# A Hybrid MaxEnt/HMM based ASR System

*Yasser Hifny*[*]*, Steve Renals, Neil D. Lawrence*

Department of Computer Science,
The University of Sheffield , 211 Portobello Street,
Sheffield S1 4DP, UK.
{y.hifny, n.lawrence}@dcs.shef.ac.uk, s.renals@ed.ac.uk

## Abstract

The aim of this work is to develop a practical framework, which extends the classical Hidden Markov Models (HMM) for continuous speech recognition based on the Maximum Entropy (MaxEnt) principle. The MaxEnt models can estimate the posterior probabilities directly as with Hybrid NN/HMM connectionist speech recognition systems. In particular, a new acoustic modelling based on discriminative MaxEnt models is formulated and is being developed to replace the generative Gaussian Mixture Models (GMM) commonly used to model acoustic variability. Initial experimental results using the TIMIT phone task are reported.

## 1. Introduction

The research in the acoustic modelling has many directions to enhance the acoustic frame scoring and to replace the Gaussian Mixture Models generative models. The generative models estimate the likelihoods but lack the discrimination since they do not give direct estimates of posterior probabilities of the classes given the acoustics. One of the most useful methods to overcome this problem was to replace GMM likelihoods by Neural Networks (NN) acoustic classifier [1, 2].

The maximum entropy (MaxEnt) principle encourages us to choose the most unbiased distribution that is simultaneously consistent with a set of constraints. Typically, the available information about the system is incomplete, and there is an infinite number of possible probability distributions that satisfy the constraints. E. T. Jaynes suggested maximizing Shannon's entropy criterion subject to the given constraints to choose a suitable distribution as follows [3]:

> When we make inferences based on incomplete information, we should draw them from that probability distribution that has the maximum entropy permitted by the information we do have.

MaxEnt has been used in the field of Natural Language Processing (NLP) as a principled way to combine multiple sources in a probabilistic framework [4]. In speech recognition, MaxEnt has been applied to language modelling [5], but there has been relatively little work in acoustic modelling: Likhododev and Gao [6] developed a rank based direct model for speech recognition whose parameters were estimated by MaxEnt, and Macherey and Ney [7] discriminatively estimated the parameters of a Gaussian model based speech recognizer using MaxEnt. In a previous work, we evaluated the importance of acoustic features using MaxEnt incremental acoustic space dimensionality selection and combination[8].

[*] Yasser Hifny is sponsored by a Motorola Studentship.

In this paper, a high dimensional acoustic space is constructed by a large number of acoustic constraints. This aims to simplify the acoustic classification problem as the high dimensional spaces are more likely to be linearly separable than low dimensional spaces. The new constraints constructed from this high dimensional feature space are combined in the MaxEnt framework to estimate the posterior probabilities over the phonetic labels given the acoustic input. Integrating these posterior probabilities with the HMM systems will lead to the a form of hybrid MaxEnt/HMM acoustic modelling.

In the next section, a mathematical treatment for the principle of MaxEnt is presented and in section 3 we introduce the parameter optimization procedure for the MaxEnt models. Section 4 discusses how to define and implement the acoustic constraints. Section 5 discusses the sparse MaxEnt modelling for better generalization performance. In Section 6, the Hybrid MaxEnt/HMM system is described. Section 7 introduces the experimental work and preliminary results on the TIMIT database. We conclude and discuss further work in section 8.

## 2. The Maximum Entropy Principle

Let $s$ be a discrete variable representing the possible output classes/states in a classification problem, and $o$ be an observation affecting the states of the system. The constrained optimization problem at hand is to maximize the conditional Shannon entropy:

$$\arg\max_{p\in\mathcal{C}} H(p) = -\sum_{o}\tilde{p}(o)\sum_{s}p_\Lambda(s\mid o)\ln p_\Lambda(s\mid o) \quad (1)$$

subject to

C1 $p_\Lambda(s\mid o)\geq 0$ for all $s, o.$ and $\sum_s p_\Lambda(s\mid o)=1$ for all $o.$

C2 $\sum_o p(o)\sum_s p_\Lambda(s\mid o)g_i(o,s) = \sum_{o,s}\tilde{p}(o,s)g_i(o,s)=\tilde{p}(g_i)$ for $i=1,2,\ldots,n.$

Where $H(p)$ is the expectation of the conditional entropy of the model with respect to the training database, $\tilde{p}(o)$ is the observed marginal probability, and $\Lambda=\{\lambda_i\}$ is the set of parameters to be optimized. Constraint C1 represents the direct constraint from probability theory. Constraint C2 represents the integration of the available prior knowledge on the random variables $o, s$ in terms of the characterizing constraints $g_i(o,s)$, which have expected value $\tilde{p}(g_i)$.

The maximum entropy formalism results in a probability distribution, which is the log linear or exponential model:

$$p_\Lambda(s\mid o)=\frac{1}{Z_\Lambda(o)}\exp\Big(\sum_i\lambda_i g_i(o,s)\Big) \quad (2)$$

where

- $\lambda_i$ is the Lagrange multiplier (weighting factor) associated to the function $g_i(o, s)$.
- $Z_\Lambda(o)$ (Zustandsumme) is a normalization coefficient resulting from the natural constraints over the probabilities summation, commonly called the partition function, and given by

$$Z_\Lambda(o) = \sum_s \exp\Big(\sum_i \lambda_i g_i(o, s)\Big)$$

The entropy is a concave function of the mean values of the characterizing constraints $\tilde{p}(g_i)$ [9]. Hence, the MaxEnt solution is unique given the empirical mean values of the constraints. Practically this means that the solution is not sensitive to the initial values of the model parameters and the constructed model is unique for a given database in the statistical learning procedure.

It should be noted that in the absence of any constraint other than the natural constraint, the maximum entropy formalism results in the flat uniform distribution:

$$p_\Lambda(s \mid o) = 1/n. \tag{3}$$

This result explains the basic philosophy behind maximizing the entropy as the uniform distribution produces the most unbiased distribution. Integrating constraints results in reduction of the entropy but the output distribution is the most unbiased distribution consistent with constraints.

Consider a maximum entropy problem with two constraints $\mu$ and $\sigma^2$ of a continuous random variable whose probability density function is square-integrable. In such a case, when the continuous entropy is maximized, its solution is the normal distribution. This explains the importance of this distribution and why it has been frequently used in the application of statistical inference and why it deserves the adjective "normal", where this distribution is the most uncertain and maximizes the entropy [10]. The strong assumption that the data is normally distributed for the two constraints $\mu$ and $\sigma^2$ is relaxed by introducing the concept of the parametric constraints in section 4.

# 3. MaxEnt Optimization

The MaxEnt model estimates the posterior probability of the states given the acoustic observations. Hence, these models are trained by discriminative methods directly. We train the MaxEnt models using the Conditional Maximum Likelihood (CML) criterion, equation (4), which maximizes the likelihood of the empirical model estimated from the training data, $D = \{(o_t, s_t)\}_{t=1}^T$, with respect to the hypothesized MaxEnt model. The optimal parameters, $\Lambda^*$ estimated by maximizing CML criterion imply minimization of the cross entropy between the data model and the hypothesized MaxEnt model.

$$\Lambda^* = \arg\max_\Lambda L_{\tilde{p}}(\Lambda) \tag{4}$$

## 3.1. Parameter Estimation

The purpose of the parameter estimation algorithm is to estimate the parameters $\lambda_1..\lambda_n$ using numerical methods. A modified version of the Improved Iterative Scaling (IIS) algorithm [11] was used to estimate the parameters. It was suggested to us by John Lafferty [12] to support constraints that may take negative values, which was a restriction of the original algorithm. Further details about the mathematical derivation are reported in [13]. The basic idea behind the IIS algorithm is to make use of an auxiliary function, which bounds the change in divergence from below after each iteration.

The Generalized Improved Iterative Scaling (GIIS) algorithm proceeds as follows:

1. Let $p^0_\Lambda(s \mid o)$=$1/n$, which is the uniform model, where $\lambda_i = 0.0$. $i = 1, 2, \ldots, n$
2. Solve the following equation using Newton's method: $\tilde{p}(g_i) = \sum_{o,s} \tilde{p}(o) p^t_\Lambda(s \mid o) g_i(o, s) \exp\left(\delta_i^{t+1} s_i(o, s) M(o, s)\right)$
3. Update the parameters: $\lambda_i^{t+1} = \lambda_i^t + \delta_i^{t+1}$
4. If a valid termination condition is achieved then stop else go to step 2.

where $M(o, s) = \sum_i |g_i(o, s)|$ and $s_i(o, s)$ is the sign of $g_i(o, s)$. Solving the equation in step 2 for each iteration, results in the value of Maximum Likelihood (ML) step $\delta_i^{t+1}$ towards obtaining the MaxEnt global solution. When $s_i(o, s)$ is positive, step 2 corresponds to the IIS algorithm. Furthermore, it can be shown that the equation has a unique solution by directly checking its convexity.

## 3.2. Efficient MaxEnt Training

In this work, the MaxEnt constraints are the acoustic matching scores evaluated using a large number of Gaussian Models. These constraints are defined as parametric or generative constraints as described in section 4 and they are expected to cover the whole acoustic space with a suitable resolution. Hence, the constraints have positive values and this will lead to efficient update equations. By dividing the constraint values $g_i(o, s)$ by the $M_i(o, s)$, the new $M_i(o, s)$ will equal 1 for each observation. This may be interpreted as calculating a score similar to a posterior probability over the constraints scores. Hence, solving the GIIS equation will reduce to

$$\delta_i^{t+1} = \log \frac{\tilde{p}(g_i)}{\sum_o \tilde{p}(o) \sum_s p^t_\Lambda(s \mid o) g_i(o, s)} \tag{5}$$

Although equation (5) does not take the full GIIS step towards the global solution, it is very efficient since it does not imply root finding using the Newton Raphson method. Hence, it was chosen for parallel computing facilities as it is very simple. Indeed, this equation is a special case of Generalized Iterative Scaling (GIS) developed by Darroach and Ratliff,[14], where $\max_T M(o, s) = 1$.

# 4. Parametric Constraints

The description of the constraining characterizing functions is an optional implementation issue in which the prior knowledge for different applications is integrated. These parametric constraints aim to model the high variability of the observed acoustic features and overcome the strong assumption that the data distribution is Gaussian if we used the acoustic features directly. The form of the parametric constraints is optional: in this work we have used finite GMMs, which are a flexible model with a strong and rich history in speech recognition.

The diagonal GMMs are estimated per state using the EM algorithm [15]. The mixture weights are then ignored as they are not related to discrimination. Hence, the resulting GM models will estimate the likelihood score for an observation, which

will take the role of MaxEnt constraints over labels classes per event. The GM constraints have the following form:

$$g_i(o,s) = g_i(o,s;\theta) = p_i(o \mid \theta) = \mathcal{N}_i(\mu,\Sigma) \qquad (6)$$

where

- $o$ is the observed continuous random variable. In acoustic space, it represents the acoustic features values per frame.
- $s$ is a discrete random variable representing output classes or states.
- $p_i(o \mid \theta)$ is the likelihood score for the GM parametric constraint.

## 5. Sparse MaxEnt Models

In the MaxEnt solution, the Lagrange multipliers, which may be interpreted as the importance of each acoustic constraint per each state, are the outcome of the training procedure. In order to model the variability of the high dimensional acoustic space, large number of parametric constraints (GM) are usually utilized during the training procedure. Training a large number of parameters will lead to the overfitting phenomenon and poor generalization.

One way to handle this problem is to use a greedy methodology to evaluate the importance of the constraints [8]. In this work, we add a penalty term to the CML criterion in order to control the model complexity as usually done in the Regularization framework as shown in equation (7).

$$\Lambda^* = \arg\max_{\Lambda} L_{\tilde{p}}(\Lambda) - \beta\Omega(\Lambda) \qquad (7)$$

The $\Omega(\Lambda)$ is usually explained as imposing a prior distribution over the model parameters in the Bayesian framework. Weight decay regularizer form, $\Omega(\Lambda)_2 = \mathrm{sum}_n\|\lambda_i\|^2$, is commonly used to control the complexity and it implies zero mean gaussian priors over the model parameters in the Bayesian setting [16]. However, the gaussian prior does not lead to a sparse solution as the parameter values do not approach zero after the training procedure. Also, the gaussian prior implies that the MaxEnt parameters can be either positive or negative.

We utilize the MaxEnt models as a combination stage for the parametric constraints scores. Hence, it may be desirable to ensure that $\lambda_i >= 0$ for this special case as the positive weighted summations of the constraints are only useful to compute the posterior probability. The Lasso regularizer, where $q = 1$, $\Omega(\Lambda)_1 = \mathrm{sum}_n\|\lambda_i\|$ is often used to increase the spareness of the model. This prior implies an independent double exponential (or Laplace) distribution for each parameter, with density $\frac{\beta}{2}\exp(-\beta|\lambda_i|)$ [17]. When $\lambda_i >= 0$, the double exponential distribution will be an exponential distribution.

Adding the complexity term to the CML criterion will lead to a minor modification to the original update equation (5) resulting in the new regularized equation (8). This definitely maintains the convexity of the objective function.

$$\delta_i^{t+1} = \log\frac{\tilde{p}(g_i) - \beta}{\sum_o \tilde{p}(o)\sum_s p_\Lambda^t(s \mid o) g_i(o,s)} \qquad (8)$$

Clearly, equation (8) suggests that the unreliable constraints in terms of their empirical expectations will be forced to zero. The additional constraint that $\lambda_i >= 0$ is imposed after each iteration.

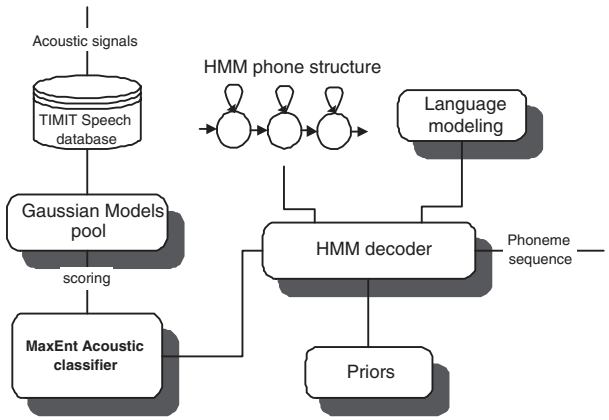


Figure 1: The proposed Hybrid MaxEnt/HMM system for TIMIT Task.

## 6. Hybrid MaxEnt/HMM Decoding

Hybrid NN/HMM systems were introduced to decode the speech signal based on the posterior probability estimation from neural network classifiers [1]. As the HMM models use class conditional densities for observation scoring, such systems compute a *scaled* likelihood score from the posteriors probabilities over the states as shown in equation (9).

$$p_\Lambda(o \mid s) \simeq \frac{P_\Lambda(s \mid o)}{P(s)} \qquad (9)$$

where $P(s)$ is estimated from the empirical data and theoretically, the estimated conditional densities $p_\Lambda(o \mid s)$ make little assumptions with respect to the GMM densities.

Here, the MaxEnt modelling is used to estimate the posterior probabilities over the states. These probabilities are associated with an HMM state for each phone. A simple bigram decoder is used to decode a phone task given the acoustic observations. The proposed system is shown in figure (1).

## 7. Experimental Work

We have performed experiments using the TIMIT database. In these experiments we have used the TIMIT phone labels as the classes in the MaxEnt model. The 61 phone classes in TIMIT were reduced to a set of 39 labels in the standard way. We used the 420 speaker training set, analyzed using a 25ms Hamming window at a 10 ms fixed frame rate, resulting in 1,410,069 frames. The MFCCs acoustic feature, along with the first and second order derivative features are extracted for each frame. The acoustic features were models with 13000 parametric GM constraints.

The normalized score over the constraints per frame are computed and saved in advance for each utterance in the TIMIT database. The empirical expectations for the MaxEnt constraints are then computed. The constraints with small $\tilde{p}(g_i)$ are removed from the initial model. The resultant constraint count was 258911 constraints. Calculating the model expectations over the constraints is the most expensive part of the training procedure. The MaxEnt model training procedure converged after 10 iterations using the algorithm described in section 3. The training frame accuracy was 69% over the training set and 63% over the test data.

The acoustic frame prior was not modelled during this work. Hence, the *scaled* likelihood is approximated $p_\Lambda(o \mid s) \simeq P_\Lambda(s \mid o)$ with the posterior probabilities calculating directly during the HMM decoding. Each phone was represented by three states left to right HMM. The three HMM states share the same scoring for each phone posterior estimated from MaxEnt model.

A bigram phone model was estimated from the TIMIT training set. The whole TIMIT test set was used in the experiment. The test data set was decoded using HMM simple decoder. The Langauge Model (LM) and the Acoustic Model (AM) scaling factors were fixed to 1 and 6, respectively. The basic decoding results are summarized in Table 1.

Table 1: *Timit decoding results.*

| Corr | Sub | Del | Ins | Acc |
|---|---|---|---|---|
| 72.9 | 16.9 | 10.2 | 6.7 | **66.2** |

The reported phone accuracy (66.2%) is comparable to many published results on TIMIT phone task. However, these results are still lower those reported by the GMM/HMM HTK system (72.3%) [18] and the Recurrent Neural Network (RNN) phone accuracy (75%) [19].

## 8. Conclusions

In this paper we present an approach to model the acoustic spaces through for the MaxEnt modelling framework. The work aims to relax the inaccurate assumptions associated with the state of art GMM/HMM based systems for continuous speech recognition. The paper addresses issues related to parameter estimation and increasing model spareness.

Currently, there is ongoing engineering work to make the system is more efficient in utilizing the parameters and selecting discriminative parametric constraints. This may lead to better recognition performance.

## 9. References


[1] N. Morgan and H. Bourlard, "Continuous speech recognition: An introduction to the hybrid hmm/connectionist approach," *IEEE Signal Processing Magazine, Invited Paper*, vol. 12, no. 3, pp. 25–42, 1995.

[2] E. Trentin and M. Gori, "a survey of hybrid ann/hmm models for automatic speech recognition," *Neurocomputing*, vol. 37, no. 1-4, pp. 91–126, April 2001.

[3] E. T. Jaynes, "On the rationale of maximum-entropy methods," *Proc. IEEE*, vol. 70, no. 9, pp. 939–952, 1982.

[4] A. L. Berger, S. A. Della Pietra, and V. J. Della Pietra, "A maximum entropy approach to natural language processing," *Computational Linguistics*, vol. 22, no. 1, pp. 39–71, 1996.

[5] R. Rosenfeld, "A maximum entropy approach to adaptive statistical language modeling," *Computer, Speech and Language*, vol. 10, pp. 187– 228, 1996.

[6] A. Likhododev and Y. Gao, "Direct models for phoneme recognition," in *Proc. IEEE ICASSP*, 2002, pp. 89–92.

[7] W. Macherey and H. Ney, "A comparative study on maximum entropy and discriminative training for acoustic modeling in automatic speech recognition," in *Proc. Eurospeech*, 2003, pp. 493–496.

[8] Y. H. Abdel-Haleem, S. Renals, and N. D. Lawrence, "Acoustic space dimensionality selection and combination using the maximum entropy principle," in *Proc. IEEE ICASSP*, 2004.

[9] J. Kapur and H. Kesavan, *Entropy Optimization Principles with Applications*. Academic Press, 1992.

[10] S. Guiasu and A. Shenitzer, "The principle of maximum entropy," *The Mathematical Intelligencer*, vol. 7, no. 1, 1985.

[11] S. Della Pietra, V. J. Della Pietra, and J. D. Lafferty, "Inducing features of random fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 4, pp. 380–393, 1997.

[12] J. Lafferty, "Personal communication," May 2002.

[13] Y. H. Abdel-Haleem, "The use of maximum entropy principle in continuous speech recognition," May 2003, http://www.dcs.shef.ac.uk/~yhifny/publications/MaxEnt-ASR.pdf.

[14] J. N. Darroch and D. Ratcliff, "Generalized iterative scaling for log-linear models," *The Annals of Mathematical Statistics*, vol. 43, pp. 1470–1480, 1972.

[15] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the *EM* algorithm," *Journal of the Royal Statistical Society B*, vol. 39, no. 1, pp. 1–38, 1977.

[16] C. M. Bishop, *Neural networks for pattern recognition*. Oxford: Clarendon Press, 1995.

[17] T. Hastie, R. Tibshirani, and J. H. Friedman, *The elements of statistical learning : data mining, inference, and prediction*. New York: Springer, 2001.

[18] S. Young and P. Woodland, "State clustering in hmm-based continuous speech recognition," *Computer Speech and Language*, vol. 8(4), pp. 369–384, 1994.

[19] A. J. Robinson, "An application of recurrent nets to phone probability estimation," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 298–305, March 1994.