On Semi-Supervised Learning of Gaussian Mixture Models for Phonetic Classification*

Jui-Ting Huang and Mark Hasegawa-Johnson
Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
Illinois, IL 61801, USA
{jhuang29,jhasegaw}@illinois.edu

Abstract

This paper investigates semi-supervised learning of Gaussian mixture models using an unified objective function taking both labeled and unlabeled data into account. Two methods are compared in this work – the hybrid discriminative/generative method and the purely generative method. They differ in the criterion type on labeled data; the hybrid method uses the class posterior probabilities and the purely generative method uses the data likelihood. We conducted experiments on the TIMIT database and a standard synthetic data set from UCI Machine Learning repository. The results show that the two methods behave similarly in various conditions. For both methods, unlabeled data improve training on models of higher complexity in which the supervised method performs poorly. In addition, there is a trend that more unlabeled data results in more improvement in classification accuracy over the supervised model. We also provided experimental observations on the relative weights of labeled and unlabeled parts of the training objective and suggested a critical value which could be useful for selecting a good weighing factor.

1 Introduction

Speech recognition acoustic models can be trained using untranscribed speech data (Wessel and Ney, 2005; Lamel et al., 2002; L. Wang and Woodland, 2007). Most such experiments begin by boostraping an initial acoustic model using a limited amount of manually transcribed data (normally in a scale from 30 minutes to several hours), and then the initial model is used to transcribe a relatively large amount of untranscribed data. Only the transcriptions with high confidence measures (Wessel and Ney, 2005; L. Wang and Woodland, 2007) or high agreement with closed captions (Lamel et al., 2002) are selected to augment the manually transcribed data, and new acoustic models are trained on the augmented data set.

The general procedure described above exactly lies in the context of semi-supervised learning problems and can be categorized as a self-training algorithm. Self-training is probably the simplest semi-supervised learning method, but it is also flexible to be applied to complex classifiers such as speech recognition systems. This may be the reason why little work has been done on exploiting other semi-supervised learning methods in speech recognition. Though not incorporated to speech recognizers yet, there has been some work on semi-supervised learning of Hidden Markov Models (HMM) for sequential classification. Inoue and Ueda (2003) treated the unknown class labels of the unlabeled data as hidden variables and used the expectation-maximization (EM) algorithm to optimize the joint likelihood of labeled and unlabeled data. Recently Ji et al. (2009) applied a homotopy method to select the optimal weight to balance between the log likelihood of labeled and unlabeled data when training HMMs.

Besides generative training of acoustic models, discriminative training is another popular paradigm in the area of speech recognition, but only when

*This research is funded by NSF grants 0534106 and 0703624.

Proceedings of the NAACL HLT Workshop on Semi-supervised Learning for Natural Language Processing, pages 75–83, Boulder, Colorado, June 2009. ©2009 Association for Computational Linguistics

the transcriptions are available. Wang and Woodland (2007) used the self-training method to augment the training set for discriminative training. Huang and Hasegawa-Johnson (2008) investigated another use of discriminative information from labeled data by replacing the likelihood of labeled data with the class posterior probability of labeled data in the semi-supervised training objective for Gaussian Mixture Models (GMM), resulting in a hybrid discriminative/generative objective function. Their experimental results in binary phonetic classification showed significant improvement in classification accuracy when labeled data are scarce. A similar strategy called "'multi-conditional learning"' was presented in (Druck et al., 2007) applied to Markov Random Field models for text classification tasks, with the difference that the likelihood of labeled data is also included in the objective. The hybrid discriminative/generative objective function can be interpreted as having an extra regularization term, the likelihood of unlabeled data, in the discriminative training criterion for labeled data. However, both methods in (Huang and Hasegawa-Johnson, 2008) and (Druck et al., 2007) encountered the same issue about determining the weights for labeled and unlabeled part in the objective function and chose to use a development set to select the optimal weight. This paper provides an experimental analysis on the effect of the weight.

With the ultimate goal of applying semi-supervised learning in speech recognition, this paper investigates the learning capability of algorithms within Gaussian Mixture Models because GMM is the basic model inside a HMM, therefore 1) the update equations derived for the parameters of GMM can be conveniently extended to HMM for speech recognition. 2) GMM can serve as an initial point to help us understand more details about the semi-supervised learning process of spectral features.

This paper makes the following contribution:

- it provides an experimental comparison of hybrid and purely generative training objectives.
- it studies the impact of model complexity on learning capability of algorithms.
- it studies the impact of the amount of unlabeled data on learning capability of algorithms.
- it analyzes the role of the relative weights of labeled and unlabeled parts of the training objective.

2 Algorithm

Suppose a labeled set $\mathcal{X}_L = (x_1, \ldots, x_n, \ldots, x_{N_L})$ has N_L data points and $x_n \in \mathcal{R}_d$. $\mathcal{Y}_L = (y_1, \ldots, y_n, \ldots, y_{N_L})$ are the corresponding class labels, where $y_n \in \{1, 2, \ldots, Y\}$ and Y is the number of classes. In addition, we also have an unlabeled set $\mathcal{X}_U = (x_1, \ldots, x_n, \ldots, x_{N_U})$ without corresponding class labels. Each class is assigned a Gaussian Mixture model, and all models are trained given $\mathcal{X}_L$ and $\mathcal{X}_U$. This section first presents the hybrid discriminative/generative objective function for training and then the purely generative objective function. The parameter update equations are also derived here.

2.1 Hybrid Objective Function

The hybrid discriminative/generative objective function combines the discriminative criterion for labeled data and the generative criterion for unlabeled data:

$$\mathcal{F}(\lambda) = \log P(\mathcal{Y}_L|\mathcal{X}_L;\lambda) + \alpha \log P(\mathcal{X}_U;\lambda), \quad (1)$$

and we chose the parameters so that (1) is maximized:

$$\hat{\lambda} = \arg\max_{\lambda} \mathcal{F}(\lambda). \quad (2)$$

The first component considers the log posterior class probability of the labeled set whereas the second component considers the log likelihood of the unlabeled set weighted by α. In ASR community, model training based the first component is usually referred to as Maximum Mutual Information Estimation (MMIE) and the second component Maximum Likelihood Estimation (MLE), therefore in this paper we use a brief notation for (1) just for convenience:

$$\mathcal{F}(\lambda) = \mathcal{F}_{\text{MMI}}^{(D_L)}(\lambda) + \alpha \mathcal{F}_{\text{ML}}^{(D_U)}(\lambda). \quad (3)$$

The two components are different in scale. First, the size of the labeled set is usually smaller than the size of the unlabeled set in the scenario of semi-supervised learning, so the sums over the data sets involve different numbers of terms; Second, the

scales of the posterior probability and the likelihood are essentially different, so are their gradients. While the weight α balances the impacts of two components on the training process, it may also implicitly normalize the scales of the two components. In section (3.2) we will discuss and provide a further experimental analysis.

In this paper, the models to be trained are Gaussian mixture models of continuous spectral feature vectors for phonetic classes, which can be further extended to Hidden Markov Models with extra parameters such as transition probabilities.

The maximization of (1) follows the techniques in (Povey, 2003), which uses auxiliary functions for objective maximization; In each iteration, a strong or weak sense auxiliary function is maximized, such that if the auxiliary function converges after iterations, the objective function will be at a local maximum as well.

The objective function (1) can be rewritten as

$$\begin{aligned}\mathcal{F}(\lambda) = &\log P(\mathcal{X}_L|\mathcal{Y}_L;\lambda) - \log P(\mathcal{X}_L;\lambda) \\ &+ \alpha \log P(\mathcal{X}_U;\lambda),\end{aligned} \tag{4}$$

where the term $\log P(\mathcal{Y}_L;\lambda)$ is removed because it is independent of acoustic model parameters.

The auxiliary function at the current parameter λ^{old} for (4) is

$$\begin{aligned}\mathcal{G}(\lambda,\lambda^{(\text{old})}) = &\mathcal{G}^{\text{num}}(\lambda,\lambda^{(\text{old})}) - \mathcal{G}^{\text{den}}(\lambda,\lambda^{(\text{old})}) \\ &+\alpha\mathcal{G}^{\text{den}}(\lambda,\lambda^{(\text{old})};\mathcal{D}_U) + \mathcal{G}^{\text{sm}}(\lambda,\lambda^{(\text{old})}),\end{aligned} \tag{5}$$

where the first three terms are strong-sense auxiliary functions for the conditional likelihood (referred to as the numerator(num) model because it appears in the numerator when computing the class posterior probability) $\log P(\mathcal{X}_L|\mathcal{Y}_L;\lambda)$ and the marginal likelihoods (referred to as the denominator(den) model likewise) $\log P(\mathcal{X}_L;\lambda)$ and $\alpha \log P(\mathcal{X}_U;\lambda)$ respectively. The last term is a smoothing function that doesn't affect the local differential but ensures that the sum of the first three term is at least a convex weak-sense auxiliary function for good convergence in optimization.

Maximization of (5) leads to the update equations for the class j and mixture m given as follows:

$$\hat{\mu}_{jm} = \frac{1}{\overline{\gamma}_{jm}}\left(\boldsymbol{x}^{\text{num}}_{jm,} - \boldsymbol{x}^{\text{den}}_{jm} + \alpha\boldsymbol{x}^{\text{den}}_{jm}(\mathcal{D}_U) + D_{jm}\mu_{jm}\right) \tag{6}$$

$$\begin{aligned}\hat{\sigma}^2_{jm} = \frac{1}{\overline{\gamma}_{jm}}&\left(\boldsymbol{s}^{\text{num}}_{jm} - \boldsymbol{s}^{\text{den}}_{jm} + \alpha\boldsymbol{s}^{\text{den}}_{jm}(\mathcal{D}_U)\right. \\ &\left.+D_{jm}\left(\sigma^2_{jm} + \mu^2_{jm}\right)\right) - \hat{\mu}^2_{jm},\end{aligned} \tag{7}$$

where for clarity the following substitution is used:

$$\overline{\gamma}_{jm} = \gamma^{\text{num}}_{jm} - \gamma^{\text{den}}_{jm} + \alpha\gamma^{\text{den}}_{jm}(\mathcal{D}_U) + D_{jm} \tag{8}$$

and γ_{jm} is the sum of the posterior probabilities of occupation of mixture component m of class j over the dataset:

$$\begin{aligned}\gamma^{\text{num}}_{jm}(X) &= \sum_{x_i\in X, y_i=j} p(m|x_i, y_i = j) \\ \gamma^{\text{den}}_{jm}(X) &= \sum_{x_i\in X} p(m|x_i)\end{aligned} \tag{9}$$

and $\boldsymbol{x}_{jm}$ and $\boldsymbol{s}_{jm}$ are respectively the weighted sum of x_i and x_i^2 over the whole dataset with the weight $p(m|x_i, y_i = j)$ or $p(m|x_i)$, depending on whether the superscript is the numerator or denominator model. D_{jm} is a constant set to be the greater of twice the smallest value that guarantees positive variances or γ^{den}_{jm} (Povey, 2003). The re-estimation formula for mixture weights is also derived from the Extended Baum-Welch algorithm:

$$\hat{c}_{jm} = \frac{c_{jm}\left\{\frac{\partial\mathcal{F}}{\partial c_{jm}} + C\right\}}{\sum_{m'} c_{jm'}\left\{\frac{\partial\mathcal{F}}{\partial c_{jm}} + C\right\}}, \tag{10}$$

where the derivative was approximated (Merialdo, 1988) in the following form for practical robustness for small-valued parameters :

$$\frac{\partial\mathcal{F}_{\text{MMI}}}{\partial c_{jm}} \approx \frac{\gamma^{\text{num}}_{jm}}{\sum_{m'}\gamma^{\text{num}}_{jm'}} - \frac{\gamma^{\text{den}}_{jm}}{\sum_{m'}\gamma^{\text{den}}_{jm'}}. \tag{11}$$

Under our hybrid framework, there is an extra term $\gamma^{\text{den}}_{jm}(D_U)/\sum_{m'}\gamma^{\text{den}}_{jm'}(D_U)$ that should exist in (11), but in practice we found that adding this term to the approximation is not better than the original form. Therefore, we keep using MMI-only update for mixture weights. The constant C is chosen such that all parameter derivatives are positive.

2.2 Purely Generative Objective

In this paper we compare the hybrid objective with the purely generative one:

$$\mathcal{F}(\lambda) = \log P(\mathcal{X}_L|\mathcal{Y}_L;\lambda) + \alpha \log P(\mathcal{X}_U;\lambda), \tag{12}$$

where the two components are total log likelihood of labeled and unlabeled data respectively. (12) doesn't suffer from the problem of combining two heterogeneous probabilistic items, and the weight α being equal to one means that the objective is a joint data likelihood of labeled and unlabeled set with the assumption that the two sets are independent. However, D_L or D_U might just be a sampled set of the population and might not reflect the true proportion, so we keep α to allow a flexible combination of two criteria. On top of that, we need to adjust the relative weights of the two components in practical experiments.

The parameter update equation is a reduced form of the equations in Section (2.1):

$$\hat{\mu}_{jm} = \frac{\boldsymbol{x}^{\text{num}}_{jm,} + \alpha \boldsymbol{x}^{\text{den}}_{jm}(\mathcal{D}_U)}{\gamma^{\text{num}}_{jm} + \alpha \gamma^{\text{den}}_{jm}(\mathcal{D}_U)} \tag{13}$$

$$\hat{\sigma}^2_{jm} = \frac{\boldsymbol{s}^{\text{num}}_{jm} + \alpha \boldsymbol{s}^{\text{den}}_{jm}(\mathcal{D}_U)}{\gamma^{\text{num}}_{jm} + \alpha \gamma^{\text{den}}_{jm}(\mathcal{D}_U)} - \hat{\mu}^2_{jm} \tag{14}$$

3 Results and Discussion

The purpose of designing the learning algorithms is for classification/recognition of speech sounds, so we conducted phonetic classification experiments using the TIMIT database (Garofolo et al., 1993). We would like to investigate the relation of learning capability of semi-supervised algorithms to other factors and generalize our observations to other data sets. Therefore, we used another synthetic dataset *Waveform* for the evaluation of semi-supervised learning algorithms for Gaussian Mixture model.

TIMIT: We used the same 48 phone classes and further grouped into 39 classes according to (Lee and Hon, 1989) as our final set of phone classes to model. We extracted 50 speakers out of the NIST complete test set to form the development set. All of our experimental analyses were on the development set. We used segmental features (Halberstadt, 1998) 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 into three regions with 3-4-3 proportion and calculated the PLP average over each region. Three averages plus the log duration of that phone gave a 40-dimensional $(13 \times 3 + 1)$ measurement vector.

Waveform: We used the second versions of the *Waveform* dataset available at the UCI repository (Asuncion and Newman, 2007). There are three classes of data. Each token is described by 40 real attributes, and the class distribution is even.

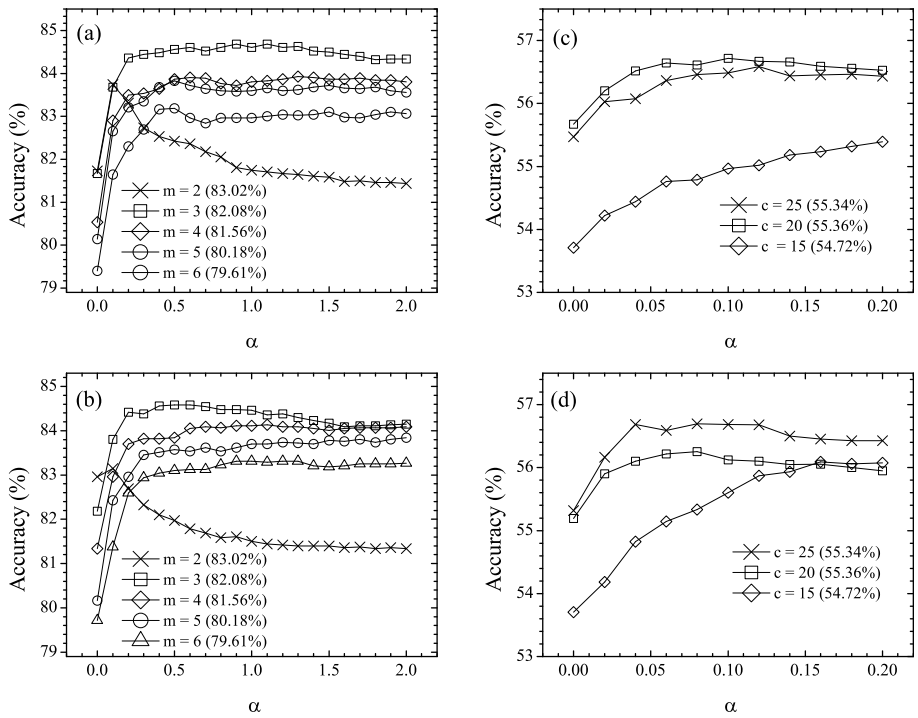
For *waveform*, because the class labels are equally distributed, we simply assigned equal number of mixtures for each class. For TIMIT, the phone classes are unevenly distributed, so we assigned variable number of Gaussian mixtures for each class by controlling the averaged data counts per mixture. For all experiments, the initial model is an MLE model trained with labeled data only.

To construct a mixed labeled/unlabeled data set, the original training set were randomly divided into the labeled and unlabeled sets with desired ratio, and the class labels in the unlabeled set are assumed to be unknown. To avoid that the classifier performance may vary with particular portions of data, we ran *five* folds for every experiment, each fold corresponding to different division of training data into labeled and unlabeled set, and took the averaged performance.

3.1 Model Complexity

This section analyzes the learning capability of semi-supervised learning algorithms for different model complexities, that is, the number of mixtures for Gaussian mixture model. In this experiment, the sizes of labeled and unlabeled set are fixed ($|D_L| : |D_U| = 1 : 10$ and the averaged token counts per class is around 140 for both data sets), as we varied the total number of mixtures and evaluated the updated model by its classification accuracy. For *waveform*, number of mixtures was set from 2 to 7; for TIMIT, because the number of mixtures per class is determined by the averaged data counts per mixture c, we set c to 25, 20 and 15 as the higher c gives less number of mixtures in total. Figure 3.1 plots the averaged classification accura-

Figure 1: Mean classification accuracies vs. α for different model complexity. The accuracies for the initial MLE models are indicated in the parentheses. (a) *waveform*: training with the hybrid objective. (b) *waveform*: purely generative objective. (c) TIMIT: training with the hybrid objective. (d) TIMIT: purely generative objective.



cies of the updated model versus the value of α with different model complexities. The ranges of α are different for *waveform* and TIMIT because the value of α for each dataset has different scales.

First of all, the hybrid method and purely generative method have very similar behaviors in both *waveform* and TIMIT; the differences between the two methods are insignificant regardless of α. The hybrid method with $\alpha = 0$ means supervised MMI-training with labeled data only, and the purely generative method with $\alpha = 0$ means extra several rounds of supervised MLE-training if the convergence criterion is not achieved. With the small amount of labeled data, most of hybrid curves start slightly lower than the purely generative ones at $\alpha = 0$, but increase to as high as the purely generative ones as α increases.

For *waveform*, the accuracies increase with α increases for all cases except for the 2-mixture model. Table 1 summarizes the numbers from Figure 3.1. Except for the 2-mixture case, the improvement over the supervised model ($\alpha = 0$) is positively correlated to the model complexity, as the largest improvements occur at the 5-mixture and 6-mixture model for the hybrid and purely generative method respectively. However, the highest complexity does not necessarily gives the best classification accuracy; the 3-mixture model achieves the best accuracy among all models after semi-supervised learning whereas the 2-mixture model is the best model for supervised learning using labeled data only.

Experiments on TIMIT show a similar behavior[1]; as shown in both Figure 3.1 and Table 2, the improvement over the supervised model ($\alpha = 0$) is also positively correlated to the model complexity,

[1]Note that our baseline performance (the initial MLE model) is much worse than benchmark because only 10% of the training data were used. We justified our baseline model by using the whole training data and a similar accuracy (74%) to other work (e.g. (Sha and Saul, 2007)) was obtained.

Table 1: The accuracies(%) of the initial MLE model, the supervised model ($\alpha = 0$), the best accuracies with unlabeled data and the absolute improvements (Δ) over $\alpha = 0$ for different model complexities for *waveform*. The bolded number is the highest value along the same column.

		Hybrid			Purely generative		
#. mix	init. acc.	$\alpha = 0$	best acc.	Δ	$\alpha = 0$	best acc.	Δ
2	**83.02**	**81.73**	83.74	2.01	**82.96**	83.14	0.18
3	82.08	81.66	**84.69**	3.03	82.18	**84.58**	2.40
4	81.56	80.53	83.93	3.40	81.34	84.13	2.79
5	80.18	80.14	83.82	3.68	80.16	83.84	**3.68**
6	79.61	79.40	83.19	**3.79**	79.71	83.31	3.60

Table 2: The accuracies(%) of the initial MLE model, the supervised model ($\alpha = 0$), the best accuracies with unlabeled data and the absolute improvements (Δ) over $\alpha = 0$ for different model complexities for TIMIT. The bolded number is the highest value along the same column.

		Hybrid			Purely generative		
c	init. acc.	$\alpha = 0$	best acc.	Δ	$\alpha = 0$	best acc.	Δ
25	55.34	55.47	56.58	1.11	**55.32**	**56.7**	1.38
20	**55.36**	**55.67**	**56.72**	1.05	55.2	56.25	1.05
15	54.72	53.71	55.39	**1.68**	53.7	56.09	**2.39**

as the most improvements occur at $c = 25$ for both hybrid and purely generative methods. The semi-supervised model consistently improves over the supervised model. To summarize, unlabeled data improve training on models of higher complexity, and sometimes it helps achieve the best performance with a more complex model.

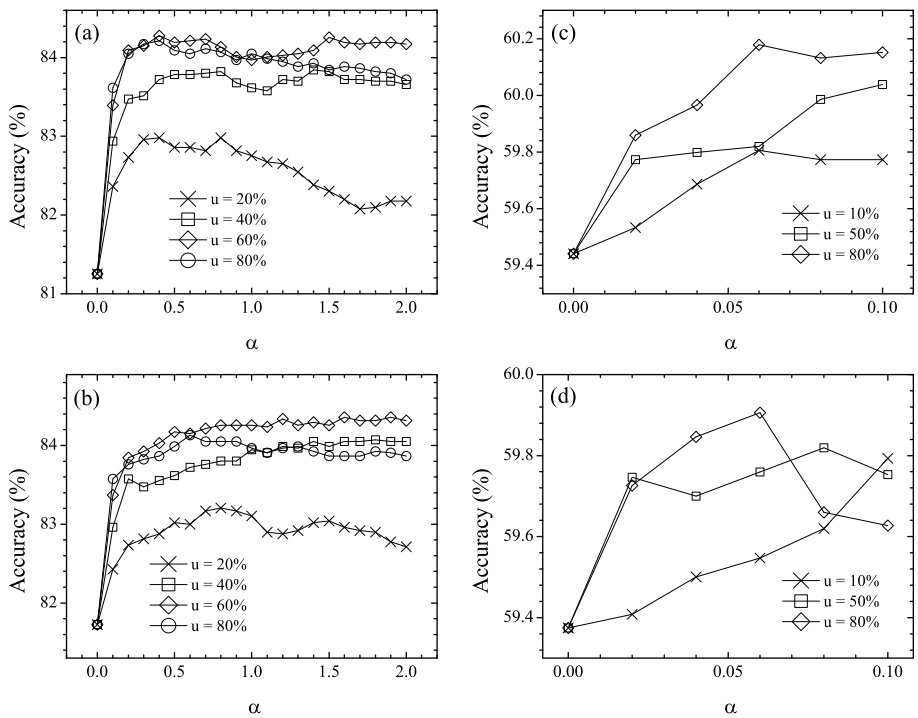
3.2 Size of Unlabeled Data

In Figure 2, we fixed the size of the labeled set (4% of the training set) and plotted the averaged classification accuracies for learning with different sizes of unlabeled data. First of all, the hybrid method and purely generative method still behave similarly in both *waveform* and TIMIT. For both datasets, the figures clearly illustrate that more unlabeled data contributes more improvement over the supervised model regardless of the value of α. Generally, a data distribution can be expected more precisely with a larger sample size from the data pool, therefore we expect the more unlabeled data the more precise information about the population, which improves the learning capability.

3.3 Discussion of α

During training, the weighted sum of $\mathcal{F}_{\text{MMI}}$ and $\mathcal{F}_{\text{ML}}$ in equation (15) increases with iterations, however $\mathcal{F}_{\text{MMI}}$ and $\mathcal{F}_{\text{ML}}$ are not guaranteed to increase individually. Figure 3 illustrates how α affects the respective change of the two components for a particular setting for *waveform*. When $\alpha = 0$, the objective function does not take unlabeled data into account, so $\mathcal{F}_{\text{MMI}}$ increases while $\mathcal{F}_{\text{ML}}$ decreases. $\mathcal{F}_{\text{ML}}$ starts to increase for nonzero α; $\alpha = 0.01$ corresponds to the case where both objectives increases. As α keeps growing, $\mathcal{F}_{\text{MMI}}$ starts to decrease whereas $\mathcal{F}_{\text{ML}}$ keeps rising. In this particular example, $\alpha = 0.05$ is the critical value at which $\mathcal{F}_{\text{MMI}}$ changes from increasing to decreasing. According to our observation, the value of α depends on the dataset and the relative size of labeled/unlabeled data. Table 3 shows the critical values for *waveform* and TIMIT for different sizes of labeled data ($5, 10, 15, 20\%$ of the training set) with a fixed set of unlabeled data (80%.) The numbers are very different across the datasets, but there is a consistent pattern within the dataset–the critical value increases as the size of labeled set increases. One possible explanation is that α contains an normal-

Figure 2: Mean classification accuracies vs. α for different amounts of unlabeled data (the percentage in the training set). The averaged accuracy for the initial MLE model is 81.66% for *waveform* and 59.41% for TIMIT. (a) *waveform*: training with the hybrid objective. (b) *waveform*: purely generative objective. (c) TIMIT: training with the hybrid objective. (d) TIMIT: purely generative objective.



ization factor with respect to the relative size of labeled/unlabeled set. The objective function in (15) can be rewritten in terms of the normalized objective with respect to the data size:

$$\mathcal{F}(\lambda) = |D_L|\overline{\mathcal{F}}_{\text{MMI}}^{(D_L)}(\lambda) + \alpha|D_U|\overline{\mathcal{F}}_{\text{ML}}^{(D_U)}(\lambda). \quad (15)$$

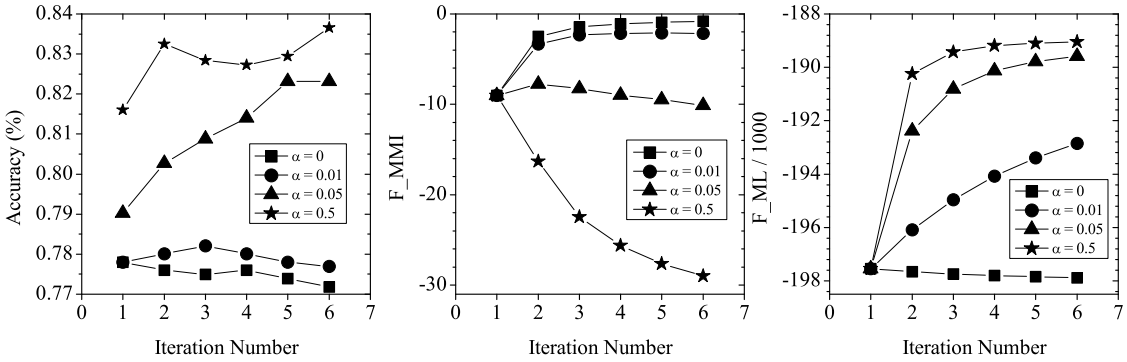
where $\overline{\mathcal{F}}^{(X)}$ means the averaged value over the data set X. When the labeled set size increases, α may have to scale up accordingly such that the relative change of the two averaged component remains in the same scale.

Although α controls the dominance of the criterion on labeled data or on unlabeled data, the fact that which dominates the objective or the critical value does not necessary indicate the best α. However, we observed that the best α is usually close to or larger than the critical value, but the exact value varies with different data. At this point, it might still be easier to find the best weight using a small development set. But this observation also provides a guide about the reasonable range to search the best α – searching starting from the critical value and it should reach the optimal value soon according to the plots in Figure 3.1.

Table 3: The critical values for *waveform* and TIMIT for different sizes of labeled data (percentage of training data) with a fixed set of unlabeled data (80 %.)

Size of labeled data	*waveform*	TIMIT
5%	0.09-0.11	0.03-0.04
10%	0.12-0.14	0.07-0.08
15%	0.5-0.6	0.08-0.09
20%	1-1.5	0.11-0.12

Figure 3: Accuracy (left), $\mathcal{F}_{\text{MMI}}$ (center), and $\mathcal{F}_{\text{ML}}$ (right) at different values of $alpha$.



3.4 Hybrid Criterion vs. Purely Generative Criterion

From the previous experiments, we found that the hybrid criterion and purely generative criterion almost match each other in performance and are able to learn models of the same complexity. This implies that the criterion on labeled data has less impact on the overall training direction than unlabeled data. In Section 3.2, we mentioned that the best α is usually larger than or close to the critical value around which the unlabeled data likelihood tends to dominate the training objective. This again suggests that labeled data contribute less to the training objective function compared to unlabeled data, and the criterion on labeled data doesn't matter as much as the criterion on unlabeled data. It is possible that most of the contributions from labeled data have already been used for training an initial MLE model, therefore little information could be extracted in the further training process.

4 Conclusion

Regardless of the dataset and the training objective type on labeled data, there are some general properties about the semi-supervised learning algorithms studied in this work. First, while limited amount of labeled data can at most train models of lower complexity well, the addition of unlabeled data makes the updated models of higher complexity much improved and sometimes perform better than less complex models. Second, the amount of unlabeled data in our semi-supervised framework generally follows 'the-more-the-better' principle; there is a trend that more unlabeled data results in more improvement in classification accuracy over the supervised model.

We also found that the objective type on labeled data has little impact on the updated model, in the sense that hybrid and purely generative objectives behave similarly in learning capability. The observation that the best α occurs after the MMI criterion begins to decrease supports the fact that the criterion on labeled data contributes less than the criterion on unlabeled data. This observation is also helpful in determining the search range for the best α on the development set by locating the critical value of the objective as a start point to perform search.

The unified training objective method has a nice convergence property which self-training methods can not guarantee. The next step is to extend the similar framework to speech recognition task where HMMs are trained and phone boundaries are segmented. It would be interesting to compare it with self-training methods in different aspects (e.g. performance, reliability, stability and computational efficiency).

References

A. Asuncion and D.J. Newman. 2007. UCI machine learning repository.

Gregory Druck, Chris Pal, Andrew McCallum, and Xiaojin Zhu. 2007. Semi-supervised classification with hybrid generative/discriminative methods. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 280–289, New York, NY, USA. ACM.

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

Andrew K. Halberstadt. 1998. *Heterogeneous Acoustic Measurements and Multiple Classifiers for Speech Recognition*. Ph.D. thesis, Massachusetts Institute of Technology.

J.-T. Huang and Mark Hasegawa-Johnson. 2008. Maximum mutual information estimation with unlabeled data for phonetic classification. In *Interspeech*.

Masashi Inoue and Naonori Ueda. 2003. Exploitation of unlabeled sequences in hidden markov models. *IEEE Trans. On Pattern Analysis and Machine Intelligence*, 25:1570–1581.

Shihao Ji, Layne T. Watson, and Lawrence Carin. 2009. Semisupervised learning of hidden markov models via a homotopy method. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(2):275–287.

M.J.F. Gales L. Wang and P.C. Woodland. 2007. Unsupervised training for mandarin broadcast news and conversation transcription. In *Proc. IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 4, pages 353–356.

Lori Lamel, Jean-Luc Gauvain, and Gilles Adda. 2002. Lightly supervised and unsupervised acoustic model training. 16:115–129.

K.-F. Lee and H.-W. Hon. 1989. Speaker-independent phone recognition using hidden markov models. *IEEE Transactions on Speech and Audio Processing*, 37(11):1641–1648.

B. Merialdo. 1988. Phonetic recognition using hidden markov models and maximum mutualinformation training. In *Proc. IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 111–114.

Daniel Povey. 2003. *Discriminative Training for Large Vocabulary Speech Recognition*. Ph.D. thesis, Cambridge University.

Fei Sha and Lawrence K. Saul. 2007. Large margin hidden markov models for automatic speech recognition. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 1249–1256. MIT Press, Cambridge, MA.

Frank Wessel and Hermann Ney. 2005. Unsupervised training of acoustic models for large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, 13(1):23–31, January.