AN ENHANCED MINIMUM CLASSIFICATION ERROR LEARNING FRAMEWORK FOR BALANCING INSERTION, DELETION AND SUBSTITUTION ERRORS

YUAN FU LIAO[1], JIA JANG TU[2], SEN CHIA CHANG[2], CHIN HUI LEE[3]

[1]Department of Electronic Engineering, National Taipei University of Technology, Taipei 106, Taiwan
[2]Advanced Technology Center of Information & Communication Research Labs, Industrial Technology Research Institute, Hsinchu 310, Taiwan
[3]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA
[1]yfliao@ntut.edu.tw, [2]{santu, chang}@itri.org.tw, [3]chl@ece.gatech.edu

ABSTRACT

In continuous speech recognition substitution, insertion and deletion errors usually not only vary in numbers but also have different degrees of impact on optimizing a set of acoustic models. To balance their contributions to the overall error, an enhanced minimum classification error (E-MCE) learning framework is developed. The basic idea is to partition acoustic model optimization into three subtasks, i.e., minimum substitution errors (MSE), insertion errors (MIE) and deletion errors (MDE), and select/generate three corresponding sets of competing hypotheses, one for each individual sub-problem. MSE, MIE and MDE are then sequentially executed to gradually reduce the overall word error rates. Experimental results on continuous Mandarin digit recognition of five different data sets collected over various acoustic conditions have consistently shown the effectiveness of the proposed E-MCE learning framework.

Index Terms— MCE, Mandarin Digit Recognition

1. INTRODUCTION

Automatic speech recognizer (ASR) usually produces three different types of errors, including insertion, deletion and substitution. These errors often pose different challenges for discriminative training. For example, the recognition result of an utterance may have many possible insertion errors, but only limited deletion and substitution errors. On the other hand, insertion or deletion errors are more destructive since they may induce neighboring substitution errors. But a substitution error may just pop up a similar word with the same starting and end times. Therefore, from the viewpoint of optimizing a set of acoustic models, it is desirable to have a training framework that can balance the contributions of these three types of errors and to further reduce the overall recognition error rate.

Table 1: A class of discriminative training criteria contained in the unifying approach (adopted from [6]).

	$f(z)$	M_r	A	$G(W,W_r)$
MMI	z	all	1	$\delta(W,W_r)$
MPE	$exp(z)$	all	Free	
MCE	$-1/\{1+exp[2exp(z)]\}$	w/o Wr	1	$A(W,W_r)$

Discriminative training algorithms, such as maximum mutual information (MMI) [1], minimum phone error (MPE) [2] and minimum classification error (MCE) [3-8], are the state-of-the-art acoustic model training techniques. In [6] a unifying view for MMI, MPE and MCE was presented and their training criteria are summarized in Table 1 and Eq. (1).

$$F(\Lambda; f, \alpha, G, M_r) = \frac{1}{R}\sum_{r=1}^{R} f\left(\log\left[\frac{\sum_{W} p_\Lambda^\alpha(X_r \mid W)\cdot p^\alpha(W)\cdot G(W,W_r)}{\sum_{W\in H_1} p_\Lambda^\alpha(X_r \mid W)\cdot p^\alpha(W)}\right]^{\frac{1}{\alpha}}\right) \quad (1)$$

where Λ, f, α and G are the parameters of acoustic model, smoothing function, weighting exponent and the gain function, respectively, and r denotes the index of a training utterance, each consisting of a sequence, X_r, of acoustic observation vectors and the corresponding word sequence, W_r, H_1 is the set of the alternative competing word sequences.

It is worth noting that from Table 1 and Eq. (1) that the choices of Kronecker $\delta(W,Wr)$ or raw word/phone accuracy $A(W, Wr)$ functions implies that these algorithms usually do not take into account the different contributions of the insertion, deletion and substitution errors.

To balance the effects of the three error types when optimizing a set of acoustic models, an enhanced minimum classification error (E-MCE) learning framework is developed in this paper. The basic idea is to partition the acoustic model optimization problem into three smaller subtasks, i.e., minimum substitution error (MSE), insertion error (MIE) and deletion error (MDE), and select/generate different sets of competing hypotheses for each subtask. In other words, three different sets of hypothesized theories, containing mainly substitution, insertion and deletion errors, respectively, are separately generated and presented to the MSE, MIE and MDE training modules in an interleaving manner to sequentially correct each corresponding type of errors. Comparing with the conventional discriminative training approaches, one benefit of our approach is that different types of competing hypotheses, especially with samples containing more insertion and deletion errors, can be observed and optimized in the training phase [8].

2. MCE TRAINING FRAMEWORK

MCE aims at minimizing the (smoothed) empirical error on the training data. The MCE formulation for hidden Markov model (HMM) with parameter set, Λ , is briefly summarized as follows.

For every training utterance X_r , a misclassification measure, $d\left(X_r \mid \Lambda\right)$ compares a discriminant function $g\left(X_r, W_r \mid \Lambda\right)$ for the known word sequence label W_r with a competing anti-discriminant function, $G\left(X_r, W_n \mid \Lambda\right)$ in Eq. (1), i.e.:

$$d\left(X_r \mid \Lambda\right) = -g\left(X_r, W_r \mid \Lambda\right) + G\left(X_r, W_n \mid \Lambda\right). \quad (2)$$

Here $G\left(X_r, W_n \mid \Lambda\right)$ is a weighted sum over the set H_1 of n competing N-best sentences, W_n [3]. Then the misclassification measure is turned into a soft error count using a sigmoid function.

$$\ell\left(X_r \mid \Lambda\right) = \frac{1}{1+\exp\left(-\lambda \cdot d\left(X_r \mid \Lambda\right) + b\right)} \quad (3)$$

where λ and b control the slope and offset, respectively.

Thirdly, given the training set, $\left\{X_r, r = 1, \ldots R\right\}$, the empirical recognition error for minimization is given by $L = \sum_{r=1}^{R} \ell\left(X_r \mid \Lambda\right)$, which is often optimized by generalized probabilistic descent [3]. Recently, improved MCE algorithms, representing competing hypotheses by a word-graph or lattice instead of an N-best list to collect more alternatives, have also been developed [6-7].

3. AN E-MCE TRAINING FRAMEWORK

The proposed E-MCE training framework is consisted of MSE, MIE and MDE, with three different sets of competing hypotheses, i.e., $\mathrm{H}_{1,\mathrm{S}}$, $\mathrm{H}_{1,\mathrm{I}}$ and $\mathrm{H}_{1,\mathrm{D}}$ .MSE, MIE and MDE all follow the MCE training framework (Eqs. (2)~(3)). But they select/generate their own sets of competing hypotheses, $\mathrm{H}_{1,\mathrm{S}}$, $\mathrm{H}_{1,\mathrm{I}}$ and $\mathrm{H}_{1,\mathrm{D}}$, which contain mainly the substitution, insertion and deletion error samples, respectively. Thus they also have their own competing anti-discriminant functions, $G_S\left(X_r, W_n \mid \Lambda\right)$, $G_I\left(X_r, W_n \mid \Lambda\right)$ and $G_D\left(X_r, W_n \mid \Lambda\right)$, when computing Eqs. (2) and (3).

3.1. Competing Hypotheses Generation

In this paper, we choose to embed some constraints into the N-best list search algorithm to directly generate three sets of competing hypotheses. In short, three different search networks, each with equal to, more than or less than the correct number of words in the training utterance, X_r , are applied to the N-best search algorithm in the training phase. This configuration facilitates the online generation of three desired sets of hypotheses, directly related to a specific requirements, either for MSE, MIE or MDE, respectively.

The generated sets, $\mathrm{H}_{1,\mathrm{S}}$, $\mathrm{H}_{1,\mathrm{I}}$ and $\mathrm{H}_{1,\mathrm{D}}$, are then used in turn to calculate the competing anti-discriminant functions, $G_S\left(X_r, W_n \mid \Lambda\right)$, $G_I\left(X_r, W_n \mid \Lambda\right)$ and $G_D\left(X_r, W_n \mid \Lambda\right)$, in the MSE, MIE and MDE modules.

4. MANDARIN DIGIT STRING RECOGNITION

The proposed E-MCE training framework is evaluated on Mandarin digit recognition [9-10]. Mandarin digits are all in monosyllables and with much shorter durations than their counterparts in Western language. Especially, Mandarin digit recognizers often produce insertion or deletion errors on some vowel-vowel sequences.

For example, “1-1” (/yi-yi/), “2-2” (/er-er/), “5-5” (/wu-wu/), “7-1” (/chi-yi/), “8-2” (/ba-er/) and “9-5” (/jiou-wu/) pairs are easily misrecognized as “1” (/yi/), “2” (/er/), “5” (/wu/), “7” (/qi/), “8” (/ba/) and “9” (/jiou/) and vice versa. According to our previous experiences, more than half of the errors could be imputed to the vowel-vowel sequence insertions or deletions [9].

.

4.1. Databases and Experimental Settings

Five Mandarin digit corpora were used in this study. All were collected through different telephony networks (PSTN and GSM). Each utterance in the training and test sets consists of 4 to 16 digits. Among them, NTUT-CONFUSION is specially designed to reflect the most challenging situation in Mandarin digit recognition, i.e., the vowel-vowel pairs and confusable pairs. A detailed description of the testing sets is listed in Table 2.

In all the following experiments, 19 context-independent phone models, trained using HTK toolkits with the maximum likelihood (ML) criteria, were used as the baseline. Thirty-nine mel-frequency cepstral coefficients (13 MFCCs and their first and second time derivatives) were computed with a window size of 20ms and a frame shift of 10ms. Moreover, Feature domain cepstrum mean subtraction, variance normalization, and ARMA filtering (MVA) were applied to partially reduce the channel, handset and background noise distortions.

To optimize the MCE, MSE, MIE and MDE loss functions, N-best lists and a generalized probabilistic gradient descent (GPD) [3] method are adopted. The length of the N-best list was empirically set to 50 for all experiments.

Table 2. The five Mandarin digit corpora used in all evaluations.

Corpus		Channel	Content	#. of utt.
Training	MAT-TR [10]	PSTN	random digit string	5080
Test	MAT-TS [10]	PSTN	random digit string	757
	ITRI-ID	PSTN	ID number	1243
	NTUT-CREDIT	GSM	credit card number	475
	NTUT-CONFUSION	GSM	vowel-vowel sequences and confusable pairs	3523

4.2. Learning Curve Analysis

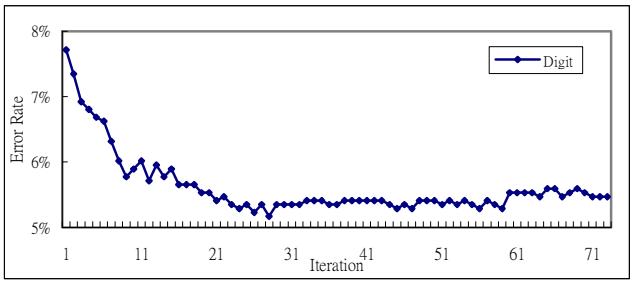
In the following subsections, the property of the conventional MCE framework on Mandarin digit recognition is first studied. Secondly, the behaviors of the MSE, MIE and MDE modules are explored. Finally, we compare the learning curves of the E-MCE and MCE approaches. In all analyses, two subsets of the MAT-TR and MAT-TS [10] corpora were used to train and evaluate these

approaches. There are 1,251 and 300 utterances in the training and test set, respectively. It is worthy noting that in this case, the ML baseline system unfortunately gives higher insertion and lower substitution and deletion errors.

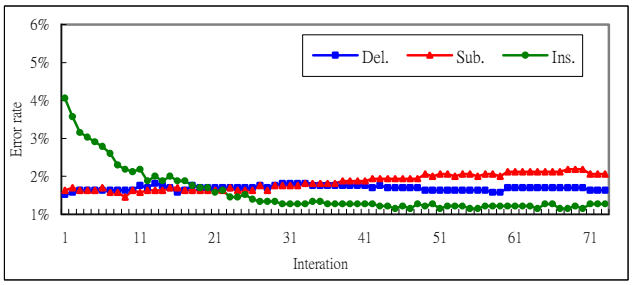
4.2.1. MCE

The learning curves of the outside test sets for the MCE methods are shown in Fig. 1 (a) and (b). Could be seen from Fig. 1 (a), fast convergence and high overall digit error reduction were achieved by applying the conventional MCE method.

However, it is also clear from Fig. 1 (b) that most of the improvements were obtained mainly from the over reduction of insertions, while little from the deletion or substitution errors. Besides, more training iterations will further cause the increasing of the substitution and deletion errors and result in instability, i.e. unbalanced insertion, deletion and subtraction errors in MCE.



(a)



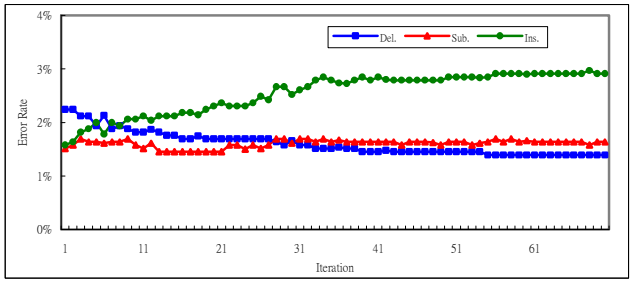
(b)

Figure 1. Learning curves of MCE on the outside test (subsets of the MAT-TS): (a) overall error, and (b) substitution, insertion and deletion error rates.

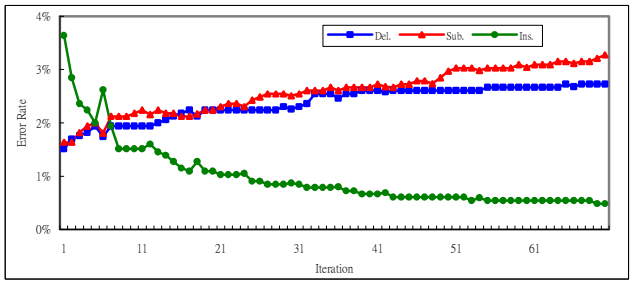
4.2.2. MSE, MIE and MDE

The learning curves of the outside test sets for MSE, MIE and MDE are shown in Fig. 2 (a), (b) and (c), respectively. It can be seen that MSE, MIE and MDE were all capable of reducing the corresponding substitution, insertion and deletion errors.

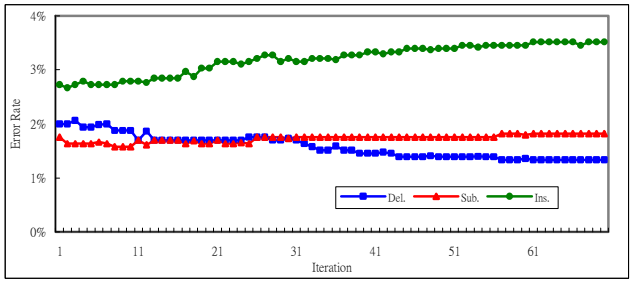
On the other hand, those learning curves in Fig. 2 (a), (b) and (c) also indicate that there are strong correlation and conflict between the corrections of the substitution, insertion and deletion errors when optimizing a set of acoustic models. Especially, Fig. 2 (a) shows that reducing the substitution errors will also increase the insertion errors at the same time. Fig. 2 (b) indicates that removing the insertion errors will raise both the substitution and deletion errors. Fig. 2 (c) reveals that reducing the deletion errors will generate more insertion errors but has only little effects on substitution errors. Therefore, a key to reduce the overall errors is to balance the contributions of MSE, MIE and MDE to the overall error reduction in the overall E-MCE training phase.



(a)



(b)



(c)

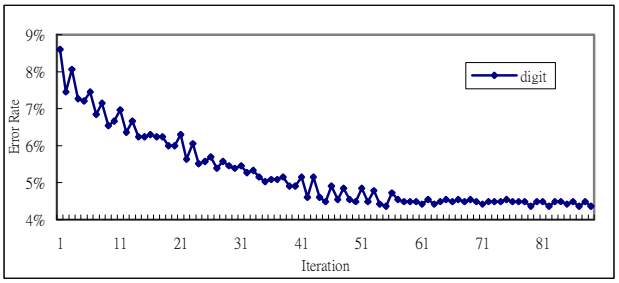
Figure 2. The learning curves of (a) MSE, (b) MIE and (c) MDE on the outside test set (subsets of the MAT-TS).

4.2.3. E-MCE

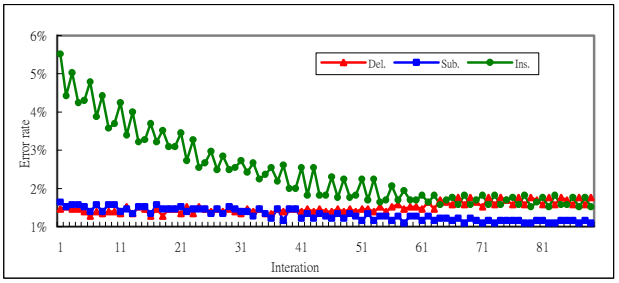
It has been observed that MSE and MDE both decrease substitution and deletion errors, (see Fig. 2 (a) and (c)), on the contrary MIE increases substitution and deletion errors, (see Fig. 2 (b)). When optimizing an acoustic model, we interleave the MSE, MIE and MDE procedures as follows:

Step 1. execute MSE and MIE in turn twice
Step 2. execute MDE 1 time
Step 3. go to Step 1 if not converged

The learning curve in Fig. 3 (a) clearly shows a zig-zag behavior caused by the interleaving of MSE, MIE and MDE. When compared with Fig. 1 (a), this behavior made E-MCE training converging slower than MCE. However the situation with over-reduction of the insertion errors in MCE was avoided. It also allowed E-MCE to reduce the substitution errors. Another benefit of E-MCE is that the insertion and deletion errors were often automatically balanced, (see Fig. 3 (b)).



(a)



(b)

Figure 3. Learning curves of E-MCE on the outside test set (subsets of the MAT-TS): (a) digit error, and (b) substitution, insertion and deletion error rates.

4.3. Experimental Results

The proposed E-MCE training framework trained with the whole MAT-TR corpus was finally evaluated on several different larger scale test corpora and compared with conventional MCE. It can be seen from Table 3, high relative error reductions (both digit and sentence errors) were achieved by applying the conventional MCE method. However, it is also clear from Table 3 that most of the improvements were obtained mainly from the over reduction of the insertion errors. Therefore, these results also demonstrate the impact of unbalanced errors in MCE training.

Table 3 Detailed performance comparisons of errors of ML, MCE and E-MCE leaning on various Mandarin connected digit corpora.

Testing set	MODEL	Ins.	Del.	Sub.	dig. acc.	sen. acc.	dig. rer.	sen. rer.
MAT-TS	ML	1.73	1.70	1.89	94.68	76.88	—	—
	MCE	0.41	1.70	1.08	96.80	85.34	39.85	36.60
	E-MCE	0.53	1.29	0.85	97.33	86.53	49.81	41.74
ITRI-ID	ML	1.19	2.55	1.98	94.27	72.73	—	—
	MCE	0.17	2.40	1.60	95.83	81.17	27.23	30.95
	E-MCE	0.25	1.96	1.45	96.34	81.74	36.13	33.04
NTUT-CARD	ML	1.76	1.74	2.22	94.28	45.26	—	—
	MCE	0.61	2.10	2.05	95.25	56.21	16.96	20.00
	E-MCE	0.74	1.48	1.72	96.05	60.42	30.94	27.69
NTUT-CONFUSION	ML	4.15	3.52	3.50	88.84	65.20	—	—
	MCE	1.28	2.78	3.23	92.72	74.79	34.77	27.56
	E-MCE	2.07	2.33	2.63	92.98	75.99	37.10	31.01

On the other hand, Table 3 also demonstrates that E-MCE can overcome the potential drawback of conventional MCE and did reduce or balance all three types of errors at the same time. It eventually achieved a better overall performance. This is true even for the case of the worst situation with the NTUT-CONFUSION corpus which has the most vowel-vowel and confusion pairs.

Since the proposed E-MCE training framework had shown consistent improvement over various testing set, we therefore conclude that E-MCE outperforms the MCE approach, and achieved a significant improvement on recognition accuracy when compared with ML training.

5. CONCLUSION

An E-MCE training framework has been successfully developed to directly control and automatically balance the insertion, deletion and subtraction errors when optimizing a set of acoustic models. Experimental results on five Mandarin digit recognition tasks have consistently shown that E-MCE outperforms the conventional MCE approach. The same concept can be applied to lattice-based MCE and other discriminative training methods.

6. ACKNOWLEDGEMENTS

This study is partially supported by Project 6352B22100 and conducted at ITRI under the sponsorship of the Ministry of Economic Affairs, Taiwan.

7. REFERENCES

[1] P. C. Woodland and D. Povey, “Large scale discriminative training of hidden Markov models for speech recognition,” Computer Speech & Language, 2002.

[2] D. Povey and P. C. Woodland, “Minimum phone error and i-smoothing for improved discriminative training,” in ICASSP, 2002.

[3] B.-H. Juang, W. Chou and C.-H Lee, “Minimum classification error rate methods for speech recognition”. IEEE Transactions on Speech and Audio Processing, IEEE-SAP, 5(3):257--265, May 1997.

[4] E. McDermott, T. J. Hazen, J. Le Roux, A. Nakamura and S. Katagiri, “Discriminative training for large vocabulary speech recognition using Minimum Classification Error”, IEEE Transactions on Audio, Speech and Language Processing, January 2007.

[5] D. Dong; D. Li, He, X. He; and A. Acero, “Large-Margin Minimum Classification Error Training for Large-Scale Speech Recognition Tasks”, in ICASSP’2007.

[6] W. Macherey, L. Haferkamp, R. Schlüter, and H. Ney, “Investigations on Error Minimizing Training Criteria for Discriminative Training in Automatic Speech Recognition,” in Interspeech 2005.

[7] Q. Fu, X. He, and L. Deng. “Phone-Discriminating Minimum Classification Error (P-MCE) Training for Phonetic Recognition,” Proceedings of Interspeech, Antweerp, Belgium, Aug. 27-31, 2007.

[8] Y. C. Tam and B. Mak, “An Alternative Approach of Finding Competing Hypotheses for Better Minimum Classification Error Training,” Proceedings of ICASSP, Orlando, Florida, USA, May 2002.

[9] Y.-F. Liao, N. Wang, M. Huang, H. Huang and F. Seide, “Improvements of the Philips 2000 Taiwan Mandarin Benchmark System”, ICSLP, pp.298-301, Beijing, 2000.

[10] H. C. Wang, F. Seide, C. Y. Tseng and L. S. Lee, “MAT2000 - Design, collection, and validation of a Mandarin 2000-speaker telephone speech database,” in ICSLP 2000.