Minimization of Utterance Verification Error Rate as a Constrained Optimization Problem

Man-hung Siu, *Senior Member, IEEE*, Brian Mak⋆, *Member, IEEE*, Wing-hei Au

***Abstract*— Since utterance verification (UV) may be treated as a 2-class classification problem, it may be improved with discriminative training such as minimum verification error training or minimum verification error rate training. However, since in practice, one usually has to pick a specific false-acceptance or false-rejection rate for one's system, it is more desirable to optimize UV performance at a particular operating point. In this paper, we show that further improvement can be achieved by treating UV at a specific operating point as a constrained optimization problem.**

***Index Terms*— utterance verification, minimum verification error, minimum verification error rate**

I. Introduction

For many practical speech applications, it is important to identify and reject out-of-vocabulary words or utterance without keywords by means of utterance verification (UV). The technique has been widely used in applications ranging from weather information retrieval, call routing application, to language learning. In general, UV is treated as hypothesis testing [1], [2], [3] using the (log) likelihood ratio test: the ratio between the null hypothesis that the required word is spoken and the alternative hypothesis that it is not. A decision is made by comparing the ratio against a pre-set threshold. In the past, the two likelihoods are computed from acoustic models which are usually Gaussian mixture models (GMM) or hidden Markov models (HMM), and they are estimated from training data using the maximum-likelihood (ML) approach. However, since the amount of training data is never unlimited in practice, and the true parametric form of these acoustic models is never known, it can be argued that these ML-estimated models may not give the best verification performance. Recently, discriminative training [4] that optimizes model parameters to achieve the minimum verification error (MVE) [3], [5], [6] or minimum verification error rate (MVER) [7] has shown to give better verification performance.

In current UV research, most of the time, the metric used to gauge different approaches is their equal-error-rate (EER) when the false-acceptance rate (FAR) equals the false-rejection rate (FRR). However, in practice, a verification application has to pick a specific operating point to work on. For example, a security check may require an FAR as low as 0.1% while a language learning application may allow an FAR as high as 10%. In [7], when we proposed the use of MVER, we had already proposed optimizing MVER at a specific operating point. Nevertheless, the approach we used in [7] only tries to minimize the total error rate and requires manual adjustment of the threshold to keep the verification system to stay at the specific operating point. In this paper, we cast the problem as a constrained optimization problem and design a cost function that incorporates both MVER as well as the constraint of operating at a specific operating point. Although there exists other algorithms such as the multi-objective evolutionary algorithm [8] that optimize a classification system by ROC analysis, we believe our method is more tractable and faster in the utterance verification context.

Dr. Man-hung Siu is with the Department of Electrical and Electronic Engineering, the Hong Kong University of Science and Technology (HKUST), Clear Water Bay, Hong Kong. E-mail: eemsiu@ee.ust.hk.

Dr. Brian Mak is with the Department of Computer Science, HKUST. Tel: +852 2358-7012; Fax: +852 2358-1477; E-mail: mak@cs.ust.hk.

Wing-hei Au was with the Department of Electrical and Electronic Engineering, HKUST.

This research is partially supported by the Research Grants Council of the Hong Kong SAR under the grant numbers CA02/03.EG04 and CA02/03.EG05. We also thank ASTRI for permitting us to use the ASTRI Kids' Corpus in our experiments.

II. Verification as a Constrained Optimization Problem

A. Log Likelihood Ratio

In utterance verification, a system is presented with a speech segment O_k of duration M_k frames, that is claimed to come from a certain model λ_0. To verify the claim, a model for the alternative hypothesis λ_a is formed and the log likelihood ratio

$$\mathrm{LLR}(O_k, \lambda_0, \lambda_a) = \frac{1}{M_k} \log \left[\frac{p(O_k|\lambda_0)}{p(O_k|\lambda_a)} \right], \qquad (1)$$

between the two models is computed. A decision is made by comparing $\mathrm{LLR}(O_k, \lambda_0, \lambda_a)$ with a threshold τ. That is,

$$\mathrm{LLR}(O_k, \lambda_0, \lambda_a) \underset{rejected}{\overset{accepted}{\gtrless}} \tau \ . \qquad (2)$$

B. Discriminative Training Methods

Let us review two discriminative training methods in speech verification.

1) Minimum-verification-error (MVE) Training: In speech recognition, discriminative training using the minimum classification error (MCE) [4], [9] or the maximum mutual information (MMI) [10] criterion has shown significant improvement over ML training of acoustic model. The idea of MCE training later led to MVE discriminative training [5], [6] of acoustic models for speech or speaker verification. As its name implies, MVE training estimates model parameters by directly minimizing the total verification errors made on the training or a development data set.

Let us denote the true label of the kth segment s_k as δ_k where

$$\delta_k = \begin{cases} +1 & \text{if } s_k \text{ is correct} \\ -1 & \text{if } s_k \text{ is incorrect} \ . \end{cases} \qquad (3)$$

Based on the MCE approach, MVE training minimizes the following soft count of the number of verification errors:

$$e(O_k, \lambda_0, \lambda_a, \tau) = \frac{1}{1 + exp(-\gamma d(O_k, \lambda_0, \lambda_a, \tau))} \quad (4)$$

where $d(O_k, \lambda_0, \lambda_a, \tau) = -\delta_k(LLR(O_k, \lambda_0, \lambda_a) - \tau)$, and γ is the slope of the sigmoid function. Thus, the total count of verification errors over K segments is

$$C_{mve}(O_1^K, \lambda_0, \lambda_a, \tau) = \sum_{k=1}^{K} e(O_k, \lambda_0, \lambda_a, \tau). \quad (5)$$

$C_{mve}(O_1^K, \lambda_0, \lambda_a, \tau)$ can be optimized by adjusting the model parameters via the GPD algorithm [4]. This approach was first proposed in [6] with τ set to zero to simplify the formulation.

2) Minimum-verification-error-rate (MVER) Training: A shortcoming of MVE training is that error rate rather than number of errors is usually the common metric in a verification task. In [7], we proposed the MVER training method that directly minimizes the verification error rates[1]. Let N_c and N_i be the number of correct and incorrect samples in the training data set respectively. Then the empirical false-acceptance rate (FAR) Ω_{FAR} and the the empirical false-rejection rate (FRR) Ω_{FRR} are given by

$$\Omega_{FAR} = \frac{1}{N_i} \sum_{k=1,\ \delta_k=-1}^{K} e(O_k, \lambda_0, \lambda_a, \tau) \ , \quad (6)$$

and

$$\Omega_{FRR} = \frac{1}{N_c} \sum_{k=1,\ \delta_k=+1}^{K} e(O_k, \lambda_0, \lambda_a, \tau) \ . \quad (7)$$

respectively.

Finally, the MVER cost function $C_{mver}(O_1^K, \lambda_0, \lambda_a, \tau)$ is the sum of FAR and FRR[2]:

$$C_{mver}(O_1^K, \lambda_0, \lambda_a, \tau) = \Omega_{FAR} + \Omega_{FRR} \ . \quad (8)$$

C. MVER Constrained Optimization Training (MVER-CO)

Although both MVE and MVER training have made improvement over ML training for UV, they do not address another important aspect of a common verification system: It is often desirable to run a verification system at a specific operating point defined by a preset FAR or FRR. In [7], we addressed the issue by initializing the threshold to its approximate value deduced from the DET curve[3] [11] of the MCE-estimated model. As the value of the threshold at a specific operating point may drift after each iteration of MVER training, one may have to re-adjust the threshold continuously after each iteration. This can be very laborious. In this paper, we try to constrain the drifting effect of the threshold by casting the problem as a constrained optimization problem on the minimum verification error rate and solve it with the common Lagrangian method.

Let us assume, with no loss of generality, that we would like our verification system to run with an FRR equal to ω_{frr}. The cost function for our new MVER-CO training is formulated as

$$C(O_1^K, \lambda_0, \lambda_a, \tau) = \Omega_{FAR} + \beta(\Omega_{FRR} - \omega_{frr})^2$$

where β is the Lagrange multiplier. Due to the non-linear nature of Ω_{FAR} and Ω_{FRR}, the constrained optimization cannot be solved analytically. Instead, it is again solved numerically by the GPD algorithm.

To optimize any system parameter θ using the GPD algorithm, one finds the gradient of $C(O_1^K, \lambda_0, \lambda_a, \tau)$ w.r.t. θ. That is,

$$\begin{aligned} &\tfrac{\partial}{\partial\theta}(C(O_1^K, \lambda_0, \lambda_a, \tau)) = \\ &\quad \tfrac{\partial}{\partial\theta}(\Omega_{FAR}) + 2\beta(\Omega_{FRR} - \omega_{frr})\tfrac{\partial}{\partial\theta}(\Omega_{FRR}) \end{aligned} \quad (9)$$

where

$$\tfrac{\partial}{\partial\theta}(\Omega_{FAR}) = \frac{1}{N_i} \sum_{k=1,\ \delta_k=-1}^{K} \tfrac{\partial}{\partial\theta}(e(O_k, \lambda_0, \lambda_a, \tau)) \ ,$$

$$\tfrac{\partial}{\partial\theta}(\Omega_{FRR}) = \frac{1}{N_c} \sum_{k=1,\ \delta_k=+1}^{K} \tfrac{\partial}{\partial\theta}(e(O_k, \lambda_0, \lambda_a, \tau)) \ ,$$

and

$$\tfrac{\partial}{\partial\theta}(e(O_k, \lambda_0, \lambda_a, \tau)) = \gamma e(\cdot)(1 - e(\cdot))\tfrac{\partial}{\partial\theta}(d(O_k, \lambda_0, \lambda_a, \tau)) \ .$$

If θ is a model parameter of λ_0 or λ_a, then

$$\begin{aligned} &\tfrac{\partial}{\partial\theta}(d(O_k, \lambda_0, \lambda_a, \tau)) = \\ &\quad -\tfrac{\delta_k}{M_k}\left[\tfrac{\partial}{\partial\theta}(\log p(O_k|\lambda_0)) - \tfrac{\partial}{\partial\theta}(\log p(O_k|\lambda_a))\right] \end{aligned}$$

where the gradients of the log likelihoods w.r.t. model means, variances, etc. can be found in [4] and will not be repeated here.

Theoretically speaking, the threshold τ and the Lagrange multiplier β can be learned as well, using the approach as described above. However, one may then have to tune the learning rates for each kind of system parameters. Instead, we chose to determine τ and β empirically.

III. Experiments

Our new MVER constrained optimization training method was evaluated on two phoneme verification experiments[4], and its performance was compared with those of MVE and MVER training. The first experiment makes use of the popular TIMIT database, and the second experiment uses a database created specifically for pronunciation learning and it represents more realistic phonemic errors made by the non-native English-speaking Hong Kong children.

[1]MVER only tries to minimize the total error rates and not the equal error rate (EER) as it is not known how to get an analytical formulation of the latter.

[2]The expression of the MVER cost function in Eqn.(8) in this letter looks different from that in Eqn.(9) of our Eurospeech 2003 paper [7], but they are actually equivalent.

[3]In a DET (Detection Error Tradeoff) curve, the FRR is plotted against the FAR as the verification threshold is varied.

[4]The verification tasks are the same as those in [7] so that one may compare the new algorithm with the old training methods. Notice that the results for some of the old experiments are different from those already reported in [7]. This is because all experiments were re-run after some bug fixes.

TABLE I
NUMBERS OF CORRECTLY AND INCORRECTLY UTTERED PHONEME SEGMENTS IN THE TIMIT AND ASTRI DATA.

Segments	TIMIT Data		ASTRI Data	
	Training	Test	Training	Test
correct	76,923	26,008	35,489	13,352
incorrect	24,605	11,076	8,453	3,182
correct/incorrect	3.126	2.348	4.196	4.198

TABLE II
COMPARISON OF DIFFERENT TRAINING METHODS FOR TIMIT PHONEME VERIFICATION AT THE 5% FRR OPERATING POINT AND EER.

System	FAR at 5% FRR	EER
MCE baseline	87.2%	39.2%
MVE, $\tau = 0$	86.3%	38.0%
MVER, $\tau = 0$	85.9%	37.8%
MVER, τ at EER	85.5%	**36.3%**
MVER, τ at 5% FRR	**84.4%**	36.3%
MVER-CO, τ at 5% FRR	**82.4%**	**35.4%**

A. Task 1: TIMIT Phoneme Verification

1) Corpus and Acoustic Models: The common 39-dimensional MFCC feature vectors were extracted at every 10ms from the speech data in the standard training and test set of the TIMIT corpus [12]. Forty-two acoustic models were trained, representing 40 monophones, silence, and a short pause. They are 3-state straightly left-to-right HMMs with a maximum number of 16 Gaussian components in each HMM state. In addition, a Gaussian mixture model (GMM) consisting of 32 mixture components was trained using all training data as the alternative model. All models were first estimated in the ML approach using the EM algorithm. Phoneme recognition was performed on the test set without using any grammar, and an accuracy of 58.7% was obtained. Using these ML-estimated HMMs as the initial models, MCE training was then performed and the resulting models gave a phoneme recognition accuracy of 59.1%.

2) Experimental Procedure: Phoneme recognition was performed on all speech data using the MCE-estimated models. The resulting phonemic transcriptions were treated as the null hypotheses for further model training and testing for phoneme verification. Table I gives the numbers of correctly and incorrectly uttered phoneme segments found in the TIMIT data sets.

Using the correct and incorrect phoneme segments in the TIMIT training set, the MCE-trained models were further enhanced by the MVE, MVER, and MVER-CO training algorithms. For MVER and MVER-CO training algorithms, we arbitrarily chose 5% FRR as the operating point for optimization. To do that, MVER and MVER-CO were carried out using the threshold τ at 5% FRR deduced from the DET curve of the MCE-trained models. As a sanity check, we also repeated MVER and MVER-CO training using different threshold settings: $\tau = 0$ and τ at the EER operating point. The threshold was not re-adjusted during MVE, MVER, or MVER-CO training iterations[5].

[5]In [7], we had investigated re-adjusting the value of τ after each training iteration. The process is very expensive and there was either no gain or very small gain.

3) Results and Discussion: The results are summarized in Table II. From the results, we have the following observations:

- Both MVE or MVER training using a threshold (τ) value of zero reduce the EER and FAR at the 5% FRR operating point; MVER training performs slightly better.
- When MVER training runs with the approximate threshold at the specific operating point deduced from the DET curve of the MCE-estimated models, significant improvement was obtained. (Compare the results on the 4th and 5th row of Table II.) It shows that it is important to use an appropriate threshold for MVER training.
- The constraint in MVER-CO training seems to have effectively prevented the verification system from drifting away from the specified operating point as it iterates; this is evidenced from the result in the last row of Table II.
- MVER-CO training reduces the FAR at the 5% FRR operating point from MCE model's 87.2% to 82.4%.
- It also turns out that when the EER is reduced by the various training methods, the FAR at the 5% FRR operating point is also reduced as a side effect, and vice versa.

B. Task 2: Pronunciation Verification

1) Corpus and Acoustic Models: The ASTRI Kids' Corpus was used in this experiment. It was collected by the Hong Kong Applied Science and Technology Research Institute (ASTRI)[6] for the development of pronunciation learning system for Hong Kong children. It contains over 42,000 English sentences and 35,000 English words uttered by 410 local and foreign children residing in Hong Kong and studying at Grade 3–5. All utterances were recorded in a quiet environment at the 16 kHz sampling rate. A subset of 8000 phonetically transcribed utterances, containing both isolated words and sentences, were used for training, and 3341 phonetically transcribed isolated words were used to evaluate the verification performance. Table I shows the numbers of correctly and incorrectly uttered phoneme segments in the ASTRI training and test sets; they are all determined by human transcribers.

A set of 41 phonemes was employed in this verification task. Since only words were used for pronunciation verification, position-dependent (PD) phone HMMs were trained. That is, each of the 41 base phonemes has three variations depending on its position in a word — at the beginning, middle, or the end of a word. Thus, theoretically there should be 123 PD phone models, but in reality only 121 of them really appear in the corpus. Feature extraction and acoustic modeling were the same as in the last TIMIT experiment except that each HMM state now has a maximum of 20 Gaussian components. The ML- and MCE-estimated models gave a phoneme recognition accuracy of 53.8% and 54.7% respectively on a separate development set.

2) Experimental Procedure: To verify the phonemes in an uttered word, its pronunciation (phoneme sequence) was looked up from a standard (American) English dictionary and was used as the null hypotheses for verification. Each phoneme was then verified independently. Instead of using a GMM as the alternative model as in the last TIMIT experiment, a phone loop consisting of all word-middle phones was used instead

[6]See http://www.astri.org/ for more information about the institute.

TABLE III
COMPARISON OF DIFFERENT TRAINING METHODS FOR THE ASTRI PRONUNCIATION VERIFICATION TASK AT THE 5% FRR OPERATING POINT AND EER.

System	FAR at 5% FRR	EER
MCE baseline	69.4%	31.3%
MVE, $\tau = 0$	63.4%	28.9%
MVER, $\tau = 0$	61.6%	22.7%
MVER, τ at EER	60.5%	**21.7%**
MVER, τ at 5% FRR	**59.3%**	21.4%
MVER-CO, τ at 5% FRR	**58.9%**	23.3%

as it was found to give better performance. MVE, MVER, and MVER-CO training were performed as in the TIMIT experiments.

3) Results and Discussion: Table III summarizes the verification results using the various training approaches. We have the following observations:

- Although the phoneme accuracy of the ASTRI task is worse than that of the TIMIT task (54.7% vs. 59.1%), the ASTRI verification EER is actually better than that of the TIMIT task (31.3% vs. 39.2%). The ASTRI data contain genuine pronunciation errors made by Hong Kong children. On the other hand, the phonemic errors in the TIMIT data are "artificial" in the sense that they are the results of acoustic confusions made by the speech recognizer, and some of them will not be made by human.
- MVE and MVER training again outperform MCE training as in the TIMIT task, but the performance gain by MVER training is much more in this ASTRI task than in the last TIMIT task. One possible reason is that the priors of correctly and incorrectly uttered segments in the training and test data sets are perfectly matched in the ASTRI data but are mismatched in the TIMIT data: from Table I, their ratios in the training and test set of the ASTRI data are 4.196 and 4.198 respectively, while the figures for the TIMIT data are 3.126 and 2.348 respectively from Table I. The mismatch in the TIMIT data implies that the threshold used in the training process is non-optimal for the test set.
- MVER training with an approximate threshold gives a further 1% drop at the EER operating point, and a further 2.3% drop in FAR at the 5% FRR operating point. (See figures in bold font in Table III.)
- MVER-CO training further reduces the FAR by 0.4% at the 5% FRR operating point when compared with the result of MVER training using an approximate threshold.
- In summary, compared with the MCE-trained models, MVER-CO training reduces FAR by an absolute 10.5% at the 5% FRR operating point; at the EER operating point, MVER training with an approximate threshold reduces the EER by an absolute 9.6%.

IV. CONCLUSIONS

This paper casts the minimization of utterance verification error rate as a constrained optimization problem, and applies standard optimization technique and discriminative training to estimate acoustic models for speech verification. In two phoneme verification tasks, our new training algorithm, the minimum-verification-error-rate constrained optimization training, gives the lowest FAR at the 5% FRR operating point than that achieved by the minimum-verification-error training or the minimum-verification-error-rate training. The 5% FRR operating point is arbitrarily chosen, and we believe the result should apply to other operating points as well. Though we only tested the new training algorithm in speech verification problems, we believe the approach should apply to other verification problems such as speaker verification.

REFERENCES

[1] M. G. Rahim, C. H. Lee, and B. H. Juang, "Discriminative utterance verification for connected digits recognition," *IEEE Trans. on SAP*, vol. 5, no. 3, pp. 266–277, May 1997.
[2] A. R. Setlur, R. A. Sukkar, and J. Jacob, "Correcting recognition errors via discriminative utterance verification," in *Proc. of ICSLP*, Oct 1996, vol. 2.
[3] R. C. Rose, B. H. Juang, and C. H. Lee, "A training procedure for verifying string hypotheses in continuous speech recognition," in *Proc. of ICASSP*, May 1995, vol. 1, pp. 281–284.
[4] W. Chou, "Discriminant-function-based minimum recognition error rate pattern-recognition approach to speech recognition," *Proc. of the IEEE*, vol. 88, no. 8, pp. 1201–1223, 2000.
[5] R. A. Sukkar, A. R. Setlur, M. G. Rahim, and C. H. Lee, "Utterance verification of keyword string using word-based minimum verification error (WB-MVE) training," in *Proc. of ICASSP*, May 1996, vol. 1, pp. 518–521.
[6] R. A. Sukkar, "Subword-based minimum verification error (SB-MVE) training for task independent utterance verification," in *Proc. of ICASSP*, May 1998, vol. 1, pp. 229–232.
[7] Wing-Hei Au and Man-Hung Siu, "A new approach to minimize utterance verification error rate for a specific operating point," in *Proc. of Eurospeech*, Sept 2003, pp. 909–912.
[8] J.E. Fieldsend and R.M. Everson, "ROC optimisation of safety related systems," in *Proc. of ROCAI 2004, part of ECAI 2004*, Valencia, Spain, August 2004, pp. 37–44.
[9] B. H. Juang and S. Katagiri, "Discriminative training for minimum error classification," *IEEE Trans. on SP*, vol. 40, no. 12, pp. 3043–3054, Dec 1992.
[10] P. C. Woodland and D. Povey, "Large scale discriminative training of hidden Markov models for speech recognition," *Journal of CSL*, vol. 16, no. 1, pp. 25–47, Jan 2002.
[11] A. Martin et al., "The DET curve in assessment of detection task performance," in *Proc. of Eurospeech*, 1997, vol. 4, pp. 1895–1898.
[12] V Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, August 1990.