Pattern Recognition Letters 30 (2009) 1228–1235

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Training data selection for improving discriminative training of acoustic models

Berlin Chen *, Shih-Hung Liu, Fang-Hui Chu

Department of Computer Science and Information Engineering, National Taiwan Normal University, Taipei 116, Taiwan

ARTICLE INFO

Article history:
Received 25 June 2008
Received in revised form 11 May 2009
Available online 18 May 2009

Communicated by Prof. R.C. Guido

Keywords:
Continuous speech recognition
Discriminative training
Acoustic models
Data selection
Phone accuracy
Entropy

ABSTRACT

This paper considers training data selection for discriminative training of acoustic models for large vocabulary continuous speech recognition (LVCSR). Three novel data selection approaches are proposed. First, the average phone accuracy over all hypothesized word sequences in the word lattice of a training utterance is utilized for utterance-level data selection. Second, phone-level data selection based on the difference between the expected accuracy of a phone arc and the average phone accuracy of the word lattice is investigated. Finally, frame-level data selection based on the normalized frame-level entropy of Gaussian posterior probabilities obtained from the word lattice is explored. The underlying characteristics of the presented approaches are extensively investigated and their performance is verified by comparison with standard discriminative training approaches. Experiments conducted on a broadcast news speech transcription task show that with the aid of phone- and frame-level data selection we can reduce more than half of the turnaround time for acoustic model training and simultaneously obtain a comparably good set of discriminative acoustic models.

1. Introduction

Discriminative training algorithms, such as the minimum classification error (MCE) training (Juang et al., 1997; McDermott et al., 2007), the maximum mutual information (MMI) training (Bahl et al., 1986; Povey and Woodland, 2002a) and the minimum phone error (MPE) training (Povey and Woodland, 2002b; Povey, 2004), aiming at estimating more accurate acoustic models, have continuously been an active focus of much research in a wide variety of large vocabulary continuous speech recognition (LVCSR) tasks in the past few years. Discriminative training is developed in an attempt to correctly discriminate the recognition hypotheses for the best recognition results rather than just to fit the model distributions. In contrast to conventional maximum likelihood (ML) training, discriminative training considers not only the correct (or reference) transcript of a training utterance, but also the competing hypotheses that are often obtained by performing LVCSR on the utterance.

Recently, the large or soft margin classifiers, motivated by the support vector machine (SVM) successfully developed in the machine learning community, have been introduced into the field of automatic speech recognition (ASR) for acoustic model training and demonstrated with good results in various speech recognition tasks (Jiang et al., 2006; Li et al., 2007; Li, 2008; Yu et al., 2007, 2008). The common concept of these margin-based methods, originating from statistical learning theory (Vapnik, 1995), is to minimize a combined score of the margin, the distance from the decision boundary to the nearest training samples, and the empirical error rate of the training data (Yu and Deng, 2007; Yu et al., 2008). Although the derivations of these margin-based approaches are conducted with a rigorous theoretical basis, the practical implementations of their corresponding training objective functions eventually can be interpreted as a kind of training data selection, i.e., selecting the training samples close to the decision boundaries for better model discrimination and generalization. For example, the large-margin hidden Markov model estimation (LME) (Jiang et al., 2006) treats each speech utterance as a whole as an sample and uses a discriminant function to select positive samples falling in a predefined margin for acoustic model training; while the soft margin estimation (SME) (Li et al., 2007; Li, 2008) conducts both frame- and utterance-level data selection, for which label matching between the reference and the recognized (or hypothesized) word sequences of the training utterance is first used to identify a candidate set of frame samples, and utterance-level data selection is then executed on the basis of the average frame-level log-likelihood ratios between correct and competing models obtained from these frames. These two approaches are not directly applicable to some discriminative training algorithms, such as the MMI and MPE training, which are commonly used in the LVCSR tasks.

Essentially, the popular discriminative training algorithms, such as MCE, MMI and MPE, have already performed some kind of utterance-level data selection (e.g., MCE and MMI) or phone-level data selection (e.g., MPE) implicitly. In more precise terms, for MMI, the

* Corresponding author. Tel.: +886 229322411; fax: +886 229322378.
E-mail addresses: berlin@csie.ntnu.edu.tw, berlin@ntnu.edu.tw (B. Chen).
URL: http://berlin.csie.ntnu.edu.tw (B. Chen).

0167-8655/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2009.05.009

training utterances whose reference (correct) word sequences have higher posterior probabilities will contribute less to the optimization of the training objective function. Conversely, the training utterances whose reference word sequences have lower posterior probabilities, i.e., the training utterances prone to being misrecognized, will play a more pronounced role in the optimization of the training objective function (Bahl et al., 1986). On the other hand, with the use of the sigmoid function, MCE emphasizes the training utterances whose loss functions have values near the medium between the two extreme values of the loss function (for example, zero and one), and deemphasizes the training utterances whose loss functions are either too high (i.e., the training utterances that are likely to be outliers) or too low (i.e., the perfectly recognized training utterances) in their values (Juang et al., 1997; Yu et al., 2008). Moreover, MPE simply neglects the phone arcs in the word lattice of a training utterance, which have expected accuracies equal to the average phone accuracy of all word sequences in the word lattice (Povey, 2004).

In this paper we investigate three novel data selection approaches for discriminative training of acoustic models for LVCSR, in an attempt to reduce the time consumed in training and simultaneously obtain a desired set of discriminative acoustic models. First, the average phone accuracy over all hypothesized word sequences in the word lattice of a training utterance is utilized for utterance-level data selection for MPE. Second, phone-level data selection based on the difference between the expected accuracy of a phone arc and the average phone accuracy of the word lattice is investigated for MPE. Finally, frame-level data selection based on the normalized frame-level entropy of Gaussian posterior probabilities obtained from the word lattice is explored for both MMI and MPE.

The remainder of this paper is organized as follows. Section 2 provides a brief introduction to two popular discriminative acoustic model training algorithms for LVCSR that are to be used in this paper. Section 3 sheds light on our proposed data selection approaches. The experimental settings and the corresponding results are described in Sections 4 and 5, respectively. Finally, Section 6 concludes this paper with future work.

2. Discriminative training approaches

2.1. Maximum mutual information (MMI) training

The MMI training, as a representative alternative to the ML training, was first proposed by Bahl et al. (1986) in the context of small vocabulary speech recognition tasks, which aims at increasing posterior probabilities of the corresponding correct transcripts given a training set of utterances (or observation vector sequences). In mathematical terms, given a training set of K observation vector sequences $O = \{O_1,\ldots,O_k,\ldots,O_K\}$, the MMI criterion for acoustic model training is designed to maximize the following objective function:

$$F_{MMI}(\lambda) = \sum_{k=1}^{K} \log P_\lambda(W_k|O_k) = \sum_{k=1}^{K} \left(\log \frac{P_\lambda(O_k|W_k)P_\lambda(W_k)}{\sum_{W' \in \mathbf{W}_k} P_\lambda(O_k|W')P_\lambda(W')} \right), \quad (1)$$

where λ is the set of parameters that needs to be estimated; W_k is the corresponding correct transcript of the observation vector sequence O_k; $P_\lambda(W_k|O_k)$ is the posterior probability of W_k given the observation vector sequence O_k; W' is one of the hypothesized word sequences in the word lattice $\mathbf{W}_k$ of O_k (Ortmanns et al., 1997); $P_\lambda(O_k|W_k)$ is the likelihood of the correct transcript W_k generating the observation vector sequence O_k, $P_\lambda(W_k)$ is the language model probability of W_k. Fig. 1 gives an schematic illustration of a word lattice. It should be noted that the optimization of Eq. (1) requires not only to maximize the numerator term $P_\lambda(O_k|W_k)P_\lambda(W_k)$, which is identical to that done by the ML training, but also to minimize the denominator term $\sum_{W' \in \mathbf{W}_k} P_\lambda(O_k|W')P_\lambda(W')$ for each training utterance. Since the denominator contains all possible word sequences (including the correct one), the objective function has a maximum value of zero. When the language model parameters are fixed during the training process, the objective function of MMI becomes equivalent to the conditional maximum likelihood (CML) proposed in (Nadas, 1983).

To go a step further, for the MMI training, the training utterances whose reference (correct) word sequences have higher posterior probabilities will contribute less to the optimization of the objective function. Conversely, the training utterances whose reference word sequences have lower posterior probabilities, i.e., the training utterances prone to being misrecognized, will play a more pronounced role in the optimization of the objective function. Therefore, the MMI training already performs some kind of utterance selection implicitly. More detailed illustrations and discussions of the MMI training formulas can be found in (Valchev, 1995).

2.2. Minimum phone error (MPE) training

The MPE criterion for acoustic model training aims to minimize the expected phone errors of the training acoustic vector sequences $O = \{O_1,\ldots,O_k,\ldots,O_K\}$ using the following objective function (Povey and Woodland, 2002b):

$$F_{MPE}(\lambda) = \sum_{k=1}^{K} \sum_{W \in \mathbf{W}_k} RawAcc(W) P_\lambda(W|O_k), \quad (2)$$

where λ is the set of parameters that needs to be estimated; $\mathbf{W}_k$ is the corresponding word lattice of O_k obtained by using LVCSR; W is one of the hypothesized word sequences in $\mathbf{W}_k$; $P_\lambda(W_k|O_k)$ is the posterior probability of hypothesis W given O_k; $RawAcc(W)$ is the "raw phone accuracy" of W in comparison with the corresponding reference transcript, which is typically computed as the sum of the phone accuracy measures of all phone hypotheses in W. The objective function defined in Eq. (2) can be maximized by applying the Extended Baum–Welch algorithm (Normandin, 1991) to update the mean μ_{hmd} and variance σ^2_{hmd} for each dimension d of a diagonal Gaussian mixture component m of a multi-state (or single-state) HMM h using the following equations:

$$\mu_{hmd} = \frac{\theta^{num}_{hmd}(O) - \theta^{den}_{hmd}(O) + D\bar{\mu}_{hmd}}{\gamma^{num}_{hm} - \gamma^{den}_{hm} + D}, \quad (3)$$

$$\sigma^2_{hmd} = \frac{\theta^{num}_{hmd}(O^2) - \theta^{den}_{hmd}(O^2) + D(\bar{\sigma}^2_{hmd} + \bar{\mu}^2_{hmd})}{\gamma^{num}_{hm} - \gamma^{den}_{hm} + D} - \mu^2_{hmd}, \quad (4)$$

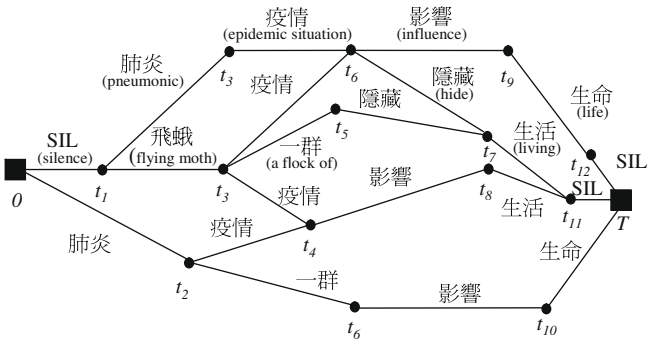

Fig. 1. An illustration of a word latice, in which each arc, together with its corresponding start and end speech frames, represents a candidate word hypothesis. A word arc can be further aligned into a sequence of phone arcs for the MMI and MPE training.

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

$$\gamma_{hm}^{den} = \sum_{k=1}^{K} \sum_{q \in \mathbf{W}_k, q=h} \sum_{t=s_q}^{e_q} \gamma_{qm}^{k}(t) \max\left(0, -\gamma_q^{k^{MPE}}\right), \tag{6}$$

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

$$\theta_{hmd}^{num}(O^2) = \sum_{k=1}^{K} \sum_{q \in \mathbf{W}_k, q=h} \sum_{t=s_q}^{e_q} \gamma_{qm}^{k}(t) \max\left(0, \gamma_q^{k^{MPE}}\right) o_t(d)^2, \tag{8}$$

$$\gamma_q^{k^{MPE}} = \gamma_q^k (c_q^k - c_{avg}^k), \tag{9}$$

where $q \in \mathbf{W}_k$, $q = h$ denotes that a phone arc q belongs to the word lattice $\mathbf{W}_k$ and refers to the physical 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 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 is 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; $\gamma_{hm}^{num}, \theta_{hmd}^{num}(O)$ and $\theta_{hmd}^{num}(O^2)$ are the training statistics for mixture component m of HMM h accumulated from those phone arcs q that physically refer to HMM h and have c_q^k larger than c_{avg}^k, and vice versa for $\gamma_{hm}^{den}, \theta_{hmd}^{den}(O)$ and $\theta_{hmd}^{den}(O^2)$; $\bar{\mu}_{hmd}$ and $\bar{\sigma}_{hmd}^2$ 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 in the lattice can be calculated in terms of the sum of the accuracy of each phone contained in W (Povey and Woodland, 2002b):

$$RawAcc(W) = \sum_{q \in W} PhoneAcc(q), \tag{10}$$

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

$$PhoneAcc(q) = \max_{z_j \in Z_k} \left\{ \begin{array}{ll} -1 + 2e(z_j, q)/l(z_j), & z_j = q \\ -1 + e(z_j, q)/l(z_j), & z_j \neq q \end{array} \right\}, \tag{11}$$

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, $l(z_j)$ is the length in frames for z_j. We can observe from Eqs. (5)–(9), for the MPE training, those phone arcs in the word lattice of a training utterance having raw phone accuracies higher than the average can provide positive contributions, and vice versa for those phone arcs with accuracies lower than the average. On the other hand, those phone arcs in the word lattice of a training utterance having expected accuracies equal to the average phone accuracy of all word sequences in the word lattice will be simply neglected from training. In our study, we experimentally observed that about 2.41% of the phone arcs of the speech utterances in the training data (cf. Section 4.2) were left out by the baseline MPE training. Interested readers can refer to (Povey, 2004; Kuo et al., 2006) for more derivation details of the MPE training.

3. Training data selection approaches

3.1. Utterance selection

Training utterance selection based on the log-likelihood ratio has been investigated previously, such as that in (Jiang et al., 2005). In this paper, we attempt an alterative approach by conducting training utterance selection directly on the phone accuracy domain for the MPE training. The word lattice (or recognition hypothesis space) $\mathbf{W}_k$ of a training utterance k, which offers the competing information for the training objective function, plays an important role in discriminative training. It can help in filtering out the training utterance whose recognition hypothesis space is devoid of discriminative information. For example, in the MPE training, the normalized average phone accuracy $\hat{c}_{avg}^k$ of each training utterance k, obtained by dividing the average phone accuracy c_{avg}^k by the phone number of the reference transcript of k, to some extent reveals the confusedness of the hypothesis space $\mathbf{W}_k$ (Liu et al., 2007b). The utterance with a too high normalized average phone accuracy implies that less competing information might be provided by it (or its hypothesized space), while the utterance with a too low normalized average phone accuracy implies that it might probably be a damaged training sample (or an outlier) and thus can be left out. Inspired by this, we conducted training utterance selection based on the normalized average phone accuracy $\hat{c}_{avg}^k$. We first estimated the mean of $\hat{c}_{avg}^k$ among all training utterances, denoted by $\bar{\hat{c}}_{avg}$, and then used it together with $\hat{c}_{avg}^k$ to select training utterances that fall in the interval defined by the following equation for the MPE training:

$$\bar{\hat{c}}_{avg} - \delta \leqslant \hat{c}_{avg}^k \leqslant \bar{\hat{c}}_{avg} + \delta, \tag{12}$$

where δ is a predefined threshold value. It is worth mentioning here that such data selection is based on the phone accuracy domain (or equivalently, the phone error rate domain) to select the more discriminative utterances. This makes a considerable distinction between our approach and the LME and SME approaches that select training sentences based on the log-likelihood ratio domain (Jiang et al., 2006; Li et al., 2007).

3.2. Phone arc selection

It is somewhat coarse to directly use an utterance as a whole as the unit for training data selection. Thus, we also propose a phone-level data selection approach, conducted on the phone accuracy domain as well, for the MPE training. As we know, in MPE, the average phone accuracy c_{avg}^k is taken as a decision boundary for accumulating the training statistics of a phone arc q into the numerator or denominator terms, as those illustrated in Eqs. (5)–(9). Thus, we can impose a margin on c_{avg}^k in order to select more critical phone arcs which are relatively close to the decision boundary on the error rate domain (Liu et al., 2007b). As a result, the final auxiliary function for the MPE training on an HMM c_{avg}^k can be defined as:

$$g_{MPE}(h) = \sum_{k=1}^{K} \sum_{q \in \mathbf{W}_k, q=h} \sum_{t=s_q}^{e_q} \sum_{m} \gamma_q^r \left[(c_q^k - c_{avg}^k) I(c_q^k \in A^k) \right] \times \gamma_{qm}^k(t) \log N(O_k(t), \mu_{qm}, \Sigma_{qm}), \tag{13}$$

$$A^k = \left\{ c_q^k | - \alpha \leqslant \kappa(c_q^k - c_{avg}^k) \leqslant \beta \right\}, \tag{14}$$

where $N(\bullet)$ is a Gaussian distribution; $I(\bullet)$ is a Kronecker delta function that will return a value of one when $c_q^k \in A^k$ and zero otherwise; the positive parameters α and β form the margin for training data selection; κ is a normalization factor that makes $\kappa(c_q^k - c_{avg}^k)$ approximately range from -1 to 1; A^k is the set of phone arcs that fall in the margin $[-\alpha, \beta]$ defined in the phone accuracy rate domain. Only those phone arcs in A^k would contribute their accumulated statistics to the MPE training.

Fig. 2 illustrates an example of the proposed phone arc selection approach. In this figure, each blue bar[1] denotes a word arc in the

[1] For interpretation of color in Fig. 2, the reader is referred to the web version of this article.

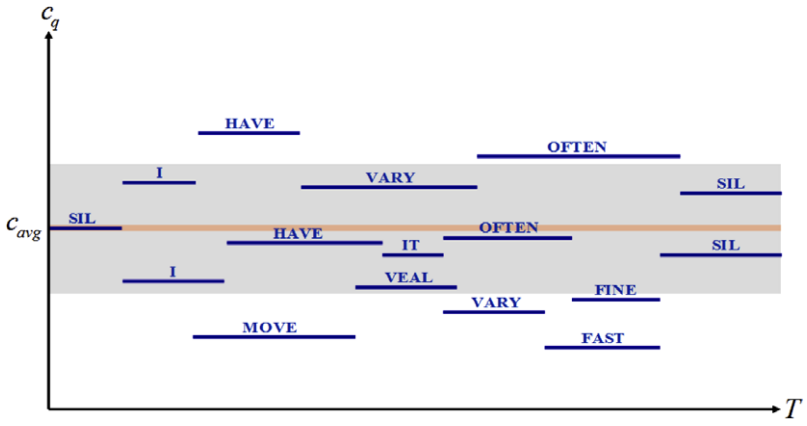

Fig. 2. An example to illustrate the phone-level selection approach.

word graph of a training sentence, and their corresponding word labels lie above them. Here the horizontal axis represents the time frame of this sentence, while the vertical axis denotes the expected phone accuracy c_q of each word arc q. In more details, the higher position a word arc is located on the figure indicates the greater the expected phone accuracy it has. Thus, the goal of phone arc selection approach is to retain the corresponding phone arcs in the word graph of the training utterance k that are located nearly around the horizontal line of the average phone accuracy c^k_{avg} (the orange bars on Fig. 2), but filter out the word arcs that are located far apart from c^k_{avg}. In other words, only the word arcs falling into the predefined margin (denoted by the gray region) can be selected for the MPE training, except for those word arcs whose expected phone accuracies are equal to c^k_{avg} of the training sentence.

3.3. Frame selection

We also propose the use of the entropy information to select the frame-level training statistics for the MMI and MPE training. The normalized entropy of a training frame sample t can be defined as (Liu et al., 2007a):

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

where $\gamma^k_{qm}(t)$ 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 the Gaussian mixtures which have nonzero posterior probabilities at frame $t(\gamma^k_{qm}(t) > 0)$; 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 Fig. 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 that belonging to C_2 (depicted as circles). In general, the decision boundary is set 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 around the decision boundary will have normalized entropies close to one, while those (hollow circles or squares) located far away the decision boundary will have normalized entropies close to zero.

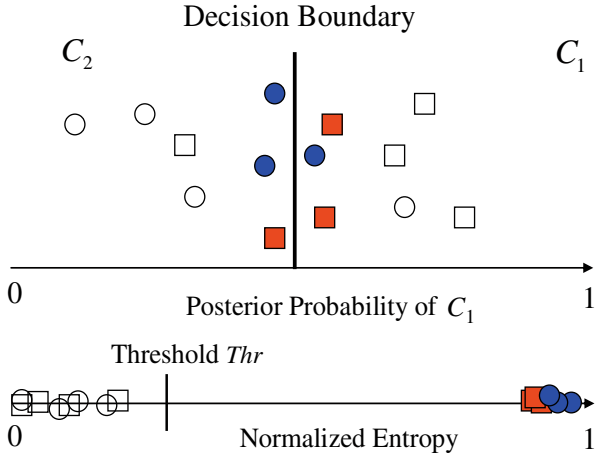

Fig. 3. A hypothetical example of binary classification illustrating the relationship between the decision boundary and the normalized entropy.

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, then 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, then it means that the posterior probabilities of mixture components tend to be uniformly distributed; that is, the frame sample i is instead located nearly around 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 nearly around the decision boundary for better sample discrimination and model generalization (Jiang et al., 2006; Li et al., 2007).

A straightforward implementation of frame-level training data selection is to define a threshold of the normalized entropy measure and 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. Another "soft version" of data selection is to emphasize the training statistics of those frame samples that are located nearly around the decision boundary according to their normalized entropy values. Fig. 4 shows the relationship between the normalized entropy and the number of training speech frame samples when "hard version" of frame-level data selection was 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 0–0.05. The large number of frame samples belonging to the left-most vertical bar also reveals that most of the training frame samples in fact are located far from the decision boundary and thus can be discarded if the threshold is appropriately set (Liu et al., 2007a, 2008). In this paper, we only implemented the "hard version" of frame-level data selection for the MMI and MPE training. Readers may refer to (Liu et al., 2008) for the implementation of the "soft-version" of frame-level data selection for the MPE training.

4. Broadcast news system

The large vocabulary continuous speech recognition system as well as the experimental speech and language data used in this paper will be described in this section (Chen et al., 2004).

4.1. Front-end signal processing

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), and then processed by MLLT (Maximum Likelihood Linear Transformation) transformation for feature de-correlation (Saon et al., 2000; Gales, 2002). The dimension of the resultant feature vectors was set to 39 (Lee and Chen, 2009). In addition, utterance-based feature mean subtraction and variance normalization were applied to all the training and test speech.

4.2. Speech corpus and acoustic model training

The speech corpus consists of about 200 h of MATBN Mandarin television news (Mandarin Across Taiwan Broadcast News) (Wang et al., 2005), which were collected by Academia Sinica and Public Television Service Foundation of Taiwan during November 2001 and April 2003. All the 200 h of speech data are equipped with corresponding orthographic transcripts, in which about 25 h of gender-balanced speech data of the field reporters collected during November 2001 to December 2002 were used to bootstrap the acoustic training. Another set of 3.0 h speech data of the field reporters collected within 2003 were reserved for the speech recognition experiments and divided into two equal parts. The first part was taken as the development set, which formed the basis for tuning the training settings. The second part was taken as the test (or evaluation) set; i.e., all the speech recognition experiments were conducted on it with the acoustic models trained on the basis of the settings optimized by the development set. Therefore, the experimental results to be presented in the following sections can validate the effectiveness of the proposed approaches on comparable real-world data. On the other hand, the acoustic models chosen here for speech recognition were 112 right-context-dependent INITIAL's and 38 context-independent FINAL's. They were selected based on consideration of the phonetic structure of Mandarin syllables. Here, INITIAL means the initial consonant of a syllable and FINAL is the vowel (or diphthong) part but also includes an optional medial or nasal ending. Each INITIAL is represented by an HMM with 3 states, while each FINAL is represented with 4 states. The Gaussian mixture number per state ranges from 2 to 128, depending on the quantity of training data.

The acoustic models were first trained at optimum settings using the ML criterion as well as the Baum–Welch training algorithm. The MMI-based and MPE-based discriminative training approaches were further applied to those acoustic models previously trained by the ML criterion. Unigram language model constraints were used in accumulating the training statistics from the word lattices for discriminative training. For the MPE training, both silence and short pause labels are also involved in the calculation of the accuracies of the hypothesized word sequences.

4.3. 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 (Saon and Padmanabhan, 2001) and added to the lexicon to form a new lexicon of 72K words. The background language models used in this paper consist of trigram and bigram models, which were estimated based on the ML criterion and using a text corpus consisting of 170 million Chinese characters collected from Central News Agency (CNA) in 2001 and 2002 (the Chinese Gigaword Corpus released by LDC) (Chiu and Chen, 2007). In implementation, the *N*-gram language models were trained using the SRI language modeling toolkit (Stolcke, 2000).

4.4. 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 (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., 2004), 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).

5. Experiments

In this section, we will present a series of experiments, performed on the test set (cf. Section 4.2), to assess the speech

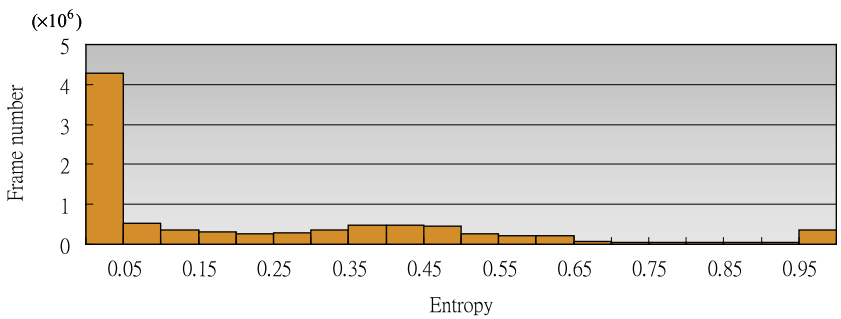

Fig. 4. A plot of the relationship between the normalized entropy and the number of training speech frame samples.

recognition performance as a function of proposed training data selection approaches and their combinations; while the acoustic models were trained on the basis of the settings tuned by the development set. As it is known that there are no explicit marks, such as the spaces or blanks, separating words in the Chinese language, the Chinese language thus often suffers from the 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), which is defined as the sum of the insertion (*Ins*), deletion (*Del*), and substitution (*Sub*) errors between the recognized and reference Chinese character strings, divided by the total number of Chinese characters in the reference string (*Ref*):

$$\mathrm{CER} = \frac{Ins + Sub + Del}{Ref}. \tag{16}$$

5.1. Baseline results

The acoustic models were trained with 24.5 h of speech utterances. The MMI and MPE training both started with the acoustic models trained by 10 iterations of the ML training, and used the information contained in the associated word lattices of the training utterances to accumulate the necessary statistics for model training. The ML-trained acoustic models yield a CER of 23.16% on the test set, while the original MMI and MPE training indeed can provide a great boost to the acoustic models initially trained by ML consistently at all training iterations, as shown in Table 1. The MPE- and MMI-trained acoustic models (at the 10th iteration) can, respectively, offer relative CER improvements of 11.65% and 6.82% over the ML-trained acoustic models. Notice that the total frame number used in the original MMI and MPE training is about 9 million frames (24.5 h).

In the following experiments, for fair comparison between our proposed methods and the baseline MMI and MPE training, the smoothing constants (i.e., the τ values of I-smoothing) (Povey and Woodland, 2002b; Povey, 2004; Kuo et al., 2006) are respectively set to be the same as those used in the baseline MMI and MPE training (which were optimally tuned using the development set). It is known that this smoothing constant makes an interpolation between the objective functions of the ML training and the MMI (or MPE) training, which can be regarded as a kind of prior information forcing the HMM parameters estimated by the MMI (or MPE) training to center around that estimated by the ML training (Povey, 2004; Kuo et al., 2006; Li, 2008).

5.2. Results on utterance selection

We first evaluate the performance of utterance-level data selection for the MPE training (denoted by MPE + US). The value of threshold δ was set to be 0.2. As can be seen from Fig. 5, MPE + US obtains slightly worse results than the baseline MPE at lower training iterations, and the difference between them is almost negligible when the acoustic models were trained with 10 iterations. As compared to the MPE baseline, the number of training utterances used by MPE + US in the training process can be reduced by about 10% without any loss of recognition performance. Moreover, we also attempted to slightly increase the predefined threshold δ through the iterations, in order to obtain more training statistics for the MPE training (denoted by MPE + USv). However, since the number of training utterances that would be additionally included by MPE + USv through the iterations is no more than 10% of the entire training utterances, this makes the performance difference between MPE + US and MPE + USv is not so significant. On the other hand, as we listened to the utterances that were left out by either MPE + US or MPE + USv, we found that many of them actually were damaged utterances, contaminated with severe noises or equipped with wrong reference transcripts.

Table 1
The CER results (%) achieved by the baseline MPE and MMI training, respectively.

Iterations	MPE	MMI
1	22.34	22.79
2	21.72	22.57
3	21.41	22.23
4	21.23	22.06
5	20.90	21.86
6	20.67	21.69
7	20.63	21.59
8	20.63	21.48
9	20.57	21.69
10	20.46	21.58

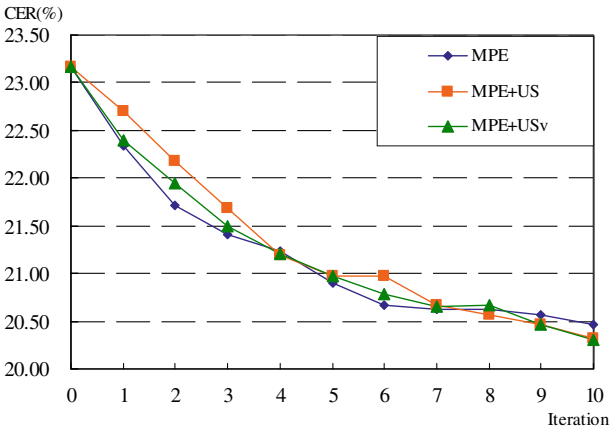

Fig. 5. The CER results (%) obtained by integrating utterance-level data selection with the MPE training.

5.3. Results on phone arc selection

We then evaluate the performance of phone-level data selection for the MPE training (denoted by MPE + PS). The parameters α and β defined in Eq. (14) that form the predefined margin were set to be 1.0 and 0.03, respectively. The number of training phone arcs thus can be reduced up to about 9%. This also means that almost all the phone arcs belonging to the denominator part in the MPE training equations (cf. Eqs. (3)–(9)) were included and contributed

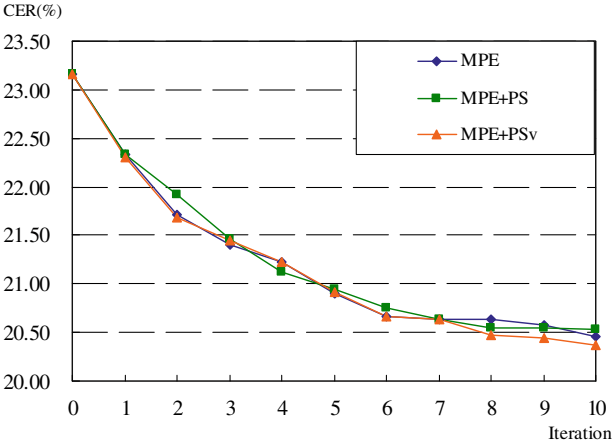

Fig. 6. The CER results (%) obtained by integrating phone-level data selection into the MPE training.

to the accumulation of training statistics. As the can be observed from Fig. 6, MPE + PS does not outperforms the baseline MPE at lower training iterations (≦7 iterations). One possible reason for this is that phone-level training data selection tends to reduce the phone arc samples that belong to the numerator part dramatically, which would probably lead to over-trained acoustic models. Therefore, we try to steadily increase the predefined margin set by the parameters α and β through the iterations, in order to include more phone arc samples for the MPE training (denoted by MPE + PSv). The results of MPE + PSv are also depicted in Fig. 6. As can be seen, MPE + PSv performs almost the same as the baseline MPE does at lower training iterations and is lightly better than the latter at higher training iterations. This therefore confirms our expectation that properly imposing a margin on the basis of the average phone accuracy can help in selecting the most confusing samples for better acoustic model discrimination.

5.4. Results on frame selection

In the third sets of experiments, we evaluate the performance levels of frame-level data selection for the MMI and MPE training (denoted by MMI + FS and MPE + FS, respectively). The corresponding results are graphically presented in Figs. 7 and 8. The threshold value *Thr* used for frame-level normalized entropy-based training data selection, as described in Section 3.3, was set to be 0.05, and the resulting number of training frame samples being used was about 4 millions (45% of the total training frame samples). The frame samples selected for the MMI and MPE training might be slightly 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 somewhat different from that calculated in the previous iteration. As shown by Figs. 7 and 8, frame-level data selection (MMI + FS and MPE + FS) will improve the performance substantially when the acoustic models are trained at lower iterations. This means that frame-level data selection can help reduce the time consumed in training but retain the same performance as that of MMI and MPE, respectively. However, when the acoustic models are trained with higher iterations (e.g., 9 and 10 iterations), the results of frame-level data selection (MMI + FS and MPE + FS) are worse than those of the original MMI and MPE training, respectively. One possible reason for this is that this data selection method, to some extent, suffers from the data sparseness problem which would make the acoustic models over-trained, especially at higher training iterations. Therefore, we alternatively attempt to not only apply frame-level data selection for the MMI and MPE training, but meanwhile also slightly decrease the threshold value *Thr* as the iteration increases (denoted by MMI + FSv and MPE + FSv, respectively), for the purpose of obtaining more training statistics and alleviating the over-training problem. The corresponding results of MMI + FSv and MPE + FSv are also depicted in Figs. 7 and 8, respectively; the performance losses at higher training iterations for the revised frame-level data selection approaches (especially for MMI + FSv) are compensated to some extent.

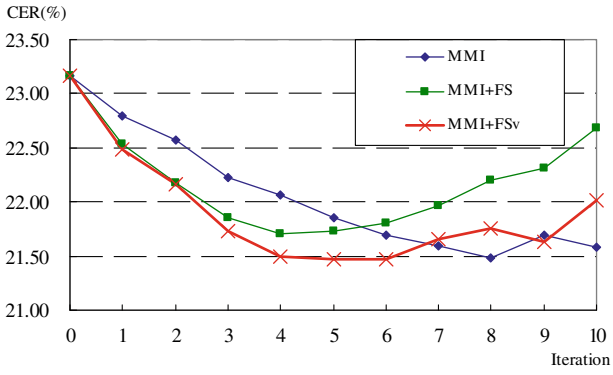

Fig. 7. The CER results (%) obtained by integrating frame-level data selection into the MMI training.

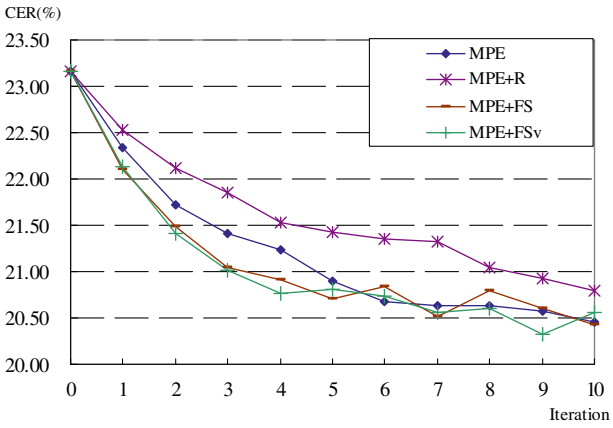

Fig. 8. The CER results (%) obtained by integrating frame-level data selection into the MPE training.

On the other hand, we additionally apply random selection to MPE training (denoted by MPE + R), 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 Fig. 8. The selecting capacity of our proposed frame-level data selection method can be verified again by comparison with random selection.

5.5. Combinations of various training data selection approaches

Finally, we investigate different combinations of the proposed data selection methods for the MPE training, for which data selection proceeds in descending order according to the granularity of the training samples being considered, i.e., utterances, phone arcs, and then frames. The corresponding results are depicted in Fig. 9. As can be seen, the combination of phone-level selection with frame-level selection (MPE + PSv + FSv) can offer additional performance gains over either phone-level selection (MPE + PSv) or frame-level selection (MPE + FSv) alone (cf. Figs. 6 and 8); and it not only can perform consistently better than the standard MPE

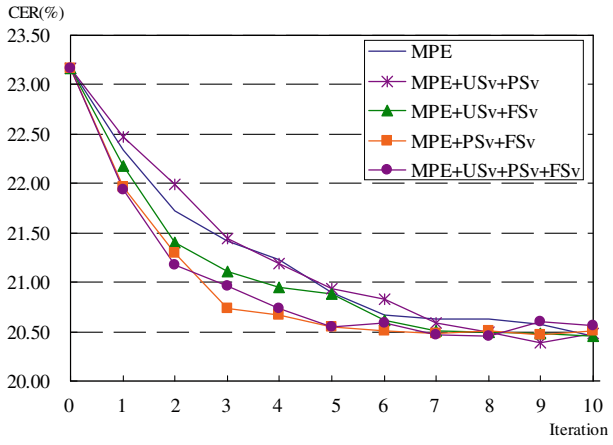

Fig. 9. The CER results (%) obtained by integrating different combinations of different-level data selection into the MPE training.

training almost at all iterations, but also can let the acoustic training converge early at the 5th training iteration compared to the result of the standard MPE at the 10th training iteration. It is noteworthy that all our experiments were conducted using an ordinary personal computer (PC), which has to take more than one day (≧24 h) to complete an iteration of the MPE training. This implies that with the aid of our training data selection approach (MPE + PSv + FSv) we can reduce more than 50% of the turnaround time (≧5 days) for acoustic model training and simultaneously obtain a comparably good set of discriminative acoustic models.

However, if we further integrate utterance-level selection with phone-level selection (MPE + USv + PSv) or frame-level selection (MPE + USv + FSv), or their combination (MPE + USv + PSv + FSv), then we can find that such an attempt in fact cannot offer additional performance gains over phone-level selection (MPE + PSv), or frame-level selection (MPE + FSv), or their combination (MPE + PSv + FSv). One possible explanation for this phenomenon is that utterance-level selection is relatively crude when compared to the other two kinds of training data selection methods.

The above results indeed justify our postulation that with the 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. For fair comparisons between our proposed approaches with the baseline discriminative training methods, all the speech recognition experiments were carefully designed to avoid "testing on training"; i.e., all the acoustic model training settings were tuned on the basis of development set. Generally speaking, the training settings tuned on the development set performed quite consistently in the test set.

6. Conclusions

In this paper, we have studied utterance-level training data selection and phone-arc-level training data selection for MPE training, and frame-level training data selection for both the MMI and the MPE training of acoustic models for LVCSR. The experimental results have demonstrated that with the use of phone- and frame-level data selection we can reduce more than half of the turnaround time for acoustic model training and simultaneously obtain a comparably good set of discriminative acoustic models. In future work, we plan to explore different ways to combine the proposed data selection methods together for the MPE and MMI training, including trying unsupervised discriminative training (Chan and Woodland, 2004; Mathias et al., 2005), investigating the joint training of feature transformation and acoustic models, etc.

Acknowledgement

This work was supported in part by the National Science Council, Taiwan, under Grants: NSC96-2628-E-003-015-MY3, NSC95-2221-E-003-014-MY3 and NSC97-2631-S-003-003.

References

Aubert, X.L., 2002. An overview of decoding techniques for large vocabulary continuous speech recognition. Comput. Speech Language 16, 89–114.

Bahl, L.R., Brown, P.F., de Souza, P.V., Mercer, R.L., 1986. Maximum mutual information estimation of hidden Markov model parameters for speech recognition. In: Proc. IEEE Internat. Conf. Acoustics, Speech, Signal Processing, pp. 49–52.

Chan, H.Y., Woodland, P.C., 2004. Improving broadcast news transcription by lightly supervised discriminative training. In: Proc. IEEE Internat. Conf. Acoustics, Speech, Signal Processing, pp.737–740.

Chen, B., Kuo, J.W., Tsai, W.H., 2004. Lightly supervised and data-driven approaches to mandarin broadcast news transcription. In: Proc. IEEE Internat. Conf. Acoustics, Speech, Signal Processing, pp. 777–780.

Chiu, H.S., Chen, B., 2007. Word topical mixture models for dynamic language model adaptation, In: Proc. IEEE Internat. Conf. Acoustics, Speech, Signal Processing, pp. 169–172.

Gales, M.J.F., 2002. Maximum likelihood multiple subspace projections for hidden Markov models. IEEE Trans. Speech Audio Process. 10 (2), 37–47.

Jiang, H., Soong, F.K., Lee, C.H., 2005. A dynamic in-search data selection method with its applications to acoustic modeling and utterance verification. IEEE Trans. Speech Audio Process. 13 (5), 945–955.

Jiang, H., Li, X.W., Liu, C.J., 2006. Large margin hidden Markov models for speech recognition. IEEE Trans. Audio, Speech Language Process. 14 (5), 1584–1595.

Juang, B.H., Chou, W., Lee, C.H., 1997. Minimum classification error rate methods for speech recognition. IEEE Trans. Speech Audio Process. 5 (3), 257–265.

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

Kuo, J.W., Liu, S.H., Wang, H.M., Chen, B., 2006. An empirical study of word error minimization approaches for mandarin large vocabulary speech recognition. Internat. J. Comput. Linguistic Chinese Language Process. 11 (3), 201–222.

Lee, H.S., Chen, B., 2009. Empirical error rate minimization based linear discriminant analysis. In: Proc. IEEE Internat. Conf. Acoustics, Speech, Signal Processing, pp. 1801–1804.

Li, J., 2008. Soft Margin Estimation for Automatic Speech Recognition. Ph.D. Dissertation, Georgia Institute of Technology.

Li, J., Ma, B., Lee, C.H., 2007. Approximate test risk bound minimization through soft margin estimation. IEEE Trans. Audio, Speech Language Process. 15 (8), 2393–2404.

Liu, S.H., Chu, F.H., Lin, S.H., Chen, B., 2007a. Investigating data selection for minimum phone error training of acoustic models. In: Proc. IEEE Internat. Conf. on Multimedia and Expo, pp. 348–351.

Liu, S.H., Chu, F.H., Lin, S.H., Lee, H.S., Chen, B., 2007b. Training data selection for improving discriminative training of acoustic models. In: Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 284–289.

Liu, S.H., Chu, F.H., Lo, Y.T., Chen, B., 2008. Improved minimum phone error based discriminative training of acoustic models for Mandarin large vocabulary continuous speech recognition. Internat. J. Comput. Linguistics Chinese Language Process. (3), 327–342.

Mathias, L., Yegnanarayanan, G., Fritsch, J., 2005. Discriminative training of acoustic models applied to domains with unreliable transcripts. In: Proc. IEEE Internat. Conf. Acoustics, Speech, Signal Processing, pp. 109–112.

McDermott, E., Hazen, T.J., Roux, J.L., Nakamura, A., Katagiri, S., 2007. Discriminative training for large vocabulary speech recognition using minimum classification error. IEEE Trans. Audio, Speech Language Process. 15 (1), 203–223.

Misra, H., Bourlard, H., 2005. Spectral entropy feature in full-combination multi-stream for robust ASR. In: Proc. European Conf. Speech Communication and Technology, pp. 2633–2636.

Nadas, A., 1983. A decision theoretic formulation of a training problem in speech recognition and a comparison of training by unconditional versus conditional maximum likelihood. IEEE Trans. Acoustics, Speech, Signal Process. 31 (4), 814–817.

Normandin, Y. Hidden, Markov Models, Maximum Mutual Information Estimation, and the Speech Recognition Problems. Ph.D. Dissertation, McGill University.

Ortmanns, S., Ney, H., Aubert, X., 1997. A word graph algorithm for large vocabulary continuous speech recognition. Comput. Speech Language 11, 43–72.

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

Povey, D., Woodland, P.C., 2002a. Large scale discriminative training of acoustic models for speech recognition. Comput. Speech Language 16, 25–47.

Povey, D., Woodland, P.C., 2002b. Minimum phone error and i-smoothing for improved discriminative training. In: Proc. IEEE Internat. Conf. Acoustics, Speech, Signal Processing, pp. 105–108.

Saon, G., Padmanabhan, M., 2001. Data-driven approach to designing compound words for continuous speech recognition. IEEE Trans. Speech Audio Process. 9 (4), 327–332.

Saon, G., Padmanabhan, M., Gopinath, R., Chen, S., 2000. Maximum likelihood discriminant feature spaces. In: Proc. IEEE Internat. Conf. Acoustics, Speech, Signal Processing, vol. 2, pp. 1129–1132.

Stolcke, A., 2000. SRI language modeling toolkit. Version 1.3.3, 2000. <http://www.speech.sri.com/projects/srilm/>.

Valchev, V., 1995. Discriminative Methods in HMM-based Speech Recognition. Ph.D. Dissertation, Peterhouse, University of Cambridge.

Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer-Verlag, New York.

Wang, H.M., Chen, B., Kuo, J.W., Cheng, S.S., 2005. MATBN: A Mandarin Chinese broadcast news corpus. Internat. J. Comput. Linguistic Chinese Language Process. 10 (1), 219–235.

Yu, D., Deng, L., 2007. Large-margin discriminative training of hidden markov models for speech recognition. In: Proc. IEEE Internat. Conf. Semantic Computing, pp. 429–438.

Yu, D., Deng, L., He, X., Acero A., 2007. Large-margin minimum classification error training for large-scale speech recognition tasks. In: Proc. IEEE Internat. Conf. Acoustics, Speech, Signal Processing, vol. 4, pp. 1137–1140.

Yu, D., Deng, L., He, X., Acero, A., 2008. Large-margin minimum classification error training: A theoretical risk minimization perspective. Comput. Speech Language 22 (4), 415–429.