# How Unlabeled Data Change the Acoustic Models For Phonetic Classification

*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

Semi-supervised learning is a class of machine learning techniques that aims to use unlabeled data to improve the performance of models trained using only labeled data. In the previous work, we have developed an integrated semi-supervised learning framework, in which speech models are trained to optimize an objective that reflects reasonable assumptions about labeled and unlabeled data. In this paper, we further investigate the model behavior when adding unlabeled data in the context of our framework. We use phonetic classification experiments to test three hypotheses: 1) With the amount of unlabeled data increases, the model parameters converge to the same point as with the amount of labeled data increases. 2) Unlabeled data will help learn more similar distributions to the true model. 3) Unlabeled data will help reduce the classification error rate on a general unseen test set. The experimental results show that the influence of unlabeled data on the estimated model is not equivalent to that of labeled data, and the training criteria control how unlabeled data change the acoustic model.

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

## 1. Introduction

The core of modern speech technology consists of a set of statistical models representing various sounds in the language to be processed. To form the statistical model for each basic speech unit, acoustic signals have to be mapped to their corresponding sound categories, according to the transcription of speech waveforms. This scheme is called *supervised learning*. However, transcribing data requires efforts of experienced human annotators, which is expensive, time-consuming, and sometimes error-prone. Since in many applications enormous amounts of unlabeled data are available with little cost, we have proposed to develop *semi-supervised learning* (SSL) algorithms [1, 2] that directly use unlabeled data, in addition to a limited amount of labeled data, to build more accurate computational models than can be achieved using only labeled data. In the previous study [1, 2], we have shown the effectiveness of our semi-supervised learning algorithms in the phone classification experiments, by showing the classification error rate reduction after adding unlabeled data to train a classifier.

While our semi-supervised phone models can have a better classification rate, we are interested in how unlabeled data change the supervised models to the semi-supervised models in the acoustic feature space. Our interested acoustic models are based on Gaussian mixtures models (GMM), which are the common probabilistic models of acoustic features in a state-of-the-art continuous density HMM based speech recognition system. Let $\theta$ denote the vector of model parameters for all classes. Now we denote that $\hat{\theta}(l)$ is the supervised estimate of $\theta$ with labeled data of size $l$, and that $\hat{\theta}(\infty)$ is the estimate with infinite number of labeled data. We also denote that $\hat{\theta}(l,u)$ is the semi-supervised estimate of $\theta$ with labeled data of size $l$ and unlabeled data of size $u$. Intuitively, we think that the semi-supervised model parameters $\hat{\theta}(l,u)$ will be closer to $\hat{\theta}(\infty)$ than the supervised parameters $\hat{\theta}(l)$. This intuition motivates the three following hypotheses about the model behaviors,

H1. From statistics, we know that $\hat{\theta}(l)$ converges to $\hat{\theta}(\infty)$ as $l \rightarrow$ increases. In other words, $\left\|\hat{\theta}(l_2)-\hat{\theta}(\infty)\right\| \leq \left\|\hat{\theta}(l_1)-\hat{\theta}(\infty)\right\|$ for any $l_2 > l_1$. Therefore, if the contribution of unlabeled data is equivalent to that of labeled data, it should follow the same fact that,

$$\left\|\hat{\theta}(l,u)-\hat{\theta}(\infty)\right\| \leq \left\|\hat{\theta}(l)-\hat{\theta}(\infty)\right\|. \quad (1)$$

H2. Let $p_l(x)$ be the probability density functions (GMMs in our case) parameterized by $\hat{\theta}(l)$, $p_{l,u}(x)$ by $\hat{\theta}(l,u)$ and $p_\infty(x)$ by $\hat{\theta}(\infty)$. The statement that the semi-supervised model can be closer to the true model [1] can be formulated via a well-defined metric between probability distributions, Kullback-Leibler divergence,

$$D\left(p_\infty(x)||p_{l,u}(x)\right) \leq D\left(p_\infty(x)||p_l(x)\right), \quad (2)$$

where $D(P||Q)$ is the K-L divergence of probability distribution $P$ and $Q$, which is always nonnegative and will be zero if and only if $P=Q$.

H3. Let $f$ be the classifier derived from $p(x)$ (as will be described in Section 2), then, a good semi-supervised classifier $f_{l,u}(x)$ should satisfy,

$$\varepsilon\left(f_\infty(x)\right) \leq \varepsilon\left(f_{l,u}(x)\right) \leq \varepsilon\left(f_l(x)\right), \quad (3)$$

where $\varepsilon(f)$ is the classification error rate on a general held-out test set.

This paper is to experimentally test the above hypotheses for our semi-supervised models. We have previously shown that our semi-supervised classifiers satisfy H3, as the classification error rate on the held out test set decreases, and we will show in this paper that H1 and H2 may or may not be true for different semi-supervised training criteria.

[1] Strictly speaking, the model assumption can be wrong, and in this case $\hat{\theta}(\infty)$ will not be the correct/true model. While we are aware of this, this paper uses the term the "true" model and $\hat{\theta}(\infty)$ interchangeably for simplicity.

## 2. Phone Classification

In the task of phone classification, we assume that the time boundary information for phone segments is available, and the classifier labels the phone identity to each segment. We first formulate our problem setting for semi-supervised phone classification. In our case, $x \in R^n$ represents the $n$-dimensional spectral feature vector associated with a phone occurrence; $y \in \{1 \cdots C\}$ is the class label, being one of $C$ phonetic classes. The classifying rule $f : R^n \to \{1 \cdots C\}$ for any test token $x$ is based on Bayes rule,

$$\hat{y} = f(x) = \arg \max_{y \in \{1 \cdots C\}} p(x|y)p(y), \tag{4}$$

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 Gaussian Mixture Models (GMM),

$$p(x|y = c) = \sum_{m=1}^{M} w_{cm} N(x; \mu_{cm}, \Sigma_{cm}), \tag{5}$$

where $w_{cm}$ is the weight for component $m$ of class $c$ satisfying $\sum_{m=1}^{M} w_{cm} = 1, w_{cm} \geq 0$.

Suppose we are given a set of points $\mathcal{X}_L = \{x_i\}_{i=1}^{l}$, for which labels $\mathcal{Y}_L = \{y_i\}_{i=1}^{l}$ are provided, and another set of points $\mathcal{X}_U = \{x_i\}_{i=l+1}^{l+u}$, of which the corresponding class labels are unknown. Our goal is to learn GMM parameters $\lambda = \{\mu_{cm}, \Sigma_{cm}\}$ for a better classification accuracy than what can be achieved using the labeled set $(\mathcal{X}_L, \mathcal{Y}_L)$ alone. We will first review the supervised training methods that are conventionally used to estimate GMM parameters, followed by the proposed semi-supervised training method.

## 3. Supervised Training Criteria

We introduce two different supervised training methods for our Gaussian mixture models. In a nutshell, MLE models (Section 3.1) aim to find a accurate description of given training data, whereas MMIE models (Section 3.2) aim to make the separation between classes as large as possible.

### 3.1. Maximum Likelihood Estimation

With only labeled data are available, GMM parameters can be estimated using generative criteria such as maximum likelihood (ML). That is, we wish to find the parameter set that maximizes the log-likelihood that the models generate the training data $\mathcal{X}_L$,

$$\mathcal{F}_{ML}(\lambda) = \log p(\mathcal{X}_L|\mathcal{Y}_L\lambda) = \sum_{i=1}^{l} \log p(x_i|y_i\lambda). \tag{6}$$

The resulting model set is conventionally called MLE models.

### 3.2. Maximum Mutual Information Estimation

It is well known that MLE can be improved by discriminative training criteria such as maximum mutual information (MMI). The basic idea is to adjust the model parameters to make hypothesis output from the classifier closer to the correct labels. MMI used in speech recognition is equivalent to maximizing the log-posterior probability of the correct labels in the classification problem,

$$\mathcal{F}_{\text{MMI}}(\lambda) = \sum_{i=1}^{l} p(y = y_i|x_i) = \sum_{i=1}^{l} \log \frac{p_\lambda(x_i|y_i)\, p(y_i)}{\sum_c p_\lambda(x_i|c)\, p(c)}. \tag{7}$$

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. By maximizing the values in (7), we make the correct label (in the numerator) $p(y_i, x_i)$ likely and all other labels unlikely, discriminating the class from all other competing classes.

## 4. Semi-Supervised Training Criteria

Our semi-supervised algorithms are essentially using certain measures on unlabeled data as regularization to the supervised training criteria. The certain regularizers are proposed according to different characteristics of classifiers, that is, generative (MLE) or discriminative (MMIE).

### 4.1. Semi-Supervised MLE

With the generative criteria such as ML, unlabeled data can be incorporate into the generative framework naturally. In particular, the model parameters will now aim to maximize the likelihood of the joint labeled and unlabeled data,

$$\begin{aligned} \mathcal{J}(\lambda) =& \log P(\mathcal{X}_L, \mathcal{Y}_L, \mathcal{X}_U; \lambda) \\ =& \log P(\mathcal{X}_L|\mathcal{Y}_L; \lambda) + \alpha \log P(\mathcal{X}_U; \lambda) \\ =& \mathcal{F}_{ML}^{(\mathcal{D}_L)}(\lambda) + \alpha \mathcal{F}_{ML}^{(\mathcal{D}_U)}(\lambda), \end{aligned} \tag{8}$$

and we choose the parameters so that (8) is maximized:

$$\hat{\lambda} = \arg \max_{\lambda} \mathcal{J}(\lambda). \tag{9}$$

The second line in Equation (8) ignores the term $p(\mathcal{Y}_L)$, as it is unrelated to change of the parameters $\lambda$. The weight $\alpha$ is set to balance the impacts of two components on the training process.

### 4.2. Semi-Supervised MMIE

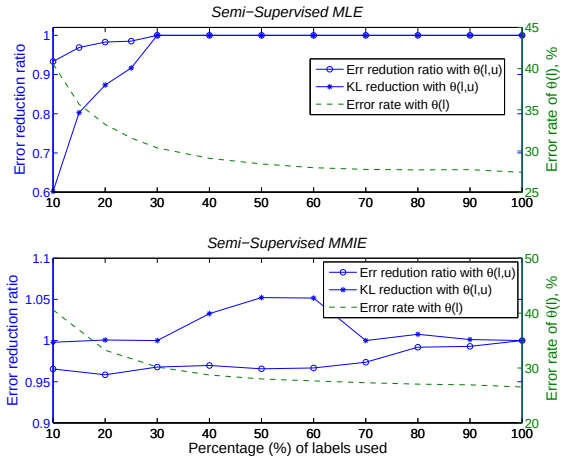
We propose to minimize the conditional entropy measured on unlabeled data, along with maximizing the averaged log posterior probability on labeled data. Intuitively, the conditional entropy regularizer encourages the model to have as great a certainty as possible about its class prediction on the unlabeled data; minimum conditional entropy is, in a sense, a discriminative training criterion for unlabeled data. This method is simple but surprisingly effective. Particularly, the estimator of GMM parameters $\lambda$ is the maximizer of the following objective,

$$\begin{aligned} \mathcal{J} =& F_{MMI}^{(D_L)}(\lambda) - \alpha H_{emp}^{(D_U)}(Y|X; \lambda) \\ =& \frac{1}{l} \sum_{i=1}^{l} \log p_\lambda(y_i|x_i) + \alpha \frac{1}{u} \sum_{i=l+1}^{l+u} \sum_{y} p_\lambda(y|x_i) \log p_\lambda(y|x_i), \end{aligned} \tag{10}$$

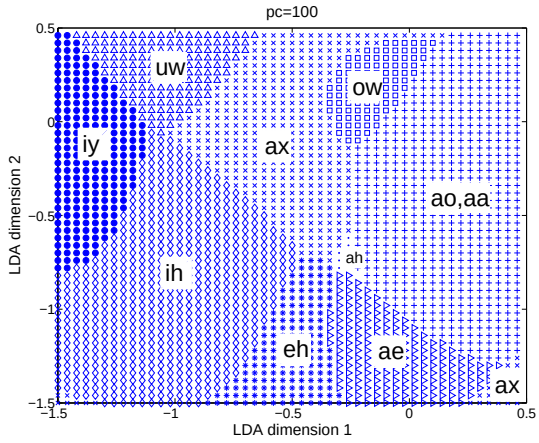
where the posterior probability is computed by

$$p_\lambda(y|x_i) = \frac{p(x|y; \lambda)p(y)}{\sum_{y'} p(x|y'; \lambda)p(y')}. \tag{11}$$

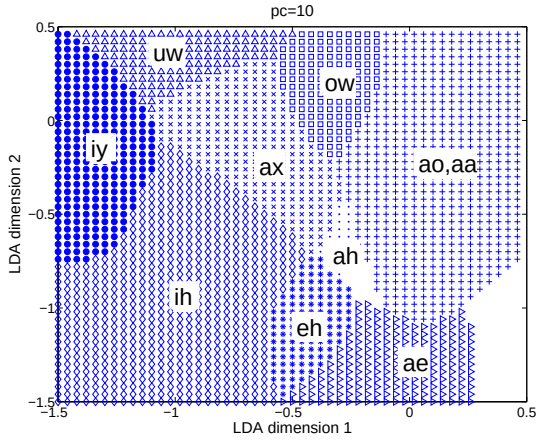
That is, we augment the original log posterior criterion on the labeled data with an empirical conditional entropy regularizer on the unlabeled data. The regularizer encourages the model to have as great a certainty as possible about its class prediction on the unlabeled data and therefore reinforces the confidence of the classifier output.



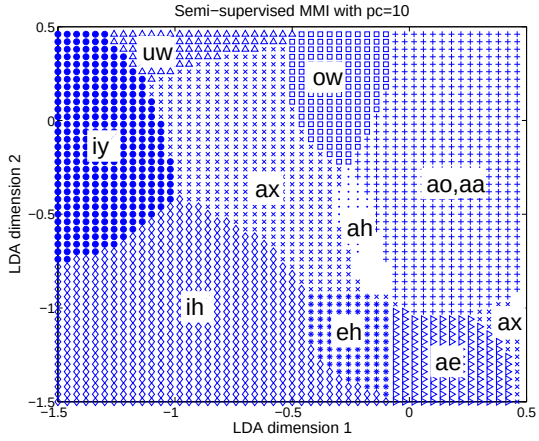

(a) Classification Error rate and KL distance reduction.




(b) The decision regions for vowels by supervised training trained using 100% of labels




(c) The decision regions for vowels by supervised training trained using 10% of labels




(d) The decision regions for vowels by semi-supervised MMIE training, using 10% of labels and the rest of unlabeled data.

Figure 1: (b)-(d):Decision regions for vowels. The white area is where the classifier assigns the feature phone classes other than the shown ones.

# 5. Experiments

## 5.1. Data

We conducted experiments on phonetic classification using the TIMIT corpus [3]. 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 [4]. We extracted 50 speakers out of the NIST complete test set to form the development set for tuning the value of $\alpha$ in Equation (8) and (10). The rest of the NIST test set formed our evaluation test set.

We used segmental features [5] 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.

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 covariance for all 48 classes.

## 5.2. Semi-Supervised MLE

It is not possible to know the true value of $\hat{\theta}(\infty)$, as we do not have infinite number of labeled training data. Therefore, we will approximate the value of $\hat{\theta}(\infty)$ with the model parameters we estimate using 100% of the training data. For convenience, we denote the approximated value as $\hat{\theta}(l = 100\%)$, and $\hat{\theta}(\infty) \cong \hat{\theta}(l = 100\%)$. Likewise, we denote the semi-supervised model that incorporates unlabeled data into training as $\hat{\theta}(l = s\%, u)$,

where $u$ is always $1 - s\%$ in our experiments.

First we examine that if the hypothesis H1 is true for the semi-supervised models trained using the MLE criteria in Equation (8). We adopt $l_2$ norm to measure the distance between mean vectors and covariance matrices of GMMs. We find that H1 is not true for our models for $l = 10\% - 100\%$; the mean vectors of the semi-supervised models are always further from the true mean vectors than the supervised models given the same amount of labeled data, which is opposite to the hypothesis of H1.

To test the hypothesis H2, we need to calculate the KL divergence between GMMs. While there is no analytically closed form, we approximate the divergence with a variational upper bound, $D_{\text{var}}$, proposed in [6]. We plot the reduction ratio of KL divergence,

$$r_{KL} = \frac{D_{\text{var}}\left(p_\infty(x)||p_{l=s\%,u}(x)\right)}{D_{\text{var}}\left(p_\infty(x)||p_{l=s\%}(x)\right)}, \tag{12}$$

for $s = 10 - 100$, as shown in the upper panel of Figure 1(a). If the ratio is lower than one, it means that the distributions in the semi-supervised models are closer to those in the true model than the supervised models. We also plot the error reduction ratio,

$$r_\varepsilon = \frac{\varepsilon\left(f_{l=s\%,u}(x)\right)}{\varepsilon\left(f_{l=s\%}(x)\right)}, \tag{13}$$

in the same plot. Likewise, a error ratio smaller than one means that the semi-supervised models decrease the error rate. We can see that both KL divergence and error rate for the semi-supervised model is smaller than the supervised model. Both H2 and H3 are true for our semi-supervised models for when label size is less than 30% of training data, and there is a good correlation of the reduction degree between KL divergence and error rate. After 30%, unlabeled data are not able to change the supervised model or to reduce the error rate.

### 5.3. Semi-Supervised MMIE

For semi-supervised MMIE, we use the supervised MMIE model using 100% of training data as an approximation of $\hat{\theta}(\infty)$. For a fair comparison, we applied I-smoothing [7] as a smoothing technique to prevent over-training. The value of the smoothing constant $\tau$ was also tuned on the development set. We test the prediction of the hypothesis H1 on semi-supervised MMIE models, and it is again false. We then plot the reduction ratio of KL divergence and error rate obtained using (12) and (13) in the lower panel of Figure 1(a). The KL divergence ratios are always larger or equal to one, meaning that the learned distributions by semi-supervised MMIE will not change the model to be be closer to the true distribution. Therefore, H2 is false for semi-supervised MMIE. The error rate, on the other hand, shows the reduction across different amounts of label sizes, even when the supervised model has achieved fairly good performances (error rates are lower than $30\%$ for $l \geq 50\%$).

The reason why H2 is not true for semi-supervised MMIE is that MMIE does not aim to find a better description of data but rather to make more correct decisions on the training data. In this sense, we expect it is class decision regions that are to be improved rather than the accuracy of probability density functions. To illustrate the idea, we plot the decision regions with respect to phone classes on the feature space in Figure 1(b), 1(c) and 1(d). We use Linear Discriminant Analysis to project the 61-dimensional features into a lower dimensional space, for the purpose of visualization. We show the feature space spanned by the first and second LDA dimensions and choose to present vowels since they are easier to visualize on the two-dimensional space. [2] We can see that the decision regions are changing, and the semi-supervised models (Figure 1(d)) seem to have more similar decision regions than the supervised models (Figure 1(c)) to the true model (Figure 1(b)), at the bottom right corner of the LDA plot.

## 6. Conclusion

In the context of our semi-supervised learning framework, the experimental results show that the influence of unlabeled data on the estimated model is not equivalent to that of labeled data, as the hypothesis H1 is rejected. Moreover, the training criteria control how unlabeled data change the acoustic model. Semi-supervised MLE models can have more similar phone distributions than supervised MLE models to the true model. Semi-supervised MMIE models do not yield more similar phone distributions but rather focus on maximize the discrimination between classes directly. As such, our semi-supervised learning framework incorporates unlabeled data in a coherent fashion, in the sense that the model behavior by adding unlabeled data faithfully reflects the characteristics of the supervised training criteria.

## 7. Acknowledgments

This research was supported by NSF 07-03624. We thank the members of the Illinois Prosody-ASR research group for helpful discussions.

## 8. References

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

[2] ——, "Semi-supervised training of gaussian mixture models by conditional entropy minimization," in *Interspeech*, 2010.

[3] 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.

[4] 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.

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

[6] J. R. Hershey and P. A. Olsen, "Approximating the kullback leibler divergence between gaussian mixture models," in *ICASSP*, vol. 4, no. 6. Ieee, 2007, pp. IV–317–IV–320.

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

[8] G. E. Peterson and H. L. Barney, "Control Methods Used in a Study of the Vowels," *Acoustical Society of America Journal*, vol. 23, pp. 148–+, 1951.

[2] Figure 1(b) looks similar to the vowel space in [8] after rotation by 45 degree counterclockwise.