Maximum Entropy Discrimination (MED) Feature Subset Selection for Speech Recognition Technical Report RR-03-086, Institut Eurecom

Fabio Valente, Christian Wellekens

October 24, 2003

Contents

Abstract

In this paper we investigate the application of *Maximum Entropy Discrimination* (MED) feature selection in speech recognition problems. We compare the MED algorithm with a classical wrapper feature selection algorithm and we propose an hybrid wrapper/MED algorithm. We experiment the three approaches on a phoneme recognition task on the TIMIT database. Results show that the MED algorithm achieves error rates comparable with the wrapper algorithm requiring a reduced computational charge. Furthmore the use of a probabilistic framework shows that the MED algorithm holds very good results even with very limited amount of data.

Chapter 1

Introduction

Speech recognition systems significantly increase their performance if several feature streams are used but on the other side computational charges increase as well; here comes the need for selecting the most significant features. Feature streams combination can be done at different levels in the system; they can be combined together after the feature calculation, after the model probability calculation or after the decoding. The case we consider in this paper is the feature combination after their calculation. A huge number of front-end techniques have been proposed with a lot of redundancy between them (MFCC, PLP, LPC, MSG, articulatory features,etc.) and many techniques for reducing this high dimensional space eliminating redundancy between features have been proposed ([8],[9],[10]). Currently speech recognition systems use discriminative feature transformation like LDA or HDA but it does not exclude a preventive feature sort for eliminating extra information. A framework for defining the theoretically optimal method for feature subset selection is presented in [5], but this approach is computationally intractable. Tractable algorithms for feature subset selection belongs to two main classes: wrappers and filters. Wrapper based algorithms are very precise but need large computational resources (see [4]): they consist in a greedy selection of best feature subset based on the calculation of an efficiency criterion. Currently used criteria are based on mutual information or classification error (see [6]). Wrapper based method usually holds very good results but they need an important quantity of data and high processing resources. On the other side feature selection algorithms based on filters are simpler and with a reduced complexity but generally they are not strongly related with the problem to solve and they are based on simpler criteria. In this paper we study the application of a feature weighting algorithm to determine the best feature subset, based on the Maximum Entropy Discrimination (MED) approach (see [1],[2]) applied to acoustic features. This is a bayesian discriminative algorithm that associates a probability with each feature: weak features receive low probability values while strong features receive high probability values; recognition is done using feature expected values w.r.t. their final distributions. The application we consider is slightly different from the original formulation of the problem: the original framework aims at weighting features to improve recognition rate, eventually weak features receive zero weight; our goal is to extract the M most important features out of the N features in the set; they may be obviously considered as the M features with the highest probability value, in the same way as in PCA space reduction just the more meaningful features are used. This condition can be incorporated in the optimal prior formulation and leads the process to find the M most essential features. This paper is organized as follows: in section 2 we describe the MED principle, in section 2.2 feature selection algorithms based on MED for gaussian distribution and HMM are described, in section 2.3 an hybrid wrapper/MED feature selection algorithm is presented and finally in section 3 we present experiments on synthetic data and speech data.

Chapter 2

Maximum Entropy Discrimination (MED)

Recently, many methods for joint generative-discriminative learning have been proposed that take advantages from the two approaches. Generative learning fits model parameters to observations, while discriminative learning produces models that can be used for efficient classification. Maximum Entropy Discrimination (introduced in [1], developed in [2],[3]) is an hybrid generative/discriminative approach to model learning. In this section we give an overview of the MED method.

Let's consider a parametric family of decision boundaries $F(X|\Theta)$ with some discrimination properties between two classes that we will call *discrimination function* with Θ parameters set. $F(X|\Theta)$ takes an input vector X and returns a scalar output; the sign ±1 of this output will determine the class to which the input vector will be assigned to. Given a training set $\{X_t\}$ and the corresponding binary labels $\{y_t\}$ with $t \; \epsilon \; [1,T]$, learning parameters means finding the $\hat{\Theta}$ that minimize some kind of classification error. Decision on an unseen input X will be taken using $\hat{y} = sign \; F(X|\hat{\Theta})$. In many classification approach the classification error measure is derived using a *classification margin* i.e. considering how large is the value of $y_t F(X_t|\Theta)$. This classification margin can be expressed in the form $y_t F(X_t|\Theta) - \gamma_t \geq 0$ where γ_t is another variable that represents the margin that $y_t F(X_t|\Theta)$ must satisfy; optimization will now consider both parameters Θ and margins γ_t. So given a certain loss function $Loss()$ non-increasing and convex, we can write the constrained solution for $\hat{\Theta}$ as:

$$
\begin{aligned}
& min_{\Theta,\gamma_t}\{\sum_t Loss(y_t F(X_t|\Theta))\} \\
& \text{subject to } y_t F(X_t|\Theta) - \gamma_t \geq 0 \;\; \forall t
\end{aligned}
\qquad (2.1)
$$

Now let's consider a bayesian framework, in which we try to estimate distributions over parameters i.e. $P(\Theta)$. Classification will be done integrating out parameters w.r.t. their optimal distribution i.e. for classifying an input X, we will use:

$$
\hat{y} = sign \int_{\Theta} P(\Theta) F(X|\Theta) d\Theta \qquad (2.2)
$$

To find the optimal $P(\Theta)$ many solutions are possible. A classical approach consists in minimizing the negative entropy of the distribution (that is equivalent to maximizing the entropy) i.e. finding the *Maximum Entropy* distribution subject to constraints in equation (2.1) i.e.

$$\begin{aligned} & min_{P(\Theta)} \ -H(P(\Theta)) \\ \text{subject to } & \int P(\Theta)[y_t F(X_t|\Theta) - \gamma_t] d\Theta \geq 0 \ \forall t \end{aligned} \tag{2.3}$$

Equation (2.3) formulates what is known as *Maximum Entropy Discrimination* (MED). Anyway negative entropy is not very flexible for optimizing distributions. If we have any prior knowledge about parameters under the form of a prior distribution $P_0(\Theta)$, it is possible to replace in equation (2.3) the negative entropy with the KL distance between $P(\Theta)$ and $P_0(\Theta)$, obtaining the so called *Minimum Relative Entropy Discrimination* (MRED) formulation i.e.

$$\begin{aligned} & min_{P(\Theta),P(\gamma)} \ KL(P(\Theta)||P_0(\Theta)) = \int P(\Theta) \, log(P(\Theta)||P_0(\Theta)) d\Theta \\ \text{subject to } & \int P(\Theta)[y_t F(X_t|\Theta) - \gamma_t] d\Theta \geq 0 \ \forall t \end{aligned} \tag{2.4}$$

Generally in literature the two approaches are referred with the same name (see [1] and [2] for details) and we will adopt this convention even in this paper. It is interesting to notice that objective function is the KL distance of two distributions that is a convex function of the argument $P(\Theta)$ and constraints are linear in $P(\Theta)$, so regardless of nonlinearities of the discriminant function, the optimization problem concerns a convex function (w.r.t. distribution $P(\Theta)$).

Up to this point we have considered fixed margin γ_t; anyway in real data problems perfect separability is almost impossible and constraints may generate an empty search space for optimal distributions. For this reason it is worth introducing a distribution on margin variables too, that gives non-zero probability for negative margins; in this way positive margins will not be penalized and negative margins will be progressively penalized depending on the distribution of the margin variables we have fixed. MED framework offers a very elegant way to combine together model parameters and margin variable parameters; in fact now the parameter set becomes $\{\Theta, \gamma\}$, and assuming the factorization $P(\Theta, \gamma) = P(\Theta) \prod_t P(\gamma_t)$, we can write the *augmented* MED formulation (see [3]) as:

$$\begin{aligned} & min_{P(\Theta)} \ KL(P(\Theta)||P_0(\Theta)) + \sum_t KL(P(\gamma_t)||P_0(\gamma_t) \\ \text{subject to } & \int P(\Theta)[y_t F(X_t|\Theta) - \gamma_t] d\Theta d\gamma_t \geq 0 \ \forall t \end{aligned} \tag{2.5}$$

where $P_0 \gamma_t$ is the prior distribution associated with the margin variables. The solution to (2.5) has the following form (see [7]):

$$P(\Theta, \gamma) = \frac{1}{Z(\lambda)} P_0(\Theta, \gamma) \, e^{\sum_t \lambda_t [y_t F(X_t|\Theta) - \gamma_t]} \tag{2.6}$$

where $Z(\lambda)$ is the normalization constant and $\lambda = \{\lambda_1, ..., \lambda_T\}$ defines a set of non-negative Lagrange multipliers. λ is found maximizing the jointly objective function:

$$J(\lambda) = -log \, Z(\lambda) \tag{2.7}$$

Given a closed form for $Z(\lambda)$, the maximum of the jointly concave objective function $J(\lambda)$ can be found using any standard convex optimization method.

In [1] the following lemma is demonstrated: any factorization of priors $P_0(\Theta, \gamma)$ across a disjoint set of variables $\{\Theta, \gamma\}$ leads to a disjoint factorization of the MED solution $P(\Theta, \gamma)$ across the same sets of variables provided that these variables appear in distinct additive components in $y_t F(X_t|\Theta) - \gamma_t$;

if we assume the following prior factorization $P_0(\Theta,\gamma) = P_0(\Theta)P_0(\gamma)$ and $P_0(\gamma) = \prod_t P_0(\gamma_t)$, the solution will be of the form $P(\Theta)\prod_t P(\gamma_t)$. As consequence the $J(\lambda)$ function can be written as sum of a term depending on marginal variables and a term depending on model parameters:

$$logZ = logZ_\Theta(\lambda) + \sum_t log\ Z_{\gamma_t}(\lambda_t) \tag{2.8}$$

$$= log(\int P_0(\Theta)e^{\sum_t \lambda_t y_t F(X_t|\Theta)}d\Theta) + \sum_t log(\int P_0(\gamma_t)e^{-\lambda_t\gamma_t}d\gamma_t)$$

Many choices for the margin variable distributions are possible. We will consider the following one with its penalty function:

$$P(\gamma_t) = c\, e^{-c(1-\gamma_t)}\ ,\gamma_t \le 1 \tag{2.9}$$

$$-log\ Z_{\gamma_t}(\gamma_t) = \lambda_t + log(1 - \lambda_t/c) \tag{2.10}$$

Each time the classification term is smaller than the margin mean value i.e. $1-1/c$, a penalty will occur, otherwise the relative Lagrange multipliers will be zero. Changing the value of the constant c will make the classification constraints more or less strict; the limit case $c \to \infty$ will lead to the case in which margin are fixed because their probability will be peaked at $\gamma_t = 1$.

2.1 Using generative models

In this section we will show how it is possible to accommodate generative models in the MED framework. Let's consider a binary classification problems where θ_+, θ_- are models parameters and $y = \{+1, -1\}$ are labels assigned to sample X. The discriminant function used in the MED solution can be chosen like following:

$$F(X|\Theta) = log\frac{P(X|\theta_+)}{P(X|\theta_-)} \tag{2.11}$$

where $\Theta = \{\theta_+, \theta_-\}$. In this way a discriminative learning framework is defined using the generative parameters of each model. Even if the discriminative function introduces some non-linearities, the optimization w.r.t. distributions $P(\Theta)$ will be always a convex optimization problem.

Obviously tractability depends on the possibility of writing the function $Z(\lambda)$ in closed form. It was found in [2] that if $P(X|\Theta)$ belongs to the exponential family a closed form for $Z(\lambda)$ can be found.

In [3] it was shown that in hybrid generative/discriminative learning a good initialization for parameters prior distributions is optimal bayesian posterior distributions because in this way the MED solution will be the solution that respects classification constraints and is as closed as possible to the generative model. If the classification problem involves more than two classes, classification constraints must be imposed for all possible couples in order to assure that the correct model will always 'dominate';

The MED framework can accommodate many currently used models. When models contain hidden variables, an EM-like algorithm is possible (see [3] for details). In [3] it was used to learn models that belong to the exponential family, mixture models (like GMM), mixture of mixtures (like HMM); basically the complete model parameter set can be learned using MED (e.g. gaussian means, variances etc.) but here we will consider just the feature selection problem. Together with classification many other problems like regression, transduction or anomaly detection can be solved using MED.

2.2 MED feature selection

A possible application of the MED learning is feature selection (see [2]). It consists in associating with each feature a switch s_i that can activate or not a certain feature. A distribution is assumed for those variables and MED learning is used to find the optimal one. Our purpose is slightly different from the original feature selection formulation: in fact we try to select the M best features out of the global N features; so M selected features will be features with the highest expected values calculated w.r.t. the MED optimal distribution.

2.2.1 Multivariate gaussian

Feature selection problem can be formulated in term of MED. We will first consider the simple case in which the competing models are multivariate gaussians with diagonal covariance matrix i.e.

$$P(X_t|\theta) = \prod_i^N \Big(\frac{1}{\sqrt{2\pi}\sigma_i} exp(-(X_{ti} - \mu_i)^2/2\sigma_i^2)\Big)^{s_i} \tag{2.12}$$

where N is the observation vector dimension, μ, σ^2 are gaussian mean and covariance, and s_i is a binary variable that indicates if a feature is selected or not i.e. $s_i = 1$ if the feature is active, else $s_i = 0$. Let's introduce a prior distribution on s_i: $P(s_i) = \rho_i^{s_i}(1-\rho_i)^{1-s_i}$ with $0 \leq \rho_i \leq 1$. Let's write the discriminative function between two different models $\{\theta_+, \theta_-\}$ (supposing to simplify notation that X_t belongs to model θ_+ and so $y_t = +1$) as:

$$F(X_t|\Theta) = log\, P(X_t|\theta_+)/P(X_t|\theta_-) = \sum_i s_i W_{it} \tag{2.13}$$

$$W_{it} = [(X_{ti} - \mu_i^+)^2/{\sigma_i^+}^2 - (X_{ti} - \mu_i^-)^2/{\sigma_i^-}^2] + log(\sigma_i^+/\sigma_i^-) \tag{2.14}$$

Eventually other models can be used e.g. model dependent coefficients s_i or joint optimization of features and model (using distributions on mean and covariance) but in this paper we will limit our investigation to the case in which s_i are common to all models.

It is now possible to find a MED solution to the feature selection problem. Expression (2.8) can be calculated in closed form:

$$J(\lambda) = \sum_t^T [\lambda_t + log(1 - \lambda_t/c)] - \sum_i^N log[1 - \rho_i + \rho_i e^{\sum_t^T \lambda_t W_{it}}] \tag{2.15}$$

Maximizing expression (2.15) will provide the optimal Lagrange multipliers set and the MED distribution can be explicitly computed. The value of discriminative function for an observation X_p can now be calculated integrating out $\Theta = \{s_i\}$ w.r.t. their distribution in equation (2.13) i.e.

$$\int P(\Theta)F(X_p|\Theta)d\Theta = \sum_i^N \frac{\rho_i W_{ip}}{(1-\rho_i)e^{-\sum_t^T \lambda_t W_{it}} + \rho_i} \tag{2.16}$$

If distribution is not a single multivariate gaussian but a gaussian mixture the previous computation is no more valid; there are two possible solutions: assigning an observation to a gaussian in the mixture (like in a K-means algorithm), falling in the mono-gaussian case, or using a probabilistic algorithm based on the reverse Jensen inequality described in [3]. Anyway those investigation are not pursuited in this paper.

2.2.2 Hidden Markov Models

In this section we consider the MED feature selection applied to an HMM with single gaussian pdf with feature weights i.e. expression (2.12)(without loss of generality because an HMM with gaussian mixture models can be interpreted as an HMM with single gaussian and more states corresponding to mixture components). In this case function $P(X_t|\Theta)$ is the likelihood of the observation sequence X_t calculated using the HMM. Let's consider two competing HMM and an observation sequence $X_t = \{X_{t0}...X_{tj}\}$. After doing forced alignment (e.g. with a Viterbi algorithm) we will obtain two sequences of states (and relative emission probability) that will give the likelihood of the sequence:

$$\begin{aligned} P(X_t|\theta_+) &= \pi^+ a_0^+ b_0^+ a_1^+ b_1^+ ... a_j^+ b_j^+ \\ P(X_t|\theta_-) &= \pi^- a_0^- b_0^- a_1^- b_1^- ... a_j^- b_j^- \end{aligned} \qquad (2.17)$$

where $a_{(.)}$ is the transition probability from state at time j to state at time $j+1$ and $b_{(.)}$ is the probability emission with an expression similar to (2.12). Now that all hidden variables are determined (i.e. the state sequence) it is possible to apply again the MED solution to the problem in the same form as the previous paragraph. A closed form for (2.8) can be obtained even in this case:

$$\begin{aligned} J(\lambda) = &\sum_t^T [\lambda_t + log(1 - \lambda_t/c)] - \sum_t^T \lambda_t \sum_{j=0}^{n_t} log \frac{a_{tj}^+}{a_{tj}^-} \\ &- \sum_i^N log[1 - \rho_i + \rho_i e^{\sum_t^T \lambda_t \sum_{j=0}^{n_t} W_{itj}}] \end{aligned} \qquad (2.18)$$

where $t = 1, ..., T$ indicates the sequence number, $j = 1, ..., n_t$, indicates the tth sequence elements,$i = 1, ...N$ indicates the feature, and W_{itj} indicates the log-likelihood difference between the jth element of the tth sequence for the ith feature. In this case the solution has a Lagrange multiplier for each data sequence.

2.2.3 Optimal prior estimation

Optimal prior estimation is always an important task in all bayesian approaches. Many criteria are possible (ML,MAP, Minimum Entropy). We decided to use a maximum likelihood approach to optimal prior estimation i.e. given data X, parameters Θ, and prior ξ, ML optimal priors are given by:

$$\xi^* = argmax_\xi \int P(X|\Theta)P(\Theta|\xi)d\Theta \qquad (2.19)$$

In our case $\Theta = \{s_i\}$ and $\xi = \{\rho_i\}$, expression (2.19) corresponds to finding ρ_i that maximize expression (2.16). To optimize (2.16), λ's values are needed; to circumvent the problem an alternate optimization w.r.t parameters (i.e. λ) and priors (i.e. ρ) can be done in the same fashion as described in [12]. Because our final goal is finding a feature subset of dimension M out of the N possible features, we impose the condition that $\sum_i^N E[s_i] = \sum_i^N \rho_i = M$ together with the other condition $a \leq \rho_i \leq b$. This is actually a suboptimal choice but experimental evidence showed its efficacy. We choose $a = 0.99$ (strong prior) and $b = 0.09$ (weak prior).

2.2.4 Practicalities

Finding MED solution of the problem corresponds in practice to maximize the function $J(\lambda)$ that is convex so any optimization techniques will bring to the global maximum of the function. Possible techniques include Newton-Raphson method, gradient descent, line search, or conjugate gradient descent. As discussed in [3] (Appendix) the most performing method consists in doing axis-parallel optimization in which a Lagrange multiplier at a time is updated. In some cases it is possible to find a closed form for the axis optimization but this is not our case. For maximizing the function we used derivatives of $J(\lambda)$ w.r.t λ_α; to find the zero of each derivate we used Brent's method (as suggested in [3]) that has the advantage of using just the function value and not its derivatives and it is more efficient compared to other methods like bisection methods. In practise the optimization process consists in iteratively solving:

$$\frac{\delta J(\lambda)}{\delta \lambda_\alpha} = 1 - \frac{1}{c - \lambda_\alpha} - \sum_i^N \frac{\rho_\alpha \, W_{i\alpha} \, e^{\Sigma_t^T \lambda_t W_{it}}}{1 - \rho_i + \rho_i \, e^{\Sigma_t^T \lambda_t W_{it}}} = 0 \tag{2.20}$$

for $\alpha = 1, ..., T$ in the case of single gaussian feature selection. Looking at equation (2.20) we can notice that the solution λ_α is in the interval $[0, c)$; when a classification error is smaller than the margin mean value, the relative Lagrange multiplier is non zero; otherwise it is zero. Another important point is that the solution is generally spare i.e. just a few out of the whole set of Lagrange multipliers are different from zero. It means that it is useless to update all multipliers with the same frequency; a stochastic solution can be used; in a first time multipliers are sorted on the base of their relevance respect to the function to optimize (i.e. eq. (2.20)) and then updates are done sampling multipliers from the sorted table (we used a gaussian distribution centered at the more relevant multiplier to sample data). In this way the more important multipliers will be updated with a higher frequency than the less significant one, reducing the number of iterations needed to converge. Particular attention must be used in determining the most relevant multipliers; using simply $\Delta J(\lambda_\alpha)$ can pose some problems; at early stages it will be large, even for irrelevant multipliers and then its value will progressively decrease. In [3] it is proposed to consider an 'adjusted' model with an exponential penalty term i.e. $\tilde{\Delta} J_t = \Delta J_t - \alpha exp(-\beta t)$ where parameters α, β can be determined in some way (e.g. with a simple least square criterion). In this way it is possible to determine the real relevant multipliers regardless the order in which they are processed.

2.3 Wrapper feature selection

A very efficient and well known approach to feature selection is the wrapper method [4]. Wrapper based methods are very precise and efficient but require very big computational resources because they consider explicitly all possible feature subsets of a given size to eliminate the weakest features in a greedy way. The iteration can be done in two ways: with a backward procedure or with a forward procedure. Generally the backward procedure is preferred because it permits to take advantage of the interaction between different features. In our experiments we will consider the backward procedure. The algorithm can be generalized as follows:

```
1. initialize the algorithm with the whole feature set $F$
2. while dimension $d$ of $F$ is larger then $N$ (desired final dimension)
3.     for all possible $S_i$ of size $d - 1$
```

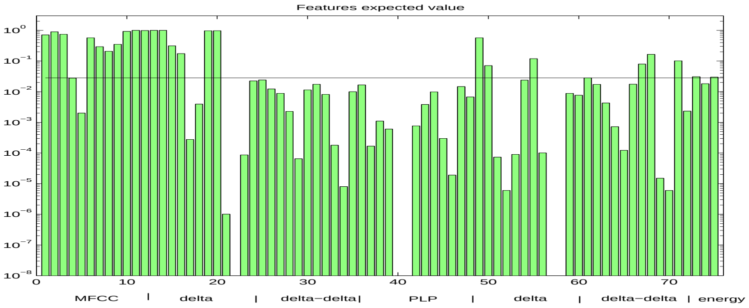

Figure 2.1: Feature expected values after MED learning; features with expected values bigger than the threshold are chosen, otherwise they are discarded; y axis is logarithmic scale

```
4.        evaluate the criterion $D(X, \Lambda_i)$
5.     set $F = S_i$ where $S_i$ minimize $D$
6. end
```

where Λ_i is the model corresponding to features in subset S_i. Many possible choices for $D(X, \Lambda_i)$ are possible: commonly used criterion are based on mutual information or classification error. To compare the method in the fairest way with the MED algorithm we choose a criterion based on the discriminative function (2.13):

$$D(X, \Lambda_i) = -\sum_t sign(log \frac{P(X_t|\theta_+, \Lambda_i)}{P(X_t|\theta_-, \Lambda_i)}) \tag{2.21}$$

where as before θ_+, θ_- are competing models. In other words the algorithm counts errors when different feature subsets are considered and eliminates features that hold the highest classification error.

2.3.1 Hybrid wrapper/MED feature selection

In order to take advantage of both methods we propose an hybrid algorithm using as criterion D the MED solution; as before the set F is initialized with the whole feature set, then MED is performed and this time the feature with the smaller expected value is taken out of the bunch. This procedure is iterated until the desired number of features is reached. This time the complexity is linear with the number of features: our initial intuition is that eliminating weakest features would increase the number of Lagrange multipliers that go to zero, converging to a different result form the one-step MED algorithm.

Chapter 3

Experiments

3.1 Synthetic data

To test the efficiency of the MED feature selection we run the following experiment on synthetic data: 1500 vectors of dimension 5 where generated using 3 gaussians with unitary diagonal covariance matrix (resulting in 3000 classification constraints) and with the following mean vectors: $m_1 = [1, 1, -1, 0, 1]^T$, $m_2 = [2, 1, 1, 0, 3]^T$ and $m_3 = [3, 1, 1, 0, 6]^T$.

Features one and five are the most discriminant between the three gaussian distributions, feature three cannot discriminate between m_2 and m_3, and features two and four cannot discriminate at all. We run MED feature selection using model (2.12), uniform priors $\rho_l = 0.009\ \forall l$ and with $c = 2$. Sorting features using their expected values w.r.t MED optimal distributions, we found that the first and the fifth features were the most significant, followed by the third and finally the second and the fourth that got almost zero as expected value.

This experiment on synthetic data shows that MED feature selection is able to learn features relevance for the classification task.

3.2 Speech recognition

The richness of speech recognition front end techniques raises the need of finding an optimal subset that permits good performance reducing the computational charges.

In [9] a wrapper method was used to select the best subset of fixed dimension (39) between MFCC features and articulatory features; in [10] an iterative algorithm based on mutual information was used to select best feature subset between MFCC,LPC and PLP features; and in [8] best feature combination is determined using conditional mutual information. A well known measure of redundancy of information between features is mutual information. To study redundancy between MFCC and PLP features we computed numerically the mutual information between an element of the MFCC feature vector x_i and an element of the PLP feature vector y_j as:

$$I(x_i, y_j) = \frac{1}{N} \sum_{t=1}^{N} p(x_{it}, y_{jt}) log(p(x_{it}, y_{jt})/p(x_{it})p(y_{it})) \qquad (3.1)$$

Probabilities $p(x_{it}, y_{jt})$, $p(x_{it})$, $p(y_{it})$ where estimated using the Parzen window density estimation (see [11]) that consists in estimating the pdf values of a vector z given M training vectors $\{z_i\}_{i=1}^{M}$ as $p(z) = \frac{1}{M}\sum_{i=1}^{M} \phi(z - z_i, h)$ where ϕ is the window function and h is the window width parameter. In our experiments ϕ is gaussian and optimal h is estimated using cross validation. To represent mutual information, we normalized scores for different features. Figure (3.1) shows mutual information between MFCC (x axis) and PLP (y axis). MI is high on the diagonal showing a certain redundancy (as expected because of the similarities in the two front end techniques).

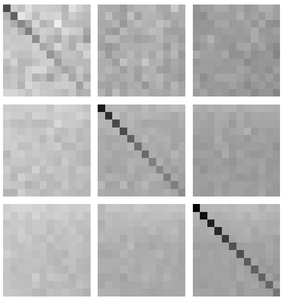

Figure 3.1: Mutual information between MFCC+Δ+$\Delta\Delta$ and PLP+Δ+$\Delta\Delta$ coefficients

In our experiments we want to compare the MED feature selection algorithm (described in 2.2.1) with a very efficient wrapper algorithm and with an hybrid wrapper/MED algorithm. In order to compare the three techniques we run context independent phoneme recognition experiments on the TIMIT database. The phoneme set is constituted by the classical 39 phoneme and each phoneme is modeled with a 3 state left-to-right HMM. Experiments are run using the HTK system. The original feature set is constituted by 75 features: 12 MFCC, their delta and delta-delta coefficients, 12 PLP, their delta and delta-delta coefficients, energy, its delta and delta-delta coefficients. In our experiments we tried to reduce the original 75 feature set to a 24 feature subset using the three different methods (wrapper,MED,hybrid wrapper/MED). At first an HMM with single gaussian distributions as emission probabilities is considered. To apply algorithms previously described all hidden variables must be estimated and all training set observations must be assigned to a class; a single gaussian HMM is trained using the full 75 feature set and forced alignment is run, associating each observation to a single state. We consider states as classes for MED and wrapper algorithm (even if other assignments are possible) and class distribution is the multivariate gaussian associated with each state. It is now possible to run feature selection algorithm. Theorically in both algorithms we should impose a classification constraint for all competing classes i.e. 38 phonemes times 3 states = 114 competitors. This results in a consistent increase of constraint number; for this reason we used the first 10 best competitors for each observation only. The value of constant c is fixed to 2. In order to increase the convergence speed we used the stochastic optimization method described in section 2.2.4 to optimize the normalization function $J(\lambda)$. Another parameter we are interested in is the robustness to the amount of data; for this reason we run the feature selection algorithm with 2 training set of respectively 20000 observations and 200000 observations. Figure (2.1) represents features expected values computed using the optimal MED distributions; features with expected value higher than the threshold are taken, others are discarded. Y axis is logarithmic scale. We selected the best 24 features. It is evident that MED seems to privilegiate MFCC feature respect to PLP features. Using the same class assignment the wrapper feature selection method was run as well the hybrid wrapper/MED algorithm. State alignment and feature subset selection is done with a single gaussian model: anyway once the optimal feature subset is found, gaussian number is increased up to 4 for each state and the feature subset is tested with the 4 gaussians model in order to verify that the subset is still significant. Table (3.1) reports recognition rate for the full feature set in column (a), while column (b) and (c) shows error rate for (b) 24 MFCC+Δ and (c) 24 PLP+Δ. Table (3.2) shows results for (d) MED feature selection, (e) hybrid wrapper/MED feature selection and (f) wrapper feature selection, for two different amount of training data: 20000 (feature I) and 200000 (feature II) training vectors. The three feature selection algorithms achieved better performance than the classical 24 features set (b) and (c). When poor amount of training data are provided (20000) MED method performs better than wrapper method, on the other side when large amount of training data is used (200000) wrapper performs better then MED. MED feature selection algorithm shows a high robustness to amount of data; we think this is due to fact that the MED framework is a bayesian framework and generally bayesian approaches are more robust to amount of training data; furthermore the final result seems to be strongly determined by prior values. It is interesting to notice that the wrapper method considers **all** possible subsets of size N to eliminate a feature i.e. in this case 2805 possible feature subsets and for each of them a computation of the criterion D must be done. On the contrary the MED just requires the optimization of an objective function and still holds interesting results. The hybrid wrapper/MED method seems to perform like MED contrarily to our initial intuition; in fact eliminating always the weakest feature does not seem to have

feature	(a) full set	(b) MFCC+Δ	(c) PLP+Δ
1 gaussian	61.2	49.5	50.2
4 gaussian	67.2	55.9	55.6

Table 3.1: Recognition rate for different features set: (a) 36 MFCC+Δ+ΔΔ+36 PLP+Δ+ΔΔ+Energy+Δ+ΔΔ; (b) 24 MFCC+Δ; (c) 24 PLP+Δ

feature I	(d) MED	(e) Wrapper/MED	(c) Wrapper
1 gaussian	54.0	53.9	51.1
4 gaussian	59.0	59.2	57.7

feature II	(d) MED	(e) Wrapper/MED	(c) Wrapper
1 gaussian	54.0	53.9	56.2
4 gaussian	59.0	59.2	61.6

Table 3.2: Recognition rate (PER) for different features subset set obtained with a training set of 20000 observation; (d) MED feature subset; (e) hybrid wrapper/MED feature subset; (f) wrapper feature subset; feature I: 20000 training vectors, feature II 200000 training vectors

any important consequence on the computation of Lagrange multipliers whose values depend on the strongest features.

It is interesting to notice that feature subset chosen are significantly different. Table (3.3) shows the 24 features selected by each of the 3 algorithms;

feature	Optimal feature subset

(c)	energy; 1,2,3,5,6,7,8,9,10,11 MFCC;
	1,2,3,4,7,8 Δ-MFCC; 12 PLP;
	1,6 Δ-PLP; 6,7,10,12 $\Delta\Delta$-PLP

(d)	energy; 1,2,3,5,6,7,8,9,10 MFCC;
	1,2,3,4,7,8 Δ-MFCC; 12 $\Delta\Delta$-MFCC;
	12 PLP; 1,6 Δ-PLP; 6,7,10,12 $\Delta\Delta$-PLP

(e)	2,3,4,9 MFCC; 5,6,9 Δ-MFCC; 4 $\Delta\Delta$-MFCC; 1,2,3,5,6,7,8,9 PLP; energy Δ,$\Delta\Delta$; 1,2,3,4,8 Δ-PLP; 8 $\Delta\Delta$-PLP

Table 3.3: Optimal feature subset determined using the algorithms (c) (d) and (e)

features selected by the MED algorithm are almost similar to the one chosen by the wrapper/MED algorithm. Wrapper feature selection seems to privilegiate the PLP coefficients while MED algorithm seems to select more MFCC coefficients. At a first look it may seem strange that two algorithms based on the same discriminative function (2.13) give such different results; but it is easy to understand the diversity of two approaches: MED is based on probabilities while wrapper (here considered) is based on pure error computation. Furthermore even if MED is based on classification error, as it is observed in [2], it has an information theoretic interpretation: MED solution minimizes the mutual information between data and parameters.

3.3 Conclusion and future works

In this paper we applied an MED approach to feature subset selection and compared it with a more classical wrapper approach. We proposed as well an hybrid MED/wrapper approach. MED feature selection does not need the same huge computational resources needed by wrapper methods. Furthermore the bayesian framework in which the MED algorithm is defined, gives a very high robustness respect to the amount of data used. Another interesting application of MED feature selection to speech recognition could be to determine the most reliable features in noisy condition. Future works may involve the joint of optimization of model parameters and features imposing a distribution on gaussian means and variances.

Bibliography

[1] Jaakkola T.,Meila M.,Jebara T., “Maximum entropy discrimination”, Neural Information Processing Systems12, NIPS 12, MIT Press,1999.

[2] Jebara T.,Jaakkola J.,” Feature selection and Dualities in Maximum Entropy Discrimination”, UAI 2000. July 2000.

[3] Jebara T., “ Discriminative, generative and imitative learning“,PhD thesis, Media Laboratory MIT, December 2001.

[4] Kohavi R.,George J., “Wrappers for Feature Subset Selection”. Artificial Intelligence journal, Vol. 97, Nos 1-2, pp. 273-324.

[5] Koller D., Sahami M. “ Toward Optimal Feature Selection“ Proceedings of the 13th Int. Conf. on Machine Learning, Bari, Italy, July 1996, pp.284-292

[6] Battiti R., “Using the mutual information for selecting features in supervised neural net learning.” IEEE Trans. on Neural Networks, 5(4):537–550, 1994

[7] Cover T., Thomas J. “ Elements of information theory“. 1991 John Wiley & Sons, Inc.

[8] Ellis D.P.W. and Bilmes J.A.,“Using mutual information to design feature combinations “ Proc. ICSLP-2000, Beijing, October 2000

[9] Kirchhoff K., Fink G.A. and Sagerer G., “Combining acoustic and articulatory feature information for robust speech recognition.” Speech Communication, May, 2002

[10] Omar M. K.,Chen K, Hasegawa-Johnson M. ”An evaluation of using mutual information for selection of acoustic-features representation of phonemes for speech recognition”,ICSLP2002 Denver, USA

[11] Fukunaga N.,” Introduction to statistical pattern recognition“. 1990 Academic Press

[12] Gauvain J. L., Chin-Hui Lee; “ Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains“,Speech and Audio Proces., IEEE Trans. on , Volume: 2 Issue: 2 , April 1994 Page(s): 291 -298