Computational Linguistics and Chinese Language Processing
Vol. 11, No. 3, September 2006, pp. 201-222

An Empirical Study of Word Error Minimization Approaches for Mandarin Large Vocabulary Continuous Speech Recognition

Jen-Wei Kuo*+, Shih-Hung Liu*, Hsin-Min Wang+, and Berlin Chen*

Abstract

This paper presents an empirical study of word error minimization approaches for Mandarin large vocabulary continuous speech recognition (LVCSR). First, the minimum phone error (MPE) criterion, which is one of the most popular discriminative training criteria, is extensively investigated for both acoustic model training and adaptation in a Mandarin LVCSR system. Second, the word error minimization (WEM) criterion, used to rescore *N*-best word strings, is appropriately modified for a Mandarin LVCSR system. Finally, a series of speech recognition experiments is conducted on the MATBN Mandarin Chinese broadcast news corpus. The experiment results demonstrate that the MPE training approach reduces the character error rate (CER) by 12% for a system initially trained with the maximum likelihood (ML) approach. Meanwhile, for unsupervised acoustic model adaptation, MPE-based linear regression (MPELR) adaptation outperforms conventional maximum likelihood linear regression (MLLR) in terms of CER reduction. When the WEM decoding approach is used for *N*-best rescoring, a slight performance gain over the conventional maximum a posteriori (MAP) decoding method is also observed.

Keywords: Broadcast News, Continuous Speech Recognition, Discriminative Training, Minimum Phone Error, Word Error Minimization

* Graduate Institute of Computer and Information Engineering, National Taiwan Normal University, Taipei, Taiwan
E-mail: rogerkuo@iis.sinica.edu.tw
+ Institute of Information Science, Academia Sinica, Taipei, Taiwan

[Received February 23, 2006; Revised August 24, 2006; Accepted September 5, 2006]

1. Introduction

Due to advances in computer technology and the growth of the Internet, large volumes of multimedia content, such as broadcast news, lectures, voice mails, and digital archives continue to grow and fill our computers, networks, and lives. It is obvious that speech is the richest source of information for the large volumes of multimedia content; thus, associated speech processing technologies will play an increasingly important role in multimedia organization and retrieval in the future. Among these technologies, automatic speech recognition (ASR) has long been the focus of research in the speech processing community.

Automatic speech recognition is a pattern classification task that classifies sound segments into different linguistic categories based on the acoustic vector sequence extracted from the speech signal. Traditionally, in most pattern classification applications, the goal of classifier design is to reduce the probability of errors by using the minimum error rate (MER) criterion [Duda *et al.* 2000]. Under this paradigm, the problems of classifier optimization are resolved by minimizing the expected loss over the training data directly. The zero-one loss function, which simply assigns no loss to a correct classification and a unit loss to an error, is often employed for this purpose. For example, in ASR, a hypothesized word sequence containing one or more word errors, or a totally different sequence, as compared to the correct sequence, will incur the same amount of loss. However, the most common performance evaluation metrics adopted in ASR often consider individual word errors, instead of merely counting the string-level errors. The use of the zero-one loss function leads to a mismatch between classifier optimization and performance evaluation. In recent years, a common practice in ASR has been to replace the zero-one loss function with alternative loss functions that consider word- or phone-level errors. In practice, such improved loss functions can be used in both model parameter estimation (i.e., classifier optimization) and speech decoding.

In this paper, we present an empirical study of word error minimization approaches for Mandarin large vocabulary continuous speech recognition (LVCSR). The minimum phone error (MPE) criterion is extensively investigated in both acoustic model training and adaptation; while the word error minimization (WEM) criterion is exploited to rescore *N*-best word strings.

The remainder of the paper is organized as follows. In Section 2, the general background of the Bayes risk and overall risk criteria is given, and their use in ASR is explained. Section 3 presents the application of the MPE criterion for acoustic model training, and Section 4 describes its extension to unsupervised linear regression based acoustic model adaptation. The use of the WEM criterion for speech decoding is discussed in Section 5. The experiment setup is detailed in Section 6 and a series of speech recognition experiments is described in Section 7. Finally, we present the conclusions drawn from the research in Section 8.

2. Bayes Risk and Overall Risk

Given an acoustic vector sequence O, the goal of an ASR system is to make a decision $\alpha_u(O)$ that identifies O as a certain word sequence u from a hypothesized space $\mathbf{W}_h$ of all possible word sequences in the language. Let $L(u,c)$ be the loss incurred by the decision $\alpha_u(O)$, where the correct (i.e., reference) transcription is c. Actually, we have no prior knowledge of the correct transcription; in other words, any arbitrary word sequence s in $\mathbf{W}_h$ could be identical to c. Consequently, for each possible decision $\alpha_u(O)$, the expected loss (or risk) is calculated as [Duda *et al.* 2000]:

$$R(\alpha_u(O)\,|\,O) = \sum_{s\in\mathbf{W}_h} L(u,s)P(s\,|\,O), \tag{1}$$

where $P(s\,|\,O)$ is the posterior probability of the word sequence s given that the acoustic vector sequence O is observed. Therefore, the Bayes decision $\alpha_{opt}(O)$ is made by selecting the action with the minimum expected loss, i.e.,

$$\begin{aligned}\alpha_{opt}(O) &= \arg\min_{u\in\mathbf{W}_h} R\left(\alpha_u(O)\,|\,O\right)\\ &= \arg\min_{u\in\mathbf{W}_h} \sum_{s\in\mathbf{W}_h} L(u,s)P(s\,|\,O)\end{aligned}. \tag{2}$$

In supervised training, on the other hand, the correct transcription of each training utterance O is known, and the overall risk $\tilde{R}_{all}$ of all possible training utterances is defined as:

$$\tilde{R}_{all} = \int R(\alpha_c(O)\,|\,O)P(O)dO, \tag{3}$$

where the integral extends over the whole acoustic space. However, in practice, we can only obtain the approximate overall risk R_{all} by summing the risks over a finite number of training utterances, i.e.,

$$\begin{aligned}R_{all} &= \sum_r R(\alpha_{c_r}(O_r)\,|\,O_r)P(O_r)\\ &= \sum_r \sum_{s\in\mathbf{W}_h^r} L(c_r,s)P(s\,|\,O_r)P(O_r)\end{aligned}, \tag{4}$$

where $\mathbf{W}_h^r$ and c_r, respectively, denote a set of likely hypothesized word sequences and the reference word sequence associated with the training utterance O_r; and the distribution $P(s\,|\,O_r)$ is always assumed to be governed by some underlying parametric distributions. To ensure that ASR is as accurate as possible, we need to design a classifier and estimate the parameters in $P(s\,|\,O_r)$ more carefully in order to minimize the overall risk R_{all}. By applying

the Bayes rule and replacing the probability $P(O_r \mid s)$ with its parameterization, $p_\lambda(O_r \mid s)$, Eq. (4) can be expressed as:

$$R_{all} = \sum_r \frac{\sum_{s \in \mathbf{W}_h^r} L(c_r, s) p_\lambda(O_r \mid s) P(s)}{\sum_{u \in \mathbf{W}_h^r} p_\lambda(O_r \mid u) P(u)} P(O_r), \tag{5}$$

where $p_\lambda(O_r \mid s)$ and $p_\lambda(O_r \mid u)$ are, respectively, the acoustic model likelihoods for s and u under the acoustic model parameter set λ; and $P(s)$ and $P(u)$ are the respective language model probabilities for s and u. The parameters of both the acoustic model and the language model can be estimated by minimizing R_{all}. However, in this study, we only focus on the discriminative estimation of the acoustic model parameters, and adopt the conventional approach for language model training. Moreover, it is assumed that the prior probability $P(O_r)$ is uniformly distributed. As a result, the overall risk becomes

$$R_{all} = \sum_r \frac{\sum_{s \in \mathbf{W}_h^r} L(c_r, s) p_\lambda(O_r \mid s) P(s)}{\sum_{u \in \mathbf{W}_h^r} p_\lambda(O_r \mid u) P(u)}, \tag{6}$$

and the optimal parameter set, λ_{opt}, can be estimated by minimizing the overall risk of the training utterances

$$\lambda_{opt} = \arg\min_\lambda \sum_r \frac{\sum_{s \in \mathbf{W}_h^r} L(c_r, s) p_\lambda(O_r \mid s) P(s)}{\sum_{u \in \mathbf{W}_h^r} p_\lambda(O_r \mid u) P(u)}. \tag{7}$$

To minimize the overall risk, as shown by Equations (4) to (7), the hypothesized word sequence with a lower loss should have a larger posterior probability, and vice versa. How to select an appropriate loss function $L(\cdot,\cdot)$ used in the above equations remains an open research issue. In most pattern classification tasks, to minimize the probability of classification errors, the loss function is often chosen based on the minimum error rate (MER) criterion. This leads directly to the following symmetrical zero-one loss function [Duda *et al.* 2000]:

$$L(u, s) = \begin{cases} 0 & , u = s \\ 1 & , u \neq s \end{cases}. \tag{8}$$

The loss function assigns no loss if $u = s$, and assigns a unit loss when a classification error occurs. In ASR, a hypothesized word sequence that is identical to the correct transcription does not introduce a loss; however, a hypothesized word sequence containing one or more

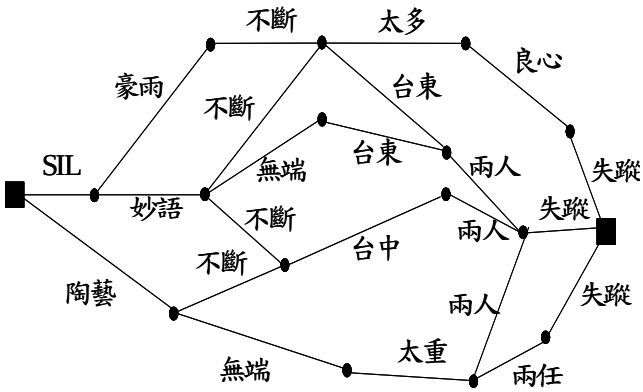

Figure 1. A word lattice can efficiently encode a large number of possible hypothesized word sequences.

word errors, or a totally different sequence, compared to the correct sequence, will incur the same unit loss. Thus, minimizing the overall risk is equivalent to minimizing the expected string error rate (SER) of the training utterances. Nevertheless, SER is not a sufficient metric for the evaluation of ASR performance because, with this metric, all incorrectly hypothesized word sequences are regarded as having the same cost of recognition risk. Instead, the loss function could be defined as the distance of the hypothesized word sequence to the correct transcription. For this purpose, the string edit or Levenshtein distance [Levenshtein 1966] associated with the word error rate (WER) can be adopted. It is believed that WER is more suitable than SER in reflecting differences in ASR results. Optimization using the Levenshtein-based loss function is often referred to as word error minimization (WEM).

However, in complicated ASR tasks, such as LVCSR, it is impossible to perform optimization over the hypothesized space $\mathbf{W}_h^r$ of each training utterance O_r without using a pruning technique because such hypothesized spaces usually contain an extremely large number of hypothesized word sequences. Recently, some practical strategies have been proposed to resolve this problem. For instance, a reduced hypothesized space in the form of an *N*-best list [Schwartz and Chow 1990] or a lattice [Ortmanns 1997] can be generated for each training utterance by only retaining recognized hypotheses with higher probabilities. The optimization process can then be applied efficiently to the reduced hypothesized space. Figure 1 illustrates an example of a word lattice.

3. Minimum Phone Error (MPE) Training

This section describes in detail the application of the minimum phone error (MPE) criterion to acoustic model training. As mentioned in the previous section, the hypothesized space $\mathbf{W}_h^r$ of a given training utterance O_r can be reduced to a smaller space represented by a number of the most likely hypothesized word sequences associated with O_r. The *N*-best list contains the *N* most likely sequences generated by applying the Viterbi algorithm, which has to retain at least *N*-best search hypotheses at both the HMM (Hidden Markov Model) acoustic model-level and word-level recombination points during the speech decoding process. For each hypothesized word sequence on the *N*-best list, it is relatively easy to compute the standard Levenshtein distance to the correct transcription directly. Based on this observation, Kaiser *et al.* proposed overall risk criterion estimation (ORCE) for acoustic model training [Kaiser *et al.* 2000, 2002; Na *et al.* 1995]. This approach takes the *N*-best list as the reduced hypothesized space to obtain training statistics, and applies the extended Baum-Welch algorithm [Gopalakrishnan *et al.* 1991; Normandin 1991] for parameter optimization. In experiments on the TIMIT database, the authors achieved a 21% word error rate reduction compared to the baseline system. However, an *N*-best list usually contains too much redundant information, i.e., two hypothesized word sequences may look very similar, which makes the training procedure inefficient. An alternative representation is the word lattice (or graph), illustrated in Figure 1, which only stores hypothesized word arcs at different segments of the time frames. Although it cannot be guaranteed that all word sequences generated from a word lattice will have higher probabilities than those not presented, it is believed that the approximation will not affect the performance significantly. Nevertheless, for the lattice structure, using the standard Levenshtein distance measure as the loss function is an issue, since it makes the implementation of computing the distance more complicated. Recently, two approaches have been proposed to deal with this problem. One focuses on how to design loss functions that approximate the Levenshtein distance measure, such as MPE training. The other concentrates on the design of algorithms to segment the word lattice so as to make the computation of the Levenshtein distance feasible, such as the minimum Bayes risk discriminative training (MBRDT) approach [Doumpiotis *et al.* 2003, 2004]. To efficiently reduce the complexity of the hypothesized space in MBRDT, a lattice segmentation algorithm is applied to divide the lattice into several non-overlapping components. It has been shown that MBRDT achieves a considerable performance improvement over the baseline system trained with the maximum likelihood (ML) criterion.

The MPE training approach, which is one of the most attractive discriminative training techniques, tries to optimize an acoustic model's parameters by minimizing the expected phone error rate. The objective function of MPE is given as [Povey 2004]:

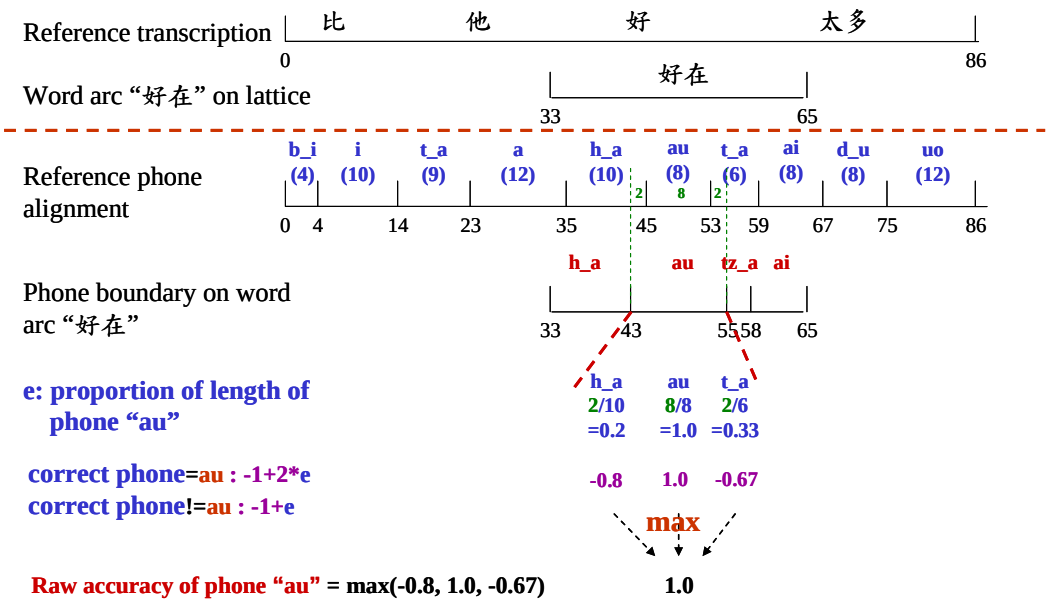

Figure 2. Raw phone accuracy calculation.

$$F_{MPE}(\lambda) = \sum_{r} \frac{\sum_{s \in \mathbf{W}_{lat}^{r}} p_{\lambda}(O_r \mid s) P(s) A(c_r, s)}{\sum_{u \in \mathbf{W}_{lat}^{r}} p_{\lambda}(O_r \mid u) P(u)}, \tag{9}$$

where $\mathbf{W}_{lat}^{r}$ is the lattice generated by the speech recognizer, used to represent a reduced hypothesized space of word sequences; and $A(c_r, s)$ is the raw accuracy of word sequence s, which is an approximation of the true accuracy computed globally using the standard Levenshtein distance. It is obvious that maximizing the objective function is equivalent to minimizing the expected phone error. The raw accuracy $A(c_r, s)$ is defined as:

$$A(c_r, s) = \sum_{q \in s} A'(c_r, q), \tag{10}$$

where q is the phone involved in s, and $A'(c_r, q)$ is a local function used to calculate the raw phone accuracy of each phone q in s. The phone accuracy is calculated locally on each phone arc of the word lattice, instead of globally on each hypothesized word sequence. Given a word arc on the word lattice, the time boundaries of the phone arcs can be determined by aligning the corresponding speech segment with its constituent HMM acoustic models. Figure 2 shows the calculation of raw phone accuracy. Notice that we adopt INITIAL/FINAL units instead of phone units as the acoustic units in our Mandarin LVCSR system. Therefore, for

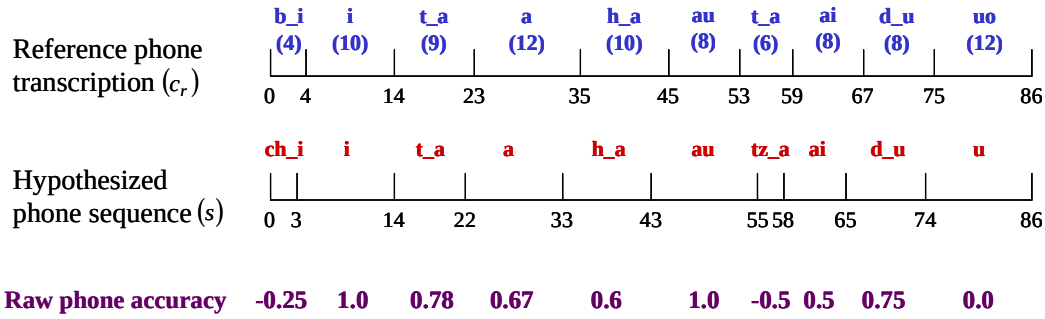

Figure 3. Approximate accuracy versus exact accuracy.

simplicity, each INITIAL or FINAL unit is regarded as a phone in the elucidation. In Figure 2, the raw phone accuracy of phone *"au"* involved in the word arc "好在" is calculated in the following steps. First, the word arc "好在" is aligned with time boundaries of a phone sequence to obtain the start and end time boundaries of the phone *"au"*. Second, for each phone q' in the correct transcription, we calculate the overlapped portion of *"au"* in time frames, and denote it as $e(q',"au")$. Finally, the raw phone accuracy of phone *"au"*, i.e., $A'(c_r,"au")$, is calculated using the following formula:

$$A'(c_r,"au") = \max_{q'} \begin{cases} -1+2e("au",q') & \text{if } q'="au" \\ -1+e("au",q') & \text{otherwise} \end{cases}. \tag{11}$$

It is obvious that $A'(c_r,"au")$ ranges from 1 to $-1+1/T_r$, where T_r is the length of observation O_r in terms of the time frames. For example, if the phone arc *"au"* overlays at least one phone q' in the correct transcription with the same identity in time, *"au"* is considered to be a correct phone, i.e., $A'(c_r,"au")=1$. Figure 3 compares the accuracy of a hypothesized word sequence obtained via the approximate function discussed here and the exact calculation using the Levenshtein distance.

According to Povey's work [Povey 2004], the auxiliary function for optimizing the objective function of MPE in Eq. (9) is

$$H_{MPE}(\lambda,\bar{\lambda}) = \sum_r \sum_{q\in \mathbf{W}_{lat}^r} \left.\frac{\partial F_{MPE}(\lambda)}{\partial \log p_\lambda(O_r \mid q)}\right|_{\lambda=\bar{\lambda}} \log p(O_r \mid q), \tag{12}$$

where $\bar{\lambda}$ is the current model parameter set, q is a specific phone arc in $\mathbf{W}_{lat}^{r}$, and $p_{\lambda}(O_r \mid q)$ is the likelihood given the phone arc q. Note that $H_{MPE}(\lambda, \bar{\lambda})$ is a weak-sense auxiliary function of $F_{MPE}(\lambda)$ around $\lambda = \bar{\lambda}$ with the following property:

$$\left.\frac{\partial F_{MPE}(\lambda)}{\partial \lambda}\right|_{\lambda=\bar{\lambda}} = \left.\frac{\partial H_{MPE}(\lambda, \bar{\lambda})}{\partial \lambda}\right|_{\lambda=\bar{\lambda}}. \tag{13}$$

In other words, both the objective and auxiliary functions have the same derivative with respect to λ when they are evaluated at the current estimate $\bar{\lambda}$. For simplicity, we only consider the MPE-based estimation of mean vectors and covariance matrices in HMMs. The state transition probabilities and mixture weights trained by the ML criterion remain unchanged. As a result, in this study, the final auxiliary function for MPE training is expressed as:

$$g_{MPE}(\lambda, \bar{\lambda}) = \sum_r \sum_{q \in \mathbf{W}_{lat}^{r}} \sum_{t=s_q}^{t=e_q} \sum_{m} \gamma_q^{rMPE} \gamma_{qm}^{r}(t) \log N(o_r(t), \mu_m, \Sigma_m), \tag{14}$$

where s_q and e_q represent the start and end times of the phone arc q, respectively; m is the mixture index of the acoustic models; μ_m and Σ_m are, respectively, the mean vector and covariance matrix for mixture m; $\gamma_{qm}^{r}(t)$ is the occupation probability for mixture m on q; $o_r(t)$ is the observation vector at time t; and γ_q^{rMPE} represents $\left.\frac{\partial F_{MPE}(\lambda)}{\partial \log p_{\lambda}(O_r \mid q)}\right|_{\lambda=\bar{\lambda}}$ in Eq. (12), which can be expressed as:

$$\begin{aligned}\left.\frac{\partial F_{MPE}(\lambda)}{\partial \log p_{\lambda}(O_r \mid q)}\right|_{\lambda=\bar{\lambda}} = & \frac{\sum_{v' \in \mathbf{W}_{lat}^{r}, q \in v'} p_{\bar{\lambda}}(O_r \mid v') P(v') A(v', s_r)}{\sum_{u' \in \mathbf{W}_{lat}^{r}, q \in u'} p_{\bar{\lambda}}(O_r \mid u') P(u')} \frac{\sum_{u' \in \mathbf{W}_{lat}^{r}, q \in u'} p_{\bar{\lambda}}(O_r \mid u') P(u')}{\sum_{u \in \mathbf{W}_{lat}^{r}} p_{\bar{\lambda}}(O_r \mid u) P(u)} \\ & - \frac{\sum_{v \in \mathbf{W}_{lat}^{r}} p_{\bar{\lambda}}(O_r \mid v) P(v) A(v, s_r)}{\sum_{u \in \mathbf{W}_{lat}^{r}} p_{\bar{\lambda}}(O_r \mid u) P(u)} \frac{\sum_{u' \in \mathbf{W}_{lat}^{r}, q \in u'} p_{\bar{\lambda}}(O_r \mid u') P(u')}{\sum_{u \in \mathbf{W}_{lat}^{r}} p_{\bar{\lambda}}(O_r \mid u) P(u)}\end{aligned}. \tag{15}$$

In Eq. (15), $\frac{\sum_{u' \in \mathbf{W}_{lat}^{r}, q \in u'} p_{\bar{\lambda}}(O_r \mid u') P(u')}{\sum_{u \in \mathbf{W}_{lat}^{r}} p_{\bar{\lambda}}(O_r \mid u) P(u)}$ is the occupation probability of phone arc q; $\frac{\sum_{v' \in \mathbf{W}_{lat}^{r}, q \in v'} p_{\bar{\lambda}}(O_r \mid v') P(v') A(v', s_r)}{\sum_{u' \in \mathbf{W}_{lat}^{r}, q \in u'} p_{\bar{\lambda}}(O_r \mid u') P(u')}$ is the weighted average accuracy of hypothesized word sequences

in $\mathbf{W}_{lat}^{r}$ that include q; and $\frac{\sum_{v\in\mathbf{W}_{lat}^{r}} p_{\bar{\lambda}}(O_r \mid v)P(v)A(v,s_r)}{\sum_{u\in\mathbf{W}_{lat}^{r}} p_{\bar{\lambda}}(O_r \mid u)P(u)}$ is the weighted average accuracy of all hypothesized word sequences in $\mathbf{W}_{lat}^{r}$. All three quantities can be calculated efficiently.

Since maximizing the weak sense auxiliary function with respect to λ does not guarantee an increase in the objective function, the auxiliary function is augmented with an extra smoothing function $g_{EB}^{smooth}(\lambda,\bar{\lambda})$ to moderate the parameter update and prevent extreme parameter values being estimated. The following is an example of a smoothing function:

$$g_{EB}^{smooth}(\lambda,\bar{\lambda}) = \sum_{m} -\frac{D_m}{2}\left[\log(|\Sigma_m|) + (\mu_m-\bar{\mu}_m)^T\Sigma_m^{-1}(\mu_m-\bar{\mu}_m) + tr(\bar{\Sigma}_m\Sigma_m^{-1})\right], \tag{16}$$

where D_m is a per-mixture level controlling constant. Note that $g_{EB}^{smooth}(\lambda,\bar{\lambda})$ is deemed a log-Gaussian prior distribution with a differential value of zero with respect to λ when it is evaluated at the current estimate $\bar{\lambda}$. Therefore, the differentials of the augmented auxiliary function with respect to μ_m and Σ_m are computed as shown, respectively, in the following equations:

$$\frac{\partial(g_{MPE}(\lambda,\bar{\lambda}) + g_{EB}^{smmoth}(\lambda,\bar{\lambda}))}{\partial\mu_m} = \sum_r \sum_{q\in\mathbf{W}_{lat}^{r}} \sum_{t=s_q}^{t=e_q} \gamma_q^{rMPE}\gamma_{qm}^{r}(t)\Sigma_m^{-1}\left(o_r(t)-\mu_m\right) - D_m\left[\Sigma_m^{-1}(\mu_m-\bar{\mu}_m)\right], \tag{17}$$

$$\frac{\partial(g_{MPE}(\lambda,\bar{\lambda}) + g_{EB}^{smooth}(\lambda,\bar{\lambda}))}{\partial\Sigma_m^{-1}} = \sum_r \sum_{q\in\mathbf{W}_{lat}^{r}} \sum_{t=s_q}^{t=e_q} \gamma_q^{rMPE}\gamma_{qm}^{r}(t)\left[\frac{1}{2}\Sigma_m^T - \frac{1}{2}\left(\left(o_r(t)-\mu_m\right)\left(o_r(t)-\mu_m\right)^T\right)\right]. + \frac{D_m}{2}\left[\Sigma_m^T - (\mu_m-\bar{\mu}_m)(\mu_m-\bar{\mu}_m)^T - \bar{\Sigma}_m^T\right] \tag{18}$$

Next, by completing the differentiations and equating the above equations to zero, the following Extended Baum-Welch (EB) update formulae [Normandin 1991] are derived:

$$\mu_m = \frac{\sum_r \sum_{q\in\mathbf{W}_{lat}^{r}} \sum_{t=s_q}^{t=e_q} \gamma_q^{rMPE}\gamma_{qm}^{r}(t)o_r(t) + D_m\bar{\mu}_m}{\sum_r \sum_{q\in\mathbf{W}_{lat}^{r}} \sum_{t=s_q}^{t=e_q} \gamma_q^{rMPE}\gamma_{qm}^{r}(t) + D_m}, \tag{19}$$

$$\Sigma_m = \frac{\sum_r \sum_{q \in \mathbf{W}_{lat}^r} \sum_{t=s_q}^{t=e_q} \gamma_q^{rMPE} \gamma_{qm}^r(t) o_r(t) o_r(t)^T + D_m \left[\bar{\Sigma}_m + \bar{\mu}_m \bar{\mu}_m^T \right]}{\sum_r \sum_{q \in \mathbf{W}_{lat}^r} \sum_{t=s_q}^{t=e_q} \gamma_q^{rMPE} \gamma_{qm}^r(t) + D_m} - \mu_m \mu_m^T . \tag{20}$$

Moreover, to incorporate the ML estimate and smooth the update, the so-called I-smoothing technique [Povey and Woodland 2002] is employed to provide a better estimate. I-smoothing is also regarded as a prior distribution for smoothing the auxiliary function, where the mode of the distribution is the same as the estimate obtained by ML training. The update equations thus become:

$$\mu_m = \frac{\sum_r \sum_{q \in \mathbf{W}_{lat}^r} \sum_{t=s_q}^{t=e_q} \gamma_q^{rMPE} \gamma_{qm}^r(t) o_r(t) + D_m \bar{\mu}_m + \frac{\tau_m}{\gamma_m^{ML}} \theta_m^{ML}(O)}{\sum_r \sum_{q \in \mathbf{W}_{lat}^r} \sum_{t=s_q}^{t=e_q} \gamma_q^{rMPE} \gamma_{qm}^r(t) + D_m + \tau_m}, \tag{21}$$

$$\Sigma_m = \frac{\sum_r \sum_{q \in \mathbf{W}_{lat}^r} \sum_{t=s_q}^{t=e_q} \gamma_q^{rMPE} \gamma_{qm}^r(t) o_r(t) o_r(t)^T + D_m \left[\bar{\Sigma}_m + \bar{\mu}_m \bar{\mu}_m^T \right] + \frac{\tau_m}{\gamma_m^{ML}} \theta_m^{ML}(O^2)}{\sum_r \sum_{q \in \mathbf{W}_{lat}^r} \sum_{t=s_q}^{t=e_q} \gamma_q^{rMPE} \gamma_{qm}^r(t) + D_m + \tau_m} - \mu_m \mu_m^T , \tag{22}$$

where τ_m is a constant, and γ_m^{ML}, $\theta_m^{ML}(O)$, and $\theta_m^{ML}(O^2)$ are further expressed, respectively, as:

$$\gamma_m^{ML} = \sum_r \sum_t \gamma_m^{r\,ML}(t), \tag{23}$$

$$\theta_m^{ML}(O) = \sum_r \sum_t \gamma_m^{r\,ML}(t) o_r(t), \tag{24}$$

and

$$\theta_m^{ML}(O^2) = \sum_r \sum_t \gamma_m^{r\,ML}(t) o_r(t) o_r(t)^T . \tag{25}$$

In each of the above equations, $\gamma_m^{r\,ML}(t)$ is the ML occupation probability for mixture m. I-smoothing can also be considered as an interpolation between the MPE estimate and the ML estimate. As $\tau_m \to \infty$, it performs like ML training. On the other hand, it behaves purely as MPE training when $\tau_m \to 0$. Basically, the technique provides better results when the value of τ_m is properly chosen (e.g., we adopted a setting of $\tau_m = 10$ in our experiments). Recently, it has been verified that using the statistics of MMI (Maximum Mutual Information) training in I-smoothing can further improve the estimate [Zheng and Stolcke 2005; Povey *et al.* 2005].

Finally, let us examine the quantity $\gamma_q^{r\,MPE}$ in more detail. To simplify the discussion, we adopt the following equations:

$$\gamma_q^r = \frac{\sum_{u' \in \mathbf{W}_{lat}^r, q \in u'} p_{\bar{\lambda}}(O_r \mid u') P(u')}{\sum_{u \in \mathbf{W}_{lat}^r} p_{\bar{\lambda}}(O_r \mid u) P(u)}, \tag{26}$$

$$c^r(q) = \frac{\sum_{v' \in \mathbf{W}_{lat}^r, q \in v'} p_{\bar{\lambda}}(O_r \mid v') P(v') A(v', s_r)}{\sum_{u' \in \mathbf{W}_{lat}^r, q \in u'} p_{\bar{\lambda}}(O_r \mid u') P(u')}, \tag{27}$$

$$c_{avg}^r = \frac{\sum_{v \in \mathbf{W}_{lat}^r} p_{\bar{\lambda}}(O_r \mid v) P(v) A(v, s_r)}{\sum_{u \in \mathbf{W}_{lat}^r} p_{\bar{\lambda}}(O_r \mid u) P(u)}, \tag{28}$$

where $c^r(q)$ is the weighted average phone accuracy of hypothesized word sequences that involve q; and c_{avg}^r is the weighted average phone accuracy of all hypothesized word sequences in $\mathbf{W}_{lat}^r$. It is clear that the three main statistics must be gathered by applying the forward-backward algorithm to the word lattice [Povey 2004]. Note that the term $c^r(q) - c_{avg}^r$ reflects the difference in the weighted average phone accuracy between the word sequences containing arc q and all word sequences in the lattice. As $c^r(q) = c_{avg}^r$, no training statistics are contributed to phone arc q in MPE training. Positive contributions are made to arc q if $c^r(q)$ is greater than c_{avg}^r, i.e., if phone arc q is more accurate than the average. Conversely, if $c^r(q)$ is smaller than c_{avg}^r, negative contributions are made to arc q and thus show the discrimination. For a reasonable combination of acoustic model likelihoods and language model probabilities, it is necessary to restrict the acoustic likelihoods by introducing an exponential scaling factor. The scaling factor is empirically set depending on the task at hand; in our experiments, we adopted a value of 1/12. Alternatively, a word unigram language model constraint can be used to improve the generalization capabilities of such discriminative

training.

4. MPE-based Linear Regression (MPELR) Adaptation

Acoustic model adaptation, which is one of the most important topics in ASR, tries to eliminate some of the spoken and environmental variations between the training and test sets. However, it is a challenging task to adjust the large number of acoustic model parameters when only a very small amount of data is available for model adaptation. To ensure a more reliable estimation of acoustic model parameters, transformation-based approaches have been developed to adapt the acoustic model indirectly by using a set of affine transforms, such as the maximum likelihood linear regression (MLLR) adaptation [Leggetter and Woodland 1995]. Similarly, word or phone error minimization approaches can be used to estimate the transformation matrices. Among these approaches, we focus on MPE-based linear regression (MPELR) adaptation [Wang and Woodland 2004], which obtains the transformation matrices by using the MPE criterion.

As in typical MLLR adaptation, Gaussian components are first clustered into several regression classes. Components in the same class share the same transformation matrix. The Gaussian mean vectors are transformed by:

$$\mu_m = A_k \bar{\mu}_m + b_k = W_k \bar{\xi}_m , \tag{29}$$

where the subscript k is the class index; $W_k = \left[b_k \; A_k \right]$ is a $d \times (d+1)$ transformation matrix; and $\bar{\xi}_m = \left[1 \; \bar{\mu}_m^T \right]^T$ is the $(d+1)$-dimensional extended mean vector based on the current estimate. Meanwhile, the covariance matrices can be updated by [Gales and Woodland 1996]

$$\Sigma_m = \bar{L}_m^{-T} H_k \bar{L}_m^{-1} , \tag{30}$$

where H_k is the linear transformation matrix to be estimated for the class k, and $\bar{L}_m$ is the Cholesky factor of $\bar{\Sigma}_m^{-1}$. Hereafter, for simplicity, the subscript k representing the cluster index is omitted. Based on Eq. (14), the auxiliary function can be derived as:

$$g_{MPE}(\{W,H\},\{\bar{W},\bar{H}\}) = \sum_m \sum_{q \in \mathbf{W}_{lat}^r} \sum_{t=s_q}^{t=e_q} \gamma_q^{MPE} \gamma_{qm}(t) \log N(o(t); W\xi_m, L_m^{-T} H L_m^{-1}) . \tag{31}$$

Like MPE training, described in Section 3, the auxiliary function in Eq. (31) can be further augmented with an extra smoothing function $g_{EBW}^{smooth}(\{W,H\},\{\bar{W},\bar{H}\})$ to derive a more reliable estimation of the transformation matrices. This is usually given by:

$$
\begin{aligned}
& g_{EBW}^{smooth}(\{W,H\},\{\bar{W},\bar{H}\}) \\
&= \sum_m -\frac{D_m}{2}\Big[\log\Big(|\,L_m^{-T}HL_m^{-1}\,|\Big) + (W\xi_m - \bar{W}\xi_m)^T L_m H^{-1} L_m^T (\bar{W}\xi_m - W\xi_m)\,, \\
& \quad + tr\Big(L_m^{-T}\bar{H}L_m^{-1}L_m H^{-1}L_m^T\Big)\Big]
\end{aligned} \tag{32}
$$

where $tr(\cdot)$ is the standard matrix trace operation. After differentiating the auxiliary function with respect to W and setting it to zero, we get the following closed-form solution:

$$
\begin{aligned}
& \sum_m \Sigma_m^{-1}\left(\sum_{q\in\mathbf{W}_{lat}^r}\sum_{t=s_q}^{t=e_q}\gamma_q^{MPE}\gamma_{qm}(t)o(t) + D_m\bar{W}\xi_m\right)\xi_m^T \\
&= \sum_m \left(\sum_{q\in\mathbf{W}_{lat}^r}\sum_{t=s_q}^{t=e_q}\gamma_q^{MPE}\gamma_{qm}(t) + D_m\right)\Sigma_m^{-1}W\xi_m\xi_m^T
\end{aligned}\,. \tag{33}
$$

The above equation can be solved row-by-row using the Gaussian elimination method to obtain the re-estimation formula for the transformation matrix of mean vectors. The re-estimation formula for the transformation matrix of covariance matrices can be derived in a similar way.

Again, to improve the generalization of the test set, extra prior information, such as the ML statistics, can be considered. Therefore, the final auxiliary function employed in this paper is augmented with the following smoothing function:

$$
g^{I-smooth}(W,H) = \sum_m \frac{\tau_m}{\gamma_m^{ML}}\sum_t \gamma_m{}^{ML}(t)\log N(o(t);W\xi_m, L_m^{-T}HL_m^{-1})\,. \tag{34}
$$

5. Word Error Minimization (WEM) Decoding

Given a speech utterance, the standard maximum a posteriori (MAP) decoding approach tries to output the hypothesized word sequence with the highest posterior probability. Actually, by substituting a zero-one loss function into Eq. (2), the MAP decoding formula can be derived. This implies that the MAP decoding approach is based on minimizing the string error rate (SER). Thus, it only provides suboptimal results when the ASR performance is measured in terms of the word error rate (WER) or the character error rate (CER). Hence, replacing the zero-one loss function in Eq. (2) with the Levenshtein distance measure leads to the WEM decoding approach, which finds the hypothesized word sequence with the minimum WER or CER. However, as mentioned in Section 3, a direct implementation of WEM decoding with the word lattice is complicated because there is still no efficient algorithm for computing the

Table 1. Detailed statistics of the training and test sets.

Gender	Training set			Test set			#Speakers in the training and test sets
	Total length (sec)	Total Syllables	#Speakers	Total length (sec)	Total Syllables	#Speakers	
Male	46,001.3	545,732	≤ 66	1,301.4	26,219	9	9
Female	46,007.2		≤ 111	3,914.0		≤ 23	≥ 13

Levenshtein distance between any two possible word sequences in the word lattice. To make the implementation of the WEM decoding approach feasible, we initially employ an *N*-best list of hypothesized word sequences. The WEM decoding approach can then be applied explicitly by choosing the hypothesized word sequence with the minimum expected risk [Stolcke *et al.* 1997]. The decision formula can thus be expressed as:

$$\alpha_{opt}(O) = \arg\min_{u \in N-\text{Nest}} \sum_{s \in N-\text{Nest}} \frac{p(O \mid s)p(s)}{\sum_{v \in N-\text{Nest}} p(O \mid v)p(v)} L(u,s), \tag{35}$$

where u, s, and v are hypothesized word sequences in the *N*-best list. Similar ideas have been proposed recently by Mangu *et al.* [Mangu *et al.* 2000] and Goel and Byrne [Goel and Byrne 2000]. As an alternative, a novel optimal Bayes decision (OBC) approach for word lattice rescoring has been developed [Chien *et al.* 2006]. It also provides a promising framework for WEM decoding.

6. Experiment Setup

In this section, we describe the large vocabulary continuous speech recognition system and the speech and text data used in this paper.

6.1 Front-End Signal Processing

Front-end processing was performed with the HLDA-based (Heteroscedastic Linear Discriminant Analysis) data-driven Mel-frequency feature extraction approach, and then processed by MLLT (Maximum Likelihood Linear Transformation) transformation for feature de-correlation. In addition, utterance-based feature mean subtraction and variance normalization were applied to all the training and test materials.

6.2 Speech Corpus and Acoustic Model Training

The speech corpus consisted of approximately 198 hours of MATBN (Mandarin Across Taiwan Broadcast News) Mandarin television news content [Wang *et al.* 2005], which was collected by Academia Sinica and the Public Television Service Foundation of Taiwan between November 2001 and April 2003. All the speech materials were manually segmented into separate stories, each of which was spoken by one news anchor, several field reporters, and interviewees. Some stories contained background noise, speech, and music. All 198 hours of speech data was accompanied by corresponding orthographic transcripts, of which about 25 hours of gender-balanced speech data of the field reporters collected from November 2001 to December 2002 was used to bootstrap the acoustic training. The training set consisted of 545,732 syllables and the average length of a word was 1.65 characters. Another set of data, 1.5 hours in length, collected during 2003 was reserved for testing. Due to the limited number of distinct field reporters in the corpus, some test data belonged to the training field reporters. The test set consisted of 26,219 syllables and the average word length was also 1.65 characters. Table 1 shows the detailed statistics of the training and test sets.

The acoustic models chosen for speech recognition were a silence model, 112 right-context-dependent INITIAL models, and 38 context-independent FINAL models. Each INITIAL model was represented by an HMM with 3 states, while each FINAL model had 4 states. Note that gender-independent models were used. The Gaussian mixture number per state ranged from 2 to 128, depending on the amount of training data. The acoustic models were first trained using the ML criterion and the Baum-Welch updating formulae. The MPE-based and MMI (Maximum Mutual Information)-based [Povey and Woodland 2002] acoustic model training approaches were further applied to acoustic models pre-trained by the ML criterion. Unigram language model constraints were used to collect the training statistics from the word lattices for these two training approaches. For MPE training, both silence and short-pause labels were involved in the calculation of the raw phone accuracy of the hypothesized word sequences.

6.3 Lexicon and N-gram Language Modeling

Initially, the recognition lexicon consisted of 67K words. A set of about 5K compound words was automatically derived using forward and backward bigram statistics and added to the lexicon to form a new lexicon of 72K words. The background language models used in this experiment were trigram and bigram models, which were estimated according to the ML criterion using a text corpus consisting of 170 million Chinese characters collected from the Central News Agency (CNA) in 2001 and 2002 (the Chinese Gigaword Corpus released by LDC). The *N*-gram language models were trained with Katz back-off smoothing technique using the SRI Language Modeling Toolkit (SRILM) [Stolcke 2000].

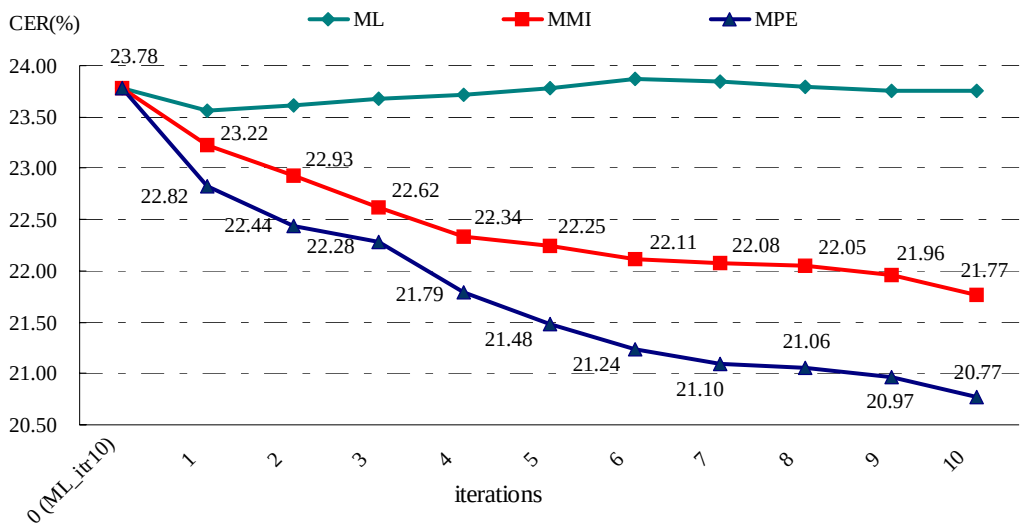

Figure 4. Recognition results, in terms of the CER, for three systems trained on ML, MMI, and MPE criteria, respectively.

6.4 Speech Recognition

The speech recognizer was implemented with a left-to-right frame-synchronous Viterbi tree-copy search and a lexical prefix tree of the lexicon. For each speech frame, a beam pruning technique, which considered the decoding scores of path hypotheses together with their corresponding unigram language model look-ahead scores and syllable-level acoustic look-ahead scores [Chen *et al.* 2005], was used to select the most promising path hypotheses. Moreover, if the word hypotheses ending at each speech frame had higher scores than a predefined threshold, their associated decoding information, such as the word start and end frames, the identities of current and predecessor words, and the acoustic score, were kept to build a word lattice for further language model rescoring. We used the word bigram language model in the tree search procedure and the trigram language model in the word lattice rescoring procedure.

7. Experiment Results and Discussions

Now, a series of experiments performed to assess speech recognition as a function of the acoustic training and adaptation approaches, as well as the speech decoding approaches will be presented.

Table 2. Recognition results of the acoustic model training and unsupervised adaptation approaches

	INITIAL/FINAL Error Rate (%)	Character Error Rate (%)
ML	13.56	23.78
(ML+) MPE	11.12	20.77
(ML+) MPE + MLLR	10.94	20.45
(ML+) MPE + MPELR	10.82	20.29

7.1 Experiments on MPE Acoustic Model Training

The acoustic models of the baseline system were first trained using the ML criterion with 10 iterations of Baum-Welch updating. Then, MPE training (with an optimum setting of $\tau_m = 10$) was applied to the ML-trained acoustic models. In the implementation, we calculated the raw accuracy of each INITIAL/FINAL, instead of each phone, i.e., we had actually performed Minimum INITIAL/FINAL Error training, not Minimum Phone Error training, in the Mandarin LVCSR system. While evaluating the ASR performance, neither the silence nor the short-pause labels were included in the calculation of CER. MMI training was also performed for comparison with MPE training. As mentioned previously, for both MPE and MMI training, unigram language model constraints were imposed when collecting the training statistics from the word lattices. The results for acoustic model training are shown in Figure 4. We observe that the ML-trained baseline system (at the 10th iteration) yields a CER of 23.78%. On the other hand, both MMI and MPE work very well, providing a great boost to the acoustic models initially trained by ML. The acoustic models trained by MPE consistently outperform those trained by MMI across all training iterations. In summary, the MPE-trained acoustic models achieve a relative CER reduction of 12.66% (at the 10th iteration) over those trained by ML. Moreover, as shown in Table 2, the improvements are consistent. The INITIAL/FINAL model error rate is reduced from 13.56% (baseline, ML training only) to 11.12% (at the 10th MPE training iteration). The 18% relative error rate reduction demonstrates the effectiveness of the Minimum INITIAL/FINIAL Error training approach, and the improvement in the acoustic models leads to a 3% absolute reduction in CER (from 23.78% to 20.77%). The use of statistical linguistic rules in MPE training still plays an important role in re-weighting the occupancy statistics, especially in an LVCSR system. In our previous work [Kuo 2005], it was found that much of the CER improvement was lost without embedding the language weight.

The question thus arises: What makes MPE superior to MMI? In Eq. (7), if the summation operator over all training utterances is replaced by the product operator and the loss function is the zero-one function in Eq. (8), one gets the following MMI criterion:

$$\lambda_{MMI} = \arg\max_{\lambda} \sum_{r} \log \frac{p_{\lambda}(O_r \mid s)p(s)}{\sum_{u \in \mathbf{W}_h^r} p_{\lambda}(O_r \mid u)p(u)}, \tag{36}$$

which maximizes the logarithmic product of the posterior probabilities of the reference transcriptions. The use of the zero-one loss function implies that MMI tends to minimize the sentence error rate. Hence, it is reasonable to say that MMI is inferior to MPE in terms of CER.

7.2 Experiments on Unsupervised MPELR Acoustic Model Adaptation

In this subsection, we evaluate the performance of the MPE-based unsupervised acoustic model adaptation approach. In these experiments, utterance-based unsupervised adaptation was used. First, each test utterance was decoded using the MPE-trained acoustic models. Then, after the forward-backward stage to gather sufficient statistics, the acoustic models were adapted according to the recognized transcriptions. All the Gaussian components of the HMM acoustic models were clustered into three broad phonetic regression classes (i.e., INITIAL, FINAL, and Silence) in advance. Only the mean vectors of each Gaussian component were adapted because it has been found that adapting the mean vectors alone yields the most improvement [Gales and Woodland 1996]. Unsupervised MLLR adaptation was performed as the baseline. In the experiment results presented in Table 2, comparing Row 4 (MPE + MLLR) to Row 3 (MPE), we observe that the CER can be reduced from 20.77% to 20.45%, which indicates that MLLR adaptation can, to some extent, effectively mitigate the degradation of ASR performance caused by different acoustic variations. Row 5 of Table 2 gives the error rate obtained by MPELR adaptation. This result, 0.16% improvement in terms of CER, shows that MPELR is slightly better than MLLR. One possible reason for the insignificant improvement over MLLR is the use of a weak-sense auxiliary function. As a result, the convergence speed of MPE-based techniques is not as fast as the strong-sense auxiliary function used in ML-based techniques. In contrast, the advantage of MPE is that it tries to achieve a lower error rate when over-training is encountered. This is why MPE training is performed after ML training and not for bootstrapping the initial models. Similarly, MPELR adaptation can be performed after MLLR adaptation. However repeated on-line adaptation causes the decoding phase to become tardy, which is why it is only performed once in the online stage.

Table 3. Recognition results (CERs) for N-best list WEM rescoring.

	CER (%)
MPE + MPELR	20.29
MPE + MPELR + WEM	20.23
50-best Error Rate	17.82
Lattice Error Rate	10.12

7.3 Experiments on WEM Decoding

For each test utterance, an *N*-best list of hypothesized word sequences was first generated from the word lattice. We limited the number of hypothesized word sequences included in the *N*-best list to 50, and the Levenshtein distance was calculated in terms of character units. The experiment results are shown in Table 3. From Row 3 (MPE + MPELR + WEM), one observes that, with the best set of acoustic models, WEM only achieves a slight reduction of 0.06% in CER compared to that obtained by conventional MAP decoding, as shown in Row 2. Row 5 (Lattice Error Rate) provides the information regarding the lattice error rate [Ortmanns *et al.* 1997], which is the best achievable lower boundary, by rescoring on the current word lattice. This can be computed by finding the best hypothesized word sequence with the minimum Levenshtein distance to the reference transcription from the corresponding word lattice. On the other hand, Row 4 (50-best Error Rate) gives the lower boundary of the best character error rate for the top 50 hypotheses with the highest scores, which is the true best achievable lower bound in our implementation. From the experiment results, the WEM algorithm seems to achieve an almost imperceptible improvement of about 0.06%. The most likely explanation is that there is a defect in the approximation of the posterior distribution. In addition, the WEM algorithm decides the word sequence with the highest posterior probability in most situations [Schlüter *et al*. 2005]. For the above reasons, we consider that the improvement in CER accuracy is insignificant.

8. Conclusions

In this paper, we have investigated the following word error minimization approaches for Mandarin large vocabulary continuous speech recognition: 1) the MPE criterion used in acoustic model training and adaptation; and 2) the WEM criterion in speech decoding. Unlike conventional techniques, these two approaches try to minimize the expected word error, rather than the string-level error. Experiments on the MATBN corpus demonstrate that MPE training can significantly improve a system initially trained with the ML criterion. Likewise, MPELR adaptation can significantly reduce the CER for the unsupervised adaptation task. This result is superior to that obtained by conventional MLLR adaptation. Finally, *N*-best rescoring using the WEM criterion achieves a slight improvement over traditional MAP decoding. We are

currently conducting an in-depth investigation of the WEM approaches to language modeling [Kuo and Chen, 2005], as well as their comparison and integration with other approaches.

References

Chen, B., J.-W. Kuo, W.-H. Tsai, "Lightly Supervised and Data-Driven Approaches to Mandarin Broadcast News Transcription," *International Journal of Computational Linguistics and Chinese Language Processing*, 10(1), 2005, pp.1-18.

Chien, J.-T., C.-H. Huang, K. Shinoda and S. Furui, "Towards Optimal Bayes Decision for Speech Recognition," in *Proc. ICASSP'06*, 2006.

Doumpiotis, V., S. Tsakalidis and W. Byrne, "Discriminative Training for Segmental Minimum Bayes Risk Decoding," in *Proc. ICASSP'03*, 2003.

Doumpiotis, V., S. Tsakalidis and W. Byrne, "Lattice Segmentation and Minimum Bayes Risk Discriminative Training," in *Proc. Eurospeech'03*, 2003.

Doumpiotis, V. and W. Byrne, "Pinched Lattice Minimum Bayes Risk Discriminative Training for Large Vocabulary Continuous Speech Recognition," in *Proc. ICSLP'04*, 2004.

Duda, R. O., P. E. Hart and D. G. Stork, *Pattern Classification*, 2nd ed. New York: John and Wiley, 2000.

Gales, M. J. F. and P. C. Woodland, "Mean and Variance Adaptation within the MLLR Framework," *Computer Speech and Language*, 10, 1996, pp.249-264.

Goel, V. and W. Byrne, "Minimum Bayes-Risk Automatic Speech Recognition," *Computer Speech and Language*, 14, 2000, pp.115-135.

Gopalakrishnan, P. S., D. Kanevsky, A. Nádas and D. Nahamoo, "An Inequality for Rational Functions with Applications to Some Statistical Estimation Problems," *IEEE Trans. Information Theory*, 37, 1991, pp.107-113.

Kaiser, J., B. Horvat and Z. Kacic, "A Novel Loss Function for the Overall Risk Criterion Based Discriminative Training of HMM Models," in *Proc. ICSLP'00*, 2000.

Kaiser, J., B. Horvat and Z. Kacic, "Overall Risk Criterion Estimation of Hidden Markov Model Parameters," *Speech Communication*, 38, 2000, pp.383-398.

Kuo, J.-W. and B. Chen, "Minimum Word Error Based Discriminative Training of Language Models," in *Proc. INTERSPEECH'05*, 2005.

Kuo, J.-W, "An Initial Study on Minimum Phone Error Discriminative Learning of Acoustic Models for Mandarin Large Vocabulary Continuous Speech Recognition," *Master Thesis, National Taiwan Normal University*, June 2005.

Leggetter, C. J. and P. C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Computer Speech and Language*, 9, 1995, pp.171-185.

Levenshtein, A., "Binary Codes Capable of Correcting Deletions, Insertions and Reversals," *Soviet Physics Doklady*, 10(8), 1966, pp.707-710.

Mangu, L., E. Brill, and A. Stolcke, "Finding consensus in speech recognition: word error minimization and other applications of confusion networks," *Computer Speech and Language*, 14, 2000, pp.373-400.

Na, K., B. Jeon, D. Chang, S. Chae, and S. Ann, "Discriminative Training of Hidden Markov Models using Overall Risk Criterion and Reduced Gradient Method," in *Proc. Eurospeech'95*, 1995.

Normandin, Y., "Hidden Markov Models, Maximum Mutual Information Estimation, and the Speech Recognition Proble," *Ph.D Dissertation, McGill University, Montreal*, 1991.

Ortmanns, S., H. Ney and X. Aubert, "A Word Graph Algorithm for Large Vocabulary Continuous Speech Recognition," *Computer Speech and Language*, 11, 1997, pp.43-72.

Povey, D. and P. C. Woodland, "Minimum Phone Error and I-smoothing for Improved Discriminative Training," in *Proc. ICASSP'02*, 2002.

Povey, D and P. C. Woodland, "Large Scale Discriminative Training of Acoustic Models for Speech Recognition," *Computer Speech and Language*, 16, 2002, pp. 25-47.

Povey, D, "Discriminative Training for Large Vocabulary Speech Recognition," *Ph.D Dissertation, Peterhouse, University of Cambridge*, July 2004.

Povey, D., B. Kingsbury, L. Mangu, G. Saon, H. Soltau and G. Zweig, "FMPE: Discriminatively Trained Features for Speech Recognition," in *Proc. ICASSP'05*, 2005.

Schlüter, R., T. Scharrenbach, V. Steinbiss and H. Ney, "Bayes Risk Minimization using Metric Loss Functions," in *Proc. Eurospeech'05*, 2005.

Schwartz, R. and Y.-L. Chow, "The N-best algorithms: an efficient and exact procedure for finding the N most likely sentence hypotheses," in *Proc. ICASSP'90*, 1990.

Stolcke, A., Y. Konig, M. Weintraub, "Explicit Word Error Minimization in N-best List Rescoring," in *Proc. Eurospeech'97*, 1997.

Stolcke, A., SRI language Modeling Toolkit, version 1.3.3, 2000. *http://www.speech.sri.com/projects/srilm/*.

Wang, H.-M., B. Chen, J.-W. Kuo, and S.-S. Cheng, "MATBN: A Mandarin Chinese Broadcast News Corpus," *International Journal of Computational Linguistics and Chinese Language Processing*, 10(2), 2005, pp.219-236.

Wang, L. and P. C. Woodland, "MPE-Based Discriminative Linear Transform for Speaker Adaptation," in *Proc. ICASSP'04*, 2004.

Zheng, J. and A. Stolcke, "Improved Discriminative Training Using Phone Lattices," in *Proc. INTERSPEECH'05*, 2005.