame-discriminative and confidence-driven adaptation for LVCSR,” in *ICASSP*, Istanbul, June 2000.

[6] J. Wu and Q. Huo, “Supervised adaptation of MCE-trained CDHMMs using minimum classification error linear regression,” in *ICASSP*, Orlando, May 2002.

[7] X. He and W. Chou, “Minimum classification error linear regression for acoustic model adaptation of continuous density HMMs,” in *ICASSP*, Hong Kong, April 2003.

[8] D. Povey, P. Woodland, and M. Gales, “Discriminative MAP for acoustic model adaptation,” in *ICASSP*, Hong Kong, April 2003.

[9] S. Fahlman, “An empirical study of learning speed in back-propagation networks,” Tech. Rep., Carnegie Mellon University, 1988.

[10] J. Gauvain and C.-H. Lee, “Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains,” *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291-298, April 1994.

[11] J. Glass, T. J. Hazen, L. Hetherington and C. Wang, “Analysis and processing of lecture audio data: Preliminary investigations,” in *HLT-NAACL 2004 Workshop on Speech Indexing and Retrieval*, Boston, May 2004.

[12] J. Glass, “A probabilistic framework for segment-based speech recognition,” *Computer, Speech, and Language*, vol. 17, no. 2-3, pp. 137-152, 2003.