Optimizing Hidden Markov Models for Chinese An-set Syllables

*Q. H. He and **S. Kwong

*Department of Electronic Engineering, South China University of Technology, China

**Department of Computer Science, City University of Hong Kong, HK

Abstract: Speech recognition for Chinese relied very much on the recognition of Chinese syllables and there are altogether 1345 syllables in it. If we take tones into considerations, the number of syllables can reduce to 408 base syllables, with different tones, in which it can further divided into 38 confused set. Among those sets, the Chinese An-set is considered as one of the major confused syllable set. Thus, the recognition of Chinese An-set syllables is very important to the Chinese recognition. In this paper, we proposed a new training approach based on maximum model distance (MMD) for HMMs to train the Chinese An-set syllables. Both the speaker-dependent and multi-speaker experiments on the confused Chinese An-set showed that significant error reduction can be achieved through the proposed approach.

1. Introduction

Recently, there are great demands in the research in the development of the computer processing of Chinese Language. One of the major area especially require researchers to put more effort onto it are the Chinese input methods. We investigated and developed many efficient input methods [8], however, there always have a major drawback is that it usually required users to remember a lot of rules and can only be used by a skillful trained user. This particular property makes it differ form other alphabetic language likes English in which users can type in the words just by directly mapping the alphabets onto the input device keyboard of the computer. Chinese input methods usually required a complicated procedure to key in the Chinese characters.

As can be seen, traditional Chinese input methods using keyboards is a problem to layman or even to some of the professionals. Therefore, a natural and easy to learn input method is desired. There are no input methods that are easier than inputting Chinese by speaking to the computer directly. Recently, Hidden Markov model is considered as one of the most successful statistical modeling methods in the area of speech recognition. The parameter sets of the HMMs for acoustic signals are usually estimated by the maximum likelihood (ML) approach [1]. There have been agreed that ML estimation for HMM parameters are better than other intuitively appealing estimation methods. However, it is experienced by many researchers during the past few years that the maximum likelihood based training approach for a given model structure may not give the best performance in terms of the recognition error rate. Therefore, there are many other alternatives of ML training criterion are developed such as the maximum mutual information (MMI) criterion [2], minimum discrimination information (MDI) criterion [3] and corrective training[4].

As mentioned in the abstract that there are 1345 toned syllables and Chinese is a tonal language. Therefore, each syllable has a tone associated with it and has its own meanings. If the tones are ignored in the recognition process, then it left about 408 tones. Among these four hundreds syllables, it can further divided into 38 sets of syllables. Among these sets of syllables, the Chinese An-set is one of the most complicated one and has most of the members in it. The Chinese An-set syllables is defined as follows: \{an, ban, pan, man, fan, dan, tan, nan, lan, gan, kan, han, jan, chan, shan, ran, tzan, tsan, san \} [7]. Thus, the successful recognition rate for the Chinese syllable sets are very important to the Chinese speech recognition system if it used syllables as the feature vectors of the system.

In this paper, we proposed a criterion based on maximum model distance (MMD) for training HMMs. The aim of the MMD is to improve the performance of HMM-based speech recognizer by maximizing the dissimilarities among all the HMMs in the
system. The performance of MMD was evaluated through two experiments on the confused An-set of Chinese syllables: one was speaker-dependent and the other was multispeaker. These two experiments demonstrated that maximum model distance training approach can significantly reduced the number of recognition errors when it is compared against ML training approach by 18.6%.

2. Maximum Model Distance Approach (MMD)

Juang and Rabiner[5] proposed a probabilistic distance measure for any pair of HMMs. Let $D(\lambda_v, \lambda_\theta)$ be the distance between two hidden Markov models, $\lambda_v$ and $\lambda_\theta$,

$$D(\lambda_v, \lambda_\theta) = \lim_{T \to \infty} \frac{1}{T} \left[ \log P(O^T | \lambda_v) - \log P(O^T | \lambda_\theta) \right]$$

where $O^T = (o^1, o^2, o^3, \ldots, o^T)$ is a sequence of observations generated by model $\lambda_v$. Petrie’s limit theorem guarantees the existence of such a distance measure and ensures that $D(\lambda_v, \lambda_\theta)$ is nonnegative. Basically, Eq. (1) is the similarity score of the observations generated by the models $\lambda_v$ and $\lambda_\theta$.

In practice, the sequence of training data $O = (o_1, o_2, o_3, \ldots, o_T)$ of a given word is always finite, the model distance can be generalized by defining $D(\lambda_v, \lambda_\theta)$ as

$$D(\lambda_v, \lambda_\theta) = \frac{1}{T} \left[ \log P(O^T | \lambda_v) - \log P(O^T | \lambda_\theta) \right]$$

Furthermore, a distance measure $D(\lambda_v, \Lambda)$ between model $\lambda_v$ and model set $\Lambda$ is defined as

$$D(\lambda_v, \Lambda) = \frac{1}{V-1} \sum_{\theta=v+1}^{V} \left[ \log P(O^T | \lambda_v) - \log P(O^T | \lambda_\theta) \right]$$

where $\Lambda = \{\lambda_v, v=1, \ldots, V\}$ is the model set. The maximum model distance (MMD) criterion is to find the entire model set $\Lambda$ such that the model distance is maximized.

$$(\Lambda)_{\text{MMD}} = \arg \max_{\Lambda} \sum_{v=1}^{V} D(\lambda_v, \Lambda)$$

Since

$$\sum_{v=1}^{V} D(\lambda_v, \Lambda) = \sum_{v=1}^{V} \left[ \frac{1}{T_v} \sum_{t=1}^{T_v} \log P(O^t | \lambda_v) - \frac{1}{V-1} \sum_{\theta=v+1}^{V} \frac{1}{T_\theta} \sum_{t=1}^{T_\theta} \log P(O^t | \lambda_\theta) \right]$$

$$= \sum_{v=1}^{V} \left( \frac{1}{T_v} \log P(O^T | \lambda_v) - \frac{1}{V-1} \sum_{\theta=v+1}^{V} \frac{1}{T_\theta} \log P(O^T | \lambda_\theta) \right)$$

The solution of Eq.(4) could be obtained by estimate the parameters of each model separately, i.e. the model parameter $\lambda_\theta$ is estimated by

$$(\lambda_\theta)_{\text{MMD}} = \arg \max_{\lambda_\theta} \left[ \frac{1}{T_v} \log P(O^T | \lambda_\theta) - \frac{1}{V-1} \sum_{\theta=v+1}^{V} \frac{1}{T_\theta} \log P(O^T | \lambda_\theta) \right]$$

It can be seen that the MMD approach emphasized on the discrimination against the tokens for the trained word and its competitive word. This consideration was taken into considerations in the training process. In this way, the MMD training utilized more information than the ML estimation and it is believed that the MMD estimation is superior to the ML estimation.

Eq.(6) can be solved by the traditional optimization procedures like the gradient scheme. The adjustment rule is

$$\pi_i = \frac{\pi_i}{\sum_{j=1}^{N} \pi_j} \left( \gamma(i)i - \frac{1}{V-1} \sum_{\theta=1}^{V} \gamma(\theta,i) \right)$$

$$a_{ij} = \frac{a_{ij} + \eta_a \left( s_{ij} - \frac{1}{V-1} \sum_{\theta=1}^{V} s_{ij} \right)}{\sum_{k=1}^{N} a_{ik} + \eta_a \left( s_{ik} - \frac{1}{V-1} \sum_{\theta=1}^{V} s_{ik} \right)}$$

$$b_{j} = \frac{b_{j} + \eta_s \left( c_{j} - \frac{1}{V-1} \sum_{\theta=1}^{V} c_{j} \right)}{\sum_{k=1}^{M} b_{j} + \eta_s \left( c_{j} - \frac{1}{V-1} \sum_{\theta=1}^{V} c_{j} \right)}$$

where $\eta_a$ is a small positive number satisfying some stochastic convergence constraints

$$\gamma^v(i) = \text{the normalized expected frequency in state } i \text{ at time } t=1 \text{ in } O^v$$

$$s_{ij} = \text{the normalized expected number of transitions from state } i \text{ to state } j \text{ in } O^v$$


\( c_{jk}^v = \) the normalized expected number of
times in state \( j \) and observing symbol \( v_k \) in
\( O^v \).

Same meaning can be attributed to
\( \gamma^v(i), s^\theta, c^\theta_{jk} \). Eq.(7) hints that the MMD
training algorithm can focused 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, the MMD training used all
training data to estimate the parameters of
model \( \lambda_v \). This training procedure has much
higher computation complexity than ML
estimation because ML estimation uses only
these data labeled for word \( v \). In order to
reduce the computation complexity, we can
combine ML training procedure and focus on
the confused data in the following way.
1) Using the training data labeled for word
\( v \), apply the forward-backward algorithm
iteratively to obtain an estimation \( \hat{\lambda}_v \);
2) Find out all the confused utterances of
word \( v \) by checking each competitive
utterance \( O^\theta \) in the training data set. If
\( \log P(O^\theta|\lambda_v) > \log P(O^v|\lambda_v) - \delta \),
word \( \theta \) is an acoustically confused word
of word \( v \). Let \( \Omega_v \) denotes the confused
word set of word \( v \), \( V_v \) denotes the
number of words in \( \Omega_v \).
3) Re-estimate \( \hat{\lambda}_v \) with MMD estimation.
Eq.(7) is still useful by replacing
\[
\frac{1}{V - 1} \sum_{\theta=1, \theta \neq v}^V \text{ with } \frac{1}{V_v} \sum_{\theta=1}^{V_v} .
\]

3. The comparison between MMD training
and Corrective training
Corrective training [4] was proposed by
Baul et al has a very similar meanings as the
MMD approach except that the corrective
training differ from the MMD in two aspects.
First, the MMD estimated each HMM
sequentially, but corrective training estimate
model set \( \Lambda \) simultaneously. The MMD used
the entire training data to train the HMM for
word \( v \), but corrective training uses a labeled
utterance \( O^v \) to re-estimate the entire model
set \( \Lambda \). Secondly, the MMD used normalized
variables, but corrective training uses un-
normalized variables to re-estimate model
parameters. If we emphasized on using each
labeled utterance \( O^v \) to improve the ability
of \( \Lambda \) to recognize \( O^v \), and apply the
maximum model distance criterion, we can
maximize the distance measure \( D(\hat{\lambda}_v, \Lambda) \)
defined in (3) in a sequential way. For each
\( O^v \), \( \Lambda \) is optimized by
(\( \Lambda \))_{MMD} = \arg \max_{\Lambda} D(\hat{\lambda}_v, \Lambda) \)  (8)
Gradient scheme is used to solve Eq.(8). The
adjustment formula are defined as:
\[
\hat{\lambda}_v^{n+1} = \hat{\lambda}_v^n + \eta_n \hat{\lambda}_v^v \)
\]
\[
\hat{\lambda}_w^{n+1} = \hat{\lambda}_w^n - \eta_n \hat{\lambda}_w^v \)
\]
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 \( O^v \). The adjustment rule
for \( b_t(k) \) in (10) is identical to the
mechanism of corrective training described by
L. Rabiner [6].

4. Experiments and Discussions
Since Chinese syllables consist of
consonant and vowel, those syllables with the
same vowel are much difficult to distinguish
from each other than those with different
vowels. In order to demonstrate the
performance of MMD, the confused An-set of
Chinese syllables was taken as the testing set,
which consists of 21 non-toned syllables or
72 toned syllables. A toned syllable could be
recognized by the un-toned syllable and its
tone, so only 21 HMMs were built for 21
non-toned syllables, but training data was
collected from 72 toned syllables for and
syllable is always pronounced with tone. Two
experiments were carried, one is speaker
dependent and the other is multi-speaker
dependent. In the speaker dependent
experiment, each toned syllable had 50
repetitions, 30 for model training and 20 for
testing. In the multispeaker experiment, the
utterances were collected from 25 talkers (13
male and 12 female), each of them provided
three tokens for per toned syllable, two for
training, and one for testing. 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: the model parameters was first estimated with ML criterion, and then re-estimated with MMD training approach. Natural logarithms was used and set $\frac{\eta}{G_{13}}=0.66$. $\eta_n$ became smaller as $n$ increases. The MMD training procedure terminates when the change of model distance is less than 1 percent of current model distance. The experimental results in terms of number of errors. In addition, the results of HMMs trained by ML and corrective training were compared. From the experiment results, it can be concluded that the maximum model distance training approach substantially reduce recognition error, compared with ML criterion. Overall, the errors were reduced for 39.3% and 18.6% for the training set and the testing set respectively. This confirmed that the ML estimates obtained via the forward-backward algorithm do not always lead to the lowest error rate in speech recognition. Furthermore, the performance of MMD is comparable with that of corrective training.

In conclusion, this paper proposed a new training approach, maximum model distance, for HMM training to improve the recognition performance. Experiments demonstrated that MMD can significantly reduced the recognition error with respect to ML. In addition, the relationship between MMD and corrective training was discussed.

Reference


