ON THE USE OF FRAME-LEVEL INFORMATION CUES FOR MINIMUM PHONE ERROR TRAINING OF ACOUSTIC MODELS

Shih-Hung Liu
Graduate Institute of Computer Science
& Information Engineering,
National Taiwan Normal University
g93470185@csie.ntnu.edu.tw

Berlin Chen
Graduate Institute of Computer Science
& Information Engineering,
National Taiwan Normal University
berlin@csie.ntnu.edu.tw

Abstract

This paper considers discriminative training of acoustic models for Mandarin large vocabulary continuous speech recognition. Two frame-level information cues were explored and integrated into the minimum phone error (MPE) training. First, the frame-level entropy of Gaussian posterior probabilities obtained from the word lattice of the training utterance was exploited to weight the frame-level statistics of the MPE training. The purpose of using entropy is to further emphasize or deemphasize the associated training statistics of plausibly correct and competing models for better discrimination. Second, we presented a new phone accuracy function based on the frame-level accuracy of hypothesized phone arcs instead of using the raw phone accuracy function of the MPE training. The underlying characteristics of the presented approaches were extensively investigated and their performance was verified by comparison with the original MPE training approach. Experiments conducted on the broadcast news collected in Taiwan showed that the presented approaches could achieve slight but consistent improvements over the baseline system.

Keywords: Discriminative training; Minimum phone error; Large vocabulary continuous speech recognition; Extended Baum-Welch algorithm; Entropy measure.

1. Introduction

Discriminative training algorithms, such as the maximum mutual information (MMI) training [1, 2, 3] and the minimum phone error (MPE) training [4, 5], which aim at estimating more accurate acoustic models, have continuously been an focus of much active 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.

In general, most discriminative training algorithms have their roots in risk minimization. For example, in [6] the overall risk criterion estimation (ORCE) was developed to minimize the risk of making speech recognition errors, which took the N-Best list as the reduced hypothesis space and used the Extended Baum-Welch algorithm (EBW) [7] for efficient parameter optimization. However, the N-Best list often contains too much redundant information. A more efficient representation is the word lattice that can compactly encode all the alternative word hypotheses occurring at different segments of speech frames. Nevertheless, for the lattice structure, it also becomes an issue that the use of the standard Levenshtein distance measure as the loss function would make the accumulation of training statistics on it much more complicated. Therefore, two main approaches have been proposed recently to tackle this problem. One is in the focus of how to design alternative loss functions to approximate the conventional Levenshtein distance measure, such as the raw phone accuracy function used in the MPE training [4]. While the other is concerned with how to design an algorithm to efficiently segment the lattice to make the Levenshtein-based alignment practical, such as the pinched lattice exploited in the minimum Bayes risk (MBR) training [8]. Moreover, for the MPE training, which is intended to maximize the expected phone accuracy and can be efficiently computed using different kinds of search structures such as word lattices or confusion networks, quite several efforts have been made recently to show its superiority in various LVCSR tasks, especially in the training [4, 5] or adaptation [9] of acoustic models and feature extraction [10, 11], as well as in the training of language models [12].

Based on these observations, in this paper we studied discriminative training of acoustic models for Mandarin large vocabulary continuous speech recognition. Two frame-level information cues were explored and integrated into the MPE training. First, the frame-level entropy of Gaussian posterior probabilities obtained from the word lattice of training utterance was exploited to weight the frame-level statistics of the MPE training. The purpose of using entropy is to further emphasize or deemphasize the associated training statistics of plausibly correct and competing models for better discrimination. Moreover, we presented a new phone accuracy function based on the frame-level accuracy of hypothesized phone arcs instead of using the raw phone accuracy function of the MPE training for sufficiently penalizing the deletion errors incurred in speech recognition.

The rest of this paper is organized as follows. In Section 2, we first briefly review the overall risk criterion estimation and the minimum phone error training framework, and then explain how to integrate the two proposed frame-level information cues into the minimum phone error training framework. The experimental setup is described in Section 3 and a series of speech recognition experiments conducted are presented in Section 4. Finally, conclusions are drawn in Section 5.

2. Acoustic Model Training

2.1 Overall Risk Criterion Estimation

In speech recognition, the recognizer may take an action $\alpha_u(O)$ to classify a given acoustic vector sequence O to a certain word sequence W_u from a hypothesis space $\mathbf{W}_h$ of all possible word sequence of the language. Let function $L(W_u, W_c)$ be the loss incurred when the recognizer take the action $\alpha_u(O)$ and the correct (or reference) transcript is W_c. Since we have no prior knowledge of the correct transcript W_c in advance, i.e., arbitrary word sequence W_s in $\mathbf{W}_h$ has the probability of being identical to W_c, for each possible action $\alpha_u(O)$, we instead calculate the expected loss (or risk) of it:

$$R(\alpha_u(O) \mid O) = \sum_{W_s \in \mathbf{W}_h} L(W_u, W_s) P(W_s \mid O). \qquad (1)$$

where $P(W_s \mid O)$ is the posterior probability of the word W_s given that the acoustic vector sequence O is observed. On the other hand, for supervised model estimation, because the corresponding correct transcript W_c of each training utterance O is known in advance, the overall risk R_{all} of all training utterances thus can be defined as [6]:

$$R_{all} = \int R(\alpha_c(O) \mid O) P(O) dO. \qquad (2)$$

where the integral extends over the whole observation sequence space. However, in a real-world application task, due to the finite amount of training data, e.g., only utterances involved in the training data, the overall risk can be approximated by:

$$R_{all} \approx \sum_{k=1}^{K} R(\alpha_c(O_k) \mid O_k) P(O_k), \qquad (3)$$

where $\alpha_c(O_k)$ is equivalent to W_{c_k}, the correct transcript of O_k. If $P(O_k)$ is further assumed to be uniformly distributed and then eliminated from the equation, it will lead to the overall risk criterion estimation (ORCE) [6]:

$$F_{ORCE}(\lambda) = \sum_{k=1}^{K} \sum_{W_s \in \mathbf{W}_k^h} L(W_{c_k}, W_s) P(W_s \mid O_k), \qquad (4)$$

where $\mathbf{W}_k^h$ is taken to be a set of likely hypothesized word sequences associated with the training utterance O_k and the distribution $P(W_s \mid O_k)$ is assumed to be governed by some underlying parameter set λ. More recently, the ORCE training had been extended in a number of publications. As one example in [6], the *N*-Best list was taken as the reduced hypothesis space and the loss function was defined as the distance of the hypothesized word sequence to the correct transcript, for which the Levenshtein distance measure associated with word error rate (WER) was adopted. As another other example in [5], the word lattice of the training utterance was instead taken as the hypothesized space and a raw accuracy function was proposed to approximate the Levenshtein distance measure for minimum phone error (MPE) training of acoustic models. In the following subsection, we will briefly review the MPE training framework that will be employed in this study.

2.2 Basic MPE formulation

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:

$$F_{MPE}(\lambda) = \sum_{k=1}^{K} \sum_{W_k \in \mathbf{W}_k^{lat}} RawAcc(W_k) P(W_k \mid O_k), \qquad (5)$$

where $\mathbf{W}_k^{lat}$ is the corresponding word lattice of O_k; 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 with 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 to update the mean μ_{qmd} and variance σ_{qmd}^2 of each dimension d of the m -th Gaussian mixture component of the phone arc q using the following equations:

$$\mu_{qmd} = \frac{\theta_{qmd}^{num}(O) - \theta_{qmd}^{den}(O) + D\overline{\mu}_{qmd}}{\gamma_{qm}^{num} - \gamma_{qm}^{den} + D}, \qquad (6)$$

$$\sigma_{qmd}^2 = \frac{\theta_{qmd}^{num}(O^2) - \theta_{qmd}^{den}(O^2) + D(\overline{\sigma}_{qmd}^2 + \overline{\mu}_{qmd}^2)}{\gamma_{qm}^{num} - \gamma_{qm}^{den} + D} - \mu_{qmd}^2,$$

(7)

$$\gamma_{qm}^{num} = \sum_{k=1}^{K} \sum_{q=1}^{Q} \sum_{t=s_q}^{e_q} \gamma_{qm}^{k}(t) \max(0, \gamma_q^{k^{MPE}}), \qquad (8)$$

$$\gamma_{qm}^{den} = \sum_{k=1}^{K} \sum_{q=1}^{Q} \sum_{t=s_q}^{e_q} \gamma_{qm}^{k}(t) \max(0, -\gamma_q^{k^{MPE}}), \qquad (9)$$

$$\theta_{qmd}^{num}(O) = \sum_{k=1}^{K} \sum_{q=1}^{Q} \sum_{t=s_q}^{e_q} \gamma_{qm}^{k}(t) \max(0, \gamma_q^{k^{MPE}}) o_t(d), \qquad (10)$$

$$\theta_{qmd}^{num}\left(O^2\right) = \sum_{k=1}^{K} \sum_{q=1}^{Q} \sum_{t=s_q}^{e_q} \gamma_{qm}^{k}(t) \max(0, \gamma_q^{k^{MPE}}) o_t(d)^2, \qquad (11)$$

$$\gamma_q^{k^{MPE}} = \gamma_q^k (c_q^k - c_{avg}^k), \qquad (12)$$

where 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 hypothesize phone sequences containing phone arc q ; $o_t(d)$ is the observation vector component at time t; s_q and e_q are the start time and end times of phone arc q ; $\gamma_{qm}^k(t)$ are the posterior probability for mixture component m of phone arc q at time t ; γ_{qm}^{num} , $\theta_{qmd}^{num}(O)$ and $\theta_{qmd}^{num}\left(O^2\right)$ 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 $\theta_{qmd}^{den}\left(O^2\right)$; $\bar{\mu}_{qmd}$ and $\bar{\sigma}_{qmd}^2$ are respectively the mean and variance estimated in the previous iteration; and D is a constant used to ensure the 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:

$$RawAcc\,(W_k) = \textstyle\sum_{q \in W_k} PhoneAcc\,(q), \qquad (13)$$

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

$$PhoneAcc(q) = \max_{z_j \in Z_k} \begin{Bmatrix} -1 + 2e(z_j, q)/l(z_j), & z_j = q \\ -1 + e(z_j, q)/l(z_j), & z_j \neq q \end{Bmatrix}, \qquad (14)$$

where Z_k 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_k and a hypothesized phone arc q in W_k, $l(z_j)$ is the length in time for z_j. More detailed derivations of the MPE training formulae also can be found in [5].

2.3 Prior Information of Training Utterances

As indicated from Sections 2.1 and 2.2, the MPE training has its roots from the ORCE training and also has the assumption that all training acoustic vector sequences have uniform priors. In this paper, we attempted to remove this assumption, and each of training acoustic vector sequences was emphasized or deemphasized by directly using its normalized prior probability or by indirectly using the entropy measure to weight its frame-level statistics. The normalized prior probability of a training utterance O_k can be defined as:

$$\overline{P}(O_r) = \frac{\sum_{W_k \in \mathbf{W}_k^{lat}} P(O_k \mid W_k) P(W_k)}{\sum_{u=1}^{K} \sum_{W_u \in \mathbf{W}_u^{lat}} P(O_u \mid W_u) P(W_u)} \qquad (15)$$

where $P(O_k \mid W_k)$ is the acoustic model probability of W_k generating O_k and $P(W_k)$ is the language model probability of W_k. Equation (15) can be efficiently computed by performing the forward search procedure on the word lattices of all training utterances. The normalized prior probability then can be used to weight the accumulated statistics of the MPE training, as those shown in Equations (8)-(12), in an utterance-wise manner.

On the other hand, we used the entropy measure to weight the frame-level statistics of the MPE training in frame-wise manner. The normalized entropy can be defined as:

$$E_k(t) = \frac{1}{\log_2 N} \sum_{q=1}^{Q} \sum_{m \in q} \gamma_{qm}^k(t) \cdot \log_2 \frac{1}{\gamma_{qm}^k(t)}, \qquad (16)$$

where $\gamma_{qm}^k(t)$ is the posterior probability for mixture component m of phone arc q at time t, which is calculated from the word lattice; N is the total Gaussian mixtures which have nonzero posterior probabilities at time t ($\gamma_{qm}^k(t) > 0$); and the value of $E_k(t)$ will range from 0 to 1 [13]. As indicated in Equations (8)-(12), for the 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. Therefore, when the entropy measure is applied to the MPE training, the accumulated statistics can be respectively modified using the following equations:

$$\hat{\gamma}_{qm}^{num} = \sum_{k=1}^{K} \sum_{q=1}^{Q} \sum_{t=s_q}^{e_q} \gamma_{qm}^k(t) \max(0, \gamma_q^{k^{MPE}} \cdot (1 - \alpha \cdot E_k(t))), \qquad (17)$$

$$\hat{\gamma}_{qm}^{den} = \sum_{k=1}^{K} \sum_{q=1}^{Q} \sum_{t=s_q}^{e_q} \gamma_{qm}^k(t) \max(0, \gamma_q^{k^{MPE}} \cdot (1 + \beta \cdot E_k(t))), \qquad (18)$$

$$\hat{\theta}_{qmd}^{num}(O) = \sum_{k=1}^{K} \sum_{q=1}^{Q} \sum_{t=s_q}^{e_q} \gamma_{qm}^k(t) \max(0, \gamma_q^{k^{MPE}} \cdot (1 - \alpha \cdot E_k(t))) o_t(d), \qquad (19)$$

$$\hat{\theta}_{qmd}^{num}\left(O^2\right) = \sum_{k=1}^{K} \sum_{q=1}^{Q} \sum_{t=s_q}^{e_q} \gamma_{qm}^k(t) \max(0, \gamma_q^{k^{MPE}} \cdot (1 + \beta \cdot E_k(t))) o_t(d)^2, \qquad (20)$$

where α and β are tunable positive parameters which are used to control the convergence rate. Notice that a higher entropy value of the observation vector indicates that it cannot be easily discriminated among the Gaussian mixtures. Therefore, its corresponding training statistics would not be helpful for those plausibly correct models it might belong to but be helpful for those plausibly competing models it might belong to, and vice versa for an observation vector with a lower entropy value.

2.4 Frame-Level Phone Accuracy Function

It is known that the standard MPE training does not sufficiently penalize deletion errors [14]. In general, the original MPE objective function discourages insertion errors more than deletion and substitution errors. Inspired by the work reported in [14, 15], in this paper we presented an alternative phone accuracy function that can look into the frame-level phone accuracies of all hypothesized word sequences in the word lattice to replace the original raw phone accuracy function for the MPE training. The frame-level phone accuracy function (FA) is defined as:

$$FrameAcc\ (q) = \frac{\sum_{t=s_q}^{e_q} \delta(q, Z(t))}{e_q - s_q + 1}, \qquad (21)$$

and

$$\delta(q, Z(t)) = \left\{ \begin{array}{ll} 1 & , if\ q = Z(t) \\ -\rho & , if\ q \neq Z(t), 0 < \rho < 1 \end{array} \right\}, \qquad (22)$$

where $Z(t)$ is the phone label of the reference transcript at time 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 time t, we thus can easily evaluate whether the phone arcs of hypothesized word sequences in the word lattice is identical to that of the reference transcript or not. Figure 1 illustrates the concept of the frame-level phone accuracy. Actually, the presented frame-level phone accuracy function emphasizes the deletion penalty on the incompletely correct phone arc. As illustrated in Figure 1, given the reference transcript "*a-b-c*", the hypothesized phone sequence "*a-c*" will be regarded as being completely correct (with a score of two) using the original raw phone accuracy function, as shown in Equations (9) and (10); however, the presented frame-level phone accuracy function will give the hypothesized phone sequence a score of 1.27 by taking into account the deletion error of it. Another frame-level phone accuracy function that used the Sigmoid function to normalize the phone accuracy value in a range between -1 and 1 was also exploited in this paper (SigFA):

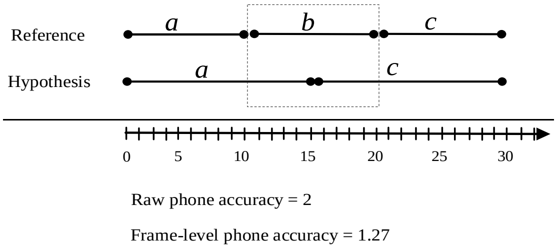
$$SigFrameAc\ c(q) = \frac{2}{1 + \exp(-net)} - 1, \qquad (23)$$

and

$$net\ = \sum_{t=s_q}^{e_q} \delta(q, z(t)), \qquad (24)$$

where $\delta(q, Z(t))$ was previously defined in Equation (22). Notice that, the purpose of the above presented phone accuracy functions are not intended to approximate the standard Levenshtein distance measure, but instead to further sufficiently penalize the deletion errors of speech recognition. Although another 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 proposed recently [16], however, no improved results but only degraded results were demonstrated.

Figure 1: An illustration of the frame-level accuracy. The shaded box indicates where the deletion errors occur.



3. Broadcast News System

The large vocabulary continuous speech recognition system [17] as well as the experimental speech and language data used in this paper will be described in this section.

3.1 Front-End Signal Processing

The front-end processing was conducted with the HLDA-based (Heteroscedastic Linear Discriminant Analysis) [18] data-driven Mel-frequency feature extraction approach and then processed by MLLT (Maximum Likelihood Linear Transformation) transformation for feature de-correlation. Utterance-based feature mean subtraction and variance normalization is applied to all the training and test materials [17].

3.2 Speech Corpus, Acoustic Model Training

The speech corpus consists of about 200 hours of MATBN Mandarin television news (Mandarin Across Taiwan Broadcast News) [12], 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 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

Table 1: The speech recognition results (CERs) for the proposed improved approaches, in comparison with the standard MPE-based training.

	CERs (%)				
Iterations	1	3	6	9	p-value
Original MPE	22.82	22.28	21.24	20.97	-
NP	22.80	21.83	21.23	20.91	0.3085
EW	22.71	21.50	20.66	20.47	<0.001
FA	22.73	22.13	20.97	20.74	0.0228
SigFA	22.88	22.06	21.05	20.58	<0.001
EW+SigFA	22.72	21.51	20.55	20.42	<0.001

pause labels are also involved in the calculation of the accuracies of the hypothesized word sequences.

3.3 Lexicon, *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 with Katz backoff smoothing using the SRI Language Modeling Toolkit (SRILM) [19].

3.4 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 [17].

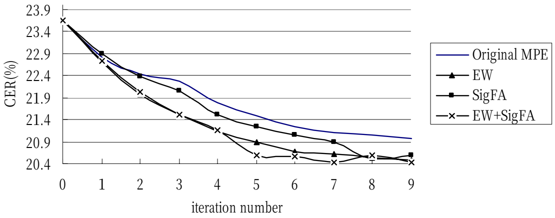
4. Experimental Results

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).

4.1 Baseline Experimental Results

The acoustic models were trained with 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.

Figure 2: The CER curves of the improved approaches in comparison with the standard MPE training at all iterations, in which Iteration 0 represents the result obtained based on the ML trained acoustic models.



The ML-trained acoustic models yields a CER of 23.64%, while the original MPE training (denoted as Original MPE) indeed can provide a great boost to the acoustic models initially trained by ML consistently at all training iterations, as shown in the third row of Table 1. For the limitation of space, only the results of the models trained at iterations 1, 3, 6, and 9 are reported in Table 1. The CER curves of the improved approaches in comparison with the standard MPE training at all iterations are also depicted in Figure 2.

4.2 Experiments on Proposed Methods

The recognition results for the MPE training that used the normalized prior probability to emphasize or deemphasize training utterances (denoted as NP), or indirectly used the entropy measure to weight their frame-level statistics (denoted as EW), are respectively shown in Rows 4 and 5 of Table 1. As can be seen, EW can provide slight but consistent performance gains over the baseline MPE training; however, only insignificant performance improvements are demonstrated by NP. One possible explanation for it is that the normalized prior probabilities of training utterances, as shown in Equation (15), are very often dominated by those utterances that have very large values of the forward search scores.

Moreover, Row 6 (FA) denotes the recognition results using the frame-level phone accuracy function that was normalized by the phone duration. The CER for FA is 20.74% (at the ninth iteration of MPE training), which is better than baseline (an absolute CER reduction of 0.23%). It also can be seen from Row 7 (SigFA) that using the sigmoid function to smooth the frame-level phone accuracy value provides additionally gains than using the frame accuracy function that was normalized by the phone duration. Actually, we have observed from a series of experiments that, using the two variants of frame-level phone accuracy functions presented in this paper with different settings of the value of ρ will give different penalties for insertions and deletions. For example, if the value of ρ is larger, the insertion errors will be discouraged; if the value of ρ is smaller, the number of deletion errors will be decrease.

Finally, Row 8 (SigFA+EW) illustrates the results of fusion of SigFA and EW. Unfortunately, fusion of these two information cues cannot come out with additional performance improvements, and the reason is still under study. Significance tests based on the standard NIST MAPSSWE [20] also have been conducted on the speech recognition results of all the improved approaches presented in this paper (for the acoustic models trained at the ninth iteration). They indicated the statistical significance of CER improvements (with *p*-value <0.001) over the orignial MPE training when either EW or SigFA was exploited in the acoustic model training, as those shown in the last column of Table 1.

In the meantime, we are extensively experimenting on the ways to improve the performance of the MPE training, including trying different sets of training settings, investigating the joint training of feature transformation, acoustic models and language models, etc.

5. Conclusions

In this paper, we have successfully explored the use of the entropy-based weighting and the frame-level phone accuracy functions for the MPE-based discriminative training of acoustic models for large vocabulary continuous speech recognition. The underlying characteristics of the proposed MPE training approaches have been investigated, and their performance was verified by comparison with the original MPE training approach as well. More in-deep investigation of the MPE-based training, as well as integration with other acoustic modeling approaches also currently undertaken.

References

[1] L. R. Bahl et al., “Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal processing*, vol. 11 (1986) 49-52

[2] Y. Normandin, “Hidden Markov models, maximum mutual information estimations and the speech recognition problem,” PhD Thesis, McGill University, Montreal (1991)

[3] D. Povey, P. C. Woodland, “Large Scale Discriminative Training of Acoustic Models for Speech Recognition,” *Computer Speech & Language*, vol. 16, no. 1 (2002) 25-47

[4] D. Povey, P.C. Woodland, “Minimum Phone Error. and I-Smoothing for Improved Discriminative Training,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal processing*, vol. 1 (2002) 105-108

[5] D. Povey “Discriminative Training for Large Vocabulary Speech Recognition,” PhD Dissertation, Cambridge, England (2004)

[6] J. Kaiser et al., “Overall Risk Criterion Estimation of Hidden Markov Model Parameters,” *Speech Communication*, vol. 38, no 3-4 (2002) 383-398

[7] P. S. Gopalakrishnan et al., “A Generalization of the Baum Algorithm to Rational Objective Functions," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal processing*, vol. 1 (1989) 631-634

[8] V. Doumpiotis, W. Byrne, “Lattice Segmentation and Minimum Bayes Risk Discriminative Training,” in *Proceedings of European Conference on Speech Communication and Technology* (2003) 1985-1988

[9] L. Wang, P. C. Woodland, “Discriminative Adaptive Training using The MPE Criterion,” in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding* (2003) 279-284

[10] D. Povey et al., “fMPE: Discriminatively Trained Features for Speech Recognition,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal processing* (2005) 961-964

[11] B. Zhang, S. Matsoukas, “Minimum Phoneme Error based Heteroscedastic Linear Discriminant Analysis for Speech Recognition,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal processing*, vol.1 (2005) 925-928

[12] J. W. Kuo, B. Chen, “Minimum Word Error Based Discriminative Training of Language Models,” in *Proceedings of European Conference on Speech Communication and Technology*, Lisbon, Portugal (2005) 1277-1280

[13] H. Misra, H. Bourlard, “Spectral Entropy Feature in Full-Combination Multi-Stream for Robust ASR,” in *Proceedings of European Conference on Speech Communication and Technology*, Lisbon (2005) 2633-2636

[14] J. Zheng, A. Stolcke, “Improved Discriminative Training using Phone Lattices,” in *Proceedings of European Conference on Speech Communication and Technology* (2005) 2125-2128

[15] F. Wessel et al., “Explicit Word Error Minimization using Word Hypothesis Posterior Probability,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal processing*, vol.1 (2001) 33-36,

[16] G. Heigold, W. Macherey, R. Schluter, H. Ney, “Minimum Exact Word Error Training,” in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, San Juan, Puerto Rico (2005) 186-190

[17] B. Chen et al., “Lightly Supervised and Data-Driven Approaches to Mandarin Broadcast News Transcription,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal processing*, Montreal, Canada, vol. 1 (2004) 777-780

[18] M. J. F. Gales, “Maximum likelihood multiple subspace projections for hidden markov models,” *IEEE Transactions on Speech and Audio Processing (SAP)*, vol. 10, no. 2 (2002) 37-47

[19] A. Stolcke, “SRI language Modeling Toolkit,” version 1.3.3, http://www.speech.sri.com/projects/srilm/.

[20] L. Gillick and S. Cox, “Some Statistical Issues in the Comparison of Speech Recognition Algorithms,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal processing*, Baltimore, USA, vol. 1 (1989) 532-535