# MAXIMUM CONDITIONAL MUTUAL INFORMATION WEIGHTED SCORING FOR SPEECH RECOGNITION

*Mohamed Kamal Omar, Ganesh N. Ramaswamy*

IBM T. J. Watson Research Center
Yorktown Heights, NY 10598, USA
mkomar,ganeshr@us.ibm.com

**ABSTRACT**

This paper describes a novel approach for extending the prototype Gaussian mixture model used in representing different classes in many recognition or classification systems and its application to large vocabulary automatic speech recognition (ASR). This is achieved by estimating weighting vectors to the log likelihood values due to different elements in the feature vector. This approach estimates the weighting vectors which maximize an estimate of the conditional mutual information between the log likelihood score and a binary random variable representing whether the log likelihood is estimated using the model of the correct label or not. It is shown in the paper that under some assumptions on the conditional probability density function (PDF) of the log likelihood score given this random variable, maximizing the differential entropy of a normalized log likelihood score is an equivalent criterion. This approach allows emphasizing different features, in the acoustic feature vector used in the system, for different hidden Markov model (HMM) states. In this paper, we apply this approach to the RT04 Arabic broadcast news speech recognition task. Compared to the baseline system, 3% relative improvement in the word error rate (WER) is obtained.

## 1. INTRODUCTION

One of the main objectives of speech signal analysis in ASR systems is to produce a parameterization of the speech signal that reduces the amount of data that is presented to the speech recognizer, and captures salient characteristics suited for discriminating among different speech units. Most ASR systems use cepstral features augmented with dynamic information from the adjacent speech frames and a dimensionality reduction technique which is a variant or an extension of linear discriminant analysis (LDA) [1]. The objective function in all these methods is not directly related to minimizing the recognition error, and therefore does not necessarily minimize the discrimination loss due to having a unified representation for all classes. LDA transformation, for example, tends to preserve distances of already well-separated classes. Conventional systems for automatic speech recognition model these features, given the HMM state, using the same prototype model of a diagonal-covariance Gaussian mixture (GMM).

In ASR, like many statistical classification and recognition problems with many classes, it is commonly the case that different classes or clusters of classes exhibit significantly different properties. This motivates using features designed to represent subsets that exhibit common properties which are not necessarily the same features used for any class outside this subset or using a prototype model which is general enough to allow emphasizing different elements of the feature vector for different classes.

The approach of using different features for different classes in speech recognition and verification has been suggested before [2]. Its main problem, due to the statistical nature of the recognizer, is how to compare *a posteriori* probabilities conditioned on different sets of features to decode a given utterance. This problem was addressed for segmental ASR systems that use different sets of features for different segments in [3] by adding extra reference models that have no physical meaning but are used to normalize the likelihoods to be comparable statistically. Other approaches, like [4] and [5], generate class-dependent features by class-dependent linear transforms from an original set of features which span the same original feature space.

In this paper, we present a novel approach for improving the performance of automatic speech recognition systems by using state-dependent weighting of the log likelihood scores due to different elements in the feature vector. Using the fact that the diagonal-covariance constraint allows the estimation of the log likelihood of an observation given a Gaussian component as the sum of the log likelihoods due to the different elements in the feature vector, this approach can be presented as a generalization of the GMM prototype model used for representing HMM states, and therefore does not have the problem of comparing likelihoods calculated using different sets of features. The weighting vectors are estimated such that the conditional mutual information of the log likelihood score and a binary random variable indicating whether the state model used in calculating the log likelihood of the frame is the correct state or not is maximized. This estimate of the mutual information is conditioned on the maximum likelihood estimated HMM model. We show that maximizing this objective function is equivalent to maximizing the differential entropy of a normalized log likelihood score under Gaussianity assumption of the log likelihood conditional PDF given the value of the binary random variable. As the weighting vectors are state-dependent, our approach improves the performance of the system by emphasizing different elements of the feature vector for different HMM states without having to change either the feature vector or the HMM model.

In the next section, we will formulate the problem and describe the objective criterion. In Section 3, the algorithm used in estimating the weighting coefficients to optimize the objective criterion is described. The experiments performed to evaluate the performance of our approach are described in Section 4. Finally, Section 5 contains a discussion of the results and future research.

We will use capital letters to represent random variables and vectors and the corresponding small letters to denote realizations of these random variables and vectors.

## 2. PROBLEM FORMULATION

In this section, we will discuss how to estimate the conditional mutual information of the log likelihood acoustic score and a binary random variable representing whether the log likelihood score is calculated using the correct state or not and then show how the problem can be reduced to a maximum differential entropy problem.

The mutual information of the log likelihood score and the binary random variable representing whether the log likelihood is calculated using the correct state is

$$I(S,B) = H(S) - H(S|B), \tag{1}$$

where $S$ is the log likelihood acoustic score of an observation vector, $B$ is the binary random variable, $H(S)$ is the differential entropy of the log likelihood acoustic score, and $H(S|B)$ is the conditional differential entropy of the acoustic log likelihood score given the value of the binary random variable.

We estimate the conditional mutual information given the HMM model trained using maximum likelihood estimation. Therefore the acoustic log likelihood values for each frame in the training data is calculated using this HMM model as

$$s_{kt}^{\rho} = \log P(O_{kt}|\rho), \tag{2}$$

where $P(O_{kt}|\rho)$ is the likelihood of the observation $O_{kt}$ at frame $t$ of the $k$th utterance given the HMM state $\rho$. Using state-dependent weighting of the contributions to the likelihood due to different feature elements and replacing the sum over the Gaussian components by the maximum,

$$\log P(O_{kt}|\rho) = \sum_{j=1}^{n} w_{\rho}^{j} \log P(O_{kt}^{j}|m_{\rho}^{*}), \tag{3}$$

where $m_{\rho}^{*} = \arg\max_{m_{\rho}} \left(H_{m_{\rho}} P(O_{kt}|m_{\rho})\right)$, $H_{m_{\rho}}$ is the weight of the Gaussian component $m_{\rho}$ of the Gaussian mixture model of state $\rho$, $w_{\rho}^{j}$ is the weight for state$\rho$ of the log likelihood corresponding to the $j$th element of the feature vector, and $n$ is the dimension of the feature vector. We note that the baseline HMM using diagonal-covariance Gaussian mixture model for each state is equivalent to using an all-one weighting vector. To be able to compare the likelihood values estimated using different HMM states and to guarantee that the likelihood function will integrate to one over all the observation space, it can be shown that the following constraints on the weighting coefficients for each state are necessary and sufficient

$$w_{\rho}^{j} > 0 \quad \text{for } 0 < \rho \leq K, 0 < j \leq n, \tag{4}$$

and

$$\sum_{m_{\rho}=1}^{M_{\rho}} H_{m_{\rho}} \prod_{j=1}^{n} \frac{\left(\sqrt{2\pi}\sigma_{jm_{\rho}}\right)^{1-w_{\rho}^{j}}}{\sqrt{w_{\rho}^{j}}} = 1 \quad \text{for } 0 < \rho \leq K, \tag{5}$$

where $K$ is the total number of HMM states, $\sigma_{jm_{\rho}}^{2}$ is the variance of the $m$th Gaussian component of state $\rho$ corresponding to the $j$th element in the feature vector, and $M_{\rho}$ is the total number of Gaussian components for the state $\rho$. The set of equality constraints in Equation 5 can be approximately satisfied by using a penalty function that is less than zero and equal zero if and only if the constraints are satisfied to modify the objective function to be maximized. We will discuss how the set of constraints in Equations 4 and 5 are imposed in the next section.

By using two Gaussian mixture models to model $P(S|B=0)$ and $P(S|B=1)$ and replacing the expectation in the expressions of the differential entropy, $H(S)$, and the conditional differential entropy, $H(S|B)$, with the sum over all possible word sequences in the lattice, we get the following estimate of the maximum conditional mutual information (MCMI) objective function

$$\hat{I} = \sum_{k=1}^{N} \sum_{t=1}^{T_k} \sum_{\rho=1}^{K} \left(\gamma_{kt}^{\rho} \left(\log P(s_{kt}^{\rho}|b_{kt}^{\rho}) - \log P(s_{kt}^{\rho})\right)\right), \tag{6}$$

where $N$ is the number of training utterances, $T_k$ is the length of the $k$th utterance in frames, $\gamma_{kt}^{\rho}$ is the sum of the *a posteriori* probabilities of the state $\rho$ at frame $t$ of utterance $k$ over all word sequences in the lattice, $b_{kt}^{\rho} = 0$ if the state $\rho$ is the correct HMM state for frame $t$ from the training utterance $k$ and $b_{kt}^{\rho} = 1$ if $\rho$ is a different HMM state,

$$P(S) = q(B=0)P(S|B=0) + q(B=1)P(S|B=1), \tag{7}$$

where $q(B)$ is the probability mass function of the binary random variable $B$.

We will investigate also an alternative approach to calculating an estimate of the objective function by noticing that if both $P(S|B=0)$ and $P(S|B=1)$ are Gaussian PDFs with mean $\mu_0$ and $\mu_1$ and variance $\sigma_0^2$ and $\sigma_1^2$ respectively and using a normalization of the log likelihood score in the form

$$\tilde{s}_{kt}^{\rho} = \frac{s_{kt}^{\rho} - \mu_{b_{kt}^{\rho}}}{\sigma_{b_{kt}^{\rho}}}, \tag{8}$$

where $b_{kt}^{\rho}$ is 0 in case $\rho$ is the correct state model for frame $t$ of the $k$th training utterance and 1 if otherwise, the conditional differential entropy of the normalized log likelihood score, $\tilde{S}$, is constant. Therefore maximizing the conditional mutual information of the normalized log likelihood score, $\tilde{S}$, and the binary random variable $B$ is equivalent to maximizing the differential entropy of $\tilde{S}$. Since the variance of $\tilde{S}$ is constant, the differential entropy of the normalized log likelihood score is maximized if and only if its probability density function (PDF) is Gaussian [6]. Hence, maximizing the differential entropy of the the normalized log likelihood score becomes a maximum likelihood problem. In which, we maximize the likelihood that the normalized log likelihood score is a Gaussian random variable. In this case, the maximum differential entropy (MDE) objective function to be maximized is

$$L = \sum_{k=1}^{N} \sum_{t=1}^{T_k} \sum_{\rho=1}^{K} \left(\gamma_{kt}^{\rho} \left(-\frac{1}{2}\log\sigma^2 - \frac{(\tilde{s}_{kt}^{\rho} - \mu)^2}{2\sigma^2}\right)\right), \tag{9}$$

where $\mu$ is the mean of the normalized log likelihood Gaussian PDF which equals 0, and $\sigma^2$ is the variance of the the normalized log likelihood Gaussian PDF which equals 1.

As discussed before, we introduce a weighting vector for each state to weight the log likelihood values corresponding to different elements of the feature vector and estimate the values of these weights which maximize the objective function in Equation 6 or Equation 9.

## 3. IMPLEMENTATION

In this section, we present our implementation of the approach described in the previous section. Our goal is to calculate the state-dependent weights of the log likelihood scores, $\{w_\rho^j\}_{\rho=1,j=1}^{\rho=K,j=n}$, where $K$ is the number of states in the HMM, which maximize the MCMI and MDE objective functions in Equations 6 and 9 respectively.

To calculate these weights which maximize the MCMI or the MDE objective function subject to the constraints in Equations 4 and 5, we can use an interior point optimization algorithm with penalized objective function [7]. Alternatively to simplify the optimization problem, we impose the constraints in Equation 5 by normalizing the weights of the Gaussian components of each state model using the relation

$$H_{m_\rho}^r = \frac{H_{m_\rho}}{\prod_{j=1}^n \frac{\left(\sqrt{2\pi}\sigma_{jm_\rho}\right)^{1-w_\rho^{jr}}}{\sqrt{w_\rho^{jr}}}}, \tag{10}$$

where $H_{m_\rho}$ is the original maximum likelihood estimate of the Gaussian component weight, and $H_{m_\rho}^r$ is the normalized Gaussian component weight at the $r$th iteration, $w_\rho^{jr}$ is the weight for state $\rho$ of the log likelihood corresponding to the $j$th element of the feature vector at the $r$th iteration. Therefore the problem is reduced to maximizing the objective function subject to the constraints in Equation 4. This is an optimization problem over a convex set and we use the conditional gradient method to calculate the weighting vectors [7].

Using the conditional gradient algorithm, the value of the weights are updated in each iteration using the relation

$$W_\rho^{r+1} = W_\rho^r + \alpha^{r+1}(\bar{W}_\rho^r - W_\rho^r), \tag{11}$$

where $W_\rho^{r+1}$ is the weighting vector at iteration $r+1$, $W_\rho^r$ is the weighting vector at iteration $r$, and $\alpha^{r+1}$ is a step size that should be chosen small enough to guarantee convergence and large enough to reduce the number of iterations required to achieve convergence,

$$\bar{W}_\rho^r = \arg\max_{W_\rho>0} (W_\rho - W_\rho^r)^T \frac{\partial \hat{I}}{\partial W_\rho}|_{W_\rho=W_\rho^r}, \tag{12}$$

where $\frac{\partial \hat{I}}{\partial W_\rho}$ is the gradient of the objective function. The Armijo rule is used to estimate the step size $\alpha^{r+1}$ at each iteration [7].

We update also the parameters of the HMM model using maximum likelihood estimation by using a likelihood function that takes into account the values of the weighting vectors and the normalized Gaussian mixture weights in the current iteration.

The gradient of the MCMI objective function in Equation 6 with respect to the state-dependent weighting vectors is

$$\begin{aligned}\frac{\partial \hat{I}}{\partial W_\rho} &= -\sum_{k=1}^N \sum_{t=1}^{T_k} \gamma_{kt}^\rho \sum_{m=1}^{M_b} \gamma_{kt}^{mb} \frac{(s_{kt}^\rho - \mu_{mb})}{\sigma_{mb}^2} V_{kt}^\rho \\ &+ \sum_{k=1}^N \sum_{t=1}^{T_k} \gamma_{kt}^\rho \left( \sum_{f=0}^1 q(f) \sum_{m=1}^{M_f} \gamma_{kt}^{mf} \frac{(s_{kt}^\rho - \mu_{mf})}{\sigma_{mf}^2} V_{kt}^\rho \right),\end{aligned} \tag{13}$$

where $W_\rho$ is the weighting vector for state $\rho$, $\gamma_{kt}^\rho$ is the sum of the posterior probabilities of the state $\rho$ at frame $t$ of utterance $k$ over all word sequences in the lattice for this utterance of the training data, $\gamma_{kt}^{mb}$ is the posterior probability of the $m$th Gaussian component of the Gaussian mixture model of $P(S|B=b)$, $\mu_{mb}$ and $\sigma_{mb}^2$ are its mean and variance respectively , $s_{kt}^\rho$ is the best log likelihood score for frame $t$ of utterance $k$ using a Gaussian component of the state $\rho$ GMM model, $q(f)$ is the prior probability mass function of the binary random variable $B$, and $V_{kt}^\rho = \left[s_{kt}^{\rho 0}, s_{kt}^{\rho 1}, \ldots, s_{kt}^{\rho j}, \ldots, s_{kt}^{\rho n}\right]$ is the vector of log likelihood values for frame $t$ of utterance $k$ using the state $\rho$ corresponding to different elements in the feature vector.

For the second implementation, we used the MDE objective function in Equation 9 instead of the MCMI objective function. The gradient of the MDE objective function with respect to the state-dependent weighting vectors is

$$\frac{\partial L}{\partial W_\rho} = -\sum_{k=1}^N \sum_{t=1}^{T_k} \gamma_{kt}^\rho \frac{s_{kt}^\rho}{\sigma_{b_{kt}^\rho}} V_{kt}^\rho. \tag{14}$$

The calculation of the gradient of both the MCMI and the MDE objective functions requires the generation of the posterior probabilities of the HMM states for each frame in the training data which is computationally expensive and therefore many approximations were used before for discriminative training of HMM models for large vocabulary ASR systems. The most successful approach that is adopted here is to use word lattices that fully encode sequential acoustic and language model constraints, [8]. Lattices are generated once using the MLE baseline acoustic HMM model and a unigram language model and then used repeatedly for several iterations. Given the word lattice, a forward-backward pass at the word lattice node/arc level is used to generate the posterior probability of a given arc. Then the Viterbi state-level segmentation for each arc was found and used with the posterior arc probabilities to estimate the gradient of the objective function.

## 4. EXPERIMENTS AND RESULTS

This section gives the experimental results of applying our approach using the two objective functions described in the last section on the Arabic DARPA 2004 Rich Transcription (RT04) broadcast news evaluation data. The raw features for the baseline ASR system used in the tests were 13-dimensional MFCC features computed every 10 ms. from 25-ms. frames with a Mel filter bank that spanned 0.125–8 kHz. The recognition features were computed from the raw features by splicing together nine frames of raw features ($\pm 4$ frames around the current frame), projecting the 117-dim. spliced features to 60 dimensions using an LDA projection, and then applying maximum likelihood linear transformation (MLLT) to the 60-dim. projected features to reduce the mismatch between the statistics of the final features and the constraints of the diagonal-covariance Gaussian mixtures that model the HMM observation densities.

The acoustic model training data were 70 hours of the foreign broadcast information services (FBIS) and the Arabic topic detection and tracking (TDT) databases provided by the Language Data Consortium (LDC). The acoustic model consisted of 5307 context-dependent states and 149K diagonal-covariance Gaussian mixtures. The states were clustered using decision trees that could ask questions about phone identity across words in a $\pm 5$-phone window. The number of Gaussian mixtures assigned to a state was

| System | WER |
|---|---|
| Baseline | 33.8 |
| MCMI Weights | 32.9 |
| MDE Weights | 32.8 |

**Table 1**. Word error rates (%) on the Arabic RT04 evaluation data using the SI systems.

| System | WER |
|---|---|
| Baseline | 28.8 |
| MCMI Weights | 27.9 |
| MDE Weights | 27.9 |

**Table 2**. Word error rates (%) on the Arabic RT04 evaluation data using the SAT systems.

chosen proportional to the logarithm of the number of observations in the training data which belongs to the state. The phonetic transcription in Arabic requires the existence of certain diacritic symbols which are usually not found in text transcriptions, and hence we use the one-to-one grapheme-to-phoneme approach [9]. The phoneme set consists of 38 phoneme and each phoneme is represented by 5 HMM states with left-to-right topology.

The decoding consists of two passes: the first pass outputs a lattice which is used to adapt the models, and the second decoding pass uses the adapted models to generate the final output of the decoder. In the context of speaker-adaptive training to produce canonical acoustic models, we use feature-space maximum likelihood linear regression (MLLR), [10]. We do also a single pass of MLLR adaptation, using a regression tree to generate transforms for different sets of mixture components, [10].

The language model is a 64K vocabulary 30M n-gram interpolated back-off trigram language model. It is built from the Arabic Giga-word text corpus distributed by LDC and the transcripts of the audio training data and some Arabic news web resources. The out of vocabulary (OOV) rate on the test data is 5.9% and the preplixity is 450.

We tested estimating the weighting vectors using both the MCMI objective criterion in Equation 6 and the MDE criterion in Equation 9 and compared the results to the baseline system trained with maximum likelihood estimation and without using any weighting of log likelihood values. The estimation of the weights using the MCMI criterion converged after six iteration of the conditional gradient algorithm, while using the MDE criterion, it converged after four iterations.

As shown in Table 1, the speaker-independent results improved by 3% relative compared the baseline for both systems using weighted log likelihood scores. The two systems with the weights estimated using MCMI and MDE performed almost the same.

The results after speaker adaptation reflects the same improvement over the baseline system as shown in Table 2. The MCMI and MDE systems gave exactly the same word error rate after speaker adaption.

## 5. DISCUSSION

In this paper, we examined an approach for state-dependent weighting of the log likelihood scores corresponding to different feature elements in the feature vector. We described two similar criteria to estimate these weights : the first maximizes the conditional mutual information of the log likelihood value and a binary random variable indicating whether the frame was scored by the correct state or not, the second maximizes the same objective function, but after normalizing the log likelihood score and under the assumption that the conditional PDF of the log likelihood score given the value of the binary random variable is Gaussian, it is equivalent to maximizing the differential entropy of the normalized log likelihood score. We applied this approach to the DARPA Arabic RT04 evaluation data. This approach decreased the word error rate by 3% relative compared to the baseline system for both the speake-independent and speaker-adapted systems. This improvement can be attributed to emphasizing scores corresponding to important features in the feature vector for each state.

Further investigation of the performance of our approach on other evaluation tasks will be our main goal. We will consider also other objective functions to estimate the state-dependent weights.

## 6. REFERENCES


[1] N. Kumar, "Investigation of silicon-auditory models and generalization of linear discriminant analysis for improved speech recognition," Ph.D. dissertation, John Hopkins Univ., Baltimore, MD, 1997.

[2] A. K. Halberstadt and J. R. Glass, "Heterogeneous measurements and multiple classifiers for speech recognition," in *Proc. of Int. Conf. of Spoken Language Processing*, Sydney, Australia, 1998, pp. 1379–1382.

[3] J. Glass, J. Chang, and M. McCandless, "A probabilistic framework for feature-based speech recognition," in *Proc. of Int. Conf. of Spoken Language Processing*, Philadelphia, PA, 1996, pp. 2277–2280.

[4] R. A. Gopinath, "Maximum likelihood modeling with Gaussian distributions for classification," in *Proc. ICASSP*, 1998.

[5] M. J. F. Gales, "Maximum likelihood multiple subspace projections for hidden Markov models" *IEEE Trans. On Speech And Audio Processing*, vol. 10, no. 2, pp. 37–47, February 2002.

[6] T. M. Cover, and J. A. Thomas, *Elements of Information Theory.* New York, NY: Wiley, 1997.

[7] D. P. Bertsekas, *Nonlinear Programming.* Belmont, MA: Athena Scientific, 1999.

[8] P. C. Woodland and D. Povey, "Large scale discriminative training for speech recognition," in *ASRU Workshop Proceedings*, Delavan, Wisconsin, December 2000, pp. 7–16.

[9] J. Billa, M. Noamny, A. Srivastava, D. Liu, R. Stone, J. Xu, J. Makhoul, F. Kubala, "Audio Indexing of Arabic Broadcast News," in *Proc. ICASSP*, Orlando, Florida, 2002.

[10] M. J. F. Gales, "Maximum likelihood linear transformation for HMM-based speech recognition," Cambridge University, Engineering Department, Technical Report CUED/F-INFENG, 1997.