# DISCRIMINATIVE TRAINING OF HIDDEN MARKOV MODELS USING A CLASSIFICATION MEASURE CRITERION

*C. Chesta ⋆ A. Girardi ⋆ P. Laface ⋆ and M. Nigra ⋄*

⋆ Dipartimento di Automatica e Informatica - Politecnico di Torino
Corso Duca degli Abruzzi 24 - I-10129 Torino, Italy
E-Mail chesta|girardi|laface@polito.it
⋄ CSELT - Centro Studi e Laboratori Telecomunicazioni
Via G. Reiss Romoli 274 - I-10148 Torino, Italy
E-Mail mario.nigra@cselt.it

## ABSTRACT

This paper proposes the optimization of a non standard objective function in the framework of Maximum Mutual Information Estimation (MMIE).
In contrast with the classical MMIE estimation, where only misrecognized training utterances contribute to the optimization process, the contributions of near-miss classifications are naturally embedded in the maximization of the proposed function because it takes into account a non linear combination of the probabilities of the competing models that can be tuned by means of a single parameter.
This corrective training procedure has been applied to an Isolated Word Recognition task leading to significant performance improvements with respect to Maximum Likelihood Estimation and MMIE.

## 1. INTRODUCTION

Maximum Mutual Information Estimation is a popular discriminative training technique for HMM parameter estimation [1] that has shown to be a very good complement to the conventional Maximum Likelihood Estimation for reducing the error rate both in small and in large vocabulary domains [5, 6].
The practical application of MMIE is Corrective MMIE [5]. It uses as its reestimation set only the misrecognized utterances because, as will be detailed in Section 2, the contribution to the adjustment of the HMM parameters given by the correct utterances is negligible. Since the bootstrap models for MMI estimation are accurate, MLE trained models, the recognition errors on the training utterances are very few. Thus a small amount of the available training data remains at disposal as a reestimation set. Moreover, the size of this set decreases at every iteration of the algorithm. This is a major drawback for this technique that is partially faced by smoothing the new models with those from the previous iteration [5].
Corrective training [2], despite its name, is an alternative to Corrective MMIE, whose goal is the minimization of the error rate on the training set. This technique, as well as other recently introduced Minimum Classification Error training methods (MCE), based on Generalized Probabilistic Descent (GPD) [4], is able to adjust the model parameters in order to reduce not only the number of errors, but also near-misses in classifying the training utterances. A training technique taking into account near-misses in addition to misrecognized utterances is appealing because

- it makes better use of the training data since the reestimation set size increases,
- it reduces the risk of introducing new, undetected, errors adjusting the model parameters,
- it tries to increase the separation between correct and incorrect models increasing the robustness of the models.

In this paper we propose the optimization of a non standard objective function in the framework of Maximum Mutual Information Estimation that allows not only to take into account the contributions of near-miss classifications in the maximization process, but also to weight the competing models by tuning a single parameter.
In the next Section the MMIE formulation is recalled to point out its above mentioned drawbacks. In Section 3 our rational objective function, referred to as the $\eta$-criterion is proposed. It is similar to the classical MMIE formulation but it has also a correlation with the MCE GPD formulation, allowing all the competing models to be taken into consideration. A frame-dependent adjustment of the models parameters is motivated and illustrated in Section 4. Finally, the database and the results of this corrective training procedure

applied to an isolated word recognition task are reported in Section 5.
In the following the discussed formulation and the reported experimentation refer to the case of isolated word recognition only, but the presented techniques can be extended to connected word or continuous speech recognition considering the competing candidates included in the N-best list of sentence hypotheses.

## 2. MMIE TRAINING

The MMI objective function is defined by the sum over the logarithms of the a posteriori probability of each training utterance. When isolated word models are trained in a system with a vocabulary of $V$ words having the same a priori probability, the MMI criterion leads to the maximization of the following objective function:

$$R_{MMI}(\Lambda) = \sum_{v=1}^{V} \sum_{r=1}^{R_v} \log \frac{P(O_r^v|\lambda_v)}{\sum_{w=1}^{W} P(O_r^v|\lambda_w)} \qquad (1)$$

where $R_v$ is the number of training utterances of word $v$, and $\Lambda$ is the set of the models.
According to [5] and using the notation introduced in [7], the reestimation formula for the mean parameters of Gaussian mixture component $k$ at state $j$ of word model $\lambda_v$ is given by

$$\widetilde{\mu}_{jk}^v = \frac{\Gamma_{jk}^v(x) + D_{jk} \cdot \mu_{jk}}{\Gamma_{jk}^v(1) + D_{jk}} \qquad (2)$$

and the discriminative average $\Gamma_{jk}^v(g(x))$ for isolated word models without tied states is defined by

$$\Gamma_{jk}^v(g(x)) = \sum_{v'=1}^{V} \sum_{r=1}^{R_v} \sum_{t=1}^{T_r} (\gamma_{r,t}^v(j,k;O_r) \cdot \delta(v',v) - \gamma_{r,t}^v(j,k;O_r) \cdot \frac{P(O_r|\lambda_v)}{\sum_{w=1}^{W} P(O_r|\lambda_w)}) \cdot g(x) \qquad (3)$$

where $\gamma_{r,t}^v(j,k;O_r)$ is the probability of occupying the $m-th$ mixture component of state $j$, and $\delta(\cdot,\cdot)$ denotes the Kronecker delta.
Similar formulae can be derived for the reestimation of the mixture weights and variance parameters [5].
It is worth noting that since the dynamic range of the a posteriori probabilities is generally very large, for correctly recognized utterances the factor

$$\widetilde{P}_{MMIE} = P(O_r|\lambda_v) / \sum_{w=1}^{W} P(O_r|\lambda_w) \qquad (4)$$

will be very close to 1 for the correct model and to 0 for the incorrect ones. As a consequence, correctly recognized utterances will not contribute to the reestimation of the HMM parameters because they produce similar positive and negative contributions to the discriminative average counts (3). Thus, the reestimation set is limited only to the misrecognized utterances, a small fraction of the available training set.
As an example, the training set for the experiments reported in Section 5 consists of 102983 utterances of a vocabulary of 68 words. Only 649 of these utterances are misrecognized by using models with a mixture of 4 Gaussians per state, while there are 7990 near-misses according to a preset threshold of the distance between the log probabilities of the correct and of second best candidate models. Unfortunately, the contribution of near-misses to the reestimation is close to null because $\widetilde{P}_{MMIE}$ is close to 1 for most of them.

## 3. THE $\eta$-CRITERION

To account for near-miss classifications in the adjustment of the parameters, and to weight the contribution of competing words, we propose to maximize the objective function defined by the sum over a non standard *classification measure* of each training utterance:

$$R_\eta(\Lambda) = \sum_{v=1}^{V} \sum_{r=1}^{R_v} \log \frac{P(O_r^v|\lambda_v)^{\eta(r)}}{\sum_{w=1}^{W} P(O_r^v|\lambda_w)^{\eta(r)}} \qquad (5)$$

This classification measure is related to a smoothed count of the recognition errors and of the near-miss classifications due to the non linear combination of the probabilities of the competing models.
The reestimation of the mean parameters of Gaussian mixture component $k$ at state $j$ is still obtained through equation (2), while it is easy to show, deriving $R_\eta(\Lambda)$ with respect to the emission probability $b_{jk}^v(t)$, that the new formulation of the discriminative average $\Gamma_{jk}^v(g(x))$ becomes:

$$\Gamma_{jk}^v(g(x)) = \sum_{v'=1}^{V} \sum_{r=1}^{R_v} \eta(r) \sum_{t=1}^{T_r} \Big( \gamma_{r,t}^v(j,k;O_r) \cdot \delta(v',v) - \gamma_{r,t}^v(j,k;O_r) \cdot \frac{P(O_r|\lambda_v)^{\eta(r)}}{\sum_{w=1}^{W} P(O_r|\lambda_w)^{\eta(r)}} \Big) \cdot g(x) \qquad (6)$$

where the corrective factor $\widetilde{P}_{\eta(r)}$ is similar to $\widetilde{P}_{MMIE}$ but with the probabilities $P(O_r|\lambda_v)$ raised to power $\eta(r)$.
Rather than using a constant value for $\eta$, as often suggested in the framework of MCE GPD training, we use instead a value $\eta(r)$ depending on the duration of the utterance. In

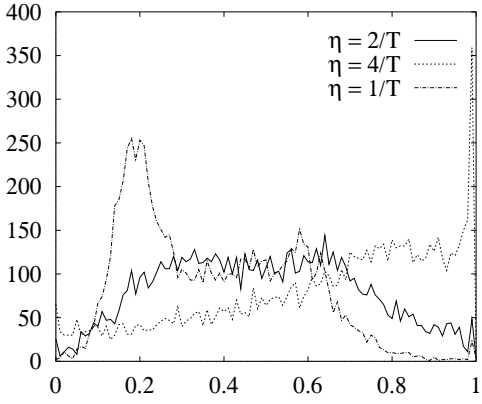


Figure 1: Distribution of the values of $\widetilde{P}_{\eta(r)}$

particular, defining $\eta(r) \equiv 1/T(r)$, where $T(r)$ is the number of frames of utterance $r$, is equivalent to maximizing an objective function similar to (1), where the log likelihood $log(P(O_r|\lambda_v))$ is replaced by the average log likelihood per frame $log(P(O_r|\lambda_v))/T(r)$.

By varying a single parameter in the definition of $\eta(r)$ it is possible to take all the competing words into consideration with greater or less extent as shown in Figure 1. The figure shows the distribution of the values of the classification measure $\widetilde{P}_{\eta(r)}$, obtained using the bootstrap models on the same reestimation set of 7990 confusable training utterances considered in the previous Section. These histograms refer to $\eta(r) \equiv 1/T(r), 2/T(r)$, and $4/T(r)$ respectively. Notice that the distribution corresponding to $\eta \equiv 2/T(r)$ is more uniform than the other ones, thus almost all the utterances in the reestimation set will give a weighted contribution to the adjustment of the parameters. In all the experiments presented in this paper $\eta(r) \equiv 2/T(r)$ has been used because the histogram of $\widetilde{P}_{4/T(r)}$, like $\widetilde{P}_{MMIE}$, presents a remarkable peak for values close to 1, and this effect is, of course, more evident after a few retraining iterations. On the other hand, the distribution of $\widetilde{P}_{1/T(r))}$ is unbalanced toward the lower values of the classification measure, weighting too much the competing words. It is possible to use an adaptive $\eta(r)$ after each retraining iteration, that counteracts the natural progressive increase of the number of utterances with confidence measure close to 1.

## 4. FRAME-DEPENDENT WEIGHTING

Looking at the discriminative average formula (6), it can be noticed that the same corrective factor $\widetilde{P}_{\eta(r)}$ is used for every frame of a given utterance. We argue, however, that each frame adds a different contribution to the log probability $log(P(O_r|\lambda_v))$, and as a consequence, to the final correct/incorrect classification. We propose, therefore, to weight differently the contribution of each frame to the discrimination average, multiplying $\widetilde{P}_{\eta(r)}$ by a frame-dependent weight $fdw_t$. To compute these frame-dependent corrective factors, rather than relying on linguistic information that are not readily available, we use the following strategy based on acoustic information only. For the correct model $\lambda_v$, we want to adjust its parameters giving more weight to "bad" frames - those contributing less than the average to $log(P(O_r|\lambda_v))$-. On the other hand, we can limit the amount of adjustment for frames that are already "good", i.e. that are well recognized by the current model. The opposite is true for an incorrect model $\lambda_w$. A frame contributing less than the average to $log(P(O_r|\lambda_w))$ can be taken into account with a small weight because we agree that it must not be well recognized by the current competing model.

It is particularly easy to obtain the contribution of frame $O_{t,r}$ to the log probability $log(P(O_r|\lambda_v))$ because it is exactly the logarithm of the so called *scaling factor* computed during the forward trellis iterations $c_t = \sum_s \widetilde{\alpha}_t(s)$, where $s$ are the states of model $\lambda_v$ and $\widetilde{\alpha}_t(s)$ are the rescaled forward probabilities.

Thus, the frame-dependent weights for the correct model is computed using the sigmoid function

$$fdw_t = \frac{1}{2} + \frac{1}{1 + \mathrm{e}^{-\beta \cdot (logc_t - log(P(O_r)|\lambda_v)/T(r))}} \qquad (7)$$

while the following (symmetric) sigmoid function is used for the incorrect models

$$fdw_t = \frac{3}{2} - \frac{1}{1 + \mathrm{e}^{-\beta \cdot (logc_t - log(P(O_r)|\lambda_v)/T(r))}} \qquad (8)$$

the sigmoid slope $\beta$ can be tuned to select a suitable extent for the corrective factors.

## 5. RESULTS

To test this approach we trained whole word HMMs for an isolated word recognition task with a vocabulary of 68 words consisting of digits, credit-card names, commands, positive and negative answers, and 26 Italian spelling names. The training set includes 102983 telephone line utterances of 1488 speakers, while testing is performed on a separate set of 38196 utterances of 512 different speakers.

The results of the experiments are summarized in Table 1. The first row reports the baseline results using a set of models obtained after 5 segmental k-means and 3 MLE iterations, respectively. These models have been used as a bootstrap for the discriminative training iterations.

The value of the iteration constant $D^v_{jk}$ in the reestimation formula (2) has been set according to

$$D^v_{jk} = \sum_{r=1}^{R_v} \eta(r) \sum_{t=1}^{T_r} \gamma^v_{r,t}(j,k;O_r) \qquad (9)$$

Table 1: Results comparing the baseline system and different reestimation approaches

| Num. of Gaussians | 1 | 2 | 4 | 16 |
|---|---|---|---|---|
| MLE | 973 (97.4%) | 757 (98%) | 649 (98.3%) | 581 (98.5%) |
| MMIE | 666 (98.2%) | 637 (98.3%) | n.a | n.a |
| $\eta$-Criterion | 556 (98.5%) | 535 (98.6%) | 479 (98.7%) | 472 (98.8%) |
| Frame-dependent | 539 (98.6%) | 520 (98.6%) | 467 (98.8%) | 463 (98.8%) |
| Reestimation set size | 14242 | 10038 | 7990 | 8042 |

here $\sum_{t=1}^{T_r} \gamma_{r,t}^{v}(j,k;O_r)$ is the occupation count of mixture $k$ at state $j$ computed during the Forward-Backward training for obtaining the MLE models. Using these settings we achieved a relatively fast convergence of the algorithms (8-14 iterations to obtain the reported results) that did not produce negative Gaussian weights, except in a few cases for models with 16 Gaussians per state. This problem was solved, as usual, increasing the value of $D$.

Variances were not reestimated because preliminary experiments showed that their value almost always decreases at the end of the reestimation process reducing the generalization capabilities of the models.

The first two experiments were performed using MMIE models with 1 and 2 Gaussians per state, their results are given in the second row of the table confirming that discriminative training is indeed a powerful approach for reducing the error rate in this small vocabulary recognition task. In parallel, we tested the corresponding models trained by means of the $\eta$-criterion with $\eta(r) \equiv 2/T(r)$. Since the obtained improvements with respect to the classical MMIE were significant with respect to a 95% confidence interval, we did not train any further MMIE model with a different number of Gaussians per state.

The results of the third rows confirm the importance of using a nonlinear combination of the competing word probabilities as a classification measure rather than simply relying on the training set errors. The relative improvements are, of course, reduced increasing the number of parameters per state. This is evident comparing the $\eta$-criterion results with 4 and 16 Gaussians per state respectively. One of the reasons for this effect is that recognizing with more precise models the utterances in the training set obviously decreases the number of errors and near-misses, reducing the size of the reestimation set as shown in the last rows of Table 1.

Finally, constant, but marginal improvements are obtained using the frame-dependent weights.

## 6. CONCLUSIONS

We have developed a corrective training technique that has been experimented using as a testbed an isolated word recognition task. We have argued that, since it accounts for errors and near-misses in the training set, it has the potential for outperforming the classical MMIE training approach. The results of our experiments confirm our findings giving a 16% relative improvement over MMIE models and from 18% to 42% with respect to MLE models. It is worth noting that discriminatively trained models with a single Gaussian per state give better results than MLE trained models even with 16 Gaussian per state.

Work is in progress to adapt the $\eta$ parameter at each iteration.

## 7. REFERENCES


[1] Bahl L.R., Brown P.F., de Souza P.V., Mercer R.L., "Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition", Proc. ICASSP-86, pp. 49–52, 1986.

[2] Bahl L.R., Brown P.F., de Souza P.V., Mercer R.L., "A New Algorithm for the Estimation of Hidden Markov Model Parameters" Proc. ICASSP-88, pp. 493–496, 1998.

[3] Gopalakrishnan P.S., Kanewsky D., Nadas A., Nahamoo D., "An Inequality for Rational Functions with Applications to Some Statistical Estimation Problems", IEEE Transaction on Information Theory, Vol. 37, N. 1, pp 107–113, 1991.

[4] Juang B.-H., Chou W., C.-H. Lee, "Statistical and Discriminative Methods for Speech Recognition" In "Automatic Speech and Speaker Recognition", C.-H. Lee, F.K. Soong, K.K. Paliwal eds., pp. 109–132, Kluwer Academic Publisher, 1996.

[5] Normandin Y., Lacouture R., Cardin R., "MMIE Training for Large Vocabulary Continuous speech Recognition", Proc. ICSLP-94, pp. 1367–1370, 1994.

[6] Normandin Y., "Maximum Mutual Information Estimation of Hidden Markov Models", In "Automatic Speech and Speaker Recognition", C.-H. Lee, F.K. Soong, K.K. Paliwal eds., pp. 57–81, Kluwer Academic Publisher, 1996.

[7] Schlüter R., Macherey W., Kantak S., Ney H., Welling L., "Comparison of Optimization Methods for Discriminative Training Criteria", Proceedings of EUROSPEECH'97, Rhodes, pp. 15–18, 1997.