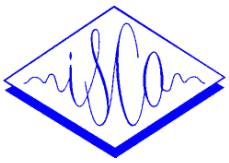

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

DISCRIMINANT LEARNING WITH MINIMUM MEMORY LOSS FOR IMPROVED NON-VOCABULARY REJECTION

Hugues Leprieur and Patrick Haffner
France Télécom, Centre National d'Etudes des Télécommunications
CNET/LAA/TSS/RCP, Technopole Anticipa, 2 avenue Pierre Marzin, 22307 LANNION, FRANCE
email: leprieur,haffner@lannion.cnet.fr

ABSTRACT

A limitation to current HMM-based Speech Recognition approaches lies in the modeling of non-vocabulary utterances. Improved rejection is a key research direction in Interactive Voice Response Services (IVR), where field evaluations show that many users do not only utter the requested keywords. This paper compares several discriminant training criteria on this problem and applies a novel optimization technique which can be used to improve rejection, without seriously disturbing HMM modeling assumptions. A 23% reduction in the error rate is observed on *field* data recorded during the operation of an IVR service.

1. INTRODUCTION

The quality of Interactive Voice Response Systems greatly relies on their ability to reject out-of-vocabulary words and to spot words in continuous utterances. Many efforts have been made to improve the performances of Automatic Speech Recognition Systems. When trained with Maximum Likelihood Estimation (MLE) techniques, HMM wordspotting systems increase their performance when an accurate modeling of non-vocabulary words is done [7].

However, two limitations appear when using MLE to train wordspotting and rejection systems. Firstly, an MLE training algorithm relies on correct data generating models, but we do not have proper acoustic models for garbage models. Secondly, when dealing with insertion/rejection problems, the most important parameter to fine-tune is an acceptation threshold, which cannot be easily fine-tuned with MLE techniques.

Gradient based discriminant training techniques can overcome those problems. They handle incorrect modeling assumptions [1]. Recently, these techniques have contributed significant improvements when used to fine-tune systems already trained to maximize the emission likelihood. In [3], on a Spanish digit recognition system, the HMM parameters were optimized by minimizing a cost function which is the weighted sum of the substitution, deletion and false alarm probability estimates. In [2], a new discriminant scheme was used to fine-tune a hybrid HMM-radial basis function system in order to directly maximize the Figure Of Merit (FOM) on Switchboard.

This paper addresses two problems: what kind of cost or error function should be minimized on the training data in order to minimize the end-user error of the IVRS? What is the best optimization technique available in order to minimize this function? In Section 2 different cost functions are examined, in Section 3 a new optimization technique is proposed in order to minimize these functions, and their performances are compared in Section 4.

2. DISCRIMINANT TRAINING CRITERIA

Whereas the Baum-Welch and Viterbi algorithms for ML estimation try to maximize the emission probability of each observation in the training corpus, discriminant training tries to maximize the recognition rate on the training data explicitly. This section examines four different "reward" functions $F(\Theta)$ of the set of parameters Θ. These functions must reach a global maximum when the wordspotting error rate is 0, and must be trainable with an algorithm based on gradient descent. The corresponding cost function which must be minimized is $E(\Theta) = -F(\Theta)$.

2.1. Sum Of Correct Class Probabilities

Given that the goal of any classification system is to minimize the classification error probability, it seems interesting to optimize the parameters of the system specifically for that criterion. The error is binary in the sense that there is or is not an error, but a differentiable criterion is needed in order to obtain a calculable gradient. The first option that comes to mind is using the sum of correct class *a posteriori* probabilities as a reward function:

$$F(\Theta) = \sum_{n=1}^{N} P_\Theta(K_n|X_n) \quad (1)$$

where K_n is the correct class, i.e. the acoustic word sequence in the utterance X_n, N is the number of utterances in the training corpus, and $P_\Theta(.)$ represents the probability estimated by the system.

2.2. Maximum Mutual Information

Another possibility is minimizing the cross-entropy between the desired and the estimated *a posteriori* class probability, and setting at 1 the desired *a posteriori* probability of the correct class (and the others at 0), thereby obtaining the maximization of the equation (2), which is equivalent to Maximize Mutual Information (MMI, see [5]).

$$F(\Theta) = \sum_{n=1}^{N} \ln P_\Theta(K_n|X_n) \quad (2)$$

2.3. Corrective Maximum Mutual Information

In order to reduce the computational burden, it is possible to modify the parameters according to equation (2) only when an error occurs. The criterion to maximize becomes:

$$F(\Theta) = \sum_{n=1}^{N} \ln P_\Theta(K_n|X_n) \times \delta_{K_n \neq \arg\max_M (P_\Theta(M|X_n))} \quad (3)$$

2.4. Figure Of Merit training

The Figure Of Merit (FOM) is a criterion commonly used in order to compare different wordspotting systems

(see [2]). The FOM represents the average correct detection rate for the false alarm rates varying from 0 to 10 false alarms-per-hour and per-keyword, and can be defined as follows:

$$FOM = (p_1 + p_2 + ... + p_N + ap_{N+1})/10T \quad (4)$$

where:
p_i : correct detection rate before the i^{th} false alarm,
T : database duration (in hours),
N : first integer $\geq 10T - 1/2$
$a = 10T - N$: an interpolation factor.

The FOM can also be directly maximized as shown in [2].

3. CHOOSING AN OPTIMIZATION TECHNIQUE

One characteristic common to the optimization of any discriminant criteria is the modification of the parameters of the models in order to maximize, more-or-less directly, the ratio:

$$P_\Theta(K_n|X_n) = \frac{P_\Theta(X_n|K_n)P_\Theta(K_n)}{\sum_M P_\Theta(X_n|M)P_\Theta(M)} \quad (5)$$

where M is any keyword or garbage model.

Unlike the MLE criterion, discriminant criteria take the denominator of (5) into account. For this reason, the Baum-Welch or Viterbi re-estimation of the HMM parameters (a version of the Expectation-Maximization algorithm) is not easily applicable, and techniques based on gradient descent are often used.

With our Gaussian density HMMs, the main problem is changing the means and variances which have already been estimated by the ML algorithm, in order to further reduce the number of errors. In this section, an initial standard gradient descent technique is applied for this optimization, and a novel technique, which appears to be more efficient, is proposed.

3.1. Gradient Descent

The general form of the gradient is

$$\frac{\partial E(\Theta)}{\partial \theta} = -\sum_n \sum_M w_n(M) \frac{\partial \ln P_\Theta(X_n|M)}{\partial \theta} \quad (6)$$

where $w_n(M)$ depends on the criterion to be maximized. As an example, this gradient is derived with respect to the mean μ_g of the Gaussian g:

$$\frac{\partial E(\Theta)}{\partial \mu_g} = -\sum_{t,n} \left[\frac{X_n(t) - \mu_g}{\sigma_g^2} \sum_M P_\Theta(g_t|M, X_n) w_n(M) \right] \quad (7)$$

where $X_n(t)$ is the input frame at time t in sample n, and $P_\Theta(g_t|M, X_n)$ is the probability of selecting Gaussian g at time t, given the input X_n and the model M. Standard gradient descent with step size η modifies the mean $\tilde{\mu_g}$ as:

$$\mu_g^* - \tilde{\mu_g} = -\eta \frac{\partial E(\Theta)}{\partial \mu_g} \quad (8)$$

Many variations exist for gradient descent, (with momentum, conjugated, etc.). Most of them attempt to improve the learning speed. However, in this study, the first expectation of this discriminant training procedure is that it be able to modify the MLE trained parameters enough to achieve its corrective goal.

3.2. A Novel Optimization Technique

A clear definition of what makes corrective fine-tuning optimal must be specified. In the context of this study, the goal is to correct certain empirical errors (by minimizing the discriminant cost function), while minimizing the modifications to the system. The HMM is defined by the distribution $P_{\tilde{\Theta}}(X, Q)$, where X is a sequence of acoustic frames, and Q the corresponding sequence of hidden states. After modification of the parameters, the distribution $P_{\tilde{\Theta}}(.)$ becomes $P_\Theta(.)$. So, the cross entropy should be minimized (also known as the Kullback-Liebler information divergence) between those two distributions

$$H\left(\Theta \left\| \tilde{\Theta}\right.\right) = \int_x P_\Theta(x) \ln \frac{P_\Theta(x)}{P_{\tilde{\Theta}}(x)} dx \quad (9)$$

The two functions were combined in order to optimize the framework of the information theory. Suppose one wants to Minimize the Description Length (noted as MDL[6]), which is the number of bits of information which are required in order to code:

1. The distribution $P_\Theta(.)$, given the a priori $P_{\tilde{\Theta}}(.)$: the number of information bits is $H\left(\Theta \left\| \tilde{\Theta}\right.\right)$.
2. The recognition errors when using the distribution $P_\Theta(.)$ instead of the true distribution: $E(\Theta) = -F(\Theta)$ as derived from the MMI criterion (eq.(2)).

The sum of these two terms has to be weighted[1]

$$L(\Theta) = H\left(\Theta \left\| \tilde{\Theta}\right.\right) + \lambda E(\Theta) \quad (10)$$

One can consider that $\lambda > 0$ is small, so the additional $\lambda E(\Theta)$ term does not decisively alter the global minimum of $H\left(\Theta \left\| \tilde{\Theta}\right.\right)$, $\tilde{\Theta}$. The new minimum of $L(\Theta)$, Θ^* is obtained in a single step, with the Newton interpolation method:

$$\theta^* - \tilde{\theta} = \frac{-\left.\frac{\partial L(\Theta)}{\partial \theta}\right|_{\tilde{\Theta}}}{\left.\frac{\partial^2 L(\Theta)}{\partial \theta^2}\right|_{\tilde{\Theta}}} \begin{cases} \tilde{\Theta} \text{ is a minimum} : \left.\frac{\partial H\left(\Theta \left\| \tilde{\Theta}\right.\right)}{\partial \theta}\right|_{\tilde{\Theta}} = 0. \\ \lambda \text{ small} : \left.\frac{\partial^2 L(\Theta)}{\partial \theta^2}\right|_{\tilde{\Theta}} \simeq \left.\frac{\partial^2 H\left(\Theta \left\| \tilde{\Theta}\right.\right)}{\partial \theta^2}\right|_{\tilde{\Theta}} \end{cases}$$

So that

$$\theta^* - \tilde{\theta} = -\frac{\lambda}{\left.\frac{\partial^2 H\left(\Theta \left\| \tilde{\Theta}\right.\right)}{\partial \theta^2}\right|_{\tilde{\Theta}}} \left.\frac{\partial E(\Theta)}{\partial \theta}\right|_{\tilde{\Theta}} \quad (11)$$

This parameter modification is a variation on gradient descent with a local learning rate which takes into account the impact of this modification over previously learned samples. This modification has certain similarities with the traditional Newton algorithm. However, the present objective is to minimize, at each training iteration, the memory loss, rather than maximizing the training speed. In this paper, the training procedure reiterates eq.(11) several times: at epoch t, the posterior parameters Θ_{t-1}^* of the previous epoch are taken as the new prior parameters $\tilde{\Theta}_t$. This optimization technique was applied even when the information theoretic formalism is not applicable to justify eq.(10), for instance in the case of the sum of probabilities (eq.(1)) or the FOM (eq.(4)).

In the case of HMMs with Gaussian densities, $\frac{\partial^2 H\left(\Theta \left\| \tilde{\Theta}\right.\right)}{\partial \theta^2}$ is very simple to compute for both the means and the variances. For instance, in the case of the mean μ_g of the

[1] In the ideal case where we could express this minimization as a constraint ($E(\Theta) > target$) (which must be linear), we could consider λ as a Lagrange Multiplier [8].

Gaussian g:

$$\frac{\partial^2 H\left(\Theta \,\|\, \tilde{\Theta}\right)}{\partial \mu_g^2} = \frac{P_\Theta(g)}{\sigma_g^2} \tag{12}$$

where $P_\Theta(g)$ is the overall probability to select g. By developing eq.(8) and eq.(12) into eq.(11),

$$\mu_g^* - \tilde{\mu}_g = \frac{\lambda}{P_\Theta(g)} \sum_{t,n} (x(t) - \mu_g) \sum_M P_\Theta(g_t|M, X_n) w_n(M) \tag{13}$$

is obtained. Using the same approach, a reestimate formula for the covariances is derived. In the rest of the paper, this new optimization technique will be denoted as Minimum Memory Loss (MML).

4. EXPERIMENTS

4.1. The "Les Baladins" Service

This 26-keyword IVR Service has provided information about the cinema programs around Lannion, since 1988. This service uses PHIL90, the CNET ASR technology, in order to propose a menu-based dialogue where the user interacts with the server by uttering isolated words. PHIL90 also provides incorrect utterance rejection and wordspotting procedures.

4.2. Database

In this study, the training and recognition experiments were performed on utterances collected over the long distance telephone network. These utterances were automatically segmented by a speech/noise system, and hand-labeled as correct when they contained an isolated keyword. The utterances were labelled as incorrect when they contained noise or out-of-vocabulary words. About half of the utterances were recorded under laboratory conditions (*lab* data), the other half were recorded during the actual operation of the service (*field* data).

4.3. Baseline HMM training

The HMM used in this study is constituted of 4 different garbage models and an allophonic network describing the 26 keywords (see [4]). The language of the application is described using a null grammar, which allows the system to detect a maximum of one keyword-per-utterance. The garbage models are standard left to right models with 30 states and 30 mono-Gaussian pdfs linked to the transitions. The acoustic input frames are computed every 16ms, and consist of 8 Mel Frequency Cepstral Coefficients and the Energy, as well as their first and second derivatives. All the models were trained using the Viterbi algorithm on both *lab* and *field* data. The keyword models were trained on correct utterances of keywords from both *lab* and *field* data. The garbage models were trained on correct utterances of out-of-vocabulary words from *lab* data and on incorrect or noisy utterances from *field* data as shown in the first line of table 1.

4.4. Discriminant training

Discriminant learning is implemented as an additional fine-tuning procedure for the CNET HMM-based system, PHIL90, which was first trained with MLE. The learning rate is set to 5.10^{-4} and the parameters are updated in a batch mode[2], after each epoch (a presentation of the whole training set). All the results presented in this article are shown after the fifth epoch, because a partial cross-validation[3] showed that it was a reasonable choice. As shown in table 1, the discriminant training set B[4] consist of the correct utterances from the *lab* data and all the ut terances from the *field* data.

Table 1: "Les Baladins" corpus composition

	# utterances	
	Keywords	Out-of-vocab.
Baseline Train. A	12013 *field* + 9918 *lab*	3279 *field* + 12000 *lab*
Discrim. Train. B	12013 *field* + 9918 *lab*	3279 *field*
Test (*field*)	12083	3309
Test (*lab*)	9968	0

4.5. Measuring the error rates

To evaluate the performance, we use the following rates the keywords substitution rate (*sub*), the keywords deletion rate (*del*), the false alarm rate (FA), and the tota error rate (Total), with:

$$sub = \frac{N_{ks}}{N_k} \; ; \; del = \frac{N_{kd}}{N_k} \; ; \; FA = \frac{N_{FA}}{N_i}$$

with the following counters:
N_k : correct keyword utterances in the database,
N_i : incorrect utterances (containing no keyword).
N_{ks}: keyword substitutions,
N_{kd}: keyword deletions,
N_{FA}: false alarms.

In order to compare the performances in terms of error, we note: $\int E = 100 - FOM$.

The following error rates are given in percent. However, the % sign will only be used to give the percentage of change when comparing the new error rate to the baseline results (table 2).

Table 2: Baseline MLE training

	sub	*del*	FA	Total	$\int E$
Train. (*lab + field*)	1.1	2.1	21.3	5.5 ± 0.3	10.4
Test (*lab + field*)	1.1	2.6	23.2	6.2 ± 0.3	-
Test (*field*)	1.6	4.1	23.2	9.5 ± 0.5	11.8

The confidence intervals mentioned in table 2 are given at 95% on both test and training.

Since we intend to improve the performance of the IVRS, all the following results presented below are obtained on *field* data, but very similar relative improvements are obtained on *lab* data.

4.6. Influence of the optimization technique

In (table 3), we compare the different optimization techniques (section 3) to the corrective MMI training criterion, as given by eq. (3). We tested this criterion first because its computational burden was the lightest.

[2]More frequent updates would make learning faster, but their implementation within a HMM software which is basically designed for batch parameter reestimation is not easy

[3]We currently use a cross-validation set to stop learning (when the error is minimum on this set). However, this set was not available for some of the experiments quoted in this paper

[4]We removed from the baseline training set A all the out-of-vocabulary *lab* data. Besides making training significantly faster, another reason that motivated this choice is that training set B has now the same distribution as the test set. Note also that, in the experiments, we found ML training to give comparable results with set B than with set A, while being significantly faster.

Table 3 : Comparison of different optimization techniques on the MMI corrective criterion

	sub	*del*	*FA*	Total	$\int E$
Gradient(μ)	1.7 +5 %	3.4 -16 %	25.2 +9 %	9.5 0 %	11.9 +1 %
MML(μ)	1.4 -14 %	4.0 -13 %	18.9 -19 %	8.0 -16 %	11.0 -6 %
MML($\mu + \sigma$)	1.1 -32 %	4.0 -4 %	15.6 -33 %	7.3 -23 %	10.0 -15 %

1. The first line in Table 3 shows that no gain is obtained by using the standard gradient descent applied to the Gaussian means (eq. (8)). Only a redistribution between FA and *del* is observed; this could have been obtained by simply adding a penalty to the garbage models.
2. The second line in Table 3 shows that a significant error reduction (16%) can be obtained on the test data by using MML based estimation of the Gaussian means (eq. (13)). This gain is evenly distributed over the different error types (*sub*, *del*, FA): this would not have been possible with the mere fine-tuning of some word entrance penalties.
3. The third line in Table 3 shows that if, in addition, MML based estimation is also applied to the Gaussian variances, the reduction in the error rate is even larger: 23%.

This subsection clearly shows that MML training gives recognition performances which are significantly better than standard gradient descent[5].

4.7. Influence of the training criterion

Since MML based estimation gave the best results, it was chosen to optimize the different training criteria proposed in section 2. In this section, MML based estimation is applied to the Gaussian means only.

Table 4 : MML optimization of the mean according to various criteria

	sub	*del*	*FA*	Total	$\int E$
Full MMI	1.4 -14 %	3.5 -14 %	18.9 -18 %	7.9 -16 %	10.9 -7 %
Sum. of Cor.	1.7 +5 %	3.5 -14 %	25.2 +9 %	9.6 +1 %	12.0 % +2 %
FOM	1.2 % -27 %	5.8 +40 %	15.4 -34 %	8.8 -8 %	9.9 -16 %

1. The first line in Table 4 shows that a 16% error reduction can be obtained on the test data with the MMI criterion (eq. (2)). This is similar to what was obtained with corrective training only in the second line in Table 3. With MMI training, it should be possible to skip the correctly recognized samples without significant loss in performance.
2. The second line in Table 4 shows that no gain over the baseline HMM is obtained after maximization of the sum of correct probabilities given by eq. (1).
3. The third line in Table 4 shows that a significant gain (16%) in the Figure of Merit ($FOM = 100 - \int E$) can be obtained on the test data after maximization of eq. (4). Explicitly maximizing the FOM, leads to make fewer FA but more deletions, which is normal as FOM is more concerned with low FA rates.

[5]Standard gradient descent was tested on several configurations (learning rate, number of epochs, learning of the variances) and never showed any significant improvement. We did not experiment more sophisticated versions of gradient descent (for instance conjugate), but they are supposed to improve learning speed, not recognition performance.

5. CONCLUSION

This paper explores different solutions used in order to apply discriminant fine-tuning to ML trained HMMs.

Firstly, this paper shows that it is important to choose a discriminant technique which minimizes modifications over the data-generating distribution of the HMM (what was optimized with MLE). Whereas standard gradient descent appears inefficient in improving wordspotting performance of an IVR service, Minimum Memory Loss (MML) training yields a 23% reduction in the error rate.

Secondly, as a cost function that must be optimized (herein, maximized) with the aforementioned technique, MMI appears to give the best results. However, many other cost functions remain to be tested (for instance, Minimum Classification Error, MCE, as in [5], [3]).

REFERENCES

[1] P.F. Brown. *The Acoustic-Modelling Problem in Speech Recognition.* PhD thesis, Carnegie Mellon University, Pittsburgh, 1987.

[2] E.I. Chang and R.P. Lippmann. Figure of merit training for detection and spotting. In *Advances in Neural Information Processing Systems 6*, pages 1019–1026, Denver, CO, 1994. Morgan Kaufmann, San Mateo.

[3] C. de la Torre and A. Acero. Discriminative training of garbage model for non-vocabulary utterance rejection. In *1994 International Conference on Spoken Language Processing*, pages 475–478, Yokohama, September 1994.

[4] D. Jouvet, K. Bartkova, and J. Monné. On the modelization of allophones in an HMM based speech recognition system. In *EUROSPEECH'91, 2nd European Conference on Speech Communication and Technology*, Genova, Italie, September 1991.

[5] H. Ney. On the probabilistic interpretation of neural network classifiers and discriminative training criteria. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(2):107–119, February 1995.

[6] J. Rissanen. Stochastic complexity and modeling. *Annals of Statistics*, 14:1080–1100, 1986.

[7] J.R. Rohlicek, P Jeanrenaud, K Ng, H Gish, B Musicus, and M Siu. Phonetic training and language modelling for word spotting. In *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume II, pages 459–462, Minneapolis, April 1993.

[8] J.E Shore and R.W. Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory*, IT-26(1), January 1980.