Computational Linguistics and Chinese Language Processing
Vol. 13, No. 3, September 2008, pp. 343-362

Improved Minimum Phone Error based Discriminative Training of Acoustic Models for Mandarin Large Vocabulary Continuous Speech Recognition

Shih-Hung Liu*, Fang-Hui Chu*, Yueng-Tien Lo*, and Berlin Chen*

Abstract

This paper considers minimum phone error (MPE) based discriminative training of acoustic models for Mandarin broadcast news recognition. We present a new phone accuracy function based on the frame-level accuracy of hypothesized phone arcs instead of using the raw phone accuracy function of MPE training. Moreover, a novel data selection approach based on the frame-level normalized entropy of Gaussian posterior probabilities obtained from the word lattice of the training utterance is explored. It has the merit of making the training algorithm focus much more on the training statistics of those frame samples that center nearly around the decision boundary for better discrimination. The underlying characteristics of the presented approaches are extensively investigated, and their performance is verified by comparison with the standard MPE training approach as well as the other related work. Experiments conducted on broadcast news collected in Taiwan demonstrate that the integration of the frame-level phone accuracy calculation and data selection yields slight but consistent improvements over the baseline system.

Keywords: Discriminative Training, Minimum Phone Error, Phone Accuracy Function, Training Data Selection, Large Vocabulary Continuous Speech Recognition

1. Introduction

Speech is the primary and the most convenient means of communication between individuals. Due to the successful development of much smaller electronic devices and the popularity of wireless communication and networking, it is widely believed that speech will possibly serve as a major human-machine interface for the interaction between people and different kinds of smart devices in the near future. On the other hand, huge quantities of multimedia information,

*Department of Computer Science & Information Engineering, National Taiwan Normal University
E-mail: {g93470185, g94470144, g96470198, berlin}@csie.ntnu.edu.tw

[Received January 6, 2008; Revised August 21, 2008; Accepted August 27, 2008]

such as that in broadcast radio and television programs, voice mails, digital archives, and so on are continuously growing and filling our computers, networks, and daily lives. Speech is obviously one of the most important information-bearing sources for the great volumes of multimedia. Based on these observations, it is expected that automatic speech recognition (ASR) technology will play a very important role in human-machine interaction, as well as in organization and retrieval of multimedia content.

When considering the development of an ASR system, acoustic modeling is always an indispensable and crucial ingredient we have to carefully manipulate. The purpose of acoustic modeling is to provide a method for calculating the likelihood of a speech utterance occurring given a word sequence. In principle, the word sequence can be decomposed into a sequence of phone-like (subword, *e.g.* INITIAL or FINAL in Mandarin Chinese) units or acoustic models, each of which is normally represented by a continuous density hidden Markov model (HMM), and the corresponding model parameters can be estimated from a corpus of orthographically transcribed training utterances using maximum likelihood (ML) training [Rabiner 1989]. The acoustic models can be alternatively trained with discriminative training algorithms, such as maximum mutual information (MMI) training [Bahl *et al.* 1986] and minimum phone error (MPE) training [Povey 2004; Kuo *et al*. 2006]. These algorithms were developed in an attempt to correctly discriminate the recognition hypotheses for the best recognition results rather than just to fit the model distributions as done by ML training; therefore, they have continuously been a focus of considerable active research in a wide variety of large vocabulary continuous speech recognition (LVCSR) tasks over the past few years. Moreover, in contrast to ML training, discriminative training considers not only the reference (or correct) transcript of a training utterance, but also the competing (or incorrect) hypotheses that are often obtained by performing LVCSR on the utterance.

In this paper, we consider minimum phone error (MPE) based discriminative training of acoustic models for Mandarin broadcast news recognition. In order to remedy the defect in the phone accuracy function of the MPE training algorithm, we present a new phone accuracy function based on the frame-level accuracy of hypothesized phone arcs. Moreover, a novel data selection approach based on the frame-level normalized entropy of Gaussian posterior probabilities obtained from the word lattice of the training utterance is explored, which has the merit of making the MPE training algorithm focus much more on the training statistics of those frame samples that center nearly around the decision boundary for better discrimination. The underlying characteristics of the presented approaches are extensively investigated and their performance is verified by comparison with the original MPE training approach as well as other related work.

The remainder of this paper is organized as follows. In Section 2, the general background of MPE based acoustic model training is briefly reviewed. Section 3 elucidates our proposed

new accuracy function for MPE training, and Section 4 presents two novel training data selection approaches based on frame-level normalized entropy information. The experimental setup is detailed in Section 5, and a series of speech recognition experiments is described in Section 6. Finally, we present the conclusions drawn from the research in Section 7.

2. Review of Minimum Phone Error (MPE) Training

Given a training set of K acoustic vector sequences $O = \{O_1,..,O_k,..,O_K\}$, the MPE criterion for acoustic model training aims to minimize the expected phone errors of these acoustic vector sequences using the following objective function [Povey and Woodland 2002]:

$$F_{MPE}(\lambda) = \sum_{k=1}^{K} \sum_{W_k \in \mathbf{W}_k^{lat}} RawAcc(W_k) P_\lambda(W_k \mid O_k), \tag{1}$$

where λ denotes a set of phone-like acoustic models; $\mathbf{W}_k^{lat}$ is the corresponding word lattice [Ortmanns *et al.* 1997] of O_k obtained using LVCSR, as graphically illustrated in Figure 1; W_k is one of the hypothesized word sequences in $\mathbf{W}_k^{lat}$; $P(W_k \mid O_k)$ is the posterior probability of hypothesis W_k given O_k; $RawAcc(W_k)$ is the "raw phone accuracy" of W_k in comparison to the corresponding reference transcript, which is typically computed as the sum of the phone accuracy measures of all phone hypotheses in W_k. Then, the objective function in Equation (1) can be maximized by applying the Extended Baum-Welch algorithm [Gopalakrishnan *et al.* 1989] to update the mean μ_{hmd} and variance σ_{hmd}^2 for each dimension d of a Gaussian mixture component m of a multi-state (or single-state)HMM h using the following equations:

$$\mu_{hmd} = \frac{\theta_{hmd}^{num}(O) - \theta_{hmd}^{den}(O) + D\bar{\mu}_{hmd}}{\gamma_{qm}^{num} - \gamma_{qm}^{den} + D}, \tag{2}$$

$$\sigma_{hmd}^2 = \frac{\theta_{hmd}^{num}(O^2) - \theta_{hmd}^{den}(O^2) + D(\bar{\sigma}_{hmd}^2 + \bar{\mu}_{hmd}^2)}{\gamma_{hm}^{num} - \gamma_{hm}^{den} + D} - \mu_{hmd}^2, \tag{3}$$

$$\gamma_{hm}^{num} = \sum_{k=1}^{K} \sum_{q \in \mathbf{W}_k^{lat}, q=h} \sum_{t=s_q}^{e_q} \gamma_{qm}^k(t) \max(0, \gamma_q^{k^{MPE}}), \tag{4}$$

$$\gamma_{hm}^{den} = \sum_{k=1}^{K} \sum_{q \in \mathbf{W}_k^{lat}, q=h} \sum_{t=s_q}^{e_q} \gamma_{qm}^k(t) \max(0, -\gamma_q^{k^{MPE}}), \tag{5}$$

$$\theta_{hmd}^{num}(O) = \sum_{k=1}^{K} \sum_{q \in \mathbf{W}_k^{lat}, q=h} \sum_{t=s_q}^{e_q} \gamma_{qm}^k(t) \max(0, \gamma_q^{k^{MPE}}) o_t(d), \tag{6}$$

$$\theta_{hmd}^{num}(O^2) = \sum_{k=1}^{K} \sum_{q \in \mathbf{W}_k^{lat}, q=h} \sum_{t=s_q}^{e_q} \gamma_{qm}^k(t) \max(0, \gamma_q^{k^{MPE}}) o_t(d)^2, \tag{7}$$

$$\gamma_q^{k^{MPE}} = \gamma_q^k (c_q^k - c_{avg}^k), \tag{8}$$

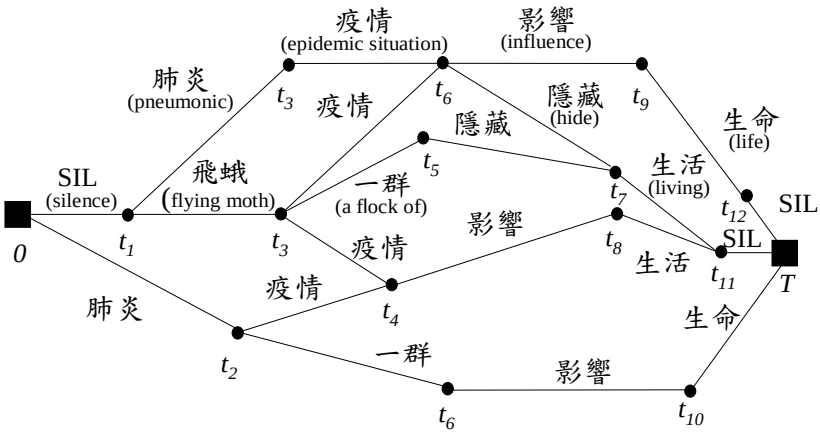


Figure 1. An illustration of a word lattice, in which each arc, together with its corresponding start and end speech frames, represents a candidate word hypothesis. A word arc can be further aligned into a sequence of phone arcs for MPE training.

where $q \in \mathbf{W}_k^{lat}, q = h$ denotes that a phone q arc belongs to the word lattice $\mathbf{W}_k^{lat}$ and physically refers to the HMM h; c_{avg}^k is the average phone accuracy over all hypothesized word sequences in the word lattice; c_q^k is the expected phone accuracy over all hypothesized word sequences containing a phone arc q; $o_t(d)$ is the observation vector component at frame t; s_q and e_q are the start and end times of phone arc q; γ_q^k the posterior probability for phone arc q of utterance k; $\gamma_{qm}^k(t)$ is the posterior probability for mixture component m of phone arc q of utterance k at frame t; γ_{qm}^{num}, $\theta_{qmd}^{num}(O)$ and $\theta_{qmd}^{num}(O^2)$ are the accumulated training statistics for mixture component m of phone arc q whose c_q^k is larger than c_{avg}^k, and vice-versa for γ_{qm}^{den}, $\theta_{qmd}^{den}(O)$ and $\theta_{qmd}^{den}(O^2)$; $\bar{\mu}_{qmd}$ and $\bar{\sigma}_{qmd}^2$ are, respectively, the mean and variance estimated in the previous iteration; and D is a constant used to ensure positive variance values. On the other hand, the calculation of c_{avg}^k and c_q^k is actually based on the phone accuracies of phone arcs in the word lattice. For example, the raw phone accuracy for each word sequence W_k in the lattice can be calculated in terms of the sum of the accuracy of each phone contained in W_k [Povey and Woodland 2002]:

$$RawAcc(W_k) = \sum_{q \in W_k} PhoneAcc(q), \tag{9}$$

where $PhoneAcc(q)$ is the raw phone accuracy for a phone arc q in W_k, which can be defined as follows:

$$PhoneAcc(q) = \max_{z_j \in Z_k} \begin{Bmatrix} -1 + 2e(z_j, q)/l(z_j), & z_j = q \\ -1 + e(z_j, q)/l(z_j), & z_j \neq q \end{Bmatrix}, \tag{10}$$

where Z_k is the set of phone labels in the corresponding reference transcript, and $e(z_j, q)$ is the overlap length in frames (or in time) for a phone label z_j in Z_k and a hypothesized phone arc q in W_k, $l(z_j)$ is the length in frames for z_j. We can observe from Equations (4)-(8), for MPE training, those hypotheses having raw phone accuracies higher than the average can provide positive contributions, and vice-versa for those hypotheses with accuracies lower than the average. Interested readers can refer to [Povey 2004; Kuo *et al*. 2006] for more derivation details of MPE training.

3. New Accuracy Functions

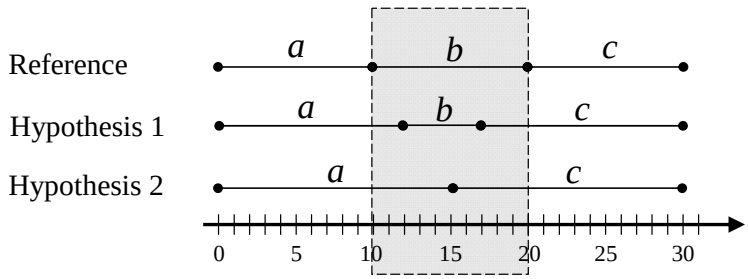
It is known that the standard MPE training approach has some drawbacks [Zheng and Stolcke 2005]. One of them is that MPE training does not sufficiently penalize deletion errors. In general, the original MPE objective function discourages insertion errors more than deletion and substitution errors. Inspired by the work of word lattice rescoring (or decoding) using frame-level accuracy information [Wessel *et al*. 2001], in this paper we present an alternative phone accuracy function that can look into the frame-level phone accuracies of all hypothesized word sequences to replace the original raw phone accuracy function for MPE training [Liu *et al*. 2007a]. The frame-level phone accuracy function (FA) is defined as:

$$FrameAcc(q) = \frac{\sum_{t=s_q}^{e_q} \delta(q, Z_k(t))}{e_q - s_q + 1}, \tag{11}$$

and

$$\delta(q, Z_k(t)) = \begin{Bmatrix} 1 & , if \ \ q = Z_k(t) \\ -\rho & , if \ \ q \neq Z_k(t), \ 0 < \rho < 1 \end{Bmatrix}, \tag{12}$$

where $Z_k(t)$ is the phone label of the reference transcript Z_k at frame t; ρ is a tunable positive parameter used to control the penalty if the phone arc q is incorrect in its label; and the value of $FrameAcc(q)$ will range from $-\rho$ to 1. For each frame t, we thus can easily evaluate whether the phone arc of each hypothesized word sequence in the word lattice is identical to that of the reference transcript or not. Actually, the presented frame-level phone accuracy function emphasizes the deletion penalty on the incompletely correct phone arc; whereas the insertion and substitution errors of the hypothesized word sequences, as well as the errors caused by inaccurate time boundaries of the phone arcs, are also taken into consideration evenly. As illustrated in Figure 2, given the reference phone transcript "*a-b-c*", the first hypothesized phone sequence "*a-b-c*" will be regarded as partially correct (with a score of two) using the original MPE raw phone accuracy function, as shown in Eq. (10);



Raw phone accuracy of Hypothesis 1 = 2

Raw phone accuracy of Hypothesis 2 = 2

Frame-level phone accuracy of Hypothesis 1 = 2.56

Frame-level phone accuracy of Hypothesis 2 = 1.27

Figure 2. An illustration of the frame-level accuracy. The shaded box indicates where the frame-level errors occur.

while the presented frame-level phone accuracy function, as shown in Eq. (11), will give it a score of 2.56 (with ρ set to 0.1) by similarly taking into account the incorrect time boundaries of the associated phone arcs. On the other hand, for the second hypothesized phone sequence "*a*-*c*", it is obvious that there exists a deletion error of the phone arc "*b*." Nevertheless, the original MPE raw phone accuracy function gives the second hypothesized phone sequence a score of two, which is equivalent to that of the first hypothesized phone sequence, and the phone arcs ("*a*" and "*c*") of it will be treated as completely correct. While using our proposed frame-level phone accuracy function, both of the two phone arcs in the second hypothesized phone sequence will instead be treated as partially correct by considering the frame-level substitution errors. Thus, the frame-level phone accuracy function will only assign a total score of 1.27 (with ρ set to 0.1) to the second hypothesized phone sequence.

Another frame-level phone accuracy function that uses the Sigmoid function to normalize the phone accuracy value in a range between -1 and 1 is also investigated in this paper (SFA):

$$SigFrameAcc(q) = \frac{2}{1+\exp(-\alpha \cdot net)} - 1, \tag{13}$$

and

$$net = \sum_{t=s_q}^{e_q} \delta(q, z_k(t)), \tag{14}$$

where $\delta(q, Z_k(t))$ was previously defined in Eq. (12), α is a positive parameter that controls the slope of the Sigmoid function (the larger the value of α, the steeper the slope of the function). Notice that, the purpose of the above two new phone accuracy functions is not to

approximate the standard Levenshtein distance measure, but instead to sufficiently penalize the frame-level substitution errors of each hypothesized phone arc that may be neglected by the original raw phone accuracy function. From now on, the proposed improved MPE training algorithms, by adopting either one of the two frame-level phone accuracy functions defined in Eqs. (11) and (13), are referred to as the maximum frame accuracy training (denoted as MFA) and the maximum Sigmoid-based frame accuracy training (denoted as MSFA), respectively.

In recent years, there also has been considerable independent research on the design of new phone accuracy functions for improving MPE training [Zheng and Stolcke 2005; Gibson *et al.* 2006; Du *et al.* 2006; Povey *et al.* 2007]. As one example, the minimum phone frame error (MPFE) criterion [Zheng and Stolcke 2005] simply counts the number of frames of the recognition hypothesis having correct phone labels in comparison to the reference transcript, which is quite similar to our proposed frame-level accuracy functions. The major differences are that MPFE gives a score of zero (but not a negative value as done by MFA and MSFA) to the frames with incorrect phone labels, and the corresponding phone accuracy value is not normalized by the phone duration or the Sigmoid function. As another example, the state-level minimum Bayes risk (sMBR) criterion [Gibson *et al.* 2006; Povey *et al*. 2007] uses the HMM state-level information to fulfill label matching. As still another example, the minimum divergence (MD) criterion [Jun Du *et al*. 2006] defines phone accuracy on the basis of the Kullback-Leibler divergence between the corresponding acoustic models of the reference and hypothesized phone labels. More detailed elucidation and comparison of these alternative phone accuracy functions can be found in [Povey *et al*. 2007].

Although a discriminative training approach using the finite state transducer, retaining the corresponding recognition hypotheses of the training acoustic vector sequence, for calculating the exact Levenshtein distance based word error rate was also proposed recently [Heigold *et al*. 2005], no improved results but only degraded results were demonstrated by the approach.

4. Frame-Level Training Data Selection

In this section, we elucidate the theoretical roots of frame-level training data selection using the entropy information, as well as two variant implementations to achieve this goal.

4.1 Normalized Frame-Level Entropy

We propose the use of the entropy information to select the frame-level training statistics for the MPE training. The normalized entropy of a training frame sample i can be defined as [Liu *et al.* 2007b]:

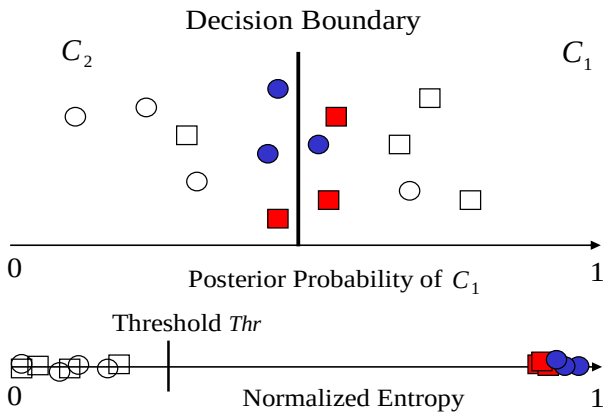


Figure 3. A hypothetical example of binary classification illustrating the relationship between the decision boundary and the normalized entropy.

$$E_k(t) = \frac{1}{\log_2 N_t} \sum_{q \in \mathbf{W}_k^{lat}} \sum_{m \in q} \gamma_{qm}^k(t) \cdot \log_2 \frac{1}{\gamma_{qm}^k(t)}, \tag{15}$$

where $\gamma_{qm}^k(t)$ is the posterior probability for mixture component m of phone arc q at frame t, which is calculated from the word lattice; N_t is the number of Gaussian mixtures which have nonzero posterior probabilities at frame t ($\gamma_{qm}^k(t) > 0$); and the value of $E_k(t)$ will range from zero to one [Misra and Bourlard 2005]. Here, we use a hypothetical example of binary classification to illustrate the relationship between the decision boundary and the normalized entropy. As shown in Figure 3, the decision boundary constructed based on the posterior probability of the class C_1 can discriminate most of the samples belonging to C_1 (depicted as squares) from those belonging to C_2 (depicted as circles). In general, the decision boundary is at the value of 0.5 for the posterior probability of C_1 and the class posterior probabilities can be used to calculate the normalized entropies of the samples. Thus, the samples (solid circles or squares) located near the decision boundary will have normalized entropies close to one, while those (hollow circles or squares) located far away from the decision boundary will have normalized entropies close to zero.

For the speech recognition task, two extreme cases are considered as follows. First, if the normalized entropy measure of a frame sample i is close to zero, it means that the corresponding frame-level posterior probabilities will be dominated by one specific mixture component. From the viewpoint of frame sample classification using posterior probabilities, the difference of probabilities between the true (correct) mixture component and the competing (incorrect) ones is larger. That is, the frame sample i is actually located far from the decision boundary. On the other hand, if the normalized entropy measure is close to one, it

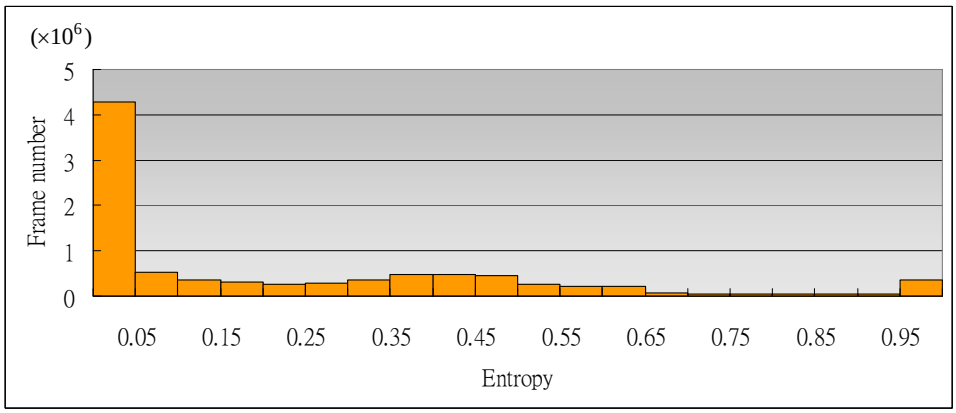


Figure 4. A plot of the relationship between the normalized entropy and the number of training speech frame samples.

means that the posterior probabilities of mixture components tend to be uniformly distributed. Then, the frame sample i is instead located near the decision boundary. In a word, the normalized entropy measure to some extent can define a kind of margin for the selection of useful training frame samples. Therefore, we may take advantage of the normalized entropy measure to make the MPE training focus much more on the training statistics of those frame samples that center near the decision boundary for better sample discrimination and model generalization [Jiang *et al.* 2006; Li *et al*. 2006].

4.2 Hard Version of Frame Sample Selection (HS)

A straightforward implementation of frame-level training data selection is to define a threshold of the normalized entropy measure then completely discard the training statistics of those frame samples whose normalized entropy values fall below it. This can be viewed as a "hard version" of data selection. Figure 4 shows a histogram describing the relationship between the normalized entropy and the number of training speech frame samples used in this study. For example, the leftmost vertical bar denotes the number of training speech frame samples whose normalized entropy values are in the range of 0 to 0.05. The large number of frame samples belonging to the leftmost vertical bar also reveals that most of the training frame samples in fact are located far from the decision boundary; thus, they can be discarded if the threshold is appropriately set.

4.3 Soft Version of Frame Sample Selection (SS)

We also attempt an alternative implementation (or a "soft version") of frame-level training data selection to emphasize the training statistics of those frame samples that are located near

the decision boundary according to their normalized entropy values using the following formula:

$$\gamma_q^{k\,MPE\prime} = \gamma_q^{k\,MPE} \cdot (1 + \omega \cdot E_k(t)), \tag{16}$$

where ω is tunable positive parameter whose value ranges from 0 to 1. As indicated by Equation (16), if the normalized entropy value $E_k(t)$ of a training frame sample i is higher, then its corresponding training statistics will be emphasized. On the contrary, for a frame sample with a lower entropy value, its training statistics will be deemphasized when compared to those of the frame samples with higher normalized entropy values.

5. Experiment Setup

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

5.1 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 were 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 more than five hundred thousand characters, and the average length of a word was 1.65 characters. Another set of about 1.5 hours of speech data of the filed reporters (more than twenty-six thousand characters) collected during 2003 was reserved for testing. The training and test data overlapped in speakers; roughly 30% of the test data was spoken by the field reporters whose previous recordings were also included in the 25-hour training data.

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 update formulas. The MPE-based acoustic model training was further applied to acoustic models pre-trained by the ML criterion. Both silence and short-pause labels were involved in the calculation of the raw

phone accuracy of the word sequence hypotheses for the MPE training.

5.2 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 [Saon and Padmanabhan 2001] 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). In implementation, the *n*-gram language models were trained with the SRI Language Modeling Toolkit [Stolcke 2000].

5.3 Speech Recognition System

The front-end processing for speech recognition was performed with the HLDA-based (Heteroscedastic Linear Discriminant Analysis) data-driven Mel-frequency feature extraction approach [Kumar 1997] then processed by MLLT (Maximum Likelihood Linear Transformation) transformation [Saon *et al.* 2000] for feature de-correlation. In addition, utterance-based feature mean subtraction and variance normalization were applied to all the training and test speech.

The speech recognizer was implemented with a left-to-right frame-synchronous Viterbi tree-copy search and a lexical prefix tree of the lexicon [Aubert 2002]. 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 [Ortmanns *et al*. 1997].

6. Experiment Results

As it is known that there are no explicit marks, such as spaces or blanks, separating words in the Chinese language, the Chinese language often suffers from word tokenization problems. The performance evaluation metric used in Mandarin speech recognition usually is the character error rate (CER) rather than the word error rate (WER).

Table 1. CER results (%) obtained for different parameter settings of the MPE training using two variant phone accuracy functions (MFA and MSFA).

Iterations	MPE	MFA ρ =0.1	MFA ρ =0.3	MFA ρ =0.5	MFA ρ =0.8	MSFA ρ =0.1 α = 0.5	MSFA ρ =0.5 α = 0.5	MSFA ρ =0.1 α = 1	MSFA ρ =0.5 α = 1
1	22.82	22.85	22.73	22.74	22.80	22.88	22.82	22.83	22.77
2	22.44	22.35	22.33	22.36	22.39	22.37	22.34	22.37	22.38
3	22.28	22.07	22.13	22.14	22.19	22.06	22.10	22.02	22.05
4	21.79	21.65	21.50	21.56	21.69	21.52	21.58	21.41	21.56
5	21.48	21.26	21.14	21.26	21.34	21.23	21.47	21.30	21.52
6	21.24	20.98	20.97	21.09	21.23	21.05	21.27	21.06	21.32
7	21.10	20.91	20.87	21.09	21.19	20.89	21.11	20.80	21.19
8	21.06	20.87	20.81	20.82	20.93	20.50	20.97	20.54	20.98
9	20.97	20.84	20.74	20.85	20.90	20.58	20.82	20.57	21.03
10	20.77	20.82	20.80	20.72	20.93	20.46	20.87	20.65	21.10

6.1 Baseline System

The acoustic models were trained with about 25 hours of speech utterances. The MPE training started with the acoustic models trained by 10 iterations of the ML training, and used the information contained in the associated word lattices of training utterances to accumulate the necessary statistics for model training. The ML-trained acoustic models yields a CER (Chinese Character Error Rate) of 23.64%, while the standard MPE training (denoted as MPE) indeed can provide a great boost to the acoustic models initially trained by ML consistently at all training iterations, as the curve "MPE" depicted in Figure 4 or the results shown in the leftmost column of Table 1.

In the following experiments, for fair comparison between our proposed methods and the baseline MPE training, the smoothing constant (*i.e.*, the τ value of I-smoothing) [Povey and Woodland 2002; Povey 2004; Kuo *et al*. 2006] is set to be the same as that used in the baseline MPE training. It is known that this smoothing constant can be regarded as a kind of prior information which forces the HMM parameters estimated by the MPE training to center around that estimated by the ML training [Povey *et al*. 2007].

6.2 Experiments on Proposed Frame-level Phone Accuracy Functions

We first evaluate the performance of our proposed two frame-level phone accuracy functions, FA (corresponding to the MFA training) and SFA (corresponding to the MSFA training), as previously described in Section 3. As can be seen from Figure 5, both MFA and MSFA

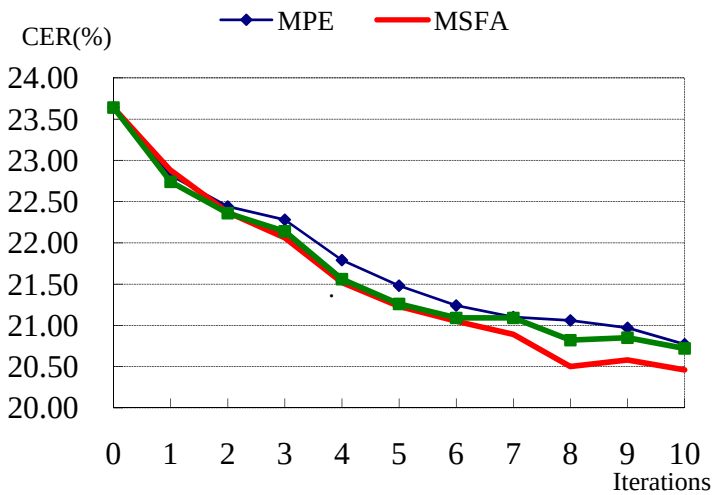


Figure 5. CER results (%) of two new phone accuracy functions in comparison with the standard MPE training.

outperform the standard MPE at higher training iterations, and MSFA is slightly better than MFA, though the difference between them is negligible at lower training iterations. On the other hand, we have observed from a series of experiments that, using the two variants of frame-level phone accuracy functions with different settings of the value of their parameter ρ will give different penalties for insertions and deletions. For example, if the value of ρ is set to be larger, insertion errors will be discouraged; while, if the value of ρ is set to be smaller, the number of deletion errors will be decreased. More concretely, we can trade off insertion and deletion errors by appropriately adjusting the penalty parameter ρ. Table 1 shows the results obtained for different parameter settings of the two variant phone accuracy functions, where the optimum setting for MFA is ρ=0.5, while for MSFA is ρ=0.1 and $\alpha = 0.5$. MFA (ρ=0.5) trained with 10 iterations (20.46%) leads to an absolute CER reduction of 0.31% over MPE trained with the same iterations (20.77%), which is equivalent to a condition where about 81 of the character recognition errors have been corrected. A significance test based on the standard NIST MAPSSWE [Gillick and Cox 1989] also indicates the statistical significance of such an improvement (p-value <0.001).

6.3 Comparison of Proposed and Other Phone Accuracy Functions

We then compare our proposed new frame-level phone accuracy function (SFA) with the other alternative modifications (*i.e.*, MPFE, sMBR and MD mentioned in Section 3) to the phone accuracy function for the MPE-based discriminative training. The corresponding recognition results are shown in Figure 5. As mentioned earlier, for MPE training, the smoothing constant

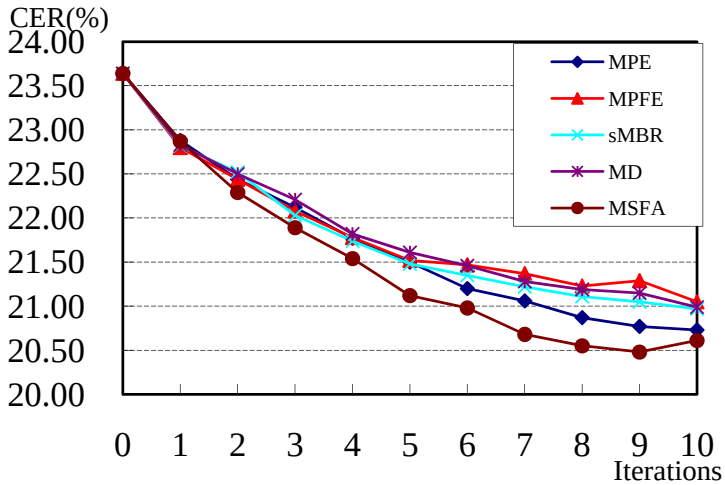


Figure 6. CER results (%) of the MPE training and various modifications using different phone accuracy functions.

(*i.e.*, the τ value of I-smoothing) is a very important factor and should be properly scaled on the basis of the ML training statistics [Povey 2004]. Owing to the different dynamic ranges of the phone accuracy values of the other three modified phone accuracy functions, the smoothing constant is suggested to be scaled accordingly when different training criteria (or phone accuracy functions) are being used. For example, the dynamic range of the phone accuracy values of MPFE training is apparently far larger than that of the standard MPE training, so the smoothing constant for the MPFE training should be empirically set to be larger than that of the standard MPE training.

As evidenced by Figure 6, the recognition results of MD training are slightly worse than the standard MPE training for most of the training iterations. One possible reason for this is that the MD objective function is not well optimized, since the statistics for computing the KL divergence between any two HMM state-level probability distributions are fixed during the training process. Similar observations were also made in [Povey *et al*. 2007]. Furthermore, the corresponding results of the MPFE and sMBR training are also worse than those of the standard MPE training, which could be analyzed as follows. The statistics $\gamma_q^{k^{MPE}}$ of MPE training mainly depend on two parts (cf. Eq. (8)). One is the posterior probability γ_q^k of a phone arc q, while the other is the difference between the expected phone accuracy c_q^k over all hypothesized phone sequences containing q and the average phone accuracy c_{avg}^k over all hypothesized word sequences in the word lattice (*i.e.*, $c_q^k - c_{avg}^k$). However, due to the larger dynamic range of phone accuracy values for the MPFE and the sMBR training, the resulting value of $\gamma_q^{k^{MPE}}$ will probably be dominated by $c_q^k - c_{avg}^k$. An extremely high value of $c_q^k - c_{avg}^k$ (either positive or negative) would make the frame-level statistics of a

Table 2. CER results (%) of the data selection approaches.

Iterations	MPE	MPE+HS	MPE+SS	MPE+Random
1	22.82	22.63	22.84	23.02
2	22.44	22.05	22.40	22.62
3	22.28	21.60	22.21	22.22
4	21.79	21.40	21.65	22.16
5	21.48	21.19	21.34	21.76
6	21.24	20.92	21.33	21.66
7	21.10	20.90	21.29	21.74
8	21.06	20.79	21.00	21.62
9	20.97	20.97	21.02	21.78
10	20.77	20.80	20.94	21.84

phone arc over-weighted even though its corresponding posterior probability is low. In contrast, the performance of our proposed method (MSFA) outperforms standard MPE training, as well as the other three modifications. This is because MSFA has a similar dynamic range of phone accuracy values to that of the standard MPE training, and all types of recognition errors (insertion, substitution, and deletion) are properly considered during the training process, unlike in standard MPE training. Actually, if the penalty ρ is set to zero, MFA and MSFA are quite analogous to MPFE. However, the phone accuracy values of MFA and MSFA are further normalized by the frame number of a phone arc and the Sigmoid function, respectively.

6.4 Experiments on Data Selection Approaches

Moreover, we evaluated the effectiveness of our proposed frame-level normalized entropy-based training data selection approaches for MPE training. The best recognition results for the two variants, *i.e.*, the hard (HS) and soft (SS) versions of frame-level data selection, are shown in Table 2 (MPE+HS and MPE+SS, respectively). The corresponding threshold value Thr for MPE+HS was empirically set to 0.05, while the weighting parameter ω for MPE+SS was empirically set to 1. It is worth mentioning that when threshold value Thr for MPE+HS is set to 0.05, the corresponding number of training frame samples used is about 4 million, which is 45.88% of the total training frame samples. Moreover, for MPE+HS, the frame samples being selected for the MPE training might be different from iteration to iteration, since the acoustic models will be updated after each training iteration, which will make the entropy value calculated for a given frame sample different from that calculated in the previous iteration.

As evidenced by Table 2, data selection (either MPE+HS or MPE+SS) will improve the performance of MPE when the acoustic models are trained at the lower iterations, and achieve comparable results to that of MPE trained at higher iterations. This means that data selection can help reduce the time consumed in training but retain the same performance. However, when the acoustic models of the frame-level data selection method are trained at higher iterations (*e.g.*, 9 and 10 iterations), the corresponding performance, especially for MPE+HS, will become slightly worse than the standard MPE training. One possible reason for this is that the normalized entropy value and the amount of data selected by the hard-version data selection method (MPE+HS) would decrease through the training iterations, which has the side effect of making the training to some extent suffer from the data sparseness problem that makes the acoustic models over-trained. Therefore, one of our future research directions is to study the analysis of such an effect in more detail and try to dynamically adjust the selection threshold value through the iterations.

On the other hand, we also apply random frame-level training sample selection to the MPE training, which randomly selects about 45% of the frame-level training samples for the MPE training at each training iteration, and the corresponding results are depicted in Table 2 (MPE+Random). The selecting capacity of our proposed frame-level data selection method can be verified again by comparison with random selection. The above results indeed justify our postulation that, with proper integration of data selection into the acoustic model training process, we can make the discriminative training algorithms focus much more on the useful training samples to achieve a better discrimination capability on the new test set.

6.5 Experiments on Combination of Frame-level Accuracy Function and Data Selection

Finally, we attempt to combine our proposed frame-level accuracy function and frame-level data selection. The two frame-level training data selection approaches, *i.e.*, HS and SS, respectively, are integrated with the MSFA training. The corresponding results are shown in Table 3. Actually, the data selection approaches are simply based on the entropy information of the Gaussian posterior probabilities of phone arcs, without taking any phone accuracy information into consideration. Thus, such a combination can be viewed as a loosely coupled approach, which to some extent would make the effect of the combination less pronounced. As can be seen from Table 3, HS can considerably boost the performance of the MSFA training at lower training iterations, while SS only demonstrates marginal improvement. We also investigate the combination of HS and SS for the MSFA training, which is achieved using SS to emphasize or deemphasize the training samples selected by HS. Such a combination also can provide additional performance gains (at lower training iterations) over that obtained by using either HS or SS alone.

Table 3. CER results (%) of various combinations of the data selection approaches with the MSFA training.

Iterations	MPE	MSFA	MSFA+HS	MSFA+SS	MSFA+HS+SS
1	22.82	22.88	22.46	22.75	22.53
2	22.44	22.37	21.87	22.25	21.72
3	22.28	22.06	21.40	21.83	21.45
4	21.79	21.52	21.38	21.45	21.38
5	21.48	21.23	21.08	21.27	21.03
6	21.24	21.05	21.03	20.94	20.90
7	21.10	20.89	21.02	20.65	21.14
8	21.06	20.50	21.15	20.78	21.14
9	20.97	20.58	20.86	20.56	21.07
10	20.77	20.46	21.43	20.86	21.37

7. Conclusions

In this paper, we have the explored the use of frame-level information for improved MPE training of acoustic models for Mandarin broadcast news recognition. A new phone accuracy function directly based on the frame-level accuracy has been presented. Moreover, a novel data selection approach using the normalized frame-level entropy of Gaussian posterior probabilities has been proposed as well. Promising and encouraging results on the recognition of Mandarin broadcast news speech were demonstrated. More in-depth investigation of the proposed training data selection, as well as its integration with other discriminative acoustic model training algorithms, is also currently being undertaken.

References

Aubert, X. L., "An Overview of Decoding Techniques for Large Vocabulary Continuous Speech Recognition," *Computer Speech and Language*, Vol.16, 2002, pp. 89-114.

Bahl, L. R., 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. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1986, Tokyo, Japan, pp. 49-52.

Chen, B., J.W. Kuo, and 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.

Du, J., P. Liu, F. K. Soong, J. L. Zhou, and R. H. Wang, "Minimum Divergence Based Discriminative Training", *in Proc. Int. Conf. Spoken Language Processing*, 2006, Pittsburgh, USA, pp. 2410-2413.

Gibson, M., and T. Hain, "Hypothesis Spaces for Minimum Bayes Risk Training in Large Vocabulary Speech Recognition", *in Proc. Int. Conf. Spoken Language Processing*, 2006, Pittsburgh, USA , pp. 2406-2409.

Gillick, L., and S. Cox, "Some Statistical Issues in the Comparison of Speech Recognition Algorithms," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1989, Glasgow, UK, pp. 532-535.

Goldberger, J., "An Efficient Image Similarity Measure based on Approximations of KL-Divergence between Two Gaussian Mixtures", *in Proc. International Conference on Computer Vision*, 2003 , Nice, France, pp. 370-377.

Gopalakrishnan, P.S., D. Kanevsky, A. Nadas, and D. Nahamoo, "A Generalization of the Baum Algorithm to Rational Objective Functions," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1989, Glasgow, UK, pp. 631-634.

Heigold, G., W. Macherey, R. Schluter, and H. Ney, "Minimum Exact Word Error Training," in *Proc. IEEE workshop on Automatic Speech Recognition and Understanding*, 2005, Cancun, Mexico, pp. 186-190.

Jiang, H., X. Li, and C. Liu, "Large Margin Hidden Markov Models for Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5), 2006, pp. 1584-1595.

Kumar, N., "Investigation of Silicon-Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition," Ph.D. Thesis, John Hopkins University, 1997.

Kuo, J.W., S. H. Liu, H.M. Wang, and B. Chen, "An Empirical Study of Word Error Minimization Approaches for Mandarin Large Vocabulary Speech Recognition," *International Journal of Computational Linguistics and Chinese Language Processing*, 11(3), 2006, pp. 201-222.

Li, J., M. Yuan, and C. H. Lee, "Soft Margin Estimation of Hidden Markov Model Parameters," in *Proc. Int. Conf. Spoken Language Processing*, 2006, Pittsburgh, USA, pp. 2422-2425.

Liu, S.H., F.H. Chu, and B. Chen, "Improved MPE Based Discriminative Training of Acoustic Models for Mandarin Large Vocabulary Continuous Speech Recognition," in *Proc. ROCLING XIX: Conference on Computational Linguistics and Speech Processing*, 2007a.

Liu, S.H., F.H. Chu, S.H. Lin, and B. Chen, "Investigating Data Selection for Minimum Phone Error Training of Acoustic Models," in *Proc. IEEE International Conference on Multimedia & Expo*, 2007b, Beijing, China, pp. 348-351.

Misra, H., and H. Bourlard, "Spectral Entropy Feature in Full-Combination Multi-Stream for Robust ASR," in *Proc. European Conf. Speech Communication and Technology*, 2005, Lisbon, Portugal, pp. 2633-2636.

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

Povey, D., and B. Kingsbury, "Evaluation of Proposed Modifications to MPE for Large Scale Discriminative Training," *in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 2007, Hawaii, USA, pp. 321-324.

Povey, D., and P. C. Woodland, "Minimum Phone Error and I-smoothing for Improved Discriminative Training," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 2002, Florida, USA, pp. 105-108.

Rabiner, L., "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, 77(2), 1989, pp. 257-286.

Saon, G., M. Padmanabhan, R. Gopinath, and S. Chen, "Maximum likelihood discriminant feature spaces," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 2000, Istanbul, Turkey, pp. 1129-1132.

Saon, G., and M. Padmanabhan, "Data-Driven Approach to Designing Compound Words for Continuous Speech Recognition," *IEEE Trans. on Speech And Audio Processing*, 9(4), 2001, pp. 327-332.

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

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.

Wessel, F., R. Schluter, and H. Ney, "Explicit Word Error Minimization Using Word Hypothesis Posterior Probabilities," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 2001, Salt Lake City, USA, pp. 33-36.

Zheng, J., and A. Stolcke, "Improved Discriminative Training using Phone Lattices," in *Proc. European Conf. Speech Communication and Technology*, 2005, Lisbon, Portugal, pp. 2125-2128.