Maximum Mutual Information Estimation with Unlabeled Data for Phonetic Classification

Jui-Ting Huang, Mark Hasegawa-Johnson

Department of Electrical and Computer Engineering,
University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
jhuang29@uiuc.edu, jhasegawa@uiuc.edu

Abstract

This paper proposes a new training framework for mixed labeled and unlabeled data and evaluates it on the task of binary phonetic classification. Our training objective function combines Maximum Mutual Information (MMI) for labeled data and Maximum Likelihood (ML) for unlabeled data. Through the modified training objective, MMI estimates are smoothed with ML estimates obtained from unlabeled data. On the other hand, our training criterion can also help the existing model adapt to new speech characteristics from unlabeled speech. In our experiments of phonetic classification, there is a consistent reduction of error rate from MLE to MMIE with I-smoothing, and then to MMIE with unlabeled-smoothing. Error rates can be further reduced by transductive-MMIE. We also experimented with the gender-mismatched case, in which the best improvement shows MMIE with unlabeled data has a 9.3% absolute lower error rate than MLE and a 2.35% absolute lower error rate than MMIE with I-smoothing.

Index Terms: unlabeled speech, Maximum mutual information, Gaussian mixture models

1. Introduction

There have been many successful discriminative training techniques applied to the parameter estimation of Hidden Markov Models (HMM) for automatic speech recognition (ASR). Examples of these include maximum mutual information estimation (MMIE) [1], minimum classification error (MCE) training [2], minimum phone error (MPE) training [3], and more recently large margin methods [4]. All of these discriminative methods require the correct transcription for the speech corpora used for training HMMs.

On the other hand, there are a large number of untranscribed corpora from the Internet and other sources, which are relatively easy and cheap to maintain compared to the transcribed ones. In the machine learning world, some researchers have started to seek "semi-supervised learning" [5] as a way of making use of those "cheap" unlabeled data together with the labeled set. Based on the similar motivation, this paper investigates how unlabeled data can be integrated into a discriminative training objective, how much improvement it can provide, and under what condition it can help most.

In recent years, researchers in ASR have been also interested in the potential benefit from unlabeled speech to the acoustic model training. Most published approaches [6, 7] use an existing speech recognizer to transcribe unlabeled speech (possibly with the help of closed captions), and then the newly transcribed data with sufficiently high confidence measures are added into the training set for training an improved recognizer. Furthermore, this procedure is applied iteratively. Starting from another angle, we seek a way to combine unlabeled and labeled data for training simultaneously, instead of iterative tagging; our approach is probably complementary to iterative tagging, though we have not tested the combination. The advantage of our approach is that we are free of worries about finding reliable recognizer outputs or defining confidence thresholds. The basic idea of our algorithm is to add the total likelihood of unlabeled data as a criterion into the original MMI objective function. In this way unlabeled data place an additional constraint in a maximum likelihood sense on the parameters estimated from labeled data. This extends the H-criterion proposed in [8], which is an interpolation of the MMI criterion and ML criterion, because the ML criterion in our objective function is for unlabeled data while that in the H-criterion is for labeled data.

Our algorithm for a combinational use of labeled and unlabeled data can be applied under at least three different scenarios, for each of which we evaluated the improvement gained from unlabeled data. One is to train an improved recognizer when the evaluation test set shares speech characteristics with the labeled training corpus. In this case, unlabeled data are expected to contribute extra information especially when labeled data are limited and insufficient for training a good model. We call it *unlabeled-smoothing* for it prevents model over-training in a similar way to I-smoothing [9]. The second scenario adapts an existing model to a new untranscribed corpus which might have different characteristics from the labeled data used for training the existing model. In this case, unlabeled data help the model adapt to the new speech characteristics, which can be regarded as an off-line unsupervised adaptation task based on a discriminative criterion. In the third scenario, evaluation test data are used as unlabeled training data; we call this scenario transductive MMIE.

The rest of the paper is organized as follows. Section 2 first describes the MMI objective function and I-smoothing. Section 3 introduces our modified objective function and the resulting re-estimation formulas. The experiment setup and results are then shown in section 4, followed by the discussion in section 5.

2. Maximum Mutual Information Estimation

Generally for a classification problem, the MMI objective function is the log sum of the posterior probability over all labeled data points:

$$\mathcal{F}_{\text{MMI}}(\lambda) = \sum_{x_i \in \mathcal{X}} \log \frac{p_\lambda(x_i|y_i)\, p(y_i)}{\sum_c p_\lambda(x_i|c)\, p(c)}, \qquad (1)$$

where λ is the model parameter set, x_i is the feature data point and y_i is the corresponding class label. Suppose the model to be trained for each class is a Gaussian mixture model, which can be extended to HMM in the context of the speech recognition system. One often-used optimization scheme is an iterative Extended Baum-Welch updating procedure, in which the mean and variance for the class j and mixture m are updated iteratively as follows:

$$\hat{\mu}_{jm} = \frac{\boldsymbol{x}_{jm}^{\text{num}} - \boldsymbol{x}_{jm}^{\text{den}} + D_{jm}\mu_{jm}}{\gamma_{jm}^{\text{num}} - \gamma_{jm}^{\text{den}} + D_{jm}} \tag{2}$$

$$\hat{\sigma}_{jm}^2 = \frac{\boldsymbol{s}_{jm}^{\text{num}} - \boldsymbol{s}_{jm}^{\text{den}} + D_{jm}\left(\sigma_{jm}^2 + \mu_{jm}^2\right)}{\gamma_{jm}^{\text{num}} - \gamma_{jm}^{\text{den}} + D_{jm}} - \hat{\mu}_{jm}^2, \tag{3}$$

where the superscript "num" represents the correct model corresponding to the numerator in eq. (1) and "den" represents the classification model in the denominator containing all possible classes. For either model, γ_{jm} is the sum of the posterior probabilities of occupation of mixture component m of class j over the dataset:

$$\begin{aligned} \gamma_{jm}^{\text{num}} &= \sum_{x_i \in X, y_i = j} p\left(m|x_i, y_i = j\right) \\ \gamma_{jm}^{\text{den}} &= \sum_{x_i \in X} p\left(m|x_i\right) \end{aligned} \tag{4}$$

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 $p\left(m|x_i, y_i = j\right)$ or $p\left(m|x_i\right)$, depending on 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_{jm}^{den} [9].

The re-estimation formula for mixture wights is also derived from Extended Baum-Welch algorithm:

$$\hat{c}_{jm} = \frac{c_{jm}\left\{\frac{\partial \mathcal{F}_{\text{MMI}}}{\partial c_{jm}} + C\right\}}{\sum_{m'} c_{jm'}\left\{\frac{\partial \mathcal{F}_{\text{MMI}}}{\partial c_{jm}} + C\right\}}, \tag{5}$$

where the derivative was suggested [10] in the following form:

$$\frac{\partial \mathcal{F}_{\text{MMI}}}{\partial c_{jm}} \approx \frac{\gamma_{jm}^{\text{num}}}{\sum_{m'} \gamma_{jm'}^{\text{num}}} - \frac{\gamma_{jm}^{\text{den}}}{\sum_{m'} \gamma_{jm'}^{\text{den}}}. \tag{6}$$

As a smoothing technique to prevent the MMI model from over-training, [9] proposed I-smoothing which backs off the MMI estimates to the ML estimates with a certain degree. It can be shown that I-smoothing is equivalent to adding a log prior distribution to the MMI objective function:

$$\begin{aligned} \mathcal{F}\left(\lambda\right) =& \mathcal{F}_{\text{MMI}}\left(\lambda\right) + \log p\left(\lambda\right) \\ =& \mathcal{F}_{\text{MMI}}\left(\lambda\right) + \sum_{j,m} \log p\left(\mu_{jm}, \sigma_{jm}\right), \end{aligned} \tag{7}$$

where the log prior for each j and m is proportional to the log likelihood of τ artificial data points with mean and variance from the ML statistics on the numerator model, $\mu_{jm\,(ML)}^{\text{num}}$ and $\sigma_{jm\,(ML)}^{\text{num}}$:

$$\log p\left(\mu_{jm}, \sigma_{jm}^2\right) = \sum_{i=1}^{\tau} \log p\left(\bar{x}_i|\mu_{jm}, \sigma_{jm}\right), \tag{8}$$

where the artificial data $\bar{x}_i$'s are generated from the distribution $\mathcal{N}\left(\mu_{jm\,(ML)}^{\text{num}}, \sigma_{jm\,(ML)}^{\text{num}}\right)$.

Incorporating the prior likelihood into the objective function amounts to changes in the numerator statistics before the mean and variance update eqs. (2) and (3):

$$\gamma_{jm}^{\text{num}\prime} = \gamma_{jm}^{\text{num}} + \tau \tag{9}$$

$$\boldsymbol{x}_{jm}^{\text{num}\prime} = \boldsymbol{x}_{jm}^{\text{num}} + \tau\mu_{jm\,(\text{ML})}^{\text{num}} \tag{10}$$

$$\boldsymbol{s}_{jm}^{\text{num}2\prime} = \boldsymbol{s}_{jm}^{\text{num}2} + \tau\left(\sigma_{jm\,(\text{ML})}^{\text{num}2} + \mu_{jm\,(\text{ML})}^{\text{num}2}\right) \tag{11}$$

3. MMIE with unlabeled data

One way of incorporating unlabeled data into the MMI training is to add the total likelihood of unlabeled data to the MMI objective function:

$$\mathcal{F}\left(\lambda\right) = \mathcal{F}_{\text{MMI}}\left(\lambda; \mathcal{D}_L\right) + \alpha\mathcal{F}_{\text{ML}}\left(\lambda; \mathcal{D}_U\right), \tag{12}$$

where

$$\mathcal{F}_{\text{MMI}}\left(\lambda; \mathcal{D}_L\right) = \sum_{x_i \in \mathcal{D}_L} \log \frac{p_\lambda\left(x_i|y_i\right) p\left(y_i\right)}{\sum_c p_\lambda\left(x_i|c\right) p\left(c\right)} \tag{13}$$

$$\begin{aligned} \mathcal{F}_{\text{ML}}\left(\lambda; \mathcal{D}_U\right) &= \sum_{x_i \in \mathcal{D}_U} \log p_\lambda\left(x_i\right) \\ &= \sum_{x_i \in \mathcal{D}_U} \log \sum_c p_\lambda\left(x_i|c\right) p\left(c\right), \end{aligned} \tag{14}$$

and α is a scaling factor ranging from 0 to 1. With the modified objective function in eq.(12), the training process will try to maximize the posterior probability for the labeled set while the likelihood of the unlabeled set given the estimated parameter set needs to be as large as possible.

Maximization of eq. (12) can be solved by maximization of its weak-sense auxiliary function [9]:

$$\begin{aligned} \mathcal{G}(\lambda, \lambda^{(\text{old})}) =& \mathcal{G}^{\text{num}}(\lambda, \lambda^{(\text{old})}; \mathcal{D}_L) - \mathcal{G}^{\text{den}}(\lambda, \lambda^{(\text{old})}; \mathcal{D}_L) \\ &+ \alpha\mathcal{G}^{\text{den}}(\lambda, \lambda^{(\text{old})}; \mathcal{D}_U) + \mathcal{G}^{\text{sm}}(\lambda, \lambda^{(\text{old})}; \mathcal{D}_L), \end{aligned} \tag{15}$$

where the first three terms are strong-sense auxiliary functions derived separately from the log-likelihoods $\log p(\mathcal{D}_L|\mathcal{M}^{\text{num}})$, $\log p(\mathcal{D}_L|\mathcal{M}^{\text{den}})$ and $\log p(\mathcal{D}_U|\mathcal{M}^{\text{den}})$; $\mathcal{M}^{\text{den}}$ and $\mathcal{M}^{\text{den}}$ refer to the numerator and denominator model introduced in the previous section. 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 optimization.

Eq. (15) differs from the original MMIE criterion only by an additional term $\alpha\mathcal{G}^{\text{den}}(\lambda, \lambda^{(\text{old})}; \mathcal{D}_U)$, which can be easily shown to ultimately lead to changing the numerator statistics before the mean and variance update eqs. (2) and (3):

$$\gamma_{jm}^{\text{num}\prime} = \gamma_{jm}^{\text{num}} + \alpha\gamma_{jm}^{\text{den}}(\mathcal{D}_U) \tag{16}$$

$$\boldsymbol{x}_{jm}^{\text{num}\prime} = \boldsymbol{x}_{jm}^{\text{num}} + \alpha\mu_{jm\,(\text{ML})}^{\text{den}}(\mathcal{D}_U) \tag{17}$$

$$\boldsymbol{s}_{jm}^{\text{num}2\prime} = \boldsymbol{s}_{jm}^{\text{num}2} + \alpha\left(\sigma_{jm\,(\text{ML})}^{\text{den}\,2}(\mathcal{D}_U) + \mu_{jm\,(\text{ML})}^{\text{den}\,2}(\mathcal{D}_U)\right), \tag{18}$$

where $\gamma_{jm}^{\text{den}}(\mathcal{D}_U)$ is the sum of the posterior probabilities of occupation of mixture component m of class j, given the denominator (classification) model, over the unlabeled set, and $\mu_{jm\,(\text{ML})}^{\text{den}}(\mathcal{D}_U)$ and $\sigma_{jm\,(\text{ML})}^{\text{den}\,2}(\mathcal{D}_U)$ are the ML estimates of mean and variance from the unlabeled set.

Consequently, the unlabeled data are incorporated to the model in an "EM-like" way, instead of being hard-classified into any class as in other iterative semi-supervised approaches. The form of the modification of the statistics in eq. (16-18) is similar to eq. (9-11) for I-smoothing. Therefore we expect it may have similar smoothing behavior, preventing over-training. A difference exists, in that unlabeled data backs off the MMI estimates to new ML estimates of the classification model using the unlabeled set, rather than the ML estimates of the numerator model from the labeled set. As an additional comparison, the H-Criterion [8] backs off the MMI estimates to the ML estimates of the denominator (classification) model using the labeled set, which is less useful in the sense that the labeled set already has the correct class label and there is no reason to discard that information.

4. Experiments

To evaluate the performance of our modified MMIE, we conducted experiments on binary phonetic classification of phones [d] vs. [t] using the TIMIT corpus [11]. For parameter tuning of τ and α in I-smoothing and unlabeled-smoothing respectively, we extracted 50 speakers out of the NIST complete test set to form the development set. 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 [12]. Table 1 summarizes the number of tokens for each phone in each of the sets. Furthermore, for training on mixed labeled/unlabeled data, 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.

Table 1: *Number of tokens for phones "d" and "t" in the TIMIT training, development and test set.*

	d	t
Train	2432	3948
Development	239	413
Test	602	954

We used segmental features [12] 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. The Gaussian mixture model for each phone had two mixture components for all of our experiments.

4.1. MMIE with Unlabeled-Smoothing

As mentioned in the beginning of this section, $r = 15, 20, 30, 100\%$ of the training set was the labeled set and the rest formed the unlabeled set. The labeled set was used to train an initialized model using MLE, based on which other MMIE algorithms continued the training iteration. We found the best values of τ and α for I-smoothing and "unlabeled smoothing" on the development set, and the model trained with that value set was tested on the evaluation test set. In our experiments, α was usually within the range from 0.01 to 0.05, τ from 20 to 50.

In addition, we also applied the algorithm in a *transductive* way. The test data are no different from unlabeled data except that their classification results are collected for the system evaluation. Therefore, the test set can be regarded as a part of the unlabeled set as well; the whole test set was added to the unlabeled set, and the same objective function in eq. (12) still held for training a *transductive-MMIE* model.

The classification error rates are listed in Table 2. Any kind of MMIE approach has significantly less error than MLE. Regardless of the proportion of unlabeled data, the error rates have a general trend of decreasing starting from MLE to transductive MMIE as seen from the table; MMIE with I-smoothing improves over MLE, MMIE with unlabeled data improves over I-smoothing, and transductive-MMIE improves over MMIE with unlabeled data. On the other hand, the improvement becomes smaller as the proportion of labeled data increases ($15\% > 20\% > 30\%$).

Table 2: *Error rates (%) of Gaussian mixture models with $r = 15, 20, 30, 100\%$ of the training set being the labeled set and the rest being the unlabeled set.*

	15%	20%	30%	100%
MLE	38.69	21.02	21.79	20.69
MMIE with I-smoothing	19.22	20.37	19.22	17.99
MMIE with U-smoothing	18.25	20.18	19.15	N/A
Transductive-MMIE	17.87	19.99	19.15	17.99

4.2. MMIE for the gender-mismatched case

To further demonstrate the benefit of unlabeled data, we evaluated our algorithm on the mismatched case where the training and testing corpus have quite different speech characteristics. We wanted to examine whether unlabeled data which share the same characteristics with the testing corpus would improve over the MMIE model trained only with the labeled set. One of example of mismatch is gender-mismatch. Therefore, instead of randomly dividing the standard training set into the labeled and unlabeled set, we divided it into two sets by the gender of the speakers, male being the labeled set and female being the unlabeled set, and the opposite setting was also tried. From the original development and test set defined in section 4.1, tokens of the same gender as the unlabeled set were extracted to form the respective development set and test set here. The speaker identities were never overlapped across any sets. Table 3 summarizes the number of data points for each of the sets in the gender-mismatched cases.

The classification error rates are listed in Table 4. In both settings, MMIE with either I-smoothing or unlabeled data improves over MLE by a large amount. MMIE with unlabeled data improves over MMIE with pure I-smoothing by absolute 2.06% and 2.35%, for the respective gender setting.

5. Discussion

In terms of the error rate reduction, unlabeled data helped most in the mismatched case, especially when labeled data were from female speakers and unlabeled data were from male speakers. Other than that, the improvements from MMIE with I-smoothing to MMIE with unlabeled data were consistent, but not statistically significant, perhaps because of the small size of the test set. We expect to see more significant differences when the task is extended to multi-class phonetic classification.

Table 3: *Number of tokens for phones "d" and "t" in the labeled, unlabeled, development and test set in the gender-mismatched cases.*

Labeled	male		female	
Unlabeled	female		male	
	d	t	d	t
Labeled	1681	2791	751	1157
Unlabeled	451	1157	1681	2791
Development	91	131	148	282
Test	219	316	383	638

Table 4: *Mismatched case: Error rates (%) of Gaussian mixture models. * indicates that entries significantly lower than their predecessors (McNemar's test, $p = 0.05$).*

Labeled	male	female
Unlabeled	female	male
MLE	26.17	29.97
MMIE with I-smoothing	18.13*	23.02*
MMIE with unlabeled data	16.07	20.67*

In the matched case, transductive-MMIE resulted in consistent (but not statistically significant) lower error rates than MMIE with unlabeled-smoothing. This shows that our framework allows transductive learning based on discriminative training, but unlike other transductive approaches such as transductive SVM [13], we don't have to worry about issues such as setting a threshold or defining a confidence score. We also notice that when there was no available unlabeled data, taking the test set as the only unlabeled data for transductive learning did not improve over MMIE with I-smoothing, perhaps because the test set alone was too small to have an impact on the model training through our proposed objective function.

6. Conclusions and Future Work

A new training criterion that integrates Maximum Mutual Information for labeled data and Maximum Likelihood for unlabeled data has been introduced. Through the new training criterion, unlabeled data can smooth MMI estimates in a semi-supervised acoustic model training scenario; it can also provide information about new speech characteristics to the unadapted model in an unsupervised model adaptation scenario. In our experiments of phonetic classification, there are consistent improvements in error rates from MLE to MMIE with I-smoothing, and then to MMIE with unlabeled-smoothing. Moreover, error rates can be further reduced by transductive-MMIE, in which the test set is also a part of the unlabeled set. Overall, the best benefit by unlabeled data was seen in the gender-mismatched case, with a reduction in classification error rates 9.3% absolute compared to MLE and a 2.35% absolute lower error rate than MMIE with pure I-smoothing.

In the future, we plan to conduct experiments on multi-class phonetic classification. Our ultimate goal is to apply the new training criterion to the task of phonetic recognition in order to investigate the impact of unlabeled data on discriminative training of the acoustic model.

7. Acknowledgements

The authors would like to thank Dr. Mari Ostendorf from University of Washington for her inspiring idea about unlabeled data during the Semi-Supervised Language Learning workshop in 2007 summer. This material is based in part upon work supported by the National Science Foundation under Grant Number IIS-0534133 to Chilin Shih and Gary Cziko.

8. References

[1] Y. Normandin, "Hidden markov models, maximum mutual information estimation, and the speech recognition problem," Ph.D. dissertation, Montreal, Que., Canada, Canada, 1991.

[2] B.-H. Juang, W. Chou, and C.-H. Lee, "Minimum Classification Error Rate Methods for Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 3, pp. 257–265, 1997.

[3] D. Povey and P. Woodland, "Minimum phone error and i-smoothing for improved discriminative training," in *Proc. IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 2002, pp. 105–108.

[4] F. Sha and L. K. Saul, "Large margin hidden markov models for automatic speech recognition," in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. Platt, and T. Hoffman, Eds. Cambridge, MA: MIT Press, 2007, pp. 1249–1256.

[5] X. Zhu, "Semi-supervised learning literature survey," Computer Sciences, University of Wisconsin-Madison, Tech. Rep. 1530, 2005, http://www.cs.wisc.edu/∼jerryzhu/pub/ssl_survey.pdf.

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

[7] 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.

[8] P. Gopalakrishnan, D. Kanevsky, A. Nadas, D. Nahamoo, and M. Picheny, "Decoder selection based on cross-entropies," in *Proc. IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 1998, pp. 20–23.

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

[10] B. Merialdo, "Phonetic recognition using hidden markov models and maximum mutualinformation training," in *Proc. IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 1988, pp. 111–114.

[11] 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.

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

[13] 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.