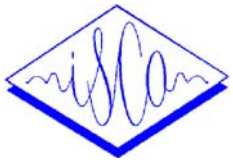

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

# A NOVEL DISCRIMINATIVE METHOD FOR HMM IN AUTOMATIC SPEECH RECOGNITION

*Jian WU, Qing GUO*
Speech Lab., Dept. of Computer Science and Technology,
Tsinghua Univ., Beijing, 100084, P.R.China
Tel: +86-10-6277 2001, FAX: +86-10-6277 2001, E-mail: jwu@sp.cs.tsinghua.edu.cn

## ABSTRACT

A novel discriminative method for estimating the parameters of Hidden Markov Models (HMMs) is described. In this method, the parameter values are chosen to ensure that the characteristics of each sound class can be maximally separated. Compared with the significant method known as the Maximum Mutual Information (MMI) estimation, the novel method represented in this paper adopts a new kind of criteria called MSDI (Maximum Samples Distinction Information). It parries many computational problems in estimating iteration. The experimental results show that the hit rate can be raised by about 6 percent compared with the MLE (Maximum Likelihood Estimation), which is similar to the other experimental results based on MMI training.

Keywords: HMM, MLE, MMIE, MSDIE



## 1. INTRODUCTION

Hidden Markov Models (HMMs) are now a standard in recent speech recognition systems [1]. The Markov model used in speech recognition is often considered as a finite state machine which makes a transition from one state to its immediate right state. The other characteristic of HMM is that only one state can be entered each time, and an observation vector is generated according to a probability density function (PDF) associated with that state. Hence, the likelihood of generating an utterance using any sequence of states can be computed. For training HMMs, the use of maximum likelihood estimation (MLE) is widespread. However, HMM is not the true distribution form for speech utterances despite its prevalence. The reason is that HMM has some fundamental assumptions:

(1) The knowledge of the form of the data distribution is well known;

(2) The training data are always sufficient for parameter estimation;

(3) The performance of the recognizer will not get worse when the trained model comes closer to the optimal model.

Unfortunately, these conditions will never be satisfied in real system. In order to reconcile the contradictions, many new approaches based on the concept of discrimination for speech recognizer design are considered. L.R.Bahl [2] proposes the Maximum Mutual Information (MMI) training, which maximizes the mutual information between the text and the acoustic samples. The experimental results show that this change makes the performance of speech recognizer more robust.

This paper begins with a review of Maximum Likelihood Estimation and Maximum Mutual Information Estimation. In the third section, our new discriminative training algorithm, which is called Maximum Sample Distinctive Information Estimation, will be proposed. Then the experimental results are reported in Section 4.

## 2. TRADITIONAL TECHNIQUES OF PARAMETER ESTIMATION

According to the information theory and the encountered problems in speech recognition [3], the optimal recognizer is the one that, given some acoustic utterances $S(=S_1S_2\cdots S_T)$, produces the text $\hat{M}$ such that:

$$P(\hat{M}|S)=\max_{M}\frac{P(S|M)\cdot P(M)}{P(S)} \qquad (1)$$

Since $P(S)$ is a prior probability that is not a function of the text M, it can be dropped from the maximization. We can design a speech recognizer consist in defining the acoustic model which is used to compute the probability $P(S|M)$ and the language model which is used to compute the probability $P(M)$. While the language model is obtained independently by statistics over the large text corpus that do not involve speech data, the acoustic modeling process is only to specify the model (distribution) λ most likely to produce the given sequence of observations. In order to get a best performance when utilizing formulation (1) in recognition process, many techniques are presented in parameter estimation.

### 2.1 Maximum Likelihood Estimation

During the training process, the parameters within the HMM system are chosen to optimise some criteria, given the training data. The most common practice is to use Maximum Likelihood Estimation which attempts to maximise the likelihood of generating the training data with the right model. The criteria can be expressed as follow,

$$O_{mle}(\lambda)=P(S|M,\lambda)\cdot P(M) \qquad (2)$$

Maximum Likelihood criteria can be achieved by using some kind of Expectation-Maximisation (EM) methods, such as B-W algorithm[4]. However, it should be noted that this optimisation criteria for MLE only considers the acoustic vectors of the class under consideration while training. Once the assumptions presented in Section 1 are not satisfied (in fact, those conditions will never be satisfied), the performance of the speech recognizer will be dropped dramatically.

### 2.2 Maximum Mutual Information Estimation

In information theory, Mutual Information is a kind of measurement units of information quantity. It can be expressed as the reduction of entropy, given the right tag of every observation. The mutual information measure is defined as

$$\begin{aligned} I(S:M) &= H(S) - H(S|M) \\ &= \sum_{S,M} P(S,M) \log \frac{P(S,M)}{P(S)P(M)} \end{aligned} \tag{3}$$

In order to maximize the mutual information, the criteria can be expressed as

$$\begin{aligned} O_{MMI}(\lambda) &= \log\left(\frac{P(S,M|\lambda)}{P(S)\cdot P(M)}\right) \\ &= \log \frac{P(S|M,\lambda)P(M)}{P(S)P(M)} \end{aligned} \tag{4}$$

Since $P(M)$ does not depend on the parameter λ, the formulation (5) can be derived from (4).

$$\begin{aligned} O_{MMI}(\lambda) &= \log P(S|M,\lambda) - \log P(S) \\ &= \log P(S|M,\lambda) - \log \sum_{M'} P(S|M',\lambda) \end{aligned} \tag{5}$$

Taking the deriative of $O_{MMI}(\lambda)$ with respect to parameter λ

$$\frac{\partial O_{MMI}(\lambda)}{\partial \lambda_i} = \frac{\partial P(S|M,\lambda) \big/ \partial \lambda_i}{P(S|M,\lambda)} - \sum_{M'} \frac{\partial P(S|M',\lambda) \big/ \partial \lambda_i}{\sum_{M'} P(S|M',\lambda)} \tag{6}$$

Compared with the criteria of MLE, the objective of MMIE is not only to increase the probability for the right word sequence, but also to decrease the probabilities for the other word sequences. Thus the MMI can be seen to be adding discrimination into the training process. The different results by MMIE and MLE is illustrated in Fig.1.

## 3. MAXIMUM SAMPLES DISTINCTION INFORMATION ESTIMATION

The MLE is both computationally efficient and gives a robust estimate when sufficient data are available. Nevertheless, it does not take into account other classes when making class distribution estimation. The MMIE can be used to maximally separate these classes, which has been proved by the experimental results [5][6]. However, since the MMI training criteria attempts to

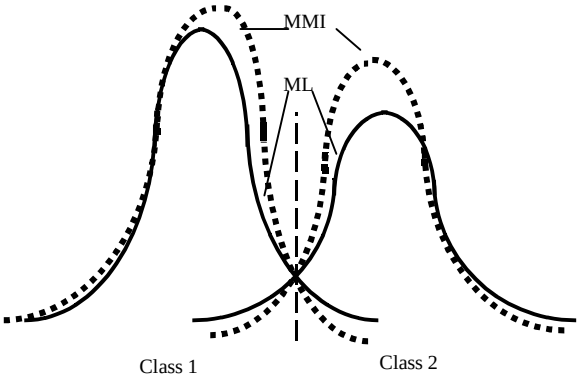


Fig.1 illustration of difference between MLE and MMIE

maximize the probability of the correct word sequence vs. all possible word sequences allowed by the language model, the application of MMI to parameter estimation requires the model parameters of all word sequences. It costs a lot of additional computation.

### 3.1 Maximum Samples Distinction Information

For the simple case illustrated in Fig.1, if the dividing line between classes were worked out, the similar effect could be achieved facilely by an efficient algorithm, which we called Maximum Samples Distinction Information Estimation (MSDIE).

First, the samples distinction information is defined as

$$SD(S,S':M) = \sum_{S,S',M} P(S,M) \log \frac{P(S,M)}{P(S',M)} \tag{7}$$

where S is the utterance of the word text M that can be classified rightly, $S'\ (= S'_1\, S'_2 \cdots S'_{T'})$ is the utterance of the word text M that would be classified incorrectly. We can prove that if the samples distinct information is maximized, the probability for the right word sequence will be increasing while the probabilities for the wrong word sequences are decreasing. According to the defination of samples distinction information, the criteria of MSDIE can be stated as

$$\begin{aligned} O_{MSDI}(\lambda) &= \log\left(\frac{P(S,M|\lambda)}{P(S',M|\lambda)}\right) \\ &= \log \frac{P(S|M,\lambda)P(M)}{P(S'|M,\lambda)P(M)} \\ &= \log P(S|M,\lambda) - \log P(S'|M,\lambda) \end{aligned} \tag{8}$$

Taking the deriative of $O_{MSDI}(\lambda)$ with respect to parameter λ, we can get the following expression

$$\frac{\partial O_{MMI}(\lambda)}{\partial \lambda_i} = \frac{\partial P(S|M,\lambda) \big/ \partial \lambda_i}{P(S|M,\lambda)} - \frac{\partial P(S'|M,\lambda) \big/ \partial \lambda_i}{P(S'|M,\lambda)} \tag{9}$$

### 3.2 Parameter Estimation

Let $a_{ij}$ denote the probability of transition from state *i*

to state *j* in our acoustic model that is based on HMM, $b_j(S_t)$ denote the probability of generating the vector $S_t$ while staying at state *j*, and $c_i$ be the initial probability of being in state *i*. we will have

$$P(S|M,\lambda)=\sum_{i_1}\cdots\sum_{i_T}c_{i_1}\prod_{t=1}^{T}a_{i_t i_{t+1}}\cdot b_{i_t}(S_t) \tag{10}$$

where the following constraints should be satisfied:

$$\sum_j a_{ij}=1,\quad \sum_y b_{ij}(y)=1,\quad \sum_i c_i=1.$$

Applying the criteria like formulation (9) in the parameter estimation of HMM, we can obtain that

$$\frac{\partial P(S\mid M,\lambda)}{\partial\lambda}=\sum_{t=1}^{T}\sum_{i,j}\alpha_i(t-1)a_{ij}\beta_j(t)\frac{\partial b_j(S_t)}{\partial\lambda} \tag{11}$$

and

$$\frac{\partial P(S'\mid M,\lambda)}{\partial\lambda}=\sum_{t=1}^{T'}\sum_{i,j}\alpha_i(t-1)a_{ij}\beta_j(t)\frac{\partial b_j(S'_t)}{\partial\lambda} \tag{12}$$

where $\alpha_i(t-1)$ and $\beta_j(t)$ are the forward and backward probabilitis in the B-W algorithm, respectively. So the formulation (9) turns into

$$\begin{aligned}\frac{\partial O_{MMI}(\lambda)}{\partial\lambda_i}&=\frac{\sum_{t=1}^{T}\sum_{i,j}\alpha_i(t-1)a_{ij}\beta_j(t)\frac{\partial b_j(S_t)}{\partial\lambda}}{P(S\mid M,\lambda)}\\&\quad-\frac{\sum_{t=1}^{T'}\sum_{i,j}\alpha_i(t-1)a_{ij}\beta_j(t)\frac{\partial b_j(S'_t)}{\partial\lambda}}{P(S'\mid M,\lambda)}\\&=\sum_{t=1}^{T}\sum_{j}\gamma_j(t)\left(\frac{\partial b_j(S_t)/\partial\lambda}{b_j(S_t)}\right)\\&\quad-\sum_{t=1}^{T'}\sum_{j}\gamma_j(t)\left(\frac{\partial b_j(S_t')/\partial\lambda}{b_j(S_t')}\right)\end{aligned} \tag{13}$$

If the pdf. at state *j* observes Mixture Gaussian Distribution (MGD), the output probability $b_j(S_t)$ can be formulated as

$$\begin{aligned}b_j(S_t)=&\sum_{i=1}^{M}p_i\cdot(2\pi)^{-d/2}\mid R_i\mid^{-1/2}\cdot\\&\exp\left(-\frac{1}{2}(S_{ti}-\overline{\mu_i})^T R_i^{-1}(S_{ti}-\overline{\mu_i})\right)\end{aligned} \tag{14}$$

where *N* is the number of mixture components, *d* is the dimensionality of the observation vectors.

### 3.3 Iterative Algorithm for HMM Training

In computing the MSDI criteria and its associated maximization, one needs to divide training data into two disjoint parts. One of them consists of those utterances that can be recognized correctly, which we call "right utterances". The other consists of corresponding "wrong utterances". The divide line between these two sets can be regarded as the boundary of the observations of the word under consideration approximately. However, the confirmation of the boundary depends very much on the accuracy of the models' parameters used. We will not get the satisfactory solution unless an efficient algorithm like gradient descent method is adopted. According to this requirement, the following steps are developed.

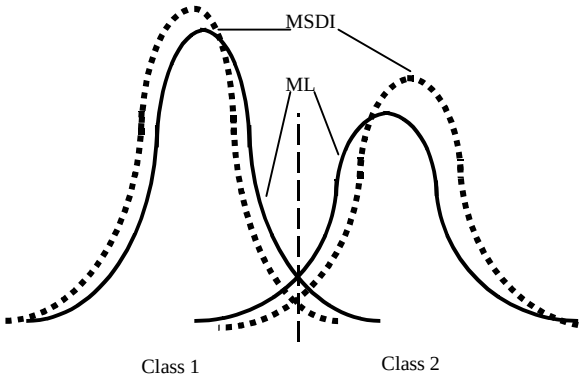


Fig.2 illustration of difference between MLE and MSDIE

Step1. Training the initial models using traditional MLE;

Step2. Dividing the utterances of each word into two disjoint sets according to the latest model;

Step3. Training models using MSDIE;

Step4. If not convergent, going to step2 and continuing, else halting this procedure.

Since the MSDIE does not need to compute the probabilities for the other models, above iteration will be ended rapidly. The Fig.2 illustrates the difference between ML and MSDI in a simple case. From the figure, the conclusion is drawn that the outcome of MSDIE is comparable with that of MMIE, while MSDIE needs less computation.

### 3.4 More Economical Algorithms for MSDIE

Though the iterative algorithm discussion in 3.3 is very useful, the establishment of the "right set" and the "wrong set" is still complex. The time for iteration can not be controlled easily. In order to reduce the computation ulteriorly, an imprecise solution is presented. Firstly, an assumption is made that the "wrong set" of one given word sequence is equivalent to all "right sets" of other word sequences and its "right set" consists of all the observations of this word sequence. Secondly, the model parameter is fixed by only one pass of MSDIE according to these two proximate sets. In practice, we can even choose only utterances of a few competitive words as the "wrong utterances". The competitors should be those which are confusable with the correct word sequence.

## 4. EXPERIMENTS AND RESULTS

The speech database is a Mandarin Continuous Speech Database recorded by 38 men. Each speaker uttered one set of sentences in a continuous mode. The database contains 250,657 Mandarin syllables totally. We used 30 men's utterances to train the HMM's parameters. The remaining part is used for testing. All the recorded materials were obtained in an officelike environment through a close-talk noise-canceling microphone. They are digitized at a sampling frequency of 16KHz. A 32ms Hamming window took the filtered speech. And then the cepstral coefficients derived from LPC of order 16 were extracted for every 16ms. The acoustic model used for experiments is the HMM with 6 states and 16 components each state.

Fig.3 shows the hit rates of top 10 candidates using MLE, MMIE and MSDIE. The results indicate that the MSDIE and MMIE both have been raised by about 6 percent compared with the traditional MLE.

Fig.4 shows the hit rates of top 10 candidates using MLE and the other two kind of approximate solutions. MSDIE (A) represents the method using all other utterances as "wrong" one; MSDIE (B) represents the method using only confusing word. The results imply that the performance of models training by approximate algorithms is also better than MLE. The reason of that MSDIE (B) is more preferable than MSDIE (A) may be that the data used for subtracting the value of criteria are excessive in MSDIE (A), which results in that the convergent result can not reach the global optimal one.

Fig.5 shows the time-consuming in the estimation using MMIE, MSDIE and MSDIE (B). The results point out that MSDIE has a great deal of reduction in computation complexity, which is the most preferable aspect of this novel discriminative method.

## 5. REFERENCES


[1] L.R.Rabiner (1989), "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Proc. IEEE*, 72(2):257-286, February 1989

[2] L.R.Bahl, P.Brown, P.de Souza, R.Mercer (1986), "Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition",*ICASSP86*, Tokyo, vol1, pp.49-52

[3] L.R.Bahl, F.Jelinek, R.Mercer (1983), "A Maximum Likelihood Approach to Continuous Speech Recognition", *IEEE Trans. on PAMI*, PAMI-5, No 2, March 1983

[4] L.E.Baum (1972), "An Inequality and Associated Maximization Technique in Statistical Estimation of Probabilistic Functions of Markov Processes," *Inequalities*, 3,1972

[5] B.Merialdo (1988), "Phonetic Recognition Using Hidden Markov Models and Maximum Mutual Information Training", *IEEE Trans. On SAP*, 1988, pp.111-114

[6] Y.L. Chow (1990), "Maximum Mutual Information Estimation of HMM Parameters for Continuous Speech Recognition using the N-Best Algorithm", *ICASSP90*, pp.701-704


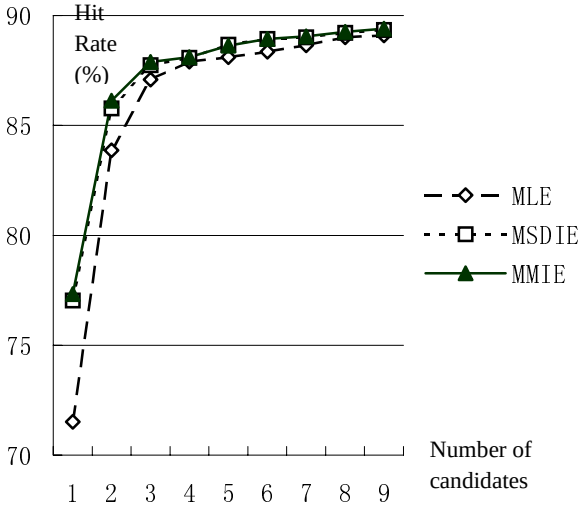


*Fig.3 comparison of three algorithms*

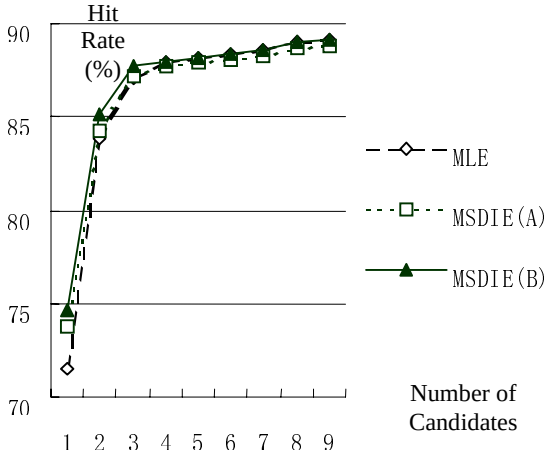


*Fig.4 comparison of approximate algorithms*

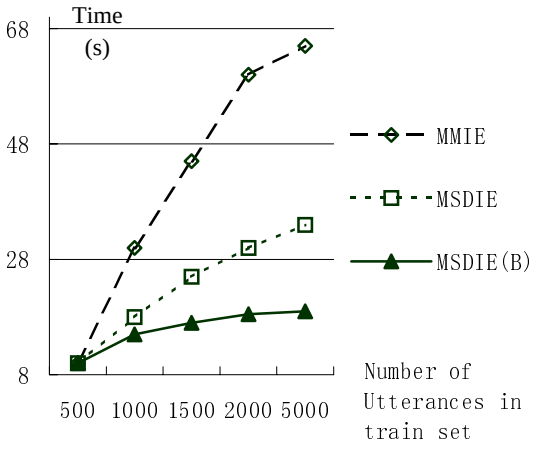


*Fig.5 comparison of time-consuming*