I.J. Information Technology and Computer Science, 2011, 2, 17-23
Published Online March 2011 in MECS (http://www.mecs-press.org/)

Pronunciation Proficiency Evaluation based on Discriminatively Refined Acoustic Models

Ke Yan, Shu Gong
USTC iFlytek Speech Laboratory, University of Science and Technology of China, Hefei, China
Email: {kenyk , shugong}@mail.ustc.edu.cn

***Abstract*— The popular MLE (Maximum Likelihood Estimation) is a generative approach for acoustic modeling and ignores the information of other phones during training stage. Therefore, the MLE-trained acoustic models are confusable and unable to distinguish confusing phones well. This paper introduces discriminative measures of minimum phone/word error (MPE/MWE) to refine acoustic models to deal with the problem. Experiments on the database of 498 people's live Putonghua test indicate that: 1) Refined acoustic models are more distinguishable than conventional MLE ones; 2) Even though training and test are mismatch, they still perform significantly better than MLE ones in pronunciation proficiency evaluation. The final performance has approximately 4.5% relative improvement.**

***Index Terms*—computer assisted language learning, MPE, MWE, posterior probability, PSC, discriminative training**

I. INTRODUCTION

PSC (Putonghua Shuiping Ceshi, Chinese mandarin test), with more than 3 minion attendances each year, plays an important role in the popularization of mandarin. However the scoring task for PSC is highly boring, time-consuming and labor-intensive. Let us suppose that each exam taker needs 12 minutes to finish his/her test and every paper needs two teachers working together. Therefore, one teacher can only finish 20 students' pronunciation quality evaluation when working 8 hours per day! The advent of automatic PSC system [1]-[3] brought about a revolution in PSC— computers can do scoring tasks as good as trained evaluators! It is now being widely used in more than ten provinces of China. However, its performance still needs improving.

Pronunciation quality evaluation plays an important role in computer assisted language learning (CALL). Frame-normalized posterior probability [4]-[8] is commonly used as promising measurement for computers. Acoustic models play an important role for the calculation of such measurements.

MLE (Maximum Likelihood Estimation) approach of model training can relax the labeling of phone boundaries and is efficient to compute, so it is widely accepted in CALL systems. However, MLE is a generative method and does not use other phones' information during training stage. As we know, some confusing pairs in mandarin, such as "zh-z", "sh-s", "in-ing", "en-eng", "c-ch" et al, are naturally similar to each other. Therefore, without seeing the difference between these pairs, MLE approach will naturally build such phonetic acoustic models similar to each other. Obviously this hampers the pronunciation quality performance.

In the field of ASR (Automatic Speech Recognition), discriminative training (DT) is commonly adopted to deal with the problem. It is a model refining method that pays more attention to the difference among acoustic models. In this way, it can make acoustic models easier to distinguish from each other. The idea of discriminative training originated from 1986 when Baul first reported the work in small vocabulary speech recognition task [9]. Until recent years, with the introduction of "Word Graph", DT has achieved better performance than MLE[10][11] in LVCSR (Large Vocabulary Continuous Speech Recognition). In 2002, D. Povey proposed DT criteria of minimum phone/word error (MPE/MWE) and they outperform traditional DT criteria in LVCSR [12].

In recent two years, there were many applications of discriminative training in error detection field and encouraging results appeared to follow hard at heel. In Feng Zhang's dissertation [13] and Xiaojun Qian's work [14], they all pointed that MPE/MWE criteria are same with the aim of error detection in some cases.

However, discriminative training has not been reported in native speakers' automatic scoring tasks from our investigations so far. Most PSC testees are native Chinese and they are able to speak Putonghua fluently. Therefore, according to "The Outline of PSC" [15], pronunciation quality evaluation is put into priority. Evaluators would pay strictly attention on the mentioned typical confusing pairs [16]. This paper introduced discriminative training measures of MPE/MWE into automatic PSC system to deal with the problem. The experimental results evidently show that DT can effectively release the confusion among acoustic models. The MPE/MWE refined acoustic models also achieve 4.5% relative improvement in pronunciation evaluation.

II. INTRODUCTION OF PHONE POSTERIOR PROBABILITY AND PHONE SCORING MODEL

A. Traditional Measurement of Frame-averaged Phone Posterior Probability

Let us suppose that $i=id(r,n)$ is the canonical phone's index for n-th phone in the r-th utterance, with its corresponding HMM (Hidden Markov Model) θ_i and acoustic feature vector $\mathbf{O}_r^n$ .Then the frame-normalized

phone posterior probability for phone θ_i is show as (1). [1][9][10].

$$\ln P\left(\theta_i \mid \mathbf{O}_r^n\right) = \frac{1}{T_r^n} \ln \frac{P\left(\mathbf{O}_r^n \mid \theta_i\right)}{\sum_{\theta_{im} \in M_i} P\left(\mathbf{O}_r^n \mid \theta_{im}\right)} \quad (1)$$

Where T_r^n is the frame length and M_i is probability space designed for phone θ_i . Phone boundaries are calculated by ASR with restrictive network generated by given text [1]. Fig.1 illustrates the way to compute phone posterior probability.

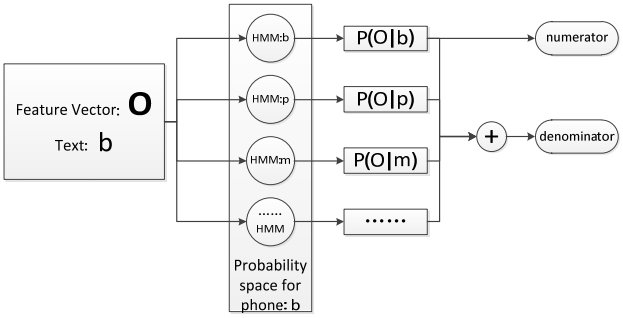

Figure 1. The Calculation of phone posterior probability

Eq.1 judges goodness of pronunciation in phone level. The speaker level measurement can be calculated via averaging all frame-normalized phone posterior probabilities as (2) shows, where N_r is the phone count for r -th utterance.

$$\sigma_{PP}(r) = \frac{1}{N_r} \sum_{n=1}^{N_r} \ln P\left(\theta_i \mid \mathbf{O}_r^n\right) \quad (2)$$

B. Phone Scoring Measurement based on Frame-normalized Phone Posterior Probability

Eq.1 and 2 measure goodness of pronunciation in log probability domain. For probabilities are seriously affected by probability spaces, our recent work proposed trainable "Phone Scoring Model" [17], which transforms frame-normalized phone posterior probabilities into phone scores. In this paper, free linear phone scoring model is adopted, as is showed in (3) and (4), where $\tilde{s}_r^n$ denotes phone machine score and $\tilde{s}_r$ is utterance machine score.

$$\tilde{s}_r^n = \alpha_i \cdot \ln P\left(\theta_i \mid \mathbf{O}_r^n\right) + \beta_i \quad (3)$$

$$\tilde{s}_r = \frac{1}{N_r} \sum_{n=1}^{N_r} \tilde{s}_r^n \quad (4)$$

The parameters for phone scoring model $\{\alpha_i, \beta_i\}$ are trained by minimizing root mean square error of human and machine scores in the development set. Our previous work in [17] reported over 24-40% relative performance gain over the popular posterior probability approach. As [17] has not been published yet, this paper will investigate discriminative training in both pronunciation quality measurements.

III. Discriminative Training for Evaluation

A. Typical Errors for Native Chinese Speakers

As most PSC testees are native Chinese speakers with fluency in mandarin, the PSC outline [15] put pronunciation accuracy into priority. Affected by different dialect, many native Chinese speakers often make mistakes on some confusing pairs such as "z-zh", "c-ch", "s-sh", "in-ing", "en-eng", "n-l" and so on. Evaluators would pay strictly attention on distinguish such confusing phones.

B. MLE-trained and DT-refined models for scoring

MLE is a generative criterion for acoustic model training. Vividly speaking, it tells the model "this is an apple" and only uses data of "apple" for models training. Therefore MLE lacks the consideration of the differences between "an apple" and "an orange".

From discussion of previous section, we can see that the mentioned confusing pairs are naturally similar to each other. Evaluators are concentrated on distinguishing tiny acoustic differences for these typical error pairs. Vividly speaking, now the task is to distinguish "a big deep red apple" and "a big light red apple". MLE approach would focus on learning characteristics like "big" , "apple" , "red" and will not pay enough attention to key differences ---- "deep red" and "light red" that can distinguish them.

Discriminative training takes more care on how to distinguish them from each other. Vividly speaking, it tells model "this is a big deep red apple, not a light red one". Fig.2 is a sketch map of the principal that how discriminative training improves scoring.

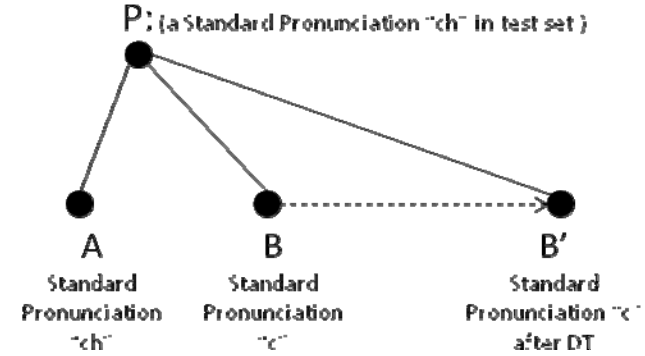

Figure 2. Schematic diagram of the principal that how discriminative training improves scoring.

Confusing pair "ch-c" are naturally similar in acoustic models. So even if "ch" is rightly pronounced in test set (Point *P*), it is still confusing (*PA*≈*PB*); after we used DT to refine the model structure, "ch-c" are much more distinguishable (*AB*'>*AB*), therefore, we can easily find that *P* is a standard pronunciation "ch"(*PA*<*PB*').

C. Acoustic Models Refinement based on MPE/MWE

MPE/MWE criteria were first proposed by D. Povey in 2002 and they outperformed other discriminative criteria in LVCSR. Therefore, this paper will investigate the

discriminative measure of MPE/MWE in pronunciation proficiency evaluation.

Let us suppose the phone set contains *I* different phones and each of them is represented by a HMM model θ_i with *s* states. Each state is represented by a GMM (Gaussian Mixture Model). Therefore, the parameters of θ_i can be denoted (5), where $(u_{isk}, \sigma_{isk}, c_{isk})$ denotes the mean vector, variance vector and component weight for *k*-th Gaussian of *s*-th state in HMM θ_i.

$$\theta_i = \left\{ (u_{isk}, \sigma_{isk}, c_{isk}), s = 1, \cdots, s_i; k = 1, \cdots, K_{s_i} \right\} \quad (5)$$

Let us use $\boldsymbol{\theta} = \{\theta_1, \theta_2, \cdots, \theta_I\}$ denotes the acoustic model collection. The objective of MPE/MWE is to minimize the phone/word errors or to maximize the phone/word correct number by adjusting $\boldsymbol{\theta}$ as (6) shows.

$$\begin{aligned} F(\boldsymbol{\theta}) &= \sum_{W \in M} P_{\boldsymbol{\theta}}^{\kappa}(W \mid \mathbf{O}) \boldsymbol{\alpha}(W, W_r) \\ \boldsymbol{\theta} &= \arg\max_{\boldsymbol{\theta}} F(\boldsymbol{\theta}) \end{aligned} \quad (6)$$

Where the word posterior probability for *W* is shown as follows:

$$P_{\boldsymbol{\theta}}^{\kappa}(W \mid \mathbf{O}) = \frac{P_{\boldsymbol{\theta}}^{\kappa}(\mathbf{O} \mid W) P(W)}{\sum_{W' \in M} P_{\boldsymbol{\theta}}^{\kappa}(\mathbf{O} \mid W') P(W')} \quad (7)$$

In (6) and (7), W is current word sequence and W_r is reference word sequence. $\boldsymbol{\alpha}(W, W_r)$ is the correctness degree for current word sequence W. $\boldsymbol{\alpha}(W, W_r)$ is phone level correctness degree for MPE criterion and word level correctness degree for MWE criterion.

Soft decision is usually applied as (8) shows, where q denotes a word/phone in current word sequence W, z denotes its corresponding word/phone in reference word sequence W_r and $e(q, z)$ denotes the overlap rate for q. This shows MPE/MWE criteria also aim at getting more accurate phone boundaries.

$$\boldsymbol{\alpha}(W, W_r) = \max_z \begin{cases} -1 + 2e(q, z) & q = z \\ -1 + e(q, z) & q \neq z \end{cases} \quad (8)$$

Extended Baum-Welch algorithm is often adopted for parameter optimization as (9) shows.

$$\begin{aligned} u_{isk} &= \frac{\Gamma_{isk}(O) + c_{isk} D u_{isk}^{(0)}}{\Gamma_{isk}(1) + c_{isk} D} \\ \sigma_{isk}^2 &= \frac{\Gamma_{isk}(O^2) + c_{isk} D\left(\left(u_{isk}^{(0)}\right)^2 + \left(\sigma_{isk}^{(0)}\right)^2 \right)}{\Gamma_{isk}(1) + c_{isk} D} - u_{isk}^2 \end{aligned} \quad (9)$$

Where *D* is step size pre-set, $\Gamma_{isk}(1)$, $\Gamma_{isk}(O)$ and $\Gamma_{isk}(O^2)$ are referred to zero-order, first-order and second-order accumulative statistics shown in (10)-(12). Symbol o_{rt} denotes the observation in *t*-th frame of *r*-th utterance. $\gamma_{tr}^{(0)}(i, s, k; W_r)$ and $\gamma_{tr}^{(0)}(i, s, k)$ are the posterior probabilities for the *k*-th Gaussian of *s*-th state in HMM model $\theta_i^{(0)}$ given reference word sequence W_r or whole word graph (lattice) generated by ASR decoding.

$$\Gamma_{ik}(1) = \frac{1}{R} \sum_{r=1}^{R} F'\left(\boldsymbol{\theta}^{(0)}\right) \left\{ \sum_{t=1}^{T_r} \left[\gamma_{tr}^{(0)}(i, s, k; W_r) - \gamma_{tr}^{(0)}(i, s, k) \right] \right\} \quad (10)$$

$$\Gamma_{ik}(O) = \frac{1}{R} \sum_{r=1}^{R} F'\left(\boldsymbol{\theta}^{(0)}\right) \left\{ \sum_{t=1}^{T_r} \left[\gamma_{tr}^{(0)}(i, s, k; W_r) - \gamma_{tr}^{(0)}(i, s, k) \right] o_{rt} \right\} \quad (11)$$

$$\Gamma_{ik}(O^2) = \frac{1}{R} \sum_{r=1}^{R} F'\left(\boldsymbol{\theta}^{(0)}\right) \left\{ \sum_{t=1}^{T_r} \left[\gamma_{tr}^{(0)}(i, s, k; W_r) - \gamma_{tr}^{(0)}(i, s, k) \right] o_{rt}^2 \right\} \quad (12)$$

For detailed algorithm of MPE/MWE, readers may refer to [12] and [13]. In this paper, we use mono-phone HMM and phonetic dictionaries, therefore MPE and MPE are the same.

IV. Database Preparation

A. Brief introduction of PSC

We carry on the experiment on PSC task. There are four parts in the test:

1) Part 1-- Characters reading: about 100 characters.

2) Part 2-- Words reading: about 50 words, mainly di-syllabled words.

3) Part 3-- Paragraph reading: a paragraph of 400 words.

4) Part 4-- Free talk: talk freely for about 3 minutes according to a given topic.

The automatic PSC system gives out scores for first three parts and leaves only the fourth part to human labor.

B. Database and exprimental settings

The database of 3685 people is collected from live PSC all over the mainland and made up of 1-3 national PSC evaluators' scores in a 100-point scale. We divided it into development set (3187 people) and test set (498 people, 2-3 evaluators' scoring) without overlapping. Table 1 shows the experimental settings in detail.

TABLE I.
Detailed Experimental Settings

Item	Settings
Wave	16kHz 16bit
Acoustic Feature	MFCC_0_D_A_Z 39 dimension
Acoustic HMM	66 Mono-phone HMM (including silence and filler), 3-states-initial and 5-states-final; 16 mixtures for each state.
Training Set	Over 100 hours; 30 people with upper first class level(Golden pronunciation)
Development set	Approximately 500 hours, 3187 people; spot PSC data collected over 10 provinces; 1-3 national experts' scoring
Test Set	Approximately 82 hours, 498 people; spot PSC data collected over 10 provinces; 2-3 experts' scoring

V. EXPERIMENTS AND RESULTS

The system structure is shown in Fig.3.

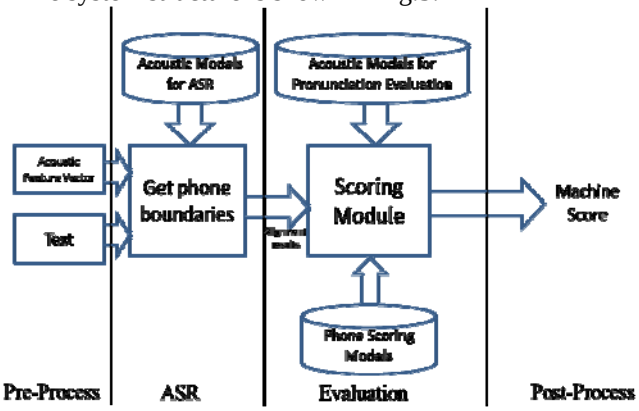

Figure 3. Overview of Automatic PSC Scoring System

It mainly consists four parts:

1) Pre-process part: This part extracts acoutic features from wave files and analyzes their corresponding text. It outputs acoustic features and recognition networks;

2) ASR part: This part aligns input speech with given text and outputs phone boundaries;

3) Evaluation part: This part calculates frame-normalized posterior probabilities as (1) or phone scores as (3) and outputs utterance level pronunciation quality measurements;

4) Post-process part: This part gives out total scores and associated pronunciation ranks [15] for the input speeches.

The acoustic models play an important role in both ASR and evaluation parts. In this way, we divide the function of the acoustic models into the following two parts:

1) ASR Function: Get phone boundaries;

2) Evaluation Function: Get utterance level machine score based on the given phone boundaries.

As the first part belongs to the field of ASR, the following experiments mainly focus on the evaluation function of acoustic models. Therefore if not specified, phone boundaries in the following experiments are the same and calculated by golden models as our previous work [1].

The performance is measured by cross correlation (*CC*) between human and machine scores in (13), where s_r and $\tilde{s}_r$ denote human and machine score for *r*- th utterance.

$$CC = \frac{\sum_{r=1}^{R}\left[\left(s_r - E(s_r)\right)\times\left(\tilde{s}_r - E(\tilde{s}_r)\right)\right]}{\sqrt{\sum_{i=1}^{R}\left(s_r - E(s_r)\right)^2 \times \sum_{i=1}^{R}\left(s_r - E(s_r)\right)^2}} \quad (13)$$

Let us consider the cases when the cross correlations rise from 0.5 to 0.6 and from 0.8 to 0.9. Although both of them increase 0.1 in cross correlation, the performance gain in the latter case is much more significant. Therefore, this paper define relative improvement (*RI*) as (14), where CC_{new} denotes the cross correlation of the improved system and CC_{old} for original one.

$$RI = \frac{CC_{new} - CC_{old}}{CC_{old}} \times 100\% \quad (14)$$

A. *"Seed Model" Selection*

Discriminative training is based on well-trained MLE models. In ASR, it is well-known that the recognition performance will seriously degrade if the training and the test are mismatch. In CALL systems for L2 learners' pronunciation evaluation tasks such as reading [18], retelling [19] and translation [20], acoustic models are trained by both native and non-native pronunciations and have achieved satisfactory performance.

In our case, the training set mismatches with the test and development set because the former is only consist of standard pronunciations while the latter are consisted of various non-standard pronunciations. The development set is not only 3.5 times greater than training set but also well matches with test set. Therefore, it may be desirable to train acoustic models from the combination of both training and development set.

In this way, we shall compare the "multi-trained models" (trained from data in both the training set and the development set) and "golden-models" (trained from data in the training set). We use the latest phone scoring approach as (3) and (4) and the experimental results are shown in Table 2.

TABLE II.
PERFORMANCE OF GOLDEN MLE MODELS AND MULTI MLE MODELS UNDER SAME PHONE BOUNDARIES

Criteria	Item	MLE Models (Baseline)	MPE/MWE Refined Models	RI
Phone Scoring Method	Characters	0.746	0.695	-20.1%
	Words	0.749	0.716	-13.1%
	Paragraph	0.760	0.743	-7.1%
	Average	**---**	**---**	**-13.4%**

Table 2 evidently indicates that the performance of acoustic models would severely degrade if we introduce non-standard data into training.

This also shows us the different objectives between speech recognition and pronunciation evaluation. The former needs to tolerant non-standard speech in order to acquire better recognition results while the latter must distinguish non-standard speeches from standard ones.

Comparing with the cases of non-native speakers' scoring tasks[18]-[20], in which there are many unpredictable mistakes, evaluators would mainly concern whether his/her speech can make others understand (like speech recognition). Therefore, they pay much less attention on confusing pronunciations. On the other hand, the rate of speech plays an important role in L2's scoring task [21] and better ASR helps to gain more accurate speech rate. In this way, the acoustic model that has better ASR (multi-trained acoustic model) performs better in L2 learners' pronunciation evaluation task.

The experiments indicate that only "Golden Models" can be used for native speakers' pronunciation proficiency evaluation task. Therefore, in the following experiments, acoustic models are all developed from the train set. In this way, it may be inevitable to face the mismatch between training and test for native speakers' pronunciation quality evaluation tasks.

B. *Experiments in the Development Set*

In this paper, we use HTK tool kit to implement the MPE/MWE refinement for acoustic models. Phone scoring models are developed from the data of development set. Since we mainly focus on the evaluation performance of acoustic models, phone boundaries re-alignment is not investigated in this section, but can be seen in [25]. The experimental results are similar.

Fig.4 shows the performance of MPE/MWE refined acoustic models each iteration **under same phone boundaries**. The phone scoring models are retrained each iteration to fit the acoustic models.

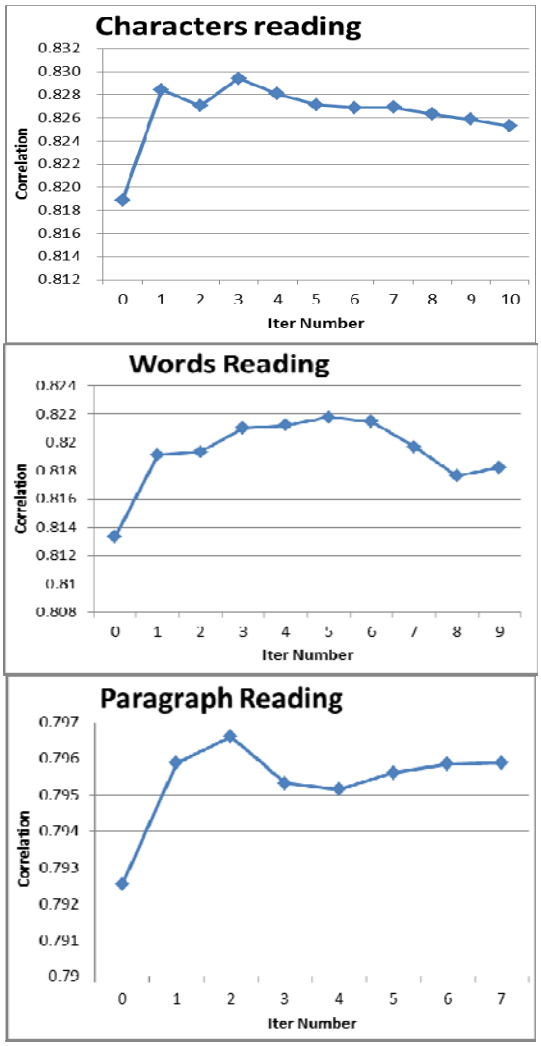

Figure 4. Performance of MPE/MWE training in the development set each iteration; phone score models are retrained each iteration

From Fig.4 we can see that the performance reached its optimum after only a few iterations. We discovered in log files that the auxiliary functions still kept rising. This shows that we need a development set to guarantee the performance rising and avoid over training for the mismatch between training and test.

In order to analyze the confusion degree after discriminative training, we investigated the KLD between some typical error patterns mentioned above. This paper adopts the symmetrical-KLD proposed in [22]. Table 3 shows the symmetrical KLDs between some typical confusing pairs before and after MPE/MWE refinement.

TABLE III.
EXAMPLE OF SYMMETRICAL KLD FOR SOME CONFUSING PAIRS BEFORE AND AFTER MPE/MWE REFINED

Characters Reading	KLDs in MLE Model (Baseline)	KLDs in MPE/MWE Model
ch-c	4.264	6.574
sh-s	8.247	8.846
z-zh	8.068	9.039
in-ing	5.021	4.840
en-eng	5.418	5.638
n-l	6.566	7.634
Average	**6.264**	**7.095**

Word Reading	KLDs in MLE Model (Baseline)	KLDs in MPE/MWE Model
ch-c	5.147	8.001
sh-s	10.416	11.904
z-zh	7.327	8.631
in-ing	4.697	4.413
en-eng	4.872	4.235
n-l	6.998	8.914
Average	**6.576**	**7.683**

Paragraph Reading	KLDs in MLE Model (Baseline)	KLDs in MPE/MWE Model
ch-c	5.624	8.568
sh-s	8.505	9.934
z-zh	5.597	7.015
in-ing	1.580	1.709
en-eng	3.645	3.224
n-l	5.652	4.882
Average	**5.100**	**5.889**

From Table 3 we can see that the average KLDs between confusing pairs rise significantly after MPE/MWE refinement. This result indicates that except for small part of confusing pairs, the refined acoustic models are much more distinguishable.

C. *Experiments in the Test Set*

The acoustic models and phone scoring models are tuned in the development set without seeing any information of the test set. Both traditional measure of frame-normalized posterior probability in (1)(2) and

phone scoring approach in (3)(4) are discussed in this section.

The performance of MLE-trained acoustic models and MPE/MWE refined acoustic models are shown in Table 4.

TABLE IV.
PERFORMANCE OF MLE MODELS AND MPE/MWE REFINED MODELS WITH SAME PHONE BOUNDARIES IN THE TEST SET

Criteria	Item	MLE Models (Baseline)	MPE/MWE Refined Models	RI
Traditional Posterior Probability	Characters	0.570	0.587	3.0%
	Words	0.532	0.575	8.1%
	Paragraph	0.591	0.610	3.2%
	Average	**---**	**---**	**4.8%**
Phone Scoring Method	Characters	0.746	0.764	7.2%
	Words	0.749	0.762	5.2%
	Paragraph	0.760	0.762	1.0%
	Average	**----**	**----**	**4.5%**

Table 4 shows that when phone boundaries are the same (same ASR results), MPE/MWE refined acoustic models perform consistently better than MLE counterpart in both traditional posterior probability measurement and phone scoring approach.

As is mentioned in section 2, MPE and MWE are criteria aiming at improving the performance of ASR. Therefore they may help to get more appropriate phone boundaries. Table 5 shows the experiment of MPE/MWE refined model used both for getting phone boundaries and pronunciation quality evaluation.

TABLE V.
PERFORMANCE OF MLE MODELS AND MPE/MWE REFINED MODELS WITH DIFFERENT PHONE BOUNDARIES IN TEST SET

Criteria	Item	MLE Models (Baseline)	MPE/MWE Refined Models	RI
Traditional Posterior Probability	Characters	0.570	0.560	-1.7%
	Words	0.532	0.536	0.8%
	Paragraph	0.591	0.614	3.9%
	Average	**---**	**---**	**1.0%**
Phone Scoring Method	Characters	0.746	0.752	2.4%
	Words	0.749	0.772	9.5%
	Paragraph	0.760	0.764	1.8%
	Average	**-----**	**-----**	**4.6%**

From Table 5 we can see that the MLE/MPE refined model still outperforms MLE model in scoring task.

But comparing Table 4 and Table 5, we can see that the performance improvements are not stable and the performance seriously degrade in characters reading part. The overall improvement remains as same in phone scoring approach and degrade significantly in traditional posterior probability method.

The experimental result shows no performance improvement when we apply MPE/MWE refined models to ASR which means the ASR-oriented refinement criteria fail to work in the experiment. This may sounds astonishing, but it is precisely the case for the inevitable mismatch between training and test in native speakers' pronunciation quality evaluation tasks.

VI. CONCLUSIONS AND DISCUSSION

This paper discovers that traditional MLE-trained acoustic models are confusable and may not suitable for native speakers' pronunciation proficiency evaluation. Aiming at the problem, this paper introduces MPE/MWE criteria to refine acoustic models. The experiment results show that MPE/MWE refined acoustic models are much more distinguishable and perform consistently better than MLE ones as evaluation models even though the training and test are mismatch.

The experimental results also reveal the controversy in acoustic modeling for native speakers' CALL system: the golden models well match with the "evaluation objective" but mismatch with the "ASR objective" causing ASR-oriented refinement fail to work; on the other hand, the "multi-trained" models well match "ASR objective" but mismatch with "evaluation objective" causing seriously degradation in performance. Therefore, it is desirable to design two separate acoustic models, one is "ASR-oriented acoustic model" and the other is "evaluation oriented acoustic model".

In the future work, we shall use two independent acoustic models in our automatic PSC system and mainly put our effort at improving the "evaluation-oriented acoustic models". As is shown in this paper, the ASR-oriented refinement of MPE/MWE criteria can significantly improve acoustic models' evaluation performance. Therefore refining acoustic models with evaluation oriented objective function must be more effective.

Readers may visit http://www.isay365.com to experience our automatic PSC scoring system (Fig.5).

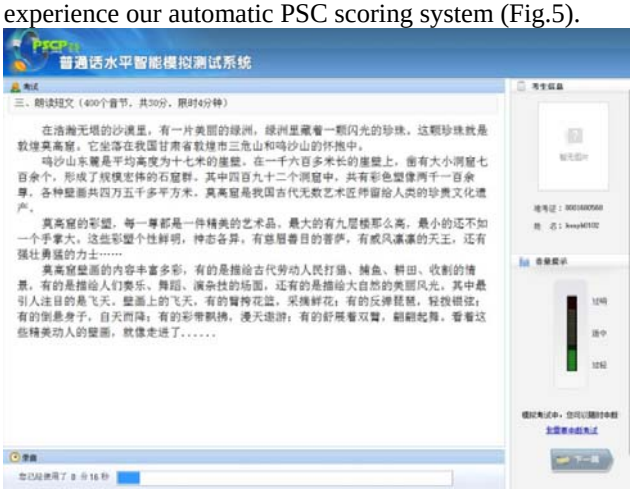

Figure 5. Screen copy of automatic PSC scoring system

VII. Acknowledgment

The authors wish to thank iFLYTEK Research for supporting this work.

References

[1] Si Wei, Yu Hu, Renhua Wang, "The Electronic PSC Testing System", *Journal of Chinese Information Processing*, Vol 20, No.6, Jun 2006, pp.89-96 (in Chinese)

[2] Qingsheng Liu, Si Wei, Yu Hu, Renhua Wang, "The Linguistic Knowledge Based Improvement in Automatic Putonghua Pronunciation Quality Assessment Algorithm", *Journal of Chinese Information Processing*, Vol 21, No.4, July 2007, pp.92-96 (in Chinese)

[3] Si Wei, et al. Putonghua Proficiency Test and Evaluation, *Advances in Chinese Spoken Language Processing*, Chapter 18: Springer Press, 2006

[4] H.L Franco, L.Neumeyer, Y.Kim, O.Ronen. "Automatic pronunciation scoring for language instruction", *ICASSP 1997*, pp 1465-146.8

[5] L. Neumeyer, H. Franco, V. Digalakis, M.Weintraub. "Automatic Scoring of Pronunciation Quality". *Speech Communication 30*, 2000, pp 83-93.

[6] L. Neumeyer, H. Franco, V. Digalakis, M.Weintraub. "Automatic Scoring of Pronunciation Quality". *Speech Communication 30*, 2000, pp 83-93.

[7] C. Cucchiarini, F.D.Wet, H.Strik, L.Boves, "Automatic Evaluation of Dutch Pronunciation by Using Speech Recognition Technology", *ICSLP* Vol.5, 1998, 1739-1742.

[8] S.M Witt, "Use of speech recognition in computer assisted language learning", *A dissertation for doctor's degree of Cambridge*, Nov 1999

[9] S.M Witt, S,J.Young, "Phone-level pronunciation scoring and assessment for interactive language learning", *Speech Communication 30*, 2000, 95-108.

[10] Bahl L R, Brown P F, Souza P V, et al, "Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition". *Proceedings of ICASSP1986*, 1986. 49-52

[11] Valtchev V, Odell J, Woodland P, et al. "Lattice-Based Discriminative Training for Large Vocabulary Speech Recognition". *Proceedings of ICASSP1996*, 1996. Vol2,605-608

[12] Valtchev V, Odell J, Woodland P, et al. "MMIE Training of Large Vocabulary Recognition Systems", *Speech Communication*, 1997. 22(4): 303-314.

[13] D. Provey and P. Woodland, "Minimum Phone Error and I-Smoothing for Improved Discriminative Training", *Proceedings of ICASSP 2002*, pp105-108.

[14] Feng Zhang, "A Research on Automatic Error Detection Based on Statistical Pattern Recognition", *A dissertation for doctor's degree at USTC*, May 2009 (in Chinese)

[15] Xiaojun Qian, Frank Soong, Helen Meng, "Discriminative Acoustic Model for Improving Mispronunciation Detection and Diagnosis in Computer-Aided Pronunciation Training(CAPT)", *Interspeech 2010*, Sep 2010.

[16] Putonghua training and testing center, "the Outline for Putonghua proficiency test and evaluation", *Commercial Press*, 2004 (in Chinese)

[17] Si Wei, "Automatic Error Detection Based on Statistical Pattern Recognition", *A dissertation for doctor's degree of USTC*, Apr. 2008 (in Chinese)

[18] Ke Yan, "Pronunciation Quality Assessment based on Phone Scoring Model", *Journal of Chinese Information Processing, accepted*, (in Chinese)

[19] www.isay365.com

[20] Ke Yan, "Research on Automatic Evaluation of English Recitation and Retelling Test", *A dissertation for master's degree at USTC*, May 23rd. 2008, (in Chinese)

[21] Ke Yan, Guoping Hu, Si Wei, Lirong Dai et al, "Automatic Evaluation of English Retelling Proficiency for Large Scale Machine Examinations of Oral English Test", *Academy Journal of TsingHua Univerisity (Nature Science Edition)*, 2009 S1. pp1356-1362 (in Chinese)

[22] Chiharu Tsurutani, "Foreign Accent Matters Most When Timing is Wrong", *Interspeech 2010*, pp1854-1857

[23] Peng Liu, Frank K. Soong, "Kullback-Leibler Divergence between Two Hidden Markov Models", *Microsoft Research Asia, Speech Group*, unpublished

[24] Ke Yan, "Evaluation Oriented Acoustic Models Training for Computer Assisted Language Learning Systems", *SMSEM 2011*,April, 2011 (in Chinese)

[25] Shu Gong, "the Implementation of Discriminative Training in Pronunciation Proficiency Evaluation based on TANDEM", *A dissertation to master's degree at USTC*, May 2010. (in Chinese)

Ke Yan was born in Chengdu, China in 1984. He is a Ph.D. candidate at USTC (university of science and technology of China) and received his master's degree on speech signal processing in 2009.

His research topics are computer assisted language learning. He did a series of pioneering researches on text-independent speech evaluation field and developed automatic recitation, retelling and translation evaluation systems for Chinese English learners (L2 learners). He also helps to improve the automatic PSC system.

Shu Gong was born in Hefei, China in 1983. He is an engineer working at ZTE (Zhongxing Telecom Equipment) and received his master's degree on speech signal processing in 2010.

During the master study, he mainly research on computer assisted pronunciation quality assessment and automatic error detection. He introduced discriminative training and TANDEM features into automatic PSC evaluation system. He also helped to develop the system for Germans learning Chinese.

