# Semi-Supervised Training of Gaussian Mixture Models by Conditional Entropy Minimization

*Jui-Ting Huang and Mark Hasegawa-Johnson*

Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign
jhuang29@illinois.edu, jhasegaw@illinois.edu

## Abstract

In this paper, we propose a new semi-supervised training method for Gaussian Mixture Models. We add a conditional entropy minimizer to the maximum mutual information criteria, which enables to incorporate unlabeled data in a discriminative training fashion. The training method is simple but surprisingly effective. The preconditioned conjugate gradient method provides a reasonable convergence rate for parameter update. The phonetic classification experiments on the TIMIT corpus demonstrate significant improvements due to unlabeled data via our training criteria.

**Index Terms**: semi-supervised learning, conditional entropy, Gaussian Mixture Models, phonetic classification

## 1. Introduction

For speech recognition, untranscribed speech data are easy to collect and free of human transcribing efforts. This motivates the research on semi-supervised learning (SSL) approaches that aims to directly use unlabeled data, in addition to a limited amount of labeled data, to make more accurate model than can be achieved using only labeled data. Limited research on semi-supervised learning has been conducted for speech applications, and most such experiments employ *self-training* methods. In particular, an initial acoustic model is bootstrapped using a limited amount of manually transcribed data, and it is used to transcribe a relatively large amount of untranscribed data. Automatic transcriptions with high confidence [1, 2, 3] are then selected to augment the training set to train new models. While self-training can sometimes be successful, it is unclear which assumption it corresponds to in theory .

As an alternative to self-training, several researchers work on developing an integrated framework, in which models are trained to optimize an objective that reflects reasonable assumptions about labeled and unlabeled data. For example, graph-based methods derive a training objective based on the assumption that the data live close to an intrinsic low-dimensional manifold, and nearby data points with respect to the underlying manifold are likely to have the same labels [4, 5]. While the approaches in [4, 5] focus on direct modeling of $p(y|x)$, where $x$ is the input data and $y$ is the output target label, by nonparametric models and multiple-layered perceptrons respectively, we investigate semi-supervised training of Gaussian mixtures models (GMM) under the discriminative training framework. Our scheme is useful because it can be naturally applied to HMM-based speech recognition systems where GMM is the common modeling of the statistical distribution of acoustic features.

GMM can be trained using generative criteria such as maximum likelihood (ML), and can be further improved by using discriminative training criteria such as maximum mutual information (MMI) and minimum classification error (MCE). The goal of our research is to keep the advantage of discriminative training while incorporating additional improvements from unlabeled data. To this end, we propose to minimize the conditional entropy measured on unlabeled data, along with maximizing the averaged log posterior probability on labeled data. Intuitively, the conditional entropy regularizer encourages the model to have as great a certainty as possible about its class prediction on the unlabeled data; minimum conditional entropy is, in a sense, a discriminative training criterion for unlabeled data. This method is simple but surprisingly effective. The improvement over the supervised baseline model is significant, and the optimization process can have a reasonable convergence rate. Moreover, there is only one tuning parameter in the training criterion, and the performance gain remains almost the same within a wide range of tuning parameter values.

In this paper, we introduce the semi-supervised training criterion based on conditional entropy minimization in the context of GMM classifier (Section 2) and discuss its relation to other work (Section 3). We then detail the optimization method (Section 4) to train GMMs using the proposed criterion. To update GMM parameters, we use the preconditioned conjugate gradient method for a faster convergence rate. We evaluate our approach on the phonetic classification task using the TIMIT corpus and demonstrate its effectiveness compared to other semi-supervised learning methods (Section 5). Here we focus on the phonetic classification problem, but the method can be naturally extended to the HMM-based recognition framework.

## 2. Training Objective

We first formulate our problem setting. Suppose we are given a set of points $X_L = \{x_1, ..., x_l\}$, for which labels $Y_L = \{y_1, ..., y_l\}$ are provided, and another set of points $X_U = \{x_{l+1}, ..., x_{l+u}\}$, of which the corresponding class labels are unknown. In our case, $x_i \in R^n$ represents the $n$-dimensional spectral feature vector associated with a phone occurrence $i$; $y_i \in \{1 \cdots C\}$ is the class label, being one of $C$ phonetic classes. The classifying rule $f : R^n \rightarrow \{1 \cdots C\}$ is based on Bayes rule,

$$\hat{y} = f(x) = \max_{y \in \{1 \cdots C\}} p(x|y)p(y), \tag{1}$$

where $p(y)$ is the class prior estimated from the labeled set of training data, and the conditional distribution $p(x|y)$, $y \in \{1 \cdots C\}$ is modeled using GMM, a mixture of Gaussians,

$$p(x|y = c) = \sum_{m=1}^{M} w_{cm} N(x; \mu_{cm}, \Sigma_{cm}), \tag{2}$$

where $w_{cm}$ is the weight for component $m$ of class $c$ satisfying $\sum_{m=1}^{M} w_{cm} = 1, w_{cm} \geq 0$. Our goal is to learn GMM parameters $\lambda = \{\mu_{cm}, \Sigma_{cm}\}$ for a better classification accuracy.

We propose that the estimator of GMM parameters $\lambda$ is the maximizer of the following objective,

$$\begin{aligned} J =& F_{MMI}(\lambda) - \alpha H_{emp}(Y|X;\lambda) \\ =& \frac{1}{l}\sum_{i=1}^{l} \log p_\lambda(y_i|x_i) + \alpha \frac{1}{u} \sum_{i=l+1}^{l+u} \sum_{y} p_\lambda(y|x_i) \log p_\lambda(y|x_i), \end{aligned} \tag{3}$$

where the posterior probability is computed by

$$p_\lambda(y|x_i) = \frac{p(x|y;\lambda)p(y)}{\sum_{y'} p(x|y';\lambda)p(y')}. \tag{4}$$

That is, we augment the original log posterior criterion on the labeled data with an empirical conditional entropy regularizer on the unlabeled data. The regularizer encourages the model to have as great a certainty as possible about its class prediction on the unlabeled data and therefore reinforces the confidence of the classifier output. Moreover, according to [6], the conditional entropy $H(Y|X)$ measures the degree of class overlap, and the tuning parameter $\alpha$ controls the weights of the contribution of unlabeled data, as the unlabeled data are most beneficial when the classes have small overlap. When $x$ represents acoustic observation sequence and $y$ represents word sequence, the first term becomes the training criterion for maximum mutual information (MMI) estimation of HMMs for speech recognition. Here we borrow the terminology, implying the potential extension to the recognition problem.

## 3. Relation to Other Work

Conditional entropy measure was first introduced in the context of semi-supervised learning in [6], specifically for discriminative classifiers such as logistic regression models. Jiao et al. [7] then extended this idea to conditional random fields. Both of the methods demonstrated encouraging improvements over the model using labeled data only, whereas self-training might give little improvement [7]. In [8], conditional entropy is used for $n$-gram language model adaptation in speech recognition and showed significant improvement. The method of [8] can be also seen as a semi-supervised learning approach, in the sense that the initial language model estimated from the transcribed data serves as prior knowledge in their training criterion. While our training criterion is in the same spirit, we extend such regularization to discriminative training of generative models.

We previously proposed a hybrid training criterion to incorporate labeled and unlabeled data [9]; our criterion was the composition of the MMI criterion on labeled data combined with the ML criterion on unlabeled data. The following study [10] further found that the choice of discriminative/generative criterion on the labeled data actually does not affect the improvement from unlabeled data. Apparently the ML criteria on unlabeled data models the missing data (missing class and mixture identity) problem in a generative way, the information from which doesn't change the discriminative nature of the labeled part. The results of [10] therefore suggest that, to have models improved in a discriminative sense, it is crucial to have a measure on unlabeled data that can reinforce the discriminative power of the GMM classifier on labeled data.

Unlabeled data do not necessarily contribute to a better learning of decision boundary. It is essential to have one or more assumptions about the connection between the marginal distribution $P(x)$ and the conditional $P(y|x)$ [11]. Two successful assumptions include the *manifold* assumption used in graph-based methods and the *low-density separation* assumption used in Transductive-SVM [12], which assumes the decision boundary lies in low-density regions. Conditional entropy encourages the class regions of unlabeled data to be as separate as possible, guiding the decision boundary away from high-density regions. In this regard, this paper exploits the use of unlabeled data in a way that shares the assumptions of transductive learning rather than the assumptions of the graph-based learning proposed in [4, 5].

## 4. Optimization Method

We optimize the training objective in Equation (3) with respect to GMM parameters by gradient methods. The steepest descent method, using directly the gradient of the function, convergences too slowly. To improve the convergence speed, we use the preconditioned conjugate gradient methods, in which the search direction is computed based on the first-order gradients of the objective. In the following, we will first show the gradients and then explain our implementation of the conjugate gradient method. In current experiments, we focus on updating GMM mean vectors only.

### 4.1. Gradients

The gradient of the objective function is

$$\nabla_\lambda J = \nabla_\lambda F_{MMI} - \alpha \nabla_\lambda H_{emp}, \tag{5}$$

consisting of the gradients of two components, shown respectively in the following.

For component $m$ of the GMM for class $c$, the gradient of the first term with respect to the mean vector $\mu_{cm}$ is,

$$\nabla_{\mu_{cm}} F_{MMI} = \frac{1}{l}\Sigma_{cm}^{-1}\left[(\theta_{cm}^{num} - \theta_{cm}^{den}) - (\gamma_{cm}^{num} - \gamma_{cm}^{den})\mu_{cm}\right] \tag{6}$$

where

$$\begin{aligned} \gamma_{cm}^{num} &= \sum_{i=1, y_i=c}^{l} p(m|x_i,\lambda_c); \quad \theta_{cm}^{num} = \sum_{i=1, y_i=c}^{l} x_i p(m|x_i,\lambda_c), \\ \gamma_{cm}^{den} &= \sum_{i=1}^{l} p(c,m|x_i,\lambda); \qquad \theta_{cm}^{den} = \sum_{i=1}^{l} x_i p(c,m|x_i,\lambda). \end{aligned} \tag{7}$$

The gradient of the conditional entropy is,

$$\nabla_{\mu_{cm}} H_{emp} = \frac{1}{u}\Sigma_{cm}^{-1}\left(\theta_{cm}^{ent} - \gamma_{cm}^{ent}\mu_{cm}\right), \tag{8}$$

where

$$\begin{aligned} \gamma_{cm}^{ent} &= \sum_{i=l+1}^{l+u} \left[\log p(c|x_i) - \sum_y p(y|x_i)\log p(y|x_i)\right] p(c,m|x_i,\lambda) \\ \theta_{cm}^{ent} &= \sum_{i=l+1}^{l+u} \left[\log p(c|x_i) - \sum_y p(y|x_i)\log p(y|x_i)\right] x_i p(c,m|x_i,\lambda). \end{aligned} \tag{9}$$

As we can see, the first and second order statistics $\gamma$ and $\theta$ for the entropy term are just the ones for the MMI term weighted by a data-dependent factor.

### 4.2. Conjugate Gradient Methods

Conjugate gradient methods are known to accelerate the convergence rate of steepest descent by using a set of conjugate directions generated from gradient vectors [13]. The convergence rate can be further improved by introducing a scaling matrix to search directions such that the transformed local quadratic form becomes more spherical [14], which is known as *preconditioned* Conjugate Gradient methods. While a perfect choice of scaling matrix is the inverse of Hessian (second-order derivatives matrix) at the local point, we found that the local Hessian of our objective function, with respect to $\mu_{cm}$, can be approximated as being proportional to $\Sigma_{cm}^{-1}$. To see this, if we assume the mixture/class occupation probabilities $p(m|x_i, \lambda_c), p(m, c|x_i)$ and $p(c|x_i)$ in Equation (7) and (9) remain roughly the same with respect to a small change in $\mu_{cm}$, the second derivative of the objective function is approximated as

$$\nabla^2_{\mu_{cm}} J \approx \left[\frac{1}{u}\gamma_{cm}^{ent} - \frac{1}{l}(\gamma_{cm}^{num} - \gamma_{cm}^{den})\right]\Sigma_{cm}^{-1}. \quad (10)$$

As a result, the preconditioned conjugate method is described by $\mu_{cm}^{k+1} = \mu_{cm}^{k} + \eta^k d^k$, where the superscript $k$ represents the $k$-th iteration, and the search directions are generated by,

$$\begin{aligned} d^0 &= \Sigma_{cm} \cdot \nabla_{\mu_{cm}} J(\mu_{cm}^0) \\ d^k &= \Sigma_{cm} \cdot \nabla_{\mu_{cm}} J(\mu_{cm}^k) - \beta^k d^{k-1}, \end{aligned} \quad (11)$$

where

$$\beta^k = \frac{\nabla J(\mu_{cm}^k)^T \Sigma_{cm} (\nabla J(\mu_{cm}^k) - \nabla J(\mu_{cm}^{k-1}))}{\nabla J(\mu_{cm}^{k-1})^T \Sigma_{cm} \nabla J(\mu_{cm}^{k-1})}. \quad (12)$$

The step size $\eta^k$ is obtained by line maximization,

$$J(\mu_{cm}^k + \eta^k d^k) = max_\eta J(\mu_{cm}^k + \eta d^k). \quad (13)$$

The Armijo rule [13] is used to do the line search, which needs the value $J(\cdot)$ at each search point. In order to limit computational complexity of the line search, we use a random subset ($10\%$ of the training set) to compute the objective function for $\eta$ selection. In our experiments, usually only one or two iterations were needed for line maximization.

## 5. Experiments

### 5.1. Data

To evaluate the performance of our approach, we conducted experiments on phonetic classification using the TIMIT corpus [15]. Here we assume the phone boundaries are given, and the task is to assign the phone identity to each phone segment. We trained models for 48 phone classes and the classifier outputs are merged into 39 classes for final evaluation according to [16]. We extracted 50 speakers out of the NIST complete test set to form the development set for tuning the value of $\alpha$. The rest of the NIST test set formed our evaluation test set. The development and evaluation test set here are the same as the development set and fulltest set defined in [17]. To create a semi-supervised learning problem, the standard NIST training set was randomly divided into the labeled and unlabeled sets with different ratios, and we assumed the phone class labels in the unlabeled set are unavailable.

We used segmental features [17] in the phonetic classification task. For each phone occurrence, a fixed-length vector was calculated from the frame-based spectral features (12 PLP coefficients plus energy) with a 5 ms frame rate and a 25 ms Hamming window. More specifically, we divided the frames for each phone segment into three regions with 3-4-3 proportion, plus the 30ms regions beyond the start and end time of the segment, and calculated the PLP average over each region. Three averages plus the log duration of that phone gave a 61-dimensional ($12 \times 5 + 1$) measurement vector.

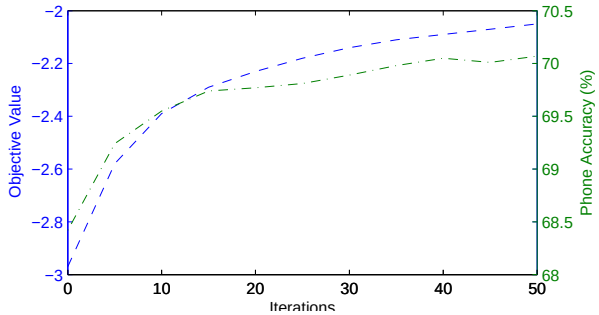


Figure 1: *Objective values (dashed line) and phone accuracies (%, dotted line) over iterations on the development set for s=25%, $\alpha = 10$.*

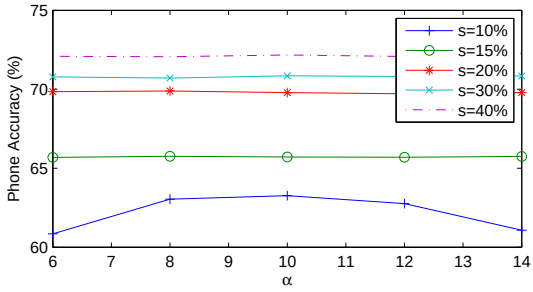


Figure 2: *Phone accuracies (%) for different values of $\alpha$ on the developmet set for $s = 10, 15, 20, 30, 40\%$. Note that all accuracies here are higher than the MMI baseline.*

### 5.2. Results

We tested our algorithm on the problems of different labeled/unlabeled ratios; labels of different percentages, varying from $s = 10\% - 100\%$, of the training set were kept. The initial model was trained using the labeled set via maximum likelihood estimation (MLE). We adopted two-component GMM with full covariances for all 48 classes. Figure 1 shows the objective function values during training over iterations, for the case of s=25%, on the development set. Regardless of the labeled to unlabeled ratio, the objective normally converges in 50 iterations, showing the effective convergence rate of the preconditioned conjugate gradient method. As a result, we used the updated parameters either after 50 iterations or at its last iteration of update, whichever comes first. The phone classification accuracy is also shown in the same figure, and it appears to correlate well with the objective value. Next, we show the insensitivity of phone accuracies to the tuning parameter $\alpha$ in Equation (3). Figure 2 plots phone accuracies versus different choices of $\alpha$, for the case of $s = 10, 15, 20, 30, 40\%$ on the development set. We can see that the accuracy is not too sensitive to the value of $\alpha$. Only when the labeled to unlabeled ratio gets

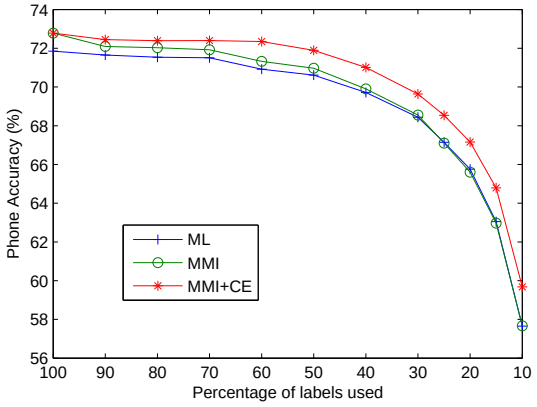


Figure 3: *Phone accuracies (%) for different percentages (s=10,15,20,25,30,40,50,60,70,80,90,100%) of labels used.*

sufficiently small ($s = 10\%$), the optimal region of $\alpha$ becomes relatively narrow.

We compared our semi-supervised training with supervised training that only uses the first term of Equation (3), a classifier version of MMI-training. For a fair comparison, we applied I-smoothing [18] as a smoothing technique to prevent over-training. The value of the smoothing constant $\tau$ was also tuned on the development set. Figure 3 shows the phone accuracies on the test set for different percentages $s$ of labels being used. Our approach (MMI+CE) provides significant improvement over MMI-training that uses only the labeled set for all $s \leq 90\%$. In particular, given enough unlabeled data ($s \leq 60\%$), the conditional entropy regularizer boosts the classification accuracy by a large margin (1-2%), even when MMI cannot improve over ML ($s \leq 30\%$).

To compare with other semi-supervised methods, we implemented a naive self-training method as it can be naturally applied to the discriminative training scheme for GMMs. We used the initial MLE model to predict labels on the unlabeled data, part of which with sufficiently high classifier confidence were added to the original labeled set, and GMMs were retrained using the enlarged set. We tried different thresholds of confidence and ran several repetitions, but there was no significant change of the result.

## 6. Conclusion

We proposed a semi-supervised discriminative training criterion for Gaussian mixture models. It is a coherently discriminative objective; the maximum mutual information on the labeled data is discriminative as well as the conditional entropy on the unlabeled data. As a result, the models trained using the MMI-CE criterion improves MMI training by a large margin. Moreover, the training objective has an efficient convergence rate by the preconditioned conjugate method.

Given the encouraging results, we are currently extending the method to phone recognition problems, where phone boundaries are not given during training and testing. The ultimate goal is to leverage on untranscribed data to improve acoustic models for continuous speech recognition. Other than simulating SSL data using TIMIT, we will also test our method on speech corpora where real untranscribed data coexist with transcribed data.

## 7. Acknowledgements

This research was supported by NSF 07-03624. Opinions and findings are those of the authors, and are not endorsed by the NSF.

## 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, Jan. 2005.

[2] L. Wang, M. Gales, and P. Woodland, "Unsupervised training for mandarin broadcast news and conversation transcription," in *Proc. IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, 2007, pp. 353–356.

[3] L. Lamel, J.-L. Gauvain, and G. Adda, "Lightly supervised and unsupervised acoustic model training," vol. 16, pp. 115–129, 2002.

[4] A. Subramanya and J. A. Bilmes, "The semi-supervised switchboard transcription project," in *Interspeech*, Brighton, UK, September 2009.

[5] J. Malkin, A. Subramanya, and J. Bilmes, "On the semi-supervised learning of multi-layered perceptrons," in *Interspeech*, Brighton, U.K., September 2009.

[6] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *Advances in Neural Information Processing Systems 17*, L. K. Saul, Y. Weiss, and L. Bottou, Eds. Cambridge, MA: MIT Press, 2005, pp. 529–536.

[7] F. Jiao, S. Wang, C.-H. Lee, R. Greiner, and D. Schuurmans, "Semi-supervised conditional random fields for improved sequence segmentation and labeling," in *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, 2006, pp. 209–216.

[8] J.-T. Huang, X. Li, and A. Acero, "Discriminative training methods for language models using conditional entropy criteria," in *Proc. IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010.

[9] J.-T. Huang and M. Hasegawa-Johnson, "Maximum mutual information estimation with unlabeled data for phonetic classification," in *Interspeech*, 2008.

[10] ——, "On the semi-supervised learning of gaussian mixture models," in *Proceedings of the NAACL HLT Workshop on Semi-supervised Learning for Natural Language Processing*, Boulder, Colorado, USA, June 2009.

[11] O. Chapelle, B. Schölkopf, and A. Zien, Eds., *Semi-Supervised Learning*. Cambridge, MA: MIT Press, 2006. [Online]. Available: http://www.kyb.tuebingen.mpg.de/ssl-book

[12] T. Joachims, "Transductive inference for text classification using support vector machines," in *Proceedings of ICML-99, 16th International Conference on Machine Learning*, Bled, SL, 1999, pp. 200–209.

[13] D. P. Bertsekas, *Nonlinear Programming*. Athena Scientific, September 1999.

[14] J. R. Shewchuk, "An introduction to the conjugate gradient method without the agonizing pain," Pittsburgh, PA, USA, Tech. Rep., 1994.

[15] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "Darpa timit acoustic phonetic continuous speech corpus," 1993.

[16] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 37, no. 11, pp. 1641–1648, 1989.

[17] A. K. Halberstadt, "Heterogeneous acoustic measurements and multiple classifiers for speech recognition," Ph.D. dissertation, Massachusetts Institute of Technology, 1998.

[18] D. Povey, "Discriminative training for large vocabulary speech recognition," Ph.D. dissertation, Cambridge University, 2003.