A Variational Approach to Robust Maximum Likelihood Estimation for Speech Recognition

Mohamed Kamal Omar

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

10.21437/Interspeech.2007-569

ABSTRACT

In many automatic speech recognition (ASR) applications, the data used to estimate the class-conditional feature probability density function (PDF) is noisy, and the test data is mismatched with the training data. Previous research has shown that the effect of this problem may be reduced by using models which take the effect of the noise into consideration, and by transforming the features or the models used in the classifier to adapt to new environments and speakers. This paper addresses the degradation in the performance of ASR systems due to small —possibly time-varying— perturbations of the training data. To approach this problem, we provide a computationally efficient algorithm for estimating the model parameters which maximize the sum of the log likelihood and the negative of a measure of the sensitivity of the estimated likelihood to these perturbations. This approach does not make any assumptions about the noise model during training. We present several large vocabulary speech recognition experiments that show significant recognition accuracy improvement compared to using the baseline maximum likelihood (ML) models.

1. INTRODUCTION

Conventional systems for ASR are based on the assumption that the training data can be represented by a finite number of hidden Markov model (HMM) states corresponding to speech units. However in most real-life problems, the training data is corrupted with different types of noise and is affected by several sources of variability not related to the identity of these units. For example, features used in speech recognition applications are affected by channel characteristics, microphone type, speaker characteristics, and speaking rate. Adaptation techniques to compensate for speaker, channel, and environment effects are usually used. Examples include vocal tract length normalization [1], maximum likelihood linear regression (MLLR), and feature-based maximum likelihood linear regression (FMLLR) [2]. However even after using noise compensation and speaker adaptation algorithms, the residual error degrades the accuracy and the quality of the models estimated from the data. In this paper, we investigate the effect of small perturbations in the training data due to sources of variability not related to the word or the phoneme identity. In other words, we assume that one or more noise compensation and speaker adaptation algorithms are already used, in case of large perturbations, before applying our approach.

Stochastic perturbations in the log likelihood objective function can produce nonlinear variations of the resulting HMM parameters estimator in ASR systems. For example, bagging under some constraints on the perturbations may cancel these effects as measured by either variance or mean-squared error and in other cases may amplify them [3]. In our work, we consider the possibility of generating an artificial local maximum of the log likelihood objective function by small perturbations in the training data. This may result in suboptimal estimation of the parameters, as they do not correspond to a maximum of the actual likelihood which we try to model. It is possible also that these small perturbations mask a local maximum of the actual likelihood function. To reduce the effect of these perturbations in the training data, we propose adding a penalty function or a regularization term to the log likelihood function which works as a measure of the sensitivity of the estimated log likelihood to small perturbations in the training data. This additional term reduces the effect of these perturbations, due to noise in the training data, on the value of the estimated model parameters, and therefore reduces the sensitivity of the estimated log likelihood of the testing data to perturbations in the training data. Our work can be linked also to the choice of the prior distribution in Bayesian inference. One of the most popular rules of selection is Jeffery's invariance rule which advocates using a special prior density. It is the one which generates a posterior density invariant to transformations of the data or the parameters [4]. Our method can be formulated as an alternative way of selecting the prior density which reduces the effect of local transformations of the data on the value of the conditional density. Alternatively our approach can be formulated as a regularized maximum likelihood estimation method. Previous approaches for regularized likelihood estimation include adding the l_1 norm of the vector formed from the entries of the sample covariance matrix as a penalty term to get a sparse covariance structure [5], or using a regularization term proportional to an estimate of the entropy of a set of random variables [6].

In this work, we introduce a variational approach to minimize the effect of small perturbations in the training data on our estimate of the log likelihood of the testing data. This is achieved by optimizing the model parameters to maximize the sum of the log likelihood objective function and the negative of a measure of the sensitivity of the log likelihood function to these perturbations in the training data.

In the next section, we formulate the problem and describe our objective criterion. In Section 3, the algorithm used in estimating the HMM parameters to optimize our 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. In this paper, a subscript is used as an index of a component of a random vector, and a superscript is used as an index of a realization of the random vector.

Capital letters are used to denote the random variables and the corresponding small letters to denote their realizations. Both vectors and matrices are in boldface to be distinguished from scalars.

2. PROBLEM FORMULATION

Motivated by the discussion of the previous section, we derive a measure of the sensitivity of the likelihood function to perturbations in the training data without making any assumptions about the sources of variability which cause these perturbations, and search for an estimate of the HMM parameters which maximize a weighted sum of the log likelihood and the negative of a function monotonically proportional to this sensitivity measure. To do that, we need the following proposition.

Proposition 1 *Let $\mathbf{y} = f(\mathbf{x})$ be a continuously differentiable map of the random vector $\mathbf{x} \in \Re^n$ to $\mathbf{y} \in \Re^n$, and let $P_{\mathbf{Y}}(\mathbf{y}, \lambda)$ be its probability density function. The following relation is satisfied for every realization of the random vector $\mathbf{x}$ at which $f(.)$ is invertible with inverse $f^{-1}(.)$*

$$\begin{aligned} &\sum_{k=1}^{K} \frac{\partial \log P_{\mathbf{Y}}(\mathbf{y}, \lambda)}{\partial \lambda_k} \frac{\partial \lambda_k}{\partial \mathbf{J}_{if}^t} |_{\mathbf{y}=\mathbf{y}_i^t} \\ &= -\frac{\partial \log P_{\mathbf{Y}}(\mathbf{y}, \lambda)}{\partial \mathbf{y}} |_{\mathbf{y}=\mathbf{y}_i^t} {\mathbf{x}^t}^T - {\mathbf{J}_{if}^t}^{T^{-1}} \\ &\qquad \forall\, 0 \le t < N, \end{aligned} \tag{1}$$

where $\lambda = [\lambda_1, \lambda_2, \ldots, \lambda_K]^T$ is the parameters vector, K is the dimension of the parameters vector, $\mathbf{y}_i^t = f_i^t(\mathbf{x}^t)$ is the ith root of ${f^t}^{-1}(.)$ at $\mathbf{x}^t$, $\mathbf{x}^t$ is the tth realization of the random vector $\mathbf{x}$, $\mathbf{J}_{if}^t$ is the Jacobian matrix of the map $f^t(.)$ at the ith root $\mathbf{y}_i^t$, and N is the number of realizations of the random vector $\mathbf{x}$.

Proof: The relation between the probability density functions of $\mathbf{x}$, $P_{\mathbf{X}}(\mathbf{x})$, and the probability density functions of $\mathbf{y} = f(\mathbf{x})$, $P_{\mathbf{Y}}(\mathbf{y}, \lambda)$, is in general [7],

$$P_{\mathbf{X}}(\mathbf{x}) = \sum_{i=1}^{R} P_{\mathbf{Y}}(\mathbf{y}, \lambda)|_{\mathbf{y}=\mathbf{y}_i} \left| det(\mathbf{J}_{if}) \right|, \tag{2}$$

where $\mathbf{y}_i = f_i(\mathbf{x})$ is the ith root of $f^{-1}(.)$ at $\mathbf{x}$, R is the number of roots at $\mathbf{x}$, and $\mathbf{J}_{if}$ is the Jacobian matrix of the map $f(.)$ at the ith root $\mathbf{y}_i$.

Differentiating both sides with respect to $\mathbf{J}_{if}$, we get

$$\begin{aligned} \mathbf{0} &= |det(\mathbf{J}_{if})| \frac{dP_{\mathbf{Y}}(\mathbf{y}, \lambda)}{d\mathbf{J}_{if}} |_{\mathbf{y}=\mathbf{y}_i} \\ &+ P_{\mathbf{Y}}(\mathbf{y}, \lambda)|_{\mathbf{y}=\mathbf{y}_i} |det(\mathbf{J}_{if})| \, {\mathbf{J}_{if}^T}^{-1}. \end{aligned} \tag{3}$$

Since $f^t(.)$ is invertible at $\mathbf{x}^t$, $|det(\mathbf{J}_{if})| \neq 0$ and therefore

$$\mathbf{0} = \frac{d \log P_{\mathbf{Y}}(\mathbf{y}, \lambda)}{d\mathbf{J}_{if}} |_{\mathbf{y}=\mathbf{y}_i} + {\mathbf{J}_{if}^T}^{-1}. \tag{4}$$

But

$$\begin{aligned} \frac{d \log P_{\mathbf{Y}}(\mathbf{y}, \lambda)}{d\mathbf{J}_{if}} &= \sum_{j=1}^{n} \frac{\partial \log P_{\mathbf{Y}}(\mathbf{y}, \lambda)}{\partial y_j} \frac{\partial y_j}{\partial \mathbf{J}_{if}} \\ &+ \sum_{k=1}^{K} \frac{\partial \log P_{\mathbf{Y}}(\mathbf{y}, \lambda)}{\partial \lambda_k} \frac{\partial \lambda_k}{\partial \mathbf{J}_{if}}, \end{aligned} \tag{5}$$

and therefore

$$\begin{aligned} &\sum_{k=1}^{K} \frac{\partial \log P_{\mathbf{Y}}(\mathbf{y}, \lambda)}{\partial \lambda_k} \frac{\partial \lambda_k}{\partial \mathbf{J}_{if}} |_{\mathbf{y}=\mathbf{y}_i} \\ &= -\sum_{j=1}^{n} \frac{\partial \log P_{\mathbf{Y}}(\mathbf{y}, \lambda)}{\partial y_j} \frac{\partial y_j}{\partial \mathbf{J}_{if}} |_{y_j=y_{ij}} - {\mathbf{J}_{if}^T}^{-1}. \end{aligned} \tag{6}$$

By writing the vector function $\mathbf{y}_i = f_i(\mathbf{x})$ at $\mathbf{x} = \mathbf{0}$ in terms of its vector Taylor series expansion

$$f_i(\mathbf{0}) = f_i(\mathbf{x}) - \mathbf{J}_{if}\mathbf{x} + O(\mathbf{x}^2), \tag{7}$$

where $O(\mathbf{x}^2)$ is the sum of terms of second order degree or higher in $\mathbf{x}$.

Taking the partial derivative of both sides with respect to $\mathbf{J}_{if}$, and substituting into Equation 6, we get

$$\begin{aligned} \sum_{k=1}^{K} \frac{\partial \log P_{\mathbf{Y}}(\mathbf{y}, \lambda)}{\partial \lambda_k} \frac{\partial \lambda_k}{\partial \mathbf{J}_{if}} |_{\mathbf{y}=\mathbf{y}_i} &= -\frac{\partial \log P_{\mathbf{Y}}(\mathbf{y}, \lambda)}{\partial \mathbf{y}} |_{\mathbf{y}=\mathbf{y}_i} \mathbf{x}^T \\ &\quad -{\mathbf{J}_{if}^T}^{-1}. \end{aligned} \tag{8}$$

Since Equation 8 is valid for every realization of the random vector $\mathbf{x}$ at which $f(.)$ is invertible, then

$$\begin{aligned} \sum_{k=1}^{K} \frac{\partial \log P_{\mathbf{Y}}(\mathbf{y}, \lambda)}{\partial \lambda_k} \frac{\partial \lambda_k}{\partial \mathbf{J}_{if}^t} |_{\mathbf{y}=\mathbf{y}_i^t} &= -\frac{\partial \log P_{\mathbf{Y}}(\mathbf{y}, \lambda)}{\partial \mathbf{y}} |_{\mathbf{y}=\mathbf{y}_i^t} {\mathbf{x}^t}^T \\ &\quad -{\mathbf{J}_{if}^t}^{T^{-1}} \qquad \forall\, 0 \le t < N. \end{aligned} \tag{9}$$

Equation 9 proves the proposition. For a nonlinear feature transformation, the Jacobian matrix of the transformation is a function of the values of the feature vectors. This makes the estimation of the Jacobian matrix at each realization of a high-dimensional input feature vector computationally expensive. A significant reduction in the computational complexity can be achieved by noticing two properties of our problem. First, we are interested in small perturbations in the training data. This is equivalent to stating that the map $f(.)$ is close to the identity map. Second, we are interested in calculating the likelihood of the original feature vector $\mathbf{x}$ not in calculating the likelihood of the observed feature vector $\mathbf{y}$. This special case motivates using a variational approach to reduce the problem to estimating the local change in the values of the likelihood when the Jacobian matrix is very close to the identity matrix, i.e. $J_f \approx I$. This important special case is covered by the following lemma

Lemma: Let $\mathbf{x}$ be a random vector in $\Re^n$, and let $\mathbf{y} = f(\mathbf{x})$ be a continuously differentiable map of the random vector $\mathbf{x} \in \Re^n$ to $\mathbf{y} \in \Re^n$ such that $J_f \approx I$, and $P_{\mathbf{Y}}(\mathbf{y}, \lambda)$ be the probability density function of $\mathbf{y}$. The following relation is satisfied for every realization of the random vector $\mathbf{x}$

$$\begin{aligned} &\sum_{k=1}^{K} \frac{\partial \log P_{\mathbf{Y}}(\mathbf{y}, \lambda)}{\partial \lambda_k} \frac{\partial \lambda_k}{\partial \mathbf{J}_f^t} |_{\mathbf{y}=\mathbf{y}^t} \\ &\approx -\frac{\partial \log P_{\mathbf{Y}}(\mathbf{y})}{\partial \mathbf{y}} |_{\mathbf{y}=\mathbf{y}^t} {\mathbf{y}^t}^T - \mathbf{I}, \end{aligned} \tag{10}$$

where $\mathbf{y}^t$ is the tth realization of the random vector $\mathbf{y}$, $\mathbf{I}$ is the identity matrix. The proof of this lemma is straightforward by taking the limit of $\mathbf{x} \rightarrow \mathbf{y}$ in Proposition 1.

In this work, we suggest estimating the parameters of the HMM model which maximize the weighted sum of the log likelihood function and the negative of the sum of the squares of the Frobenius norm, the l_2 norm of the vector representation, of the matrix $\mathbf{D^t} = [d_{ij}^t]$ where

$$d_{ij}^t = \frac{\partial \log P_{\mathbf{Y}}(\mathbf{y})}{\partial y_i}|_{\mathbf{y}=\mathbf{y}^t} y_j^t + \delta_{ij}, \forall\, 0 \leq i < n, 0 \leq j < n, \tag{11}$$

where n is the length of the random vector $\mathbf{x}$,

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

So our objective function is

$$O = \sum_{t=1}^{N} L^t + \zeta \sum_{t=1}^{N} \parallel vec[\mathbf{D}^t] \parallel_2^2, \tag{12}$$

where L^t is an estimate of the log likelihood of the tth realization, N is the number of realizations, $vec[\mathbf{D}^t]$ is the vector representation of the matrix $\mathbf{D}^t$, and ζ is a negative constant.

3. IMPLEMENTATION

In the previous section, we showed that by using a variational approach, the sensitivity of the model to the small perturbations in the training data can be reduced by adding a penalty term to the log likelihood objective function. This section therefore develops an algorithm which estimates the model parameters to maximize the modified objective function in Equation 12. The resulting algorithm may be considered a generalization of the expectation-maximization (EM) algorithm. Experiments in Section 4 show that the algorithm generates models which outperform those estimated using a maximum likelihood objective function. The variational approach presented in the last section is a good compromise between the two extremes of classical maximum likelihood estimation with its simplicity and computational efficiency but inadequacy in many applications, and general approaches taking Proposition 1 into consideration with the computational complexity associated with calculating the Jacobian matrix for every realization.

We explain in this section how the parameters of the HMM model can be optimized to maximize the objective function in Equation 12. We use the expectation maximization (EM) algorithm to estimate the parameters after adding the penalty term in Equation 12 to the auxiliary function in the maximization step, i.e.

$$Q(\theta|\theta^k) = E[\log P(\mathbf{s}, \mathbf{y}|\theta)|\mathbf{y}, \theta^{\mathbf{k}}] + \zeta \sum_{t=1}^{T} \parallel vec[\mathbf{D}^t] \parallel_2^2, \tag{13}$$

where $Q(\theta|\theta^k)$ is the modified auxiliary function after the kth iteration, and $\log P(\mathbf{s}, \mathbf{y}|\theta)$ is the complete data log likelihood. Initially the values of the model parameters θ are unknown, so we choose an initial value of the model parameters and then we estimate the modified auxiliary function, and the parameters are updated to maximize the modified auxiliary function. This sequence is repeated until a local maximum of the empirical objective function is found.

The update equations of the mean and the variance vectors of each Gaussian component of the HMM Gaussian mixture model can be shown to be

$$\mu_{ik} = \frac{\sum_{t=1}^{T} \gamma_k^t \left(y_i^t + \zeta r_i^t\right)}{\sum_{t=1}^{T} \gamma_k^t}, \tag{14}$$

$$\sigma_{ik}^2 = \frac{\sum_{t=1}^{T} \gamma_k^t y_i^t \left(y_i^t + 2\zeta r_i^t\right)}{\sum_{t=1}^{T} \gamma_k^t} - \mu_{ik}^2, \tag{15}$$

where

$$r_i^t = \sum_{k=1}^{K} \sum_{j=1}^{n} \gamma_k^t \frac{\left(\mu_{ik} - y_i^t\right)}{\sigma_{ik}^2} {y_j^t}^2, \tag{16}$$

μ_{ik} is the ith component of the mean vector of the kth Gaussian component, γ_k^t is the posterior probability of the kth Gaussian component at time t, y_i^t is the ith component of the observation vector at time t, and ζ is a negative constant.

4. EXPERIMENTS

We apply the variational approach to robust maximum likelihood (VARML) estimation described in the previous section to three large vocabulary broadcast news ASR systems. The first system is a speaker-independent English broadcast news system and the other two are speaker-adapted Arabic broadcast news systems. The main difference between the two Arabic systems is the phoneme set used. The phonetic transcription in Arabic requires the existence of certain diacritic symbols which are usually not found in text transcriptions, and correspond to short vowels. The unvowelized system models only the graphemes and does not model the diacritic symbols, while the vowelized system models the diacritic symbols and therefore models the short vowels in Arabic explicitly. For the three systems, each phoneme is represented by 3 HMM states with left-to-right topology with the exception of modeling short vowels with 2 states in the Arabic vowelized system. For the three systems, the raw features are 13-dimensional PLP features computed every 10 ms. from 25-ms. frames. The recognition features are computed from the raw features by splicing together nine frames of raw features (± 4 frames around the current frame), projecting the 117-dim. spliced features to 40 dimensions using a linear discriminant analysis (LDA) projection, and then applying maximum likelihood linear transformation (MLLT) to the 40-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.

For the English broadcast news system, the acoustic model consists of 6000 context-dependent states and 250K diagonal-covariance Gaussian mixtures. The baseline system is trained using maximum likelihood estimation. We use this baseline system as an initialization to the VARML training. The language model is a 72K vocabulary interpolated back-off 4-gram language model. We test the performance of our system on the DARPA rich transcription English broadcast news evaluation data for 2003 (RT03) and 2004 (RT04) and development data for 2004 (DEV04).

As shown in Table 1, the word error rate (WER) has been reduced significantly by using the VARML objective function to

System	RT03	RT04	DEV04
Baseline	12.9	20.5	23.5
VARML	12.4	19.8	22.7

Table 1. Word error rates (%) on the English Broadcast news RT03, RT04, and DEV04 evaluation data.

System	RT04	BNAT05
Baseline	17.2	18.9
VARML	16.6	17.8

Table 2. Word error rates (%) on the Arabic RT04 and BNAT05 evaluation data using the unvowelized system

estimate the model parameters and consistently over the three test sets.

The decoding for both speaker-adapted Arabic broadcast news systems consists of two passes: the first speaker-independent pass output is used to adapt the models, and the second decoding pass uses the adapted models to generate the lattices used to generate the final output of the decoder. In the context of speaker-adaptive training to produce canonical acoustic models, we use vocal tract length normalization (VTLN), and feature-space MLLR. We do also a single pass of MLLR adaptation, using a regression tree to generate transforms for different sets of mixture components.

For the unvowelized Arabic broadcast news system, the acoustic model consists of 5032 context-dependent states and 400K diagonal-covariance Gaussian mixtures. The language model is a 617K vocabulary interpolated back-off 4-gram language model. The vowelized acoustic model consists of 4006 context-dependent states and 400K diagonal-covariance Gaussian mixtures. We report results of the baseline and the VARML systems on both the Arabic DARPA 2004 Rich Transcription (RT04) evaluation data and the 2005 Arabic broadcast news tune test set (BNAT05).

As shown in Table 2, the WER results for the unvowelized system improved by 4.7% relative for the RT04 evaluation data and by 4.8% for the BNAT05 evaluation data compared to the baseline by using the VARML estimation of the model parameters. As shown in Table 3, the WER results for the vowelized system improved by 4.7% relative for the RT04 evaluation data and by 4.9% for the BNAT05 evaluation data compared to the baseline by using the VARML estimation of the model parameters.

System	RT04	BNAT05
Baseline	14.9	16.4
VARML	14.2	15.6

Table 3. Word error rates (%) on the Arabic RT04 and BNAT05 evaluation data using the vowelized system

5. RESULTS AND DISCUSSION

This work proposes a variational approach to regularized maximum likelihood estimation of the HMM parameters for speech recognition systems. This approach subtracts a measure of the sensitivity of the model to small variations in the training data from the log likelihood objective function. The proposed VARML objective function is used by an EM-based algorithm to estimate the parameters of the model. Our approach is applied to three large vocabulary broadcast news recognition tasks: a speaker-independent English broadcast news, a speaker-adapted unvowelized Arabic broadcast news, and a speaker-adapted vowelized Arabic broadcast news. These experiments show consistent improvement in recognition accuracy achieved by VARML estimation of the parameters compared to ML estimation. Our experiments show that the VARML objective function tends to increase the variance of the Gaussian components with small number of observations compared to the ML objective function. Most of the Gaussian components with sufficiently high number of observations have smaller values of variance when estimated using a VARML objective function compared to an ML objective function.

6. REFERENCES

[1] Puming Zhan, and Martin Westphal, "Speaker Normalization Based On Frequency Warping," *IEEE Proceedings of ICASSP*, Munich, Germany, 1997.

[2] Mark J. F. Gales, *Maximum Likelihood Linear Transformation For HMM-Based Speech Recognition,* Technical Report, Cambridge University Engineering Department, May 1997.

[3] Jerome H. Friedman, and Peter Hall, "On Bagging and Nonlinear Estimation," *Journal of Statistical Planning and Inference*, 2007, pp. 669–683.

[4] J. A. Hartigan, *Bayes Theory,* Springer-Verlag, New York, 1983.

[5] Onureena Banerjee, Alexandre d'Aspermont, and Laurent El Ghaoui, "Sparse Covariance Selection via Robust Maximum Likelihood Estimation," *ArXiv Computer Science e-prints*, June 2005.

[6] Stella Seitz, Peter Schnider, and Matthias Batelmann, "Entropy-Regularized Maximum Likelihood Cluster Mass Reconstruction," *ArXiv Computer Science e-prints*, March 2003.

[7] Athanasios Papoulis, *Probability, Random Variables, and Stochastic Processes,* McGraw-Hill, New York, 1991.