A NEW HMM TRAINING APPROACH FOR SPEECH RECOGNITION

Qian-hua HE, Gang WEI
South China University of Technology, Guangzhou, China
Tel. +86-20-87113540,E-mail:{eeqhhe,ecgwei}@scut.edu.cn

ABSTRACT

This paper proposed a *maximum model distance* (MMD) approach for HMM-based speech recognition. MMD uses the whole training set to estimate the parameters of each HMM while the traditional maximum likelihood (ML) uses only those data labeled for the model. Theoretical and practical issues concerning this approach in speech recognition are investigated. Both speaker-dependent and multi-speaker experiments on confusable Chinese An-set showed that significant error reduction can be achieved through the proposed approach. In addition, the relationship between MMD and corrective training [5] was discussed and it is proved that the corrective training is a special case of MMD approach.

1. INTRODUCTION

Hidden Markov Model (HMM) is one of the most successful statistical modeling methods in the area of speech recognition. The parameter sets of the HMM is usually estimated by the maximum likelihood (ML) approach [1]. Although strong arguments have been made in favor of ML estimation as opposed to other intuitively appealing estimation methods [2], as experienced by many researchers during the past few years, the maximum likelihood based training approach for a given model structure may not give the best performance in terms of the recognition error rate. Alternatives of ML training criterion have been proposed by some researchers [3-5], such as the *maximum mutual information* (MMI) criterion [3], *minimum discrimination information* (MDI) criterion [4] and *corrective training* [5]. All these algorithms have its own advantages and shortcomings.

This paper propose a *maximum model distance* (MMD) criterion for training HMMs. MMD aims at improving the performance of HMM-based speech recognizer by maximizing the dissimilarities among HMMs. The performance of MMD was evaluated through two experiments on confusable An-set of Chinese syllables, one was speaker-dependent and the other was multi-speaker based. These two experiments demonstrated that maximum model distance training approach can significantly reduce the number of recognition error, compared against ML training approach. Overall, the number of errors fell by 18.6% on test set.

2. MAXIMUM MODEL DISTANCE APPROACH (MMD)

Juang and Rabiner [6] proposed a probabilistic distance measure for any pair of HMMs. Let $D(\lambda_\nu,\lambda_\theta)$ be the distance between two hidden Markov models, λ_ν and λ_θ,

$$D(\lambda_\nu,\lambda_\theta)=\lim_{T_\nu\to\infty}\frac{1}{T_\nu}\left\{\log P(\mathbf{O}^\nu|\lambda_\nu)-\log P(\mathbf{O}^\nu|\lambda_\theta)\right\} \tag{1}$$

where $\mathbf{O}^\nu=(\mathbf{o}_1^\nu\mathbf{o}_2^\nu\mathbf{o}_3^\nu\cdots\mathbf{o}_{T_\nu}^\nu)$ is a sequence of observations generated by model λ_ν. Petrie's limit theorem [7] guarantees the existence of such a distance measure and ensure that $D(\lambda_\nu,\lambda_\theta)$ is nonnegative. Basically, Eq. (1) is measure of how well model λ_θ matches observations generated by model λ_ν, relative to how well model λ_ν matches observations generated by itself.

Although the distance measure of Eq.(1) was originally defined for ergodic HMMs, in fact, it works quite reliably for other types of Markov models, such as left-to-right models[6]. In practice, the sequence of training data $\mathbf{O}=(\mathbf{o}_1\mathbf{o}_2\mathbf{o}_3\cdots\mathbf{o}_T)$ of a given word is always finite, we generalize the concept of model distance by defining $D(\lambda_\nu,\lambda_\theta)$ as

$$D(\lambda_\nu,\lambda_\theta)=\frac{1}{T_\nu}\left\{\log P(\mathbf{O}^\nu|\lambda_\nu)-\log P(\mathbf{O}^\nu|\lambda_\theta)\right\} \tag{2}$$

Furthermore, a distance measure $D(\lambda_\nu,\Lambda)$ between model λ_ν and model set Λ is defined as

$$D(\lambda_\nu,\Lambda)=\frac{1}{V-1}\sum_{\theta=1,\theta\neq\nu}^{V}\frac{1}{T_\nu}$$

$$\left\{\log P(\mathbf{O}^{\nu}|\lambda_{\nu}) - \log P(\mathbf{O}^{\nu}|\lambda_{\theta})\right\}$$
$$= \frac{1}{T_{\nu}}\left\{\log P(\mathbf{O}^{\nu}|\lambda_{\nu}) - \frac{1}{V-1}\sum_{\theta=1,\theta\neq\nu}^{V}\log P(\mathbf{O}^{\nu}|\lambda_{\theta})\right\} \quad (3)$$

where $\Lambda = \{\lambda_{\nu}, \nu = 1,\cdots,V\}$ is the model set. $D(\lambda_{\nu},\Lambda)$ can be explained as the average measure of how well the competitive models of model λ_{θ} matches observations generated by model λ_{ν}, relative to how well model λ_{ν} matches observations generated by itself. The *maximum model distance* (MMD) criterion is to find the entire model set Λ such that the model distance is maximized.

$$(\Lambda)_{\mathrm{MMD}} = \arg\max_{\Lambda}\sum_{\nu=1}^{V} D(\lambda_{\nu},\Lambda) \quad (4)$$

since

$$\sum_{\nu=1}^{V} D(\lambda_{\nu},\Lambda) = \sum_{\nu=1}^{V}\frac{1}{T_{\nu}}\left\{\log P(\mathbf{O}^{\nu}|\lambda_{\nu}) - \frac{1}{V-1}\sum_{\theta=1,\theta\neq\nu}^{V}\log P(\mathbf{O}^{\nu}|\lambda_{\theta})\right\}$$
$$= \sum_{\nu=1}^{V}\left\{\frac{1}{T_{\nu}}\log P(\mathbf{O}^{\nu}|\lambda_{\nu}) - \frac{1}{V-1}\sum_{\theta=1,\theta\neq\nu}^{V}\frac{1}{T_{\theta}}\log P(\mathbf{O}^{\theta}|\lambda_{\nu})\right\} \quad (5)$$

The ν*th* item in the sum of Eq.(5) depends only on the parameters of the ν*th* HMM, so the solution of Eq.(4) could be equally gotten by estimating separately the parameters of each model, i.e. the model parameter λ_{ν} is estimated by

$$(\lambda_{\nu})_{MMD} = \arg\max_{\lambda}\left\{\frac{1}{T_{\nu}}\log P(\mathbf{O}^{\nu}|\lambda) - \frac{1}{V-1}\sum_{\theta=1,\theta\neq\nu}^{V}\frac{1}{T_{\theta}}\log P(\mathbf{O}^{\theta}|\lambda)\right\} \quad (6)$$

It is evident that MMD approach emphasizes on discrimination, both tokens of the trained word and its competitives are taken into consideration for training. Thus MMD training uses more information than ML estimation and it is true to believe that MMD estimation will be superior to ML estimation.

Eq.(6) can be solved by the traditional optimization procedures like the gradient scheme. Let λ_{n+1} be the parameter set of νth HMM at time n+1, the adjustment rule for obtaining λ_{n+1} is

$$\tilde{\lambda}_{n+1} = \lambda_n + \eta_n \mathbf{U}_n \nabla D_{\nu}(\mathbf{O},\lambda_n) \quad (7)$$

$$D_{\nu}(\mathbf{O},\lambda) = \frac{1}{T_{\nu}}\log P(\mathbf{O}^{\nu}|\lambda) - \frac{1}{V-1}\sum_{\theta=1,\theta\neq\nu}^{V}\frac{1}{T_{\theta}}\log P(\mathbf{O}^{\theta}|\lambda) \quad (8)$$

where $\mathbf{O} = \{\mathbf{O}^{\nu}, \nu = 1,2,\cdots,V\}$ is the training set, η_n is a small positive number satisfying certain stochastic convergence constraints [8]. $\mathbf{U}_n$ can be an identity matrix or a properly designed positive definite matrix, $\nabla D_{\nu}(\mathbf{O},\lambda)$ is the gradient vector of the target function with respect to the parameter set λ. By using a properly designed positive definite matrix sequence $\mathbf{U}_n$ and considering the probability constraints on model parameters, we got the following adjustment rules:

$$\pi_i^{n+1} = \frac{\pi_i^n + \eta_n\left(\gamma_1^{\nu}(i) - \frac{1}{V-1}\sum_{\theta=1,\theta\neq\nu}^{V}\gamma_1^{\theta}(i)\right)}{\sum_{i=1}^{N}\left[\pi_i^n + \eta_n\left(\gamma_1^{\nu}(i) - \frac{1}{V-1}\sum_{\theta=1,\theta\neq\nu}^{V}\gamma_1^{\theta}(i)\right)\right]}, \quad i = 1,2,\cdots,N \quad (9a)$$

$$a_{ij}^{n+1} = \frac{a_{ij}^n + \eta_n\left(s_{ij}^{\nu} - \frac{1}{V-1}\sum_{\theta=1,\theta\neq\nu}^{V}s_{ij}^{\theta}\right)}{\sum_{j=1}^{N}\left[a_{ij}^n + \eta_n\left(s_{ij}^{\nu} - \frac{1}{V-1}\sum_{\theta=1,\theta\neq\nu}^{V}s_{ij}^{\theta}\right)\right]}, \quad i,j = 1,2,\cdots,N \quad (9b)$$

$$b_j^{n+1}(k) = \frac{b_j^n(k) + \eta_n\left(c_{jk}^{\nu} - \frac{1}{V-1}\sum_{\theta=1,\theta\neq\nu}^{V}c_{jk}^{\theta}\right)}{\sum_{k=1}^{M}\left[b_j^n(k) + \eta_n\left(c_{jk}^{\nu} - \frac{1}{V-1}\sum_{\theta=1,\theta\neq\nu}^{V}c_{jk}^{\theta}\right)\right]}, \quad j = 1,2,\cdots,N,\ k = 1,2,\cdots,M \quad (9c)$$

where

$\gamma_1^{\nu}(i)$ = the expected frequency in state *i* at time *t*=1 in $\mathbf{O}^{\nu}$, normalized by the length T_{ν} of $\mathbf{O}^{\nu}$;

$s_{ij}^{\nu} = \frac{1}{T_{\nu}}\sum_{t=1}^{T_{\nu}-1}\xi_t^{\nu}(i,j)$ the expected number of transitions from state i to state j in $\mathbf{O}^{\nu}$, normalized by the length T_{ν} of $\mathbf{O}^{\nu}$;

$c_{jk}^{v} = \frac{1}{T_v}\sum_{t=1}^{T_v-1}\gamma_t^{v}(j)\delta(\mathbf{o}_t^{v}, v_k)$ the expected number of times in state j and observing symbol v_k in $\mathbf{O}^{v}$, normalized by the length T_v of $\mathbf{O}^{v}$.

Same meaning can be attributed to $\gamma_1^{\theta}(i)$, s_{ij}^{θ}, c_{jk}^{θ}. Eq.(9) hints that the MMD training algorithm can automatically focuses on those training data which are important for discriminating between acoustically similar words; because the attribution of similar part of two tokens are canceled out. This is the most obvious difference between MMD training and maximum likelihood estimation.

In principle, MMD training uses all training data to estimate the parameters of model λ_v. This training procedure has much higher computation complexity than ML estimation because ML estimation uses only these data labeled for word *v*. To reduce the computation complexity, we can combine ML training procedure and focus on the confusable data in the following way.

1) Using the training data labeled for word *v*, apply the forward-backward algorithm iteratively to obtain an estimation λ_v;
2) Find out all confusable utterances of word *v* by checking each competitive utterance $\mathbf{O}^{\theta}$ in the training data set. If $\log P\left(O^{\theta}|\lambda_v\right) > \log P\left(O^{v}|\lambda_v\right) - \delta$, word θ is an acoustically confusable word of word *v*. Let Ω_v denote the confusable word set of word *v*, V_v denote the number of words in Ω_v.
3) Reestimate λ_v with MMD estimation. Eq.(9) is still useful by replacing $\frac{1}{V-1}\sum_{\theta=1,\theta\neq v}^{V}$ with $\frac{1}{V_v}\sum_{\theta\in\Omega_v}$.

3. THE COMPARISON BETWEEN MMD AND CORRECTIVE TRAINING

In the corrective training [5], a model is described by a transition probability $p(a_l|s_i)$ and an output probability distribution $p(v_k|a_l)$, where $a_l \in \{a_1, a_2, \cdots, a_L\}$, the transition arc inventory, $s_i \in \{s_1, s_2, \cdots, s_N\}$, the state inventory, and $v_k \in \{v_1, v_2, \cdots, v_M\}$, the label alphabet. The mapping from $p(v_k|a_l)$, $p(a_l|s_i)$ to a_{ij}, $b_j(k)$ is

$$\hat{a}_{ij} = \sum_{l:a_l\in\Gamma_i}\hat{p}(a_l|s_i)\delta(j, R(s_i, a_l)) \quad (10)$$

$$b_j(k) = \sum_{l:a_l\in\Gamma_i}\hat{p}(a_l|s_i)\hat{p}(v_k|a_l) \quad (11)$$

where Γ_i denote the set of arcs which originate from state s_i and $R(s_i, a_l)$ is the state label of the successor state of arc a_l starting from state s_i.

$$\delta(j, R(s_i, a_l)) = \begin{cases} 1, & if \quad j = R(s_i, a_l) \\ 0, & if \quad j \neq R(s_i, a_l) \end{cases}$$

Based on the principle of corrective training proposed by Bahl et al [5], and consider the total effect of all acoustically confusable words of a labeled utterance $\mathbf{O}$. The adjustment formula of a_{ij} and $b_j(k)$ are

$$\hat{a}_{ij} = \frac{\hat{s}_{ij} + \eta\left[\hat{s}_{ij}^{v} - \frac{1}{V_v}\sum_{w\in\Omega_v}\hat{s}_{ij}^{w}\right]}{\sum_{j\in R(s_i)}\left[\hat{s}_{ij} + \eta\left[\hat{s}_{ij}^{v} - \frac{1}{V_v}\sum_{w\in\Omega_v}\hat{s}_{ij}^{w}\right]\right]} \quad (12a)$$

$$\hat{b}_j(k) = \frac{\hat{c}_{jk} + \eta\left[\hat{c}_{ij}^{v} - \frac{1}{V_v}\sum_{w\in\Omega_v}\hat{c}_{ij}^{w}\right]}{\sum_{k=1}^{M}\left[\hat{c}_{jk} + \eta\left[\hat{c}_{ij}^{v} - \frac{1}{V_v}\sum_{w\in\Omega_v}\hat{c}_{ij}^{w}\right]\right]} \quad (12b)$$

where

$\hat{s}_{ij}$ = expected number of transition from state *i* to state *j*

$\hat{c}_{jk}$ = expected number of transition from state *j* and observing symbol v_k

It is noted that Eq.(12) is very similar to Eq. (9), and $\hat{s}_{ij}$, $\hat{c}_{jk}$ have the same characteristics of s_{ij}, c_{jk} of (9). If we replace the variables in Eq.(10) with corresponding normalized ones, (12) will become identical to (9) in form. From (9) and (12), we note that corrective training differ form MMD in two aspects:

1) MMD sequentially estimated each HMM, using all training data, but corrective training simultaneously estimate model set Λ, using a labeled utterance $\mathbf{O}^{v}$. This means that MMD improves the performance of recognizer by utilizing the discriminating information including in the database to strength the ability of each HMM to distinguish those utterances

labeled for the model from those unlabeled for the model. But corrective training focus to improve the ability of the model set Λ to classify each utterance. Conclusively, the main difference between MMD and corrective training lies in the way how to utilize the useful information of the training data, MMD use all training data to train the HMM for word *v*, but corrective training uses a labeled utterance $\mathbf{O}^v$ to re-estimate the entire model set Λ.

2) MMD uses normalized variables, but corrective training uses unnormalized variables to re-estimate model parameters.

If we emphasize on taking full use of each labeled utterance $\mathbf{O}^v$ to improve the ability of Λ to recognize $\mathbf{O}^v$, and still use maximum model distance criterion, we can maximize the distance measure $D(\lambda_v, \Lambda)$ defined in (3) in a sequential way, i.e. training data of different word is presented in a balanced way. For each $\mathbf{O}^v$, Λ is optimized by

$$(\Lambda)_{MMD} = \arg\max_{\Lambda} D(\lambda_v, \Lambda) \qquad (13)$$

Gradient scheme is still used to solve Eq.(11). And the adjustment formula are

$$\tilde{\lambda}_v^{n+1} = \lambda_v^n + \eta_n \hat{\lambda}_v^v \qquad (14a)$$

$$\tilde{\lambda}_w^{n+1} = \lambda_w^n - \eta_n \hat{\lambda}_w^v \qquad (14b)$$

where $\hat{\lambda}_v^v$, $\hat{\lambda}_w^v$ are the estimates for the correct word and incorrect word through the forward-backward algorithm, using the labeled utterance $\mathbf{O}^v$.

The adjustment rule for $b_j(k)$ in (14) is identical to the mechanism of corrective training described by L. Rabiner [9]. This result indicates that corrective training is a special case of maximum model distance criterion.

4. EXPERIMENTS AND DISCUSSION

The performance of maximum model distance training approach was evaluated through two experiments: speaker dependent and multi-speaker dependent. And compared with that of the HMM-based recognizers trained by ML and corrective training, these HMMs had the same structure of that trained by MMD.

The database used in the experiments was the confusable An-set of Chinese syllables, which consists of 21 untoned syllables or 72 toned syllables. 21 HMMs were built for 21 untoned syllables (an, ban, can, can, dan, fan, gan, han, kan, lan, man, nan, pan, ran, san, shan, tan, wan, yan, zan, zhan), but the training data was sampled from toned syllables for some practical consideration. For the speaker dependent experiment, each toned syllable had 50 repetitions. On average, there were 171 utterances per model. The test set had 30 repetitions of each toned syllables. For the multi-speaker experiment, the utterances were collected from 25 talkers, (13 male and 12 female), each of them provided three tokens per toned syllable, two for training, and one for test. The feature vectors consists of 12 weighted cepstrum coefficients and 12 delta-cepstrum coefficients.

In the experiments, discrete left-to-right whole word model was used. The model parameters were initialized from a uniform segmentation, then adjusted in two stages. In the first stage, the statistics was obtained via the forward-backward algorithm. In the second stage, the conventional estimates were modified using MMD training approach. We use natural logarithms and parameter values η_0=0.66. η_n becomes smaller as n increases. The MMD training procedure stops when the change of model distance is less than 1 percent of current model distance. Let D_n, D_{n-1} be the model distance measure at time *n* and *n-1*, if $|D_n - D_{n-1}| < 0.01 * D_n$, the training procedure stops. Table I list the experimental results in terms of number of error recognition. For comparison, the results of HMMs trained by ML and corrective training were given in table I too. From table I, it can be concluded that *maximum model distance* training approach substantially reduce recognition error, compared with conventional ML training. Overall, on the training set, the number of errors fell by 39.3%, on the test set, the number of errors fell by 18.6%. This result confirm that ML estimates as obtained via the forward-backward algorithm, do not always lead to the lowest error rate in speech recognition.

In table I, the performance of MMD is comparable with that of corrective training. As mentioned before, both *maximum model distance* and *corrective training* aim at taking full use of the given training data. MMD uses all training data to estimate parameters of each model. The training procedure makes each trained model optimally fit the training set. On the other hand, corrective training uses each utterance to train all models, make the model set recognize it in some optimal sense. As a result, the final models should be optimal in terms of classifying the training set. Therefore, MMD and corrective training are two

ways to arrive the same target, taking full use of the discriminative information included in the training data, and making the model set optimally fit the statistics of the training set.

In conclusion, this paper proposed a new training approach, *maximum model distance*, for HMM-based speech recognition. Experiments demonstrated that MMD can significantly reduce the recognition error with respect to ML. In addition, the relationship between MMD and corrective training had been discussed in detail and it is found that corrective training can be induced as a special case of MMD. In some sense, this work provided some theoretical support to the corrective training, which is complete lack of theoretical underpinnings [5].

Table I Experimental results (number of recognition errors)

	Speaker-dependent		Multi-speaker recognition	
	training set (3600 tokens)	test set (2160 tokens)	training set (3600 tokens)	test set (1800 tokens)
MMD	282	382	477	380
ML	456	460	795	477
error reduction	38.1%	16.9%	40%	20.3%
CT	272	391	463	380

REFERENCE

[1]. A. Liporace, Maximum likelihood estimation for multivariate observations of Markov Sources. IEEE Trans. *Inform. Theory*, Vol. IT-28, No. 5, pp. 729-734, Sept. 1982.

[2]. A. Nadas, A decision theoretic formulation of a training problem in speech recognition and a comparison of training by unconditioned versus conditioned maximum likelihood. IEEE Trans. on *Acoustics, Speech, and Signal Processing*. vol. 31, pp. 814-817. 1983.

[3]. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, Maximum mutual information estimation of hidden Markov model parameters for speech recognition, in *Proc. 1986 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Tokyo, Japan, pp. 49-52, Apr. 1986.

[4]. Yariv Ephraim, Amir Dembo and L.R. Rabiner. A Minimum Discrimination Information Approach for Hidden Markov Modeling. *IEEE Trans. on Information Theory*, Vol.35, No.5, Sept. 1989.

[5]. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, Estimating Hidden Markov Model Parameters So As To Maximize Speech Recognition Accuracy. *IEEE Trans. on Speech and Audio Processing*, Vol. 1, No. 1 , Jan. 1993.

[6]. H. Juang and L. R. Rabiner, A Probabilistic Distance Measure for Hidden Markov Models. *AT&T Technical Journal*, Vol. 64, No. 2, February 1985.

[7]. Petrie, Probabilistic Functions of Finite State Markov Chains, *Ann. Math. Statist*., 40, No. 1 (1969), pp.97-115.

[8]. P. C. Chang and B. H. Juang, Discriminative template training for dynamic programming speech recognition. Proc. *ICASSP-92*, Vol. I pp.493-496. San Francisco, March 1992.

[9]. Rabiner & B. H. Juang. Fundamentals of speech recognition. Chapter 6, Prentice-Hall, Inc. 1993.