On a Model-Robust Training Method for Speech Recognition

ARTHUR NÁDAS, DAVID NAHAMOO, MEMBER, IEEE, AND MICHAEL A. PICHENY, MEMBER, IEEE

Abstract—We are interested in comparing training methods for designing better decoders. We treat the training problem as a statistical parameter estimation problem. In particular, we consider the conditional maximum likelihood estimate (CMLE)—the value of unknown parameters which maximizes the conditional probability of words given acoustics during training. We compare it to the maximum likelihood parameter which maximizes the conditional probability of words given estimate (MLE)—the estimate obtained by maximizing the joint probability of the words and acoustics. For minimizing the decoding error rate of the ("optimal") maximum a posteriori probability (MAP) decoder, we show that the CMLE (or maximum mutual information estimate, MMIE) may be preferable when the model is incorrect and, in this sense, the CMLE/MMIE appears more robust than the MLE.

INTRODUCTION

The maximum likelihood estimator (MLE) is a well-known general-purpose statistical parameter estimator. The maximum mutual information estimator (MMIE), proposed by Mercer [4], is a special-purpose estimator designed for estimating the statistical parameters of a decoder (classifier) assuming that a completely specified probabilistic model of the language is available. Under this assumption, the MMIE is equivalent to the conditional maximum likelihood estimator (CMLE) which we discuss below. The CMLE is an example of estimates referred to in the statistical literature as partial likelihood estimates; see, e.g., Cox [2], and Kotz and Johnson [3]. Bahl, Brown, de Souza, and Mercer [1] give empirical speech recognition results in which the CMLE/MMIE leads to fewer errors in decoding than the MLE does. In Nadas [5], we compared mathematically the MLE and the CMLE/MMIE under the hypothesis that the functional form produced by a talker while uttering the word w. It is assumed that the function $p_a$ is completely known except for the value of the parameter vector $\theta$. This knowledge is to be used extensively both during training, i.e., estimating the parameter $\theta$ from training data $x$, and during decoding, i.e., choosing a word $w(a)$ in deciding which word $w$ gave rise to a given acoustic feature vector $a$. We regard the training data $X = (w, a)_{i \in 1, \ldots, m}$ as the realization of a random sample $X$ which consists of a sample of i.i.d. pairs $X = (W, A)_{i \in 1, \ldots, m}$.

For training, the form of the function $p_a(w, a)$, together with some statistical principle such as least squares or maximum likelihood, or Bayes' or moment methods, etc., will determine the form of the estimator $\hat{\theta}$ of $\theta$. In our case, the principles are MLE and CMLE/MMIE. Having obtained an estimate $\hat{\theta}$, we may define the estimated MAP decoder:

$$\hat{w}(a) = \arg \max_w p_F(w | a) = \arg \max_w p_F(w, a),$$

i.e., we choose a most probable word given the acoustic information where probabilities are computed either from the estimated conditional or estimated joint probability elements (the factor $p_F(a)$ does not effect the maximization over $w$). While the exact MAP decoder, based on $p_F(w | a)$, minimizes decoding error rate, the estimated MAP decoder $p_F(w | a)$ is, in general, suboptimal.

Let $W$ be a randomly chosen word and let $A$ denote an utterance of that word. For simplicity, we consider speech
recognition for isolated words where the training data \( (w_i, a_i) \), \( i=1, \ldots, N \) may be regarded as a sequence of \( N \) independent samples from the joint distribution \((I, a)\), but the thrust of our remarks applies to the general case of continuous speech recognition.

We assume that the marginal distribution of words does not depend on the unknown \( \theta \) and is therefore a known distribution. Then (1) can be rewritten as

\[
p_b(w, a) = p_b(a | w) p(w)
\]

so that \( \theta \) from here on parametrizes only the acoustic channel model. In this special case, \( p_b(w) = p(w) \) is a constant with respect to a maximization over \( \theta \) so the CMLE/MMIE can be written as

\[
\hat{\theta} = \arg\max_{\theta} \prod_{i=1}^{N} p_b(w_i | a_i) = \arg\max_{\theta} \sum_{i=1}^{N} ML_b(w_i, a_i)
\]

(3)

where

\[
ML_b(w, a) = \log \frac{p_b(w | a)}{p(w)}, \quad (4)
\]

the reason for the information theoretic name MMIE in this special case. The maximum likelihood estimate is defined in the usual way as

\[
\hat{\theta} = \arg\max_{\theta} \prod_{i=1}^{N} p_b(w_i, a_i) = \arg\max_{\theta} \prod_{i=1}^{N} p_b(a_i | w_i),
\]

(5)

the latter equality follows from the constancy of \( p(w) \) in \( \theta \) for our special case.

We note that the CMLE and the MAP decoder maximize the same quantity, however, maximization is done over different arguments and by using different data. Observe also that MMIE is different from CMLE in the general case.

**ROBUSTNESS**

Suppose that, unknown to the experimenter, the probability element of \((W, A)\) is not \( p_b(w, a) \) for any \( \theta \) whatsoever, but it is some other unknown probability element \( q(w, a) \). An estimator of \( q \) based on the form of \( p_b(w, a) \) is considered to be robust if its performance is acceptable in spite of the fact that \( q \) is not one of the \( p_b \). Let \( \Theta \) denote the parameter space and let

\[
P = \{ p_b : \theta \in \Theta \}
\]

(6)

and we shall compare the MLE to the CMLE/MMIE with respect to language model robustness in the context of a multinomial model for acoustics, i.e., assuming a finite alphabet for the acoustic utterance \( a \).

We shall see that under these assumptions for sufficiently large sample sizes, the CMLE/MMIE is robust against errors in language modeling but the MLE is not.

We shall study language model robustness as follows. There is available to us a language model, i.e., a completely known probability distribution \( p(w) \) so that the unknown parameter \( \theta \) is needed only for defining the acoustic channel probabilities \( p_b(a | w) \). Thus, we redefine the set of all possible distributions as

\[
P = \{ p_b(a | w) p(w) : \theta \in \Theta \}
\]

(6′)

We shall obtain the estimated MAP decoders based first on the MLE \( \hat{\theta} \) and then on the CMLE/MMIE \( \hat{\theta} \). We then imbed \( P \) in \( Q \) by extending the parameter space \( \Theta \) to the Cartesian product of \( \Theta \) and \( \Gamma \) where

\[
\Gamma = \{ \gamma : \gamma \text{ is a positive probability distribution for words} \}
\]

(7)

\( Q \) is then constructed as

\[
Q = \{ p_\theta \gamma : \theta \in \Theta, \gamma \in \Gamma \}
\]

(8)

In this setup, we have \( p(w) = \gamma_0(w) \) for some \( \gamma_0 \in \Gamma \). Finally, we shall drop the hypothesis that \( p(w) = \gamma_0 \) is the true language model and compare the performances of decoders based on the MLE and on the CMLE/MMIE when the true language model is \( p(w) = \gamma(w) \) for an arbitrary \( \gamma \in \Gamma \).

**LANGUAGE MODEL ROBUSTNESS OF MLE AND CMLE/MMIE**

Consider the \( m \times n \) contingency table of words and acoustics which arises when the acoustic utterances for words are quantized to an alphabet of size \( n \) with \( n > m \). Since we know that the language model is \( \gamma_0 \), we have \( p_b(a | w) p(w) = \theta(a | w) \gamma_0(w) \) where we have put \( p_b(a | w) = \theta(a | w) \). For the acoustic outcome \( a \), the MAP decoder prefers word \( w \) to word \( w' \) if

\[
p_b(w, a) > p_b(w', a), \quad (9)
\]

i.e., if

\[
\frac{\theta(a | w)}{\theta(a | w')} > \frac{\gamma_0(w)}{\gamma_0(w')}. \quad (10)
\]

Thus, an estimated MAP decoder using estimator \( \hat{\theta} \) prefers \( w \) to \( w' \) if

\[
\frac{\hat{\theta}(a | w)}{\hat{\theta}(a | w')} > \frac{\gamma_0(w)}{\gamma_0(w')}, \quad (11)
\]

where the right-hand side of (10) is a known constant. Clearly, all parameter vectors \( \theta \) for whose components \( \theta(a | w) \) satisfy the inequalities (10) will choose the same decoder. We call all parameter vectors \( \theta \) that choose the same decoder **equivalent**. The critical property of any es-
timator is its ability to find an estimated decoder close to the optimal decoder; in the case of sufficiently large sample size, this means actually finding the exact optimal MAP decoder, i.e., producing an estimate of \( \theta \) which falls in the same equivalence class as the true \( \theta \).

Let \( N_{w,a} \) be the number of occurrences of the pair \((w, a)\) among the \( N \) pairs in the training data; these \( mn \) counts are a sufficient statistic for \( \theta \). Taking derivatives of \( \Sigma_{k=1}^{N} \log p_{\theta}(a \mid w) + \log \gamma_{0}(w) \), with respect to \( \theta \), it is easy to find that the MLE \( \theta' = \hat{\theta} \) is given by

\[
\hat{\theta}(a \mid w) = \frac{N_{w,a}}{N_{w}} \gamma_{0}(w) \tag{12}
\]

where \( N_{w} = \sum_{a=1}^{m} N_{w,a} \) is assumed positive. Observe that the only use the MLE makes of the assumed language model \( p(w) = \gamma_{0}(w) \) is the fact that the language model is known; the known value \( \gamma_{0} \) is not utilized. The constraint that for fixed \( w \) the \( \theta(a \mid w) \) be nonnegative and add to unity is obviously satisfied, and the MLE of the joint probability \( p_{\theta}(a \mid w) p(w) \) is

\[
\frac{N_{w,a}}{N_{w}} \gamma_{0}(w). \tag{13}
\]

Thus, the MLE chooses a decoder [see (11)] making all comparison of the form

\[
\frac{N_{w,a}}{N_{w}} \gamma_{0}(w) \equiv \frac{N_{w,a}}{N_{w}} \gamma_{0}(w'), \tag{14}
\]

using the "known" language model \( \gamma_{0} \) in a significant way.

The CMLE/MMIE \( \theta' = \hat{\theta} \) does not have such a simple closed form. Let \( \lambda_{w} \) be Lagrange multipliers for \( w = 1, \ldots, m \). The Lagrangian for maximizing the conditional likelihood is

\[
\sum_{w=1}^{m} \sum_{a=1}^{n} N_{w,a} \log \left( \frac{\theta(a \mid w) \gamma_{0}(w)}{\sum_{k=1}^{m} \gamma_{0}(k) \theta(a \mid k)} \right) + \sum_{w=1}^{m} \lambda_{w} \left( 1 - \sum_{a=1}^{n} \theta(a \mid w) \right). \tag{15}
\]

Setting the derivative of (15) with respect to \( \theta \) equal to zero, we find that \( \lambda_{w} \theta(a \mid w) = 0 \), hence, that if \( \lambda_{w} \equiv 0 \) is a solution, then the solution for the CMLE/MMIE \( \hat{\theta} \) may be obtained by unconstrained maximization just as the MLE was, and it must satisfy the linear equations

\[
\hat{\theta}(a \mid w) = \frac{N_{w,a}}{N_{w}} \gamma_{0}(w) \tag{16}
\]

where \( N_{w} = \sum_{a=1}^{m} N_{w,a} \). The Lagrange multipliers will vanish so long as the training counts \( N_{w,a} \) are not grossly inconsistent with \( \gamma_{0} \), and we regard any solution of (16) as the CMLE/MMIE.

This time, the CMLE/MMIE chooses a decoder making all comparisons of the form

\[
\frac{N_{w,a}}{N_{w}} \sum_{k=1}^{m} \gamma_{0}(k) \frac{\theta(a \mid k)}{\gamma_{0}(w)} \gamma_{0}(w) \tag{17}
\]

i.e., after cancelations the comparisons

\[
N_{w,a} \equiv N_{w,a}. \tag{18}
\]

Observe that here the decoder is chosen without reference to the "known" language model. Note also that while the estimate \( \hat{\theta} \) may not be uniquely defined, there is nevertheless no ambiguity in the decoder chosen by any \( \hat{\theta} \) which satisfies (16).

It is clear from the foregoing and the law of large numbers that if the true \( \gamma \) is not equivalent to the "known" language model \( \gamma_{0} \), then for sufficiently large sample size the MLE will choose a suboptimal decoder, but in contrast the CMLE/MMIE will, with probability approaching unity, find the correct (optimal) MAP decoder.

The decoder based on the estimate \( \hat{\theta} \) is a function of the observable acoustic data \( a \) defined by

\[
d_{\hat{\theta}}(a) = \arg\max_{w} p_{\theta}(a \mid w). \tag{19}
\]

The probability that the estimated decoder decodes correctly, i.e., the decoding rate associated with an estimator \( \hat{\theta} \) is the random variable

\[
r(\hat{\theta}, \theta, \gamma_{0}, \gamma) = \text{Prob} \{ \text{all pairs } (w, a) \text{ such that } w = d_{\hat{\theta}}(a) \} = \sum_{a=1}^{n} \theta(a \mid d_{\hat{\theta}}(a)) \gamma(d_{\hat{\theta}}(a)). \tag{20}
\]

Averaging over the possible training data \( X = (w_{1}, a_{1})_{i=1}^{n} \), we have that the expected decoding rate is

\[
R(\hat{\theta}, \theta, \gamma_{0}, \gamma, N) = E_{\gamma_{0}, N} r(\hat{\theta}, \theta, \gamma_{0}, \gamma) \tag{21}
\]

where \( E_{\gamma_{0}, N} \) means averaging with respect to the joint distribution of the training data \( \Pi^{{\gamma_{0}}}(a_{i} \mid w_{i}) \gamma(w_{i}) \) with \( \theta \) the true acoustic channel model, \( \gamma \) the true language model, and \( N \) the number of words uttered.

Let \( S_{\theta}(\gamma_{0}, \gamma) \) be the subset of the parameter space \( \Theta \) where the CMLE/MMIE is better than the MLE for an available language model \( \gamma_{0} \) and a true but unknown language model \( \gamma \)

\[
S_{\theta}(\gamma_{0}, \gamma) = \{ \theta: \theta \in \Theta \text{ and } R(\hat{\theta}, \theta, \gamma_{0}, \gamma, N) > R(\tilde{\theta}, \tilde{\theta}, \gamma_{0}, \gamma, N) \}. \tag{22}
\]

In the correct language model case when \( \gamma_{0} = \gamma \), the efficiency of the MLE can be rephrased by saying that the
The decoding rate versus sample size for selected values of the acoustic channel probabilities is shown in Fig. 1. The decoding rate versus the first acoustic channel probability for selected values of the second acoustic channel probability and the language model limit, \( \gamma \), is illustrated in Fig. 2. The decoding rate versus the language model for selected values of the acoustic channel probabilities is depicted in Fig. 3. The limits as \( \gamma \to \infty \) vanish. On the other hand, in case of an incorrect language model, i.e., whenever \( \gamma_0 \) and \( \gamma \) are not equivalent, the above infinite sample size robustness result can be rephrased by saying that \( \lim_{\gamma \to \infty} S_N(\gamma_0, \gamma) \) is all of \( \Theta \). We have studied a small numerical example in detail. The graphs plotted below summarize some numerical results that suggest that, for a finite sample size \( N \), the set \( S_N(\gamma_0, \gamma) \) is a nonempty proper subset of \( \Theta \) whenever \( \gamma_0 \) and \( \gamma \) are not equivalent; moreover, in this case, the numerical evidence suggests that \( S_N(\gamma_0, \gamma) \) converges to \( \Theta \) as the sample size \( N \) increases.

Here is the example. Suppose there are only two words and only two acoustic outcomes \( m = n = 2 \). Assume that the known language model is \( \gamma_0 = (0.5, 0.5) \) and let the correct language model be denoted by \( \gamma = (G, 1 - G) \).

Denote the acoustic channel probabilities by \( \theta(1|1) = p_1, \theta(1|2) = p_2 \). This is the smallest nontrivial example possible. By "rate" we mean the expected probability of correctly decoding via the trained decoder. In Fig. 1, the rate is plotted against sample size for selected values of the acoustic channel probabilities while the true language model is \( \gamma = (G, 1 - G) = (0.9, 0.1) \). In Fig. 2, the rate is plotted against the first acoustic channel parameter for various second acoustic channel parameters and language models while the sample size is fixed at 45. In Fig. 3, the rates are plotted against the true language model parameter. To avoid trivialities, all computations were conditioned on a count of at least one for each of the four possible outcomes so that the smallest interesting sample size is five. The results were obtained by integrating with respect to this true conditional four-category multinomial distribution.

**CONCLUSION**

We have constructed a simple example in which the CMLE/MMIE is more robust against errors of language modeling than the MLE. While the MLE may be preferable for use with well-fitting models, the CMLE/MMIE appears safer for the training of speech recognizers when the model fit is poor.

Observe that even if the model is correct, it is not known that the MLE is superior to CMLE/MMIE for a fixed finite sample size and all possible true parameters; this is so in spite of our small numerical example and in spite of the asymptotic result in Nadas [5]. Note that when the model is correct, the MLE decoder (14) uses the exact correct language model probabilities while the CMLE/MMIE does not. This is another heuristic in favor of the MLE in the correct model case.

When the model is incorrect, our results indicate that the MLE decoder converges with increasing sample size to a suboptimal decoder but the CMLE/MMIE converges to the optimal decoder. In the case of incorrect model, our small numerical example also hints that the CMLE/MMIE gains over the MLE with increasing sample size.

**REFERENCES**


Arthur Nadas was born in Budapest, Hungary, in 1934. He received the B.A. degree in mathematics from Alfred University, Alfred, NY, in 1959, the M.A. degree in mathematics from the University of Oregon, Eugene, in 1961, and the Ph.D. degree in mathematical statistics from Columbia University, New York, NY, in 1967. He joined the IBM Corporation in 1961 at the Product Testing Laboratory in Poughkeepsie, NY. At this time he is working in speech recognition as a Research Staff Member at the IBM T. J. Watson Research Center, Yorktown Heights, NY.

David Nahamoo (S’78–M’81) was born in Hamadan, Iran, in 1953. He received the B.S. degree from Tehran University, Iran, the M.S. degree from Imperial College of London, England, and the Ph.D. degree from Purdue University, West Lafayette, IN, all in electrical engineering, in 1975, 1976, and 1982, respectively, working on algorithms for X-ray tomography, ultrasonic diffraction, and echo imaging. Since 1982 he has been a member of the Continuous Speech Recognition Group at the IBM Thomas J. Watson Research Center, Yorktown Heights, NY, where he is currently the Manager of the Speech Recognition Modeling Group. His recent work includes the development of algorithms for improving the training of hidden Markov models and extracting robust acoustic feature vectors for speech recognition. His current interests are speaker adaptation, improved parameterization, and context dependent modeling.

Michael A. Picheny (S’73–M’81) was born in New York, NY, on July 2, 1954. He received the S.B., S.M., and Sc.D. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, in 1975, 1978, and 1981, respectively, working on aids for people with hearing impairments. Since 1981 he has been a member of the Continuous Speech Recognition Group at the IBM Thomas J. Watson Research Center, Yorktown Heights, NY. He has worked on improving both the acoustic processing and modeling components of speech recognition systems, and is particularly interested in robust acoustic processing and context-dependent acoustic modeling. He is currently the Manager of the Acoustic Processing Group. Dr. Picheny is the Associate Editor for Speech Processing for the ASSP Transactions.