TRAINING DATA SELECTION FOR IMPROVING DISCRIMINATIVE TRAINING OF ACOUSTIC MODELS

Shih-Hung Liu, Fang-Hui Chu, Shih-Hsiang Lin, Hung-Shin Lee, Berlin Chen

Graduate Institute of Computer Science & Information Engineering,
National Taiwan Normal University, Taiwan

{g93470185, g94470144, berlin}@csie.ntnu.edu.tw, 69308027@cc.ntnu.edu.tw

ABSTRACT

This paper considers training data selection for discriminative training of acoustic models for broadcast news speech recognition. Three novel data selection approaches were proposed. First, the average phone accuracy over all hypothesized word sequences in the word lattice of a training utterance was 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 was investigated. Finally, frame-level data selection based on the normalized frame-level entropy of Gaussian posterior probabilities obtained from the word lattice was explored. The underlying characteristics of the presented approaches were extensively investigated and their performance was verified by comparison with the standard discriminative training approaches. Experiments conducted on the Mandarin broadcast news collected in Taiwan shown that both phone- and frame-level data selection could achieve slight but consistent improvements over the baseline systems at lower training iterations.

Index Terms—speech recognition, discriminative training, acoustic models, data selection, entropy

1. INTRODUCTION

Discriminative training algorithms, such as the maximum mutual information (MMI) training [1, 2] and the minimum phone error (MPE) training [3, 4], 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 was 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 the training utterance, but also the competing hypotheses that are often obtained by performing LVCSR on the utterance.

On the other hand, owing to the availability of huge quantities of speech data, such as that in broadcast radio, television programs, lecture notes, and so on, it is expected that we can utilize more training data than before to reduce recognition errors. However, manual transcripts may not always accompany these speech data, and thus how to obtain reliable automatic transcriptions for these speech data for unsupervised (or lightly supervised) discriminative training of acoustic models would become another important issue. Most of the previous research work on unsupervised discriminative training merely used word posterior probability to filter out the unreliable recognized transcriptions in either utterance-, or word-, or frame-levels [5, 6].

Recently, the large or soft margin classifiers, motivated by the support vector machine (SVM) successfully developed in the machine learning community, have been introduced in the field of speech recognition and demonstrated with good results in small-vocabulary recognition tasks [7, 8]. The concept of margin-based methods is to select useful samples, i.e., the support vectors, to train the classifiers for better model discrimination and generalization. The large margin hidden Markov model (HMM) [7] treated each speech utterance as a sample and used a discriminant function to select positive samples falling in a predefined margin for acoustic model training; while in [8], the authors performed both frame- and utterance-level data selection, for which label matching of the reference and recognized word sequences of the training utterance was first used to identify a candidate set of frame samples and utterance-level data selection was then applied based on the average frame-level log-likelihood ratios obtained from these frames.

With these observations in mind, in this paper we investigated three data selection approaches for discriminative training of acoustic models for LVCSR. First, the average phone accuracy over all hypothesized word sequences in the word lattice of the training utterance was utilized for utterance-level data selection for MPE training. 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 was investigated for MPE training. Finally, frame-level data selection based on the normalized frame-level entropy of Gaussian posterior probabilities obtained from the word lattice was explored for both MMI and MPE training.

The rest of this paper is organized as follows. Section 2 briefly describes two popular discriminative training algorithms that were used in this paper. The proposed data selection approaches are elucidated in Section 3. The experimental settings and the corresponding results are described in Sections 4 and 5, respectively. Finally, conclusions are drawn in Section 6.

2. DISCRIMINATIVE TRAINING APPROACHES

2.1. Basic MMI Formulation

Given a training set of R observation vector sequences $\mathbf{O}=\{O_1,..O_r,..O_R\}$, the MMI criterion for acoustic model training aims to maximize the posterior probability of these observation vector sequences using the following objective function:

$$F_{MMI}(\lambda)=\sum_{r=1}^{R}\log P(W_r \mid O_r), \tag{1}$$

where W_r is the corresponding correct transcription of O_r. MMI attempts not only to make the correct hypothesis more probable, but also to make incorrect hypotheses less probable at the same time. More detailed derivations of the MMI training formulas can be found in [2].

2.2. Basic MPE Formulation

The MPE criterion for acoustic model training aims to minimize the expected phone errors of these observation vector sequences using the following objective function:

$$F_{MPE}(\lambda)=\sum_{r=1}^{R}\sum_{W\in \mathbf{W}_r} P(W \mid O_r)RawAcc(W), \tag{2}$$

where $\mathbf{W}_r$ is the corresponding word lattice of O_r; W is one of the hypothesized word sequences in $\mathbf{W}_r$; $P(W \mid O_r)$ is the posterior probability of hypothesis W given the observation O_r; $RawAcc(W)$ is the "raw phone accuracy" of W in comparison with the corresponding reference transcript W_r, which is typically computed as the sum of the phone accuracy measures of all phone hypotheses in W. The objective function in Eq. (2) can be maximized by applying the Extended Baum-Welch algorithm to update the mean μ_{qmd} and variance σ^2_{qmd} of each dimension d of the m-th Gaussian mixture component of a phone arc q in the word lattice $\mathbf{W}_r$ using the following equations:

$$\mu_{qmd}=\frac{\theta^{num}_{qmd}(O)-\theta^{den}_{qmd}(O)+D\overline{\mu}_{qmd}}{\gamma^{num}_{qm}-\gamma^{den}_{qm}+D}, \tag{3}$$

$$\sigma^2_{qmd}=\frac{\theta^{num}_{qmd}(O^2)-\theta^{den}_{qmd}(O^2)+D(\overline{\sigma}^2_{qmd}+\overline{\mu}^2_{qmd})}{\gamma^{num}_{qm}-\gamma^{den}_{qm}+D}-\mu^2_{qmd}, \tag{4}$$

$$\gamma^{num}_{qm}=\sum_{r=1}^{R}\sum_{q=1}^{Q}\sum_{t=s_q}^{e_q}\gamma^r_{qm}(t)\max(0,\gamma_q^{r^{MPE}}), \tag{5}$$

$$\gamma^{den}_{qm}=\sum_{r=1}^{R}\sum_{q=1}^{Q}\sum_{t=s_q}^{e_q}\gamma^r_{qm}(t)\max(0,-\gamma_q^{r^{MPE}}), \tag{6}$$

$$\theta^{num}_{qmd}(O)=\sum_{r=1}^{R}\sum_{q=1}^{Q}\sum_{t=s_q}^{e_q}\gamma^r_{qm}(t)\max(0,\gamma_q^{r^{MPE}})o_t(d), \tag{7}$$

$$\theta^{num}_{qmd}(O^2)=\sum_{r=1}^{R}\sum_{q=1}^{Q}\sum_{t=s_q}^{e_q}\gamma^r_{qm}(t)\max(0,\gamma_q^{r^{MPE}})o_t(d)^2, \tag{8}$$

$$\gamma_q^{r^{MPE}}=\gamma^r_q(c^r_q-c^r_{avg}), \tag{9}$$

where c^r_{avg} is the average phone accuracy over all hypothesized word sequences in the word lattice; c^r_q is the expected phone accuracy over all hypothesized word sequences containing phone arc q; $o_t(d)$ is the observation vector component at time t; s_q and e_q are the start and end times of phone arc q; $\gamma^r_{qm}(t)$ are the posterior probability for Gaussian mixture component m of phone arc q at time t; γ^{num}_{qm}, $\theta^{num}_{qmd}(O)$ and $\theta^{num}_{qmd}(O^2)$ are the accumulated training statistics for mixture component m of phone arc q whose c^r_q is larger than c^r_{avg}, and vice versa for γ^{den}_{qm}, $\theta^{den}_{qmd}(O)$ and $\theta^{den}_{qmd}(O^2)$; $\overline{\mu}_{qmd}$ and $\overline{\sigma}^2_{qmd}$ are respectively the mean and variance estimated in the previous iteration; and D is a constant used to ensure the positive variance values. On the other hand, the calculation of c^r_{avg} and c^r_q 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:

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

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

$$PhoneAcc(q)=\max_{z_j\in Z_r}\begin{Bmatrix}-1+2e(z_j,q)/l(z_j), & z_j=q\\ -1+e(z_j,q)/l(z_j), & z_j\neq q\end{Bmatrix}, \tag{11}$$

where Z_r is the set of phone labels in the corresponding reference transcript, and $e(z_j,q)$ is the overlap length in time for a phone label z_j in Z_r and a hypothesized phone arc q in W, $l(z_j)$ is the length in time for z_j. More detailed derivations of the MPE training formulas also can be found in [4].

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 [10]. In this paper, we attempted an alterative approach by conducting training utterance selection directly on the error rate domain for MPE training. The word lattice (or hypothesized space) $\mathbf{W}_r$ of a training utterance r, 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 hypothesized space is devoid of discrimination for discriminative training. For example, in MPE training, the normalized average phone accuracy $\hat{c}^r_{avg}$ of each training utterance r, obtained by dividing the average phone accuracy c^r_{avg} by the phone number of the reference transcription of r, to some extent reveals the confusedness of the hypothesis space $\mathbf{W}_r$. The utterance with a too high normalized average phone accuracy implies that less competing information might be provided by it (or its hypothesis space), while 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 $\hat{c}^r_{avg}$. We first estimated the mean of $\hat{c}^r_{avg}$ among all training utterances, denoted as $\bar{\hat{c}}_{avg}$, and then used it together with $\hat{c}^r_{avg}$ to select training utterance that falls in the interval defined by the following equation for MPE training:

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

where δ is a predefined threshold value.

3.2. Phone Selection

In this paper, we proposed a phone-level data selection approach for MPE training that was conducted as well on the error rate domain. As we know, in MPE training, the average phone accuracy c^r_{avg} 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 Eq. (5)-(9). Thus, we can impose a margin on c^r_{avg} in order to select more critical phone arcs which are relatively close to the decision boundary on the error rate domain. As a result, the final auxiliary objective function for MPE training can be defined as:

$$g_{MPE}(\lambda) = \sum_{r=1}^{R}\sum_{q=1}^{Q}\sum_{t=s_q}^{e_q}\sum_{m}\gamma^r_q\left[\left(c^r_q - c^r_{avg}\right)I\left(c^r_q \in A^r\right)\right]\gamma^r_{qm}(t)\log N\left(O_r(t);\mu_{qm},\Sigma_{qm}\right), \tag{13}$$

$$A^r = \left\{c^r_q \middle| -\alpha \le \kappa\left(c^r_q - c^r_{avg}\right) \le \beta\right\}, \tag{14}$$

where $N(\bullet)$ is a Gaussian distribution; $I(\bullet)$ is an indication function; the positive parameters α and β form the margin for training data selection; κ is a normalization factor that makes $\kappa\left(c^r_q - c^r_{avg}\right)$ approximately range from -1 to 1; A^r 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^r would contribute their accumulated statistics for MPE training.

3.3. Frame Selection

We also proposed the use of the entropy information to select the frame-level training statistics for both MMI and MPE training. The normalized entropy of a frame sample t of a given training utterance r can be defined as:

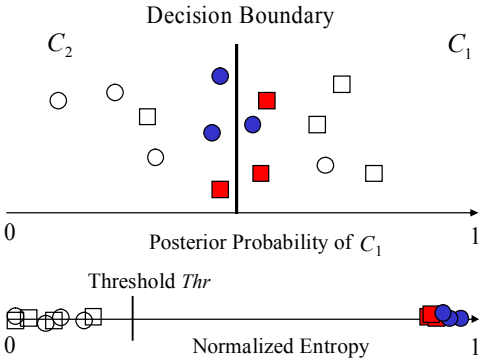

Figure 1. A hypothetical example of binary classification illustrating the relationship between the decision boundary and the normalized entropy.

$$E_r(t) = \frac{1}{\log_2 N(t)}\sum_{q=1}^{Q}\sum_{m\in q}\gamma^r_{qm}(t)\cdot\log_2\frac{1}{\gamma^r_{qm}(t)}, \tag{15}$$

where $\gamma^r_{qm}(t)$ is the posterior probability for mixture component m of phone arc q at frame t, which is calculated from the word lattice $\mathbf{W}_r$; $N(t)$ is the total Gaussian mixtures which have nonzero posterior probabilities at frame t ($\gamma^r_{qm}(t) > 0$); and the value of $E_r(t)$ will range from zero to one. 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 1, 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 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.

For the speech recognition task, two extreme cases are considered as follows. First, if the normalized entropy measure of a frame sample t 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 t is actually located far from the decision boundary. On the other hand, if the normalized entropy measure is close to one, it means

that the posterior probabilities of mixture components tend to be uniformly distributed. Then, the frame sample t 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 algorithm 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 [7, 9].

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 [11]. Figure 2 shows 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 and thus can be discarded if the threshold is appropriately set. In this paper, only the experimental results on the "hard version" of frame-level data selection were reported.

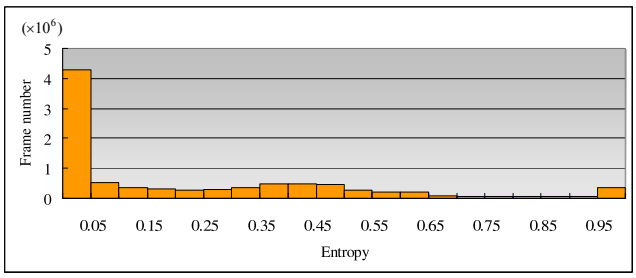

Figure 2. A plot of the relationship between the normalized entropy and the number of training speech frame samples.

4. BROADCAST NEWS SYSTEM

The large vocabulary continuous speech recognition system [12] as well as the experimental speech and language data used in this paper will be described in this section.

4.1. Front-End Signal Processing

The front-end processing was conducted with the HLDA(Heteroscedastic Linear Discriminant Analysis)-based data-driven Mel-frequency feature extraction approach and then processed by MLLT (Maximum Likelihood Linear Transformation) for feature de-correlation.

4.2. Speech Corpus and Acoustic Model Training

The speech corpus consists of about 200 hours of MATBN Mandarin television news (Mandarin Across Taiwan Broadcast News) [13], which were collected by Academia Sinica and Public Television Service Foundation of Taiwan during November 2001 and April 2003. All the 200 hours of speech data are equipped with corresponding orthographic transcripts, in which about 25 hours 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 1.5 hour speech data of the field reporters collected within 2003 were reserved for testing. On the other hand, the acoustic models chosen here for speech recognition are 112 right-context-dependent INITIAL's and 38 context-independent FINAL's.

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

The recognition lexicon consists of 72K words. The 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). The *n*-gram language models were trained using the SRI Language Modeling Toolkit (SRILM).

4.4. Speech Recognition

The speech recognizer was implemented with a left-to-right frame-synchronous Viterbi tree search as well as a lexical prefix tree organization of the lexicon. The recognition hypotheses were organized into a word lattice for further language model rescoring. In this study, the word bigram language model was used in the tree search procedure while the trigram language model was used in the word lattice rescoring procedure.

5. EXPERIMENTS RESULTS

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).

5.1. Baseline Experimental Results

The acoustic models were trained with 24.5 hours 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 training utterances to accumulate the necessary statistics for model training. The ML-trained acoustic models yields a CER of 23.64% 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 depicted in Figures 3 and 4. The total frame number used in the original MMI and MPE training is about 9 millions (24.5 hrs). In the following experiments, for fair comparison between our proposed methods and the baseline MMI and MPE training, the τ values of I-smoothing [3, 4] are set to be the same as that used in the baseline MMI and MPE training, respectively.

5.2. Experiments on Proposed Methods

The recognition results for our proposed methods, including utterance-, phone- and frame-level training data selection, are depicted in Figures 3 and 4, respectively. We first evaluate the performance of the utterance- and phone-level selection methods for MPE training, denoted as MPE+US and MPE+PS, respectively. As can be seen from Figure 3, MPE+PS outperforms the baseline MPE at most of the training iterations. Thus, our argument that imposing margin based on the average phone accuracy (or in the error rate domain) to select confused samples for better discriminative training is tenable. On the other hand, MPE+US is only slightly better than the baseline MPE, though the difference between them is almost negligible at the lower training iterations. However, the training samples used in the training process can be reduced by at least an amount of 10% without any loss of performance in recognition.

We then evaluate the effectiveness of frame-level training data selection for MMI and MPE training, denoted as MMI+FS and MPE+FS, respectively. The corresponding results are shown in Figure 4. The threshold value *Thr* of the frame-level normalized entropy-based training data selection method is set to be 0.05, and the number of training frame samples used is about 4 millions (45.88% of

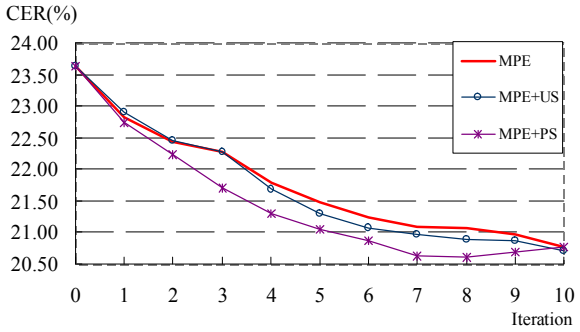

Figure 3. The best experimental results on proposed methods, in comparison with standard MPE training.

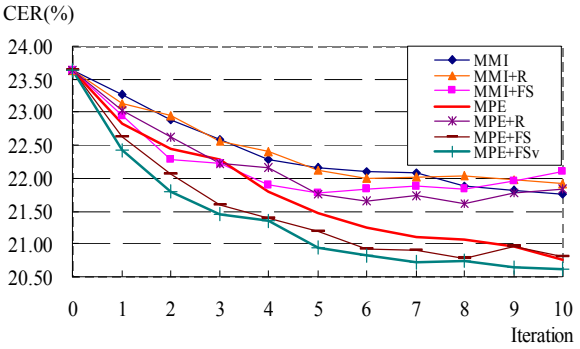

Figure 4. Experimental results on frame-level data selection and random selection approach, in comparison with standard MMI and MPE training.

the total training frame samples). As shown by the preliminary experimental results in Figure 4, frame-level data selection will improve the performance when the acoustic models are trained with lower iterations. However, when the acoustic models are trained with higher iterations (e.g., 9 and 10 iterations), the performance of frame-level data selection (MPE+FS) is slightly worse than the original MPE training. One possible reason for this is that the data selection method to some extent suffers from the data sparseness problem which would make the acoustic models over-trained. Therefore, we alternatively attempt to not only apply the frame-level data selection method for MPE training but meanwhile also decrease the threshold value *Thr* as the iteration increases (denoted as MPE+FSv), for the purpose of obtaining more training statistics and alleviating the over-training problem. The corresponding results of MPE+FSv are depicted in Figure 4, and they signify the superiority over the baseline MPE. On the other hand, we also apply random frame-level training sample selection to both MMI and MPE (denoted as MMI+R and MPE+R, respectively), which randomly selects about 45% of the frame-level training samples for MMI and MPE training at each training iteration, and the corresponding results are depicted in Figure 4. 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 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.

Significance tests based on the standard NIST MAPSSWE [14] also have been conducted on the speech recognition results of the improved methods presented in this paper (for the acoustic models trained at all iterations). Due to the length constraint, only the results using frame-level data selection are shown in Table 1. They indicate the statistical significance of CER improvements (with *p*-value <0.001) over the baseline MMI and MPE training when MMI+FS and MPE+FS, respectively, were exploited at the lower training iterations.

In the meantime, we are extensively experimenting on the ways to combine the proposed data selection methods together for MPE training, including trying different training settings, investigating the joint training of feature transformation and acoustic models, etc.

Table 1. The speech recognition results (CERs) for the proposed improved approaches.

	MMI+FS		MPE+FSv	
Iteration	CER(%)	*p*-value	CER(%)	*p*-value
1	23.28	<0.001	22.43	<0.001
2	22.89	<0.001	21.8	<0.001
3	22.58	<0.001	21.45	<0.001
4	22.28	<0.001	21.34	<0.001
5	22.16	<0.001	20.94	<0.001
6	22.10	<0.001	20.82	<0.001
7	22.08	<0.001	20.73	<0.001
8	21.88	-	20.74	<0.001
9	21.81	-	20.65	<0.001
10	21.75	-	20.63	-

6. CONCLUSIONS

In this paper, we have studied utterance and phone selection for MPE training, and frame-level selection for both MMI and MPE training of acoustic models for LVCSR. Promising and encouraging results on the recognition of Mandarin broadcast news speech have been initially demonstrated. More in-deep investigation of the proposed training data selection methods, as well as their integration with other discriminative acoustic model training algorithms, is currently undertaken. Meanwhile, we are also investigating the possibility of applying the proposed training data selection approaches to unsupervised discriminative training tasks.

7. ACKNOWLEDGEMENTS

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

8. REFERENCES

[1] L. R. Bahl et al., "Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition," in *Proc. ICASSP 1986.*

[2] D. Povey and P. C. Woodland, "Large Scale Discriminative Training of Acoustic Models for Speech Recognition," *Computer Speech & Language*, Vol. 16, 2002.

[3] D. Povey and P.C. Woodland, "Minimum Phone Error and I-Smoothing for Improved Discriminative Training," in *Proc. ICASSP 2002*.

[4] D. Povey "Discriminative Training for Large Vocabulary Speech Recognition," Ph.D. Dissertation, Cambridge, 2004.

[5] H.Y. Chan and P.C. Woodland, "Improving Broadcast News Transcription by Lightly Supervised Discriminative Training" in *Proc. ICASSP 2004*.

[6] Lambert Mathias et al., "Discriminative Training of Acoustic Models Applied to Domains with Unreliable transcripts" in *Proc. ICASSP 2005.*

[7] H. Jiang et al., "Large Margin Hidden Markov Models for Speech Recognition," *IEEE Trans. ASLP 14(5), 2006*.

[8] Jinyu Li et al., "Soft Margin Estimation of Hidden Markov Model Parameters" in *Proc. ICSLP 2006.*

[9] Jinyu Li et al., "Approximate Test Risk Minimization Through Soft Margin Estimation" in *Proc. ICASSP 2007.*

[10] Hui Jiang et al., "A Dynamic In-Search Data Selection Method With Its Applications to Acoustic Modeling and Utterance Verification" *IEEE Trans. SAP 13(5) 2005.*

[11] S. H. Liu et al., "Investigating Data Selection for Minimum Phone Error Training of Acoustic Models," in *Proc. ICME 2007*.

[12] B. Chen et al., "Lightly Supervised and Data-Driven Approaches to Mandarin Broadcast News Transcription," in *Proc. ICASSP 2004*.

[13] H. S. Chiu and B. Chen, "Word Topical Mixture Models for Dynamic Language Model Adaptation," in *Proc. ICASSP 2007*.

[14] L. Gillick and S. Cox, "Some Statistical Issues in the Comparison of Speech Recognition Algorithms," in *Proc. ICASSP 1989*.