DISCRIMINATIVE LEARNING FOR OPTIMIZING DETECTION PERFORMANCE IN SPOKEN LANGUAGE RECOGNITION

Donglai Zhu[1], *Haizhou Li*[1], *Bin Ma*[1], *Chin-Hui Lee*[2]

[1] Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
[2]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA. 30332 USA
(E-mails: {dzhu,hli,mabin}@i2r.a-star.edu.sg chl@ece.gatech.edu)

ABSTRACT

We propose novel approaches for optimizing the detection performance in spoken language recognition. Two objective functions are designed to directly relate model parameters to two performance metrics of interest, the detection cost function and the area under the detection-error-tradeoff curve, respectively. Both metrics are approximated with differentiable functions of model parameters by using a smoothing function based on a class misclassification measure. The model parameters are optimized by using the generalized probabilistic descent algorithm. We conduct experiments on the NIST 2003 and 2005 Language Recognition Evaluation corpora. Results show that the proposed approaches effectively improve the performance over the maximum likelihood training approach.

Index Terms— detection error tradeoff, detection cost function, discriminative learning, Gaussian mixture model, spoken language recognition

1. INTRODUCTION

A good spoken language recognition (SLR) system is usually optimized to have low miss-detect and false-alarm rates as well as a good tradeoff between the two error types. In NIST Language Recognition Evaluations (LRE) [1], a detection error tradeoff (DET) curve is used to illustrate explicitly the overall performance and the possible error rate tradeoffs between the two error types at each of the operational points [2]. For any point on the DET curve, which is determined by a decision threshold, a detection cost function (DCF) can be obtained to indicate the system performance with respect to the threshold [1].

In the SLR literature, little attention has been given to directly optimizing the tradeoff between the two types of decision errors and the actual operation point. One of the successful approaches for SLR is phonotactic modeling, where a speech utterance is transcribed by phoneme recognizers and the scoring is performed on phoneme strings, e.g., parallel phoneme recognizer followed by either language modeling (PPRLM) [3] or by vector space modeling (PPR-VSM) [4]. In these methods, phoneme recognizers, language models and classifiers are usually trained with common criteria in pattern recognition, e.g., maximum likelihood (ML) and large margin. These criteria don't correlate with DET directly, therefore they may not lead to an optimal DET performance.

Inspired by the research in optimizing discrete error rates through differentiable functions and the receiver operating characteristic curve with the Wilcoxon Mann Whitney (WMW) statistic [5, 6, 7], we directly optimize the language detection performance by minimizing two metrics: the value of DCF and the area under the DET curve. These two discrete metrics are approximated with continuous and differentiable objective functions by using a smoothing function based on a class misclassification measure, which is commonly adopted in the minimum classification error (MCE) framework [5] denoting the degree of separation between the desired class and the competing classes. These two objective functions are optimized in training of Gaussian mixture models (GMMs) in our PPR-VSM system [4]. The GMM parameters are reestimated using the generalized probabilistic descent (GPD) algorithm. Experimental results on the NIST 2003 and 2005 LRE tasks show that the two approaches effectively improve the SLR performance over the conventional ML training approach.

2. OPTIMIZATION OF DETECTION PERFORMANCE

Optimization is performed on the backend GMM language detectors of the PPR-VSM system [4]. A PPR-VSM system adopts a collection of parallel phone recognizers as the frontend that converts an input utterance into multiple phone sequences. With the VSM backend, the n-gram statistics from each phone sequence form a high-dimensional feature vector, and a single composite vector is generated by stacking all the feature vectors. We further design an ensemble of binary SVM classifiers. The outputs of these SVM classifiers constitute a discriminative vector to represent the phonotactic features of a composite vector [8]. For each target language i, we build a language detector consisting of two GMMs with the discriminative vectors: a positive GMM λ_i^+ models the target language, and a negative GMM λ_i^- models its competing languages. The confidence of a test sample x is defined as the likelihood ratio as follows:

$$f_i(x) = \log p(x|\lambda_i^+) - \log p(x|\lambda_i^-) \ . \tag{1}$$

The higher the $f_i(x)$ is, the more confidence the sample x gains. The likelihood ratio in Eq. (1) is used for the final language recognition decision. Parameters of the GMMs are commonly estimated using the ML estimation: $\hat{\lambda}_i^c = \arg\max_{\lambda_i^c} p(\mathcal{X}_i^c|\lambda_i^c)$, where $c \in \{+,-\}$. $\mathcal{X}_i^c$ is the collection of training data for λ_i^c. $\mathcal{X}_i^+$ consists of examples of language i, while $\mathcal{X}_i^-$ consists of examples of competing languages. Since the ML estimation attempts to maximize the likelihood of training data against the models, the resulted GMMs may not yield optimal DET curve in the SLR task. In this paper we study to retrain the GMM parameters to optimize the DET curve.

2.1. Minimizing DCF

The detection cost function (DCF) in the NIST language recognition evaluation plan is defined as follows [1]:

$$\begin{aligned} C_{Det} &= (\sum_{i=1}^{M} C_{Det}(i))/M \ , & (2) \\ C_{Det}(i) &= C_{miss}P_{tgt}P_{miss}(i) \\ &\quad + \frac{1}{N-1}\sum_{j\neq i} C_{fa}(1-P_{tgt})P_{fa}(i|j) \ , & (3) \end{aligned}$$

where C_{miss} and C_{fa} represent the relative costs of a miss and a false alarm, respectively. P_{tgt} is a priori probability that a trial is a target trial. N is the number of possibly appeared languages, which may be larger than the number of target languages M. $P_{miss}(i)$ is the miss probability for the ith target language and $P_{fa}(i|j)$ is the false alarm probability of the jth language's examples incorrectly labeled as the ith target language. For simplicity of the evaluation, both C_{miss} and C_{fa} are set to be 1, and P_{tgt} is set to be 0.5.

Direct optimization of cost in Eq. (2) is difficult because it is not a continuous and differentiable function of the classification parameters. Therefore, it needs to be approximated with a smoothed function. Similar to the MCE approach [5], let us define a class misclassification measure $d_i(x;\Lambda)$ where x is a feature vector and Λ is the set of model parameters. $d_i(x;\Lambda) > 0$ implies a misclassification and $d_i(x;\Lambda) \leq 0$ means a correct decision. It is then embedded in a loss function by using the following sigmoid function: $l_i(x;\Lambda) = 1/[1+\exp(-\gamma d_i(x;\Lambda)+\theta)]$, where γ is a positive constant that controls the size of the learning window and the learning rate, and θ is a constant measuring the offset of $d_i(x;\Lambda)$ from 0. The miss and false alarm probabilities can then be approximated by summing over training samples, as follows:

$$\begin{aligned} P_{fa}(i|j) &= \frac{1}{|\Omega_j|}\sum_{x\in\Omega_j}[1-l_i(x;\Lambda)] \ , & (4) \\ P_{miss}(i) &= \frac{1}{|\Omega_i|}\sum_{x\in\Omega_i} l_i(x;\Lambda) \ , & (5) \end{aligned}$$

where Ω_i is the set of training data belonging to i-th language. As presented previously, for each target language i, two GMMs are respectively used to model the positive data set Ω_i and the negative data set $\bar{\Omega}_i = \{\Omega_j\}|_{j\neq i}$. We define $d_i(x;\Lambda) = -\log p(x|\lambda_i^+) + \log p(x|\lambda_i^-)$, where $p(x|\lambda_i^c) = \sum_m w_{im}^c \mathcal{N}(x;\mu_{im}^c,\Sigma_{im}^c)$, and $c \in \{+,-\}$, m denotes a Gaussian component, $w_{im}^c, \mu_{im}^c, \Sigma_{im}^c$ are respectively the weight, mean vector, and covariance matrix of the m-th component in the c-th GMM for the language i.

With above approximations, the DCF function Eq. (2) is reformed to a smoothed function $L(\mathcal{X};\Lambda)$ where $\mathcal{X}$ means the whole training set. In this paper, we minimize $L(\mathcal{X};\Lambda)$ using the generalized probabilistic descent (GPD) algorithm which iteratively updates parameters in the form of $\Lambda_{t+1} = \Lambda_t - \epsilon_t \nabla L(\mathcal{X};\Lambda)|_{\Lambda=\Lambda_t}$, where t denotes the iteration number, the learning rate ϵ_t needs to satisfy the conditions for the Robbins-Monro theorem [9], and

$$\begin{aligned} \nabla L(\mathcal{X};\Lambda) &= \frac{1}{M}\Bigg[-\frac{C_{fa}(1-P_{tgt})}{(N-1)}\sum_{x\in\Omega_j}\frac{1}{|\Omega_j|}\frac{\partial l_i(x;\Lambda)}{\partial\Lambda} \\ &\quad + C_{miss}P_{tgt}\sum_{x\in\Omega_i}\frac{1}{|\Omega_i|}\frac{\partial l_i(x;\Lambda)}{\partial\Lambda}\Bigg] \ . \end{aligned} \tag{6}$$

In this paper we derive the updating formula for mean vectors in the GMMs. In order to eliminate updating bias among dimensions, the mean vectors are first transformed as: $\tilde{\mu}_{imd}^c = \mu_{imd}^c/\sigma_{imd}^c$ [5]. Then terms to be estimated in Eq. (6) are derived as follows:

$$\begin{aligned} &\frac{\partial l_i(x;\Lambda)}{\partial\tilde{\mu}_{imd}^c} = -\gamma(1-l_i(x;\Lambda))l_i(x;\Lambda)\mathrm{sgn}(c)\cdot \\ &\frac{w_{im}^c\mathcal{N}(x;\mu_{im}^c,\Sigma_{im}^c)}{\sum_{m'} w_{im'}^c\mathcal{N}(x;\mu_{im'}^c,\Sigma_{im'}^c)}\left(\frac{x_d}{\sigma_{imd}^c}-\tilde{\mu}_{imd}^c\right) \ . \end{aligned} \tag{7}$$

2.2. Minimizing area under DET curve

The area under the DET curve can be denoted by the value of the normalized Wilcoxon-Mann-Whitney (WMW) statistic[6]:

$$A_{WMW}(\mathcal{X};\Lambda) = \frac{\sum_{u=1}^{U}\sum_{v=1}^{V} I(f_u,f_v)}{UV} \ , \tag{8}$$

where $\{f_u, u = 1,\ldots,U\}$ are the outputs of the classifier on positive examples, $\{f_v, v = 1,\ldots,V\}$ are the outputs on negative examples, U and V are respectively the numbers of positive and negative examples, and

$$I(f_u,f_v) = \begin{cases} 1 & f_u > f_v \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

The non-differentiable function $I(f_u, f_v)$ needs to be approximated by a sigmoid function for deriving updating formulae. Furthermore, the cost coefficients in Eq. (3) can be embedded into the sigmoid function. In Eq. (3), miss probabilities $P_{miss}(i)$ are multiplied by $C_{miss}P_{tgt}$ while the false alarm probabilities $P_{fa}(i|j)$ multiplied by $C_{fa}(1 - P_{tgt})$. $P_{miss}(i)$ are caused by small values in $\{f_u\}$ that are lower than the threshold. $P_{fa}(i|j)$ are caused by large values in $\{f_v\}$ that are higher than the threshold. Therefore, the sigmoid function is defined as follows: $S(f_u, f_v) = 1/[1 + \exp(-\gamma(c_u f_u - c_v f_v) + \theta)]$, where $c_u = C_{miss}P_{tgt}$, $c_v = C_{fa}(1 - P_{tgt})$. Eq. (8) is defined for a binary decision problem. For our multi-language detection problem, similarly to Eq. (2), Eq. (8) can be rewritten as an average of the WMW statistics over the target languages:

$$A(\mathcal{X};\Lambda) = \frac{1}{M}\sum_{i=1}^{M} A_i(\mathcal{X};\Lambda) \ , \qquad (10)$$

where $A_i(\mathcal{X};\Lambda)$ is the WMW statistic for the ith language and is computed as follows:

$$\begin{aligned} A_i(\mathcal{X};\Lambda) \quad &= \quad \frac{1}{|\Omega_i|}\frac{1}{N-1}\sum_{j\neq i}\frac{1}{|\Omega_j|} \\ &\sum_{u\in|\Omega_i|}\sum_{v\in|\Omega_j|} S(f_i(u), f_i(v)) \ . \qquad (11) \end{aligned}$$

Because the total area in the DET plot is normalized to 1, the area under the DET curve, which is to be minimized, is $L(\mathcal{X};\Lambda) = 1 - A(\mathcal{X};\Lambda)$. Differentiating it with Λ we get:

$$\begin{aligned} \nabla L(\mathcal{X};\Lambda) = -\frac{1}{M}\frac{1}{|\Omega_i|}\frac{1}{N-1}\sum_{j\neq i}\frac{1}{|\Omega_j|} \cdot \\ \sum_{u\in\Omega_i}\sum_{v\in\Omega_j} \gamma S(1-S)(c_u\frac{\partial f_i(u)}{\partial\Lambda} - c_v\frac{\partial f_i(v)}{\partial\Lambda}) \ , \qquad (12) \end{aligned}$$

where $\partial f_i(x)/\partial\Lambda = -\partial d_i(x)/\partial\Lambda$. The same derivation follows as that in Section 2.1.

3. EXPERIMENTS

We conduct SLR experiments on our PPR-VSM system [4]. The backend GMMs were trained on the CallFriend corpus that consists of 12 languages: English, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil, Arabic, Farsi, French, German and Vietnamese [10]. For each speech utterance, a discriminative vector of 110 dimensions was generated [8]. The discriminative vectors are then modeled by GMM classifiers. For each target language, two GMMs were defined: a positive GMM consisting of 32 Gaussian components and a negative GMM consisting of 256 Gaussian components. NIST 1996 LRE data, which consists of a total of 1492 samples from 12 languages [1], were used as development set to provide a threshold for the hard decision in DCF calculation. The NIST 2003 and 2005 LRE data were used as two testing sets [1]. The NIST 2003 data are composed of 1200 utterances from 12 target languages and 80 utterances of an out-of-target language (OOL) – Russian. The NIST 2005 data are composed of 3403 utterances from 7 target languages (English, Hindi, Japanese, Korean, Mandarin, Spanish and Tamil) and 84 German utterances.

The GMMs in the baseline PPR-VSM system were trained with the ML estimation. Then the GMM parameters were re-estimated using the two proposed approaches: minimum of area under the DET curve (minDET) and minimum of the DCF point (minDCF). The parameters were updated in a sample-by-sample iterative procedure, i.e., the training samples were sequentially processed and the GMM parameters were updated after each training sample. To achieve the best convergence, the training procedure can be repeated for several epochs over the whole training set.

The value of the learning rate ϵ was empirically set to decrease in time from an initial value ϵ_0 to 0. The decreasing step size was set to be $\epsilon_0/(E \cdot |\Omega|)$, where $|\Omega|$ is the number of training samples in the training set, and E is the number of epochs and set to 50. ϵ_0 depends on the size of training data and was set to $8.0E4$ according to performance on the development data. In the smoothing function, the parameter γ affects the slope of the function curve and was set to satisfy $E(\gamma \cdot f_i(x)) \propto 1$.

Figure 1 illustrates equal error rates (EERs) of the two approaches after each of 50 epochs. Curves show that the two approaches can quickly reduce EERs and minDCF exhibits slightly better than minDET in the first several epochs (less than 5). The curves display either a fluctuation or descending tendency in around 30 epochs. Afterward both approaches achieved convergence and yielded similar EERs.

In Table 1, we report the error rates in three different categories when system is at overall EER decision point: the miss detect and false alarm rates of in-target languages (IL-Miss and IL-Fa), the false-alarm rate of OOL (OOL-Fa). We observe that minDCF offers a better OOL rejection while minDET achieves better balance between IL-Miss and IL-Fa. Both minDCF and minDET optimization substantially reduced the OOL-Fa rate.

Table 1. Error rate (in %) in 3 categories (IL-Miss/IL-Fa/OOL-Fa) on NIST 03 and 05 LRE tasks.

	NIST 03	NIST 05
Baseline	3.75/3.39/24.48	5.41/5.35/9.86
minDET	2.83/2.85/20.31	4.91/4.42/5.95
minDCF	3.33/2.47/18.75	5.26/4.28/5.10

Table 2 summarizes EERs and DCFs of the baseline and the proposed approaches on NIST 2003 and 2005 LRE tasks.

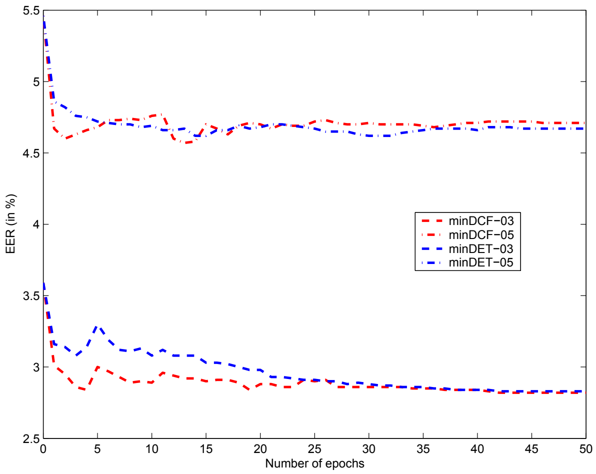

Fig. 1. EERs with increase of epochs in the training processes of the minDET and minDCF approaches on NIST 03 and 05 tasks.

Both approaches yielded similar improvement over the baseline. For minDET, relative EER and DCF improvements were 21.17% and 17.18%, respectively on the NIST 2003 task; they were 14.29% and 12.56% on the NIST 2005 task. Figure 2 illustrates DET curves of the two approaches. These curves show that the proposed approaches effectively moved down the baseline DET curves.

Table 2. EER/DCF (in %) comparison of Baseline, minDET and minDCF on NIST 03 and 05 LRE tasks.

EER/DCF	NIST 03	NIST 05
Baseline	3.59/3.55	5.46/6.69
minDET	2.83/2.94	4.68/5.85
minDCF	2.82/3.00	4.71/5.65

4. CONCLUSIONS

In this paper we proposed approaches for integrating performance metrics, the detection cost function and the area under the DET curve, into the model training. This strategy is attractive because it offers a way to directly optimize the language detection performance with evaluation measures of interest. The two objective functions are optimized in training of backend GMMs in our PPR-VSM system. The GMM parameters are embedded into the objective functions by using smooth approximations of the discrete metrics and reestimated with the GPD algorithm. Experimental results on NIST 2003 and 2005 LRE tasks show that the two approaches effectively improve the detection performance over the ML training approach and the optimization of the two metrics achieves competitive results. Ongoing and future works include 1) a comparison with optimization of other performance metrics, 2) applying proposed approaches to other classifiers, and 3) a study of simultaneous optimization of different performance metrics.

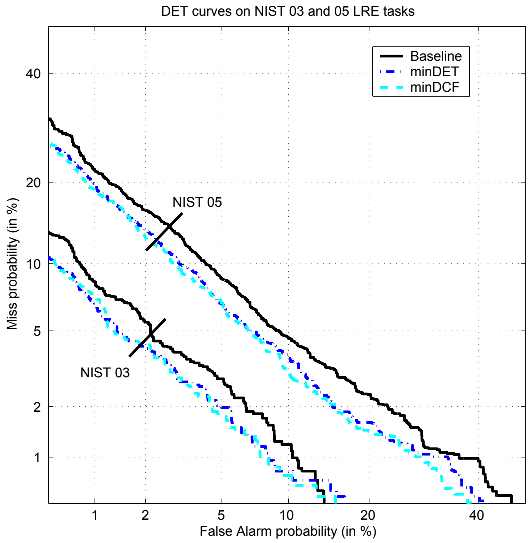

Fig. 2. DET curves of Baseline and two proposed approaches (minDET and minDCF) on the NIST 03 and 05 LRE tasks.

5. REFERENCES

[1] “The NIST language recognition evaluation plan,” *http://www.nist.gov /speech/tests/lang/index.htm*.

[2] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, “The DET curve in assessment of detection task performance,” *Proc. Eurospeech*, vol. 4, pp. 1895–1898, 1997.

[3] M. A. Zissman, “Comparison of four approaches to automatic language identification of telephone speech,” *IEEE Trans. Speech and Audio Processing*, vol. 4, no. 1, pp. 31–44, 1996.

[4] H. Li, B. Ma, and C.-H. Lee, “A vector space modeling approach to spoken language identification,” *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 271–284, 2007.

[5] B.-H. Juang, W. Chou, and C.-H. Lee, “Minimum classification error rate methods for speech recognition,” *IEEE Trans. Speech and Audio Processing*, vol. 5, no. 3, pp. 257–265, 1997.

[6] L. Yan, R. Dodier, M. C. Mozer, and R. Wolniewicz, “Optimizing classifier performance via an approximation to the wilcoxon-mann-whitney statistic,” *Proc. ICML*, 2003.

[7] S. Gao, W. Wu, C.-H. Lee, and T.-S. Chua, “A maximal figure-of-merit (MFoM)-learning approach to robust classifier design for text categorization,” *ACM Trans. on Information Systems*, vol. 24, no. 2, pp. 190–218, 2006.

[8] B. Ma, R. Tong, and H. Li, “Discriminative vector for spoken language recognition,” in *Proc. ICASSP*, 2007, pp. 15–20.

[9] C. M. Bishop, *Neural networks for pattern recognition*, Clarendon Press, 1995.

[10] “Callfriend corpus, telephone speech of 15 different languages or dialects,” *http://www.ldc.upenn.edu/Catalog/byType.jsp#speech. telephone*.