A FRAMEWORK FOR UNSUPERVISED TRANSFER LEARNING AND APPLICATION TO DIALOG DECISION CLASSIFICATION

Etienne Marcheret, Om D Deshmukh, Vaibhava Goel, Jiří Navrátil

IBM T.J. Watson Research Center
Yorktown Heights, NY 10598, USA
{etiennem,vgoel,jiri}@us.ibm.com, odeshmuk@in.ibm.com

ABSTRACT

We propose a framework for transfer learning in the unsupervised condition, and show its usefulness in addressing the problem of mismatch in test time dialog state decision classifier, which is presented here as a binary hypothesis problem. We are asked to either accept or reject the ASR output. The framework encompasses a two step process, the first step culminates in the discriminative retraining of the test time classifier using the results of an EM solution to the joint optimization between the original labelled training data and observed unlabelled test data for enhanced test time discrimination of the binary classes. The second step is optimization of the performance of this classifier in a specific operating range. This extends previous results in Bayes error reshaping to the unsupervised condition which favor a particular false alarm operating range. We show a total relative reduction in error rate of up to 15%, 12.5% from the first step, with an additional 2.5% from step 2 along with the added knowledge of the threshold needed to operate at a specific false alarm operating range.

Index Terms— Unsupervised Transfer Learning, Covariate Shift, Bayes Error Reshaping, MCE, ASR Confidence.

1. INTRODUCTION

When applying the results of supervised machine learning to the classification or regression problems the assumption that training and test points are drawn from the same distribution is often violated in practice. In [1] it was shown that under certain restrictive assumptions about how the training and test distribution varied this shift could be compensated in the training domain objective function. This result known as covariate shift correction and the closely related sample selection bias correction [2], resulted in numerous variations and applications in the machine learning literature [5], [11], [3], [4]. In this paper we extend previous work in two directions, a relaxation of the constraints implied by the covariate shift assumption to the classification problem, and Bayes error reshaping for detection error tradeoff [6], [7] under the unsupervised condition. The subsequent framework is then applied to the problem of ASR dialog decision.

In transactional ASR dialog systems, the decision to accept or reject (or reconfirm) a user utterance is often carried out by a decision function that takes as input a set of features extracted during the recognition process. The decision function may be implemented by a classifier discriminatively trained with data collected from a mixture of dialog states, with each state generally having multiple grammars or language models (LMs) operating in parallel. The features extracted during the recognition process have a dependence on the grammars (or LMs), and the state of acoustic model adaptation making it difficult to sufficiently cover all conditions with a labelled training corpus. This often places the decision classifier in a mismatched condition. The mismatch is not only evident through the usual loss in class discriminative power but in the expected miss (M), false accept (FA) ROC curve operating point.

The paper is organized as follows. In section 2 the underlying modeling framework is introduced, section 3 discusses an EM solution. Section 3.3 discusses the unsupervised detection error tradeoff. Experimental setup, results and conclusions are presented in sections 4, 5 and 6.

2. MODELING FRAMEWORK

Let $\mathbf{x}$ be the feature vector and y the class label, in this discussion a binary label. From n i.i.d. training samples $((\mathbf{x}_i, y_i) : i = 1, ..., N)$ we may learn the the model parameters θ used to model the probability $p(y|\mathbf{x}, \theta)$. In practice the assumption that the same underlying distribution generates both the training and test instances is often violated. This leads to sub-optimal class discrimination at test time, which can be reflected in a drop of equal error rate performance and unknown classifier operating range, defined here as the probability of false accept (P_{FA}), probability of miss (P_M) point on the detection error curve. For the binary hypothesis test this operating point is determined by the log likelihood ratio test and threshold t:

$$\log \frac{p(H_2|\mathbf{x})}{p(H_1|\mathbf{x})} \overset{H_2}{>} t, \tag{1}$$

Where the binary labels y are represented by H_1 and H_2. The threshold t defines the decision regions Z_1 and Z_2 from which we may determine the false accept and miss probabilities:

$$P_{FA}=\int_{Z_2} p(\mathbf{x}|H_1)d\mathbf{x}, \quad P_M=\int_{Z_1} p(\mathbf{x}|H_2)d\mathbf{x}. \tag{2}$$

To learn the parameters of the classifier θ we minimize the expected loss on the training data for some loss function $l(f(\mathbf{x}), y)$. To address the test time condition where the underlying distribution generating the features is not the same as that generating the training time features we have the covariate shift result [1]:

$$E_T\left[l(f(\mathbf{x})), y)\right] = E_D\left[\frac{p(\mathbf{x}|\theta_T)}{p(\mathbf{x}|\theta_D)}\, l(f(\mathbf{x}), y)\right], \tag{3}$$

where T denotes the test set and D the training set. This result follows from application of Bayes rule, $p(\mathbf{x}, y|\theta_D) = p(\mathbf{x}|\theta_D)p(y|\mathbf{x}, \theta_D)$

and the assumption that $p(y|\mathbf{x},\theta_T) = p(y|\mathbf{x},\theta_D) = p(y|\mathbf{x})$, we have for $w_{\mathbf{x}} = p(\mathbf{x}|\theta_T)/p(\mathbf{x}|\theta_D)$ and $l(f(\mathbf{x}),y) = L_{\mathbf{x},y}$:

$$E_D\left[w_{\mathbf{x}}L_{\mathbf{x},y}\right] = \sum_{\mathbf{x},y}\frac{p(\mathbf{x}|\theta_T)}{p(\mathbf{x}|\theta_D)}p(\mathbf{x}|\theta_D)p(y|\mathbf{x},\theta_D)L_{\mathbf{x},y} = E_T\left[L_{\mathbf{x},y}\right] \tag{4}$$

With this assumption minimization of the weighted loss function on the training data will result in minimization of the loss function on the test data. In practice the assumption that $p(y|\mathbf{x})$ remains constant from training to testing environments is too strict. It is also noted that directly modeling $p(y|x)$ would seem to circumvent this problem entirely, (as opposed to the generative approach which attempts to model $p(y|\mathbf{x})$ with $p(\mathbf{x}|y)$ and Bayes rule). Directly modeling $p(y|\mathbf{x})$ may not fully avoid the shift in the underlying feature distribution when there is a misspecification of the underlying model [2].

To avoid the strict assumption needed to satisfy the covariate shift solution, we assume the distribution $p(\mathbf{x},y)$ depends on some unobserved variable S, to get at the distribution of the observable variables we marginalize over S, $p(\mathbf{x},y) = \int p(\mathbf{x},y|S)P(S)dS$ [5]. We assume here that S captures the differences between the training and testing environments, if $p_D(S) = p_T(S)$, then $p_D(\mathbf{x},y) = p_T(\mathbf{x},y)$. Specifically here we may assume S captures the mix of language models and state of the acoustic model at a specific dialog state. To put this into practice we make the assumption that the training feature space is generated by two sources, one of those sources is responsible for both the generation of the training and test time feature spaces, the other source is responsible for features that are unique to the training feature space. Associated with each of these sources is a classifier, $p(y|\mathbf{x},S_i)$, $i = 1,2$. From these definitions the training distribution can be written:

$$p_D(\mathbf{x},y) = \sum_S p(\mathbf{x},y,S) = \sum_S p(y|\mathbf{x},S)p(\mathbf{x}|S)p(S) \tag{5}$$

We assume the underlying feature distribution is generated by a Gaussian mixture model. Based on the definition of the two sources we have the model for the training distribution:

$$p_D(\mathbf{x}) = p(s_1)\sum_{k=1}^{M_1}\alpha_{1k}^D p_{1k}(\mathbf{x}) + p(s_2)\sum_{k=1}^{M_2}\alpha_{2k}^D p_{2k}(\mathbf{x}). \tag{6}$$

Where this distribution is generated from $p(s_1)$ and $p(s_2)$ proportions of source sets 1 and 2 respectively, and test distribution:

$$p_T(\mathbf{x}) = \sum_{k=1}^{M_1}\alpha_{1k}^T p_{1k}(\mathbf{x}). \tag{7}$$

From equation 7 we note that the test distribution shares the parameterization of the Gaussian components associated with training distribution source set 1, but differs in the mixture weights $p_1^T(k)$. In practice this definition can be achieved by building separate GMMs on the training and test data and merging them to form the source set 1 GMM, the mixture weights are then re-estimated to form the corresponding P_D and P_T distributions. The source set 2 GMM would be derived from the training data, and would be composed of Gaussian components with maximum KL distance to the test data GMM. More details about this step will be presented in section 4.2. Equations 5, 6, and 7 define the modeling framework, we now present an EM solution for the parameter learning.

3. EM SOLUTION

With the description of the previous section we may write the full generative model of the observed data. For training data points $u \in D$, and test data points $\nu \in T$, given by the joint distribution: $P((\mathbf{x}^u,y^u|u\in D),(\mathbf{x}^\nu|\nu\in T)|\Theta,\mathbf{w})$, where Θ represents the parameters associated with the underlying feature generation GMMs, the component mixture weights and source proportions and $\mathbf{w}$ are the parameters of the function modeling $p(y^u|\mathbf{x}^u)$. We have the log likelihood of the observed training and test data:

$$\begin{aligned}\log P((\mathbf{x}^u,y^u|u\in D),(\mathbf{x}^\nu|\nu\in T)\,|\,\Theta,\mathbf{w}) = \\ \sum_{u\in D}\log\left(\sum_j p(s_j)\sum_{k=1}^{M_j}\alpha_{jk}^D p_{jk}(\mathbf{x}^u|\theta_{jk})p_j(y^u|\mathbf{x}^u,\mathbf{w}_j)\right) + \\ \sum_{\nu\in T}\log\sum_{k=1}^{M_1}\alpha_{1k}^T p_{1k}(\mathbf{x}^\nu|\theta_{1k}),\end{aligned} \tag{8}$$

where the unknown parameters $(\Theta,\mathbf{w})$ can be estimated with the use of the EM algorithm. The unobserved variables are the source and corresponding mixture responsible for the generation of the training features and the mixture responsible for test feature generation. The E-step for iteration t, results in computation of the posterior probablility for a given training sample $\mathbf{x}^u$, $P(s,k|\mathbf{x}^u,y^u,\Theta^t,\mathbf{w}^t)$. We have:

$$\begin{aligned}\gamma_{sk}^u &= \frac{p(\mathbf{x}^u,y^u|s,k,\Theta^t,\mathbf{w}^t)p(k|s)p(s)}{\sum_{\hat{s},\hat{k}}p(\mathbf{x}^u,y^u|\hat{s},\hat{k},\Theta^t,\mathbf{w}^t)p(\hat{k}|\hat{s})p(\hat{s})} \\ &= \frac{p(\mathbf{x}^u|s,k,\Theta^t)p(y^u|\mathbf{x}^u,\mathbf{w}_s^t)\alpha_{sk}^D p_s}{\sum_{\hat{s},\hat{k}}p(\mathbf{x}^u|\hat{s},\hat{k},\Theta^t)p(y^u|\mathbf{x}^u,\mathbf{w}_{\hat{s}}^t)\alpha_{\hat{s}\hat{k}}^D p_{\hat{s}}}\end{aligned} \tag{9}$$

The posteriors for the hidden variables associated with the test vectors $\mathbf{x}^\nu$ follow a similar equation. Note the absence of labels and source set 1 responsibility for the generation of the test vectors:

$$\gamma_{1k}^\nu = P(s_1,k|\mathbf{x}^\nu,\Theta_t) = \frac{p(\mathbf{x}^\nu|s_1,k,\Theta^t)\alpha_{1k}^T}{\sum_{\hat{k}}p(\mathbf{x}^\nu|s_1,\hat{k},\Theta^t)\alpha_{1\hat{k}}^T}, \tag{10}$$

The maximization step results in updates:

- Source set weights and source GMM weights:

$$p_s{=}\frac{1}{N^D}\sum_{i=1}^{N^D}\sum_{k=1}^{M_s}\gamma_{sk}^{u_i},\ \alpha_{sk}^D{=}\frac{1}{p_s N^D}\sum_{i=1}^{N^D}\gamma_{sk}^{u_i},\ \alpha_{1k}^T{=}\frac{1}{N^T}\sum_{i=1}^{N^T}\gamma_{1k}^{\nu_i}. \tag{11}$$

- Source model Gaussian parameters mean, and covariance (μ,Σ), with $T_A = \sum_{i=1}^{N^D}\gamma_{1k}^{u_i} + \sum_{i=1}^{N^T}\gamma_{1k}^{\nu_i}$, and $\xi_{j,k}^u = (\mathbf{x}^u - \mu_{j,k})$:

$$\begin{aligned}\mu_{1,k} &= \frac{\sum_{i=1}^{N^D}\gamma_{1k}^{u_i}\mathbf{x}^{u_i} + \sum_{i=1}^{N^T}\gamma_{1k}^{\nu_i}\mathbf{x}^{\nu_i}}{T_A} \\ \Sigma_{1,k} &= \frac{\sum_{i=1}^{N^D}\gamma_{1k}^{u_i}\xi_{1,k}^{u_i}(\xi_{1,k}^{u_i})^T + \sum_{i=1}^{N^T}\gamma_{1k}^{\nu_i}\xi_{1,k}^{\nu_i}(\xi_{1,k}^{\nu_i})^T}{T_A} \\ \mu_{2,k} &= \frac{\sum_{i=1}^{N^D}\gamma_{2k}^{u_i}\mathbf{x}^{u_i}}{\sum_{i=1}^{N^D}\gamma_{2k}^{u_i}},\ \Sigma_{2,k} = \frac{\sum_{i=1}^{N^D}\gamma_{2k}^{u_i}\xi_{2,k}^{u_i}(\xi_{2,k}^{u_i})^T}{\sum_{i=1}^{N^D}\gamma_{2k}^{u_i}}\end{aligned} \tag{12}$$

- Source classifier updates: This is dependent on the form of the underlying classifier, the fundamental result is an optimization of a weighted loss function which is analogous to equation 4 and can be seen as a relaxation of the assumption $p(y|\mathbf{x},\theta_T) = p(y|\mathbf{x},\theta_D) = p(y|\mathbf{x})$ through the posteriors γ_{sk}^u. We have the source set 1 and 2 classifier optimization

$$\frac{\partial}{\partial\mathbf{w}_s}\sum_{i=1}^{N^D}\left[\sum_{k=1}^{M_s}\gamma_{sk}^u\right]\log(p(y_i^u|\mathbf{x}_i^u,\mathbf{w}_s)). \tag{13}$$

3.1. Classifier Modeling: Weighted MCE

The model used for source classifiers $p(y|\mathbf{x})$ will be based on a GMM. The smoothed 0/1 loss function for a binary classifier when the correct class label is H_c, competing class (incorrect) H_{ic}, is given by

$$L_c(\mathbf{x}) = \frac{1}{1+\exp(\eta(\log\frac{p(\mathbf{x}|H_c)}{p(\mathbf{x}|H_{ic})}))} \tag{14}$$

For equal class priors $p(y)$, we have $p(y|\mathbf{x}) = p(\mathbf{x}|y)/\sum_{y'} p(\mathbf{x}|y')$, and for the binary class case $\max\log p(H_i|\mathbf{x}) = \max\log p(\mathbf{x}|H_i) - \min\log p(\mathbf{x}|H_j)$, therefore minimizing the loss of equation 14 is a proxy for maximizing the likelihood of the source classifiers given by equation 13. Optimization of the weighted empirical loss function results in the source S1 and S2 classifier parameter updates. For training data $((\mathbf{x}_i, y_i) : i = 1, ..., N)$, binary classes $y_i \in (H_1, H_2)$ and weights (eq. 9) we have:

$$\mathcal{L}_S = \sum_i \sum_j \left[\sum_{k=1}^{M_s} \gamma_{sk}^{u_i}\right] L_j(\mathbf{x}_i) I(\mathbf{x}_i \in H_j), \tag{15}$$

where $I()$ is the indicator function. Weighted gradients of the loss function for the correct and incorrect class for sample $\mathbf{x}$ follow directly:

$$\begin{aligned}\frac{\partial L_c(\mathbf{x})}{\partial \mathbf{w}_c} &= -\left[\sum_{k=1}^{M_s} \gamma_{sk}^{u}\right] \eta L_c(1-L_c)\frac{\partial \log p(\mathbf{x}|H_c)}{\partial \mathbf{w}_c} \\ \frac{\partial L_c(\mathbf{x})}{\partial \mathbf{w}_{ic}} &= \left[\sum_{k=1}^{M_s} \gamma_{sk}^{u}\right] \eta L_c(1-L_c)\frac{\partial \log p(\mathbf{x}|H_{ic})}{\partial \mathbf{w}_{ic}}\end{aligned} \tag{16}$$

Where it is understood that the classifier parameters of Gaussian means, variances and mixture priors are represented by the parameter vector $\mathbf{w}_c$ and $\mathbf{w}_{ic}$ for the correct and incorrect class, implicitly tied to the class labels H_c, H_{ic}. In practice a fraction of the maximum likelihood statistics are added to equation 15 to prevent overfitting.

3.2. Test Time Usage

The ultimate goal of sections 3 and 3.1 is to estimate the classifier to be used on the test set: $p(y|\mathbf{x}, \mathbf{w}_{s1})$, as this model is associated with source set 1 and is intended to capture the underlying distribution shared between training and test environments.

3.3. Unsupervised P_{FA}, P_M Trade-off

For the binary classifier plotting the false alarm vs. probability of miss (2) on a standard deviate scale, yields the detection error trade-off (DET) curve. In [6] the idea of feature space DET analysis criterion (fDETAC) was proposed. This result shows that we may estimate a feature space transform to rotate the DET curve to enhance the P_M performance in one P_{FA} region at the expense of another. This follows from assuming Gaussian score distributions for P_{FA}, P_M which results in a linear relationship. Denoting class i score mean and standard deviation by μ_i and σ_i we have the relationship for normal inverse error function Φ^{-1}:

$$\Phi^{-1}(P_M) = -\frac{\sigma_1}{\sigma_2}\Phi^{-1}(P_{FA}) + \frac{\mu_1 - \mu_2}{\sigma_2}. \tag{17}$$

By estimating a feature space transform A we may rotate and shift this DET curve. This optimization results in fDETAC, given by:

$$A^* = \underset{A}{\operatorname{argmin}} \left[\frac{\sigma_1(A)}{\sigma_2(A)}\right] w_R + \left[\frac{\mu_1(A) - \mu_2(A)}{\sigma_2(A)}\right] w_D, \tag{18}$$

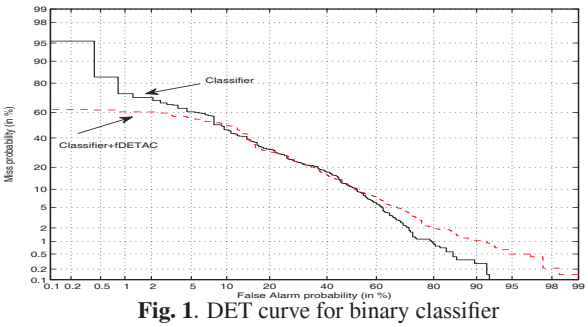

Fig. 1. DET curve for binary classifier

Where w_r and w_d can be used to regulate the slope and delta optimizations. Figure 1 illustrates a DET curve for a classifier before and after fDETAC optimization. Note that we have pivoted approximately around 10% P_{FA}, reducing the P_M from the mid 70% range to the upper 60% range at the expense of an elevated P_M in the $P_{FA} > 60\%$ range.

We extend this result to perform unsupervised optimization of equation 18, where we rely on the original training labels but a reweighting of the scores with the posteriors from equation 9. The weighted mean score for class label i is given by

$$\mu_i = \frac{1}{N}\sum_{\mathbf{x}_d \in i}\left[\sum_{k=1}^{M_{s_1}} \gamma_{s_1 k}^{\mathbf{x}_d}\right] \log\frac{\sum_k p(k|H_2)p(\mathbf{x}_d|k, H_2)}{\sum_k p(k|H_1)p(\mathbf{x}_d|k, H_1)} \tag{19}$$

Where the posteriors are specific to the underlying source set 1 models (section 3.2).

The gradient of score mean and standard deviation results in terms of the form for hypothesis j:

$$\sum_k p(k|\mathbf{x}_d, H_j)\Sigma_k^{-1}(A\mathbf{x}_d - \mu_k)\mathbf{x}_d^T \tag{20}$$

where we note that the Jacobian of the transform cancels. Solution for the transform A can be arrived at through a line search.

4. EXPERIMENTAL SETUP

The ASR system is our standard telephony enterprise installation, this is MFCC based with 250k Gaussians, 20k context dependent states. The training setup includes 48k sentences covering 11 dialog states composed of grammars varying in vocabulary size from 10 to 20k words, and perplexities in the range of 10 to 90k. The in grammar average semantic error rate is 5.76% and average out of vocabulary rate (OOV) of 12.89%. The test dialog state includes 5.5k sentences, a vocabulary size of 100, perplexity of 399k and in grammar semantic error rate of 6.0% with OOV rate of 9.14%. To estimate the initial underlying source 1 GMM, 1650 unlabelled test dialog samples were used along with the training dialog samples (discussed in section 4.2). The subsequent EM procedure (3) used these same 1650 unlabelled samples and the 48k labelled training set. fDETAC (3.3) was performed with the 48k labelled training instances and the posteriors γ_{sk}^u (9) from the EM step. The remaining 3850 samples (from the 5.5k test dialog) were used in test.

4.1. Confidence Feature Space

The confidence feature space is 11 dimensional, composed of observations extracted from the ASR process. There has been extensive studies of features that are effective, including lattice based [8], background model based [9], and heuristic search and acoustic

based [10]. The mutual information between features and the binary classes of accept/reject varies between 0.01 and 0.2 bits for the 48k utterance training set.

4.2. EM Initialization and Model Sizes

The underlying source 1 and 2 classifiers $p(y|\mathbf{x}, \mathbf{w}_{s,1,2})$ are initialized to the same globally trained MCE model. Careful initialization of the underlying source generating GMMs (eqs. 6, 7) is more critical. As discussed in section 2, source set 1 is responsible for modeling what is common between the training and testing feature spaces, source set 2 is responsible for the feature space specific to training. With that definition a reasonable technique for initialization of the source set 1 GMM parameters is based on a KL metric between a GMM built on the training data, and a GMM built on the test data. Minimum KL distance components form source set 1 model parameters, maximum KL distance form source set 2 model parameters. Mixture weights are then re-estimated with a standard EM step keeping Gaussian parameters fixed and updating training data mixture weights (α_{1k}^D), (α_{2k}^D) and test data mixture weights (α_{1k}^T).

Source set 1 and 2 models (eqs. 6,7) are 4 mixture diagonal GMMs. The underlying source set classifiers are 2 mixture diagonal GMMs (as is the initializing global classifier).

5. EXPERIMENTS

Table 1 shows the component maximum symmetric KL distance between the unlabelled test data GMM G_T (as described in (4.2)), and the source set 1 GMM after the learning $G_{S1,EM}$ (3). The value of the KL distance doesn't tell us much, but the large (minimum) KL between source set 1 and 2 $G_{S1,EM}||G_{S2,EM}$ after learning relative to $G_T||G_{S1,EM}$ illustrates a mismatch in some portion of the feature space. To measure the test time classifier performance $p(y|\mathbf{x}, \mathbf{w}_{s1})$,

KL dist.	g_i	g_j	g_l	g_m	Avg.
$G_T\|\|G_{S1,EM}$	4.17	4.03	2.32	1.21	2.93
$G_{S1,EM}\|\|G_{S2,EM}$	12.71	16.76	21.48	42.86	23.45

Table 1. Maximum KL distance between components of the test data GMM G_T and learned source set 1 GMM $G_{S1,EM}$. Also shown is Minimum KL between components of source set 1 and 2 GMMs.

(section 3.2) we look at two metrics: the average probability of miss for false alarm probabilites of 2, 3 and 4% and the area under the ROC curve (AUC). The AUC reflects the classifiers intrinsic ability to discriminate between the two classes [7] a value of 1.0 and the classes are separable, 0.5 and the classification is random. In practice the only metric that matters is the low false alarm miss probability, but the classifiers intrinsic ability to separate is important here as we expect the DET rotation to be able to take advantage of this. Figure 2 illustrates the S1 and S2 classifier low FA performance as a function of EM iteration. The degradation of the S2 classifier illustrates the mismatch. From this figure we see the S1 classifier results in a 12.5% relative reduction in miss probability, and including fDETAC we have 14.98% total relative reduction. Figure 3 shows the AUC classifier performance as a function of EM iteration, the S2 classifier mismatch performance is evident. Treating the area above the ROC curve as that contributing to the overall error we see S1 classifier training by the EM step results in an 11.2% relative reduction in error, and including fDETAC we have 15.54% total relative reduction. As a fair comparison on the number of parameters (S1+S2), a 4 mixture GMM global classifer $p(y|\mathbf{x})$ has a low FA result=0.7534 and

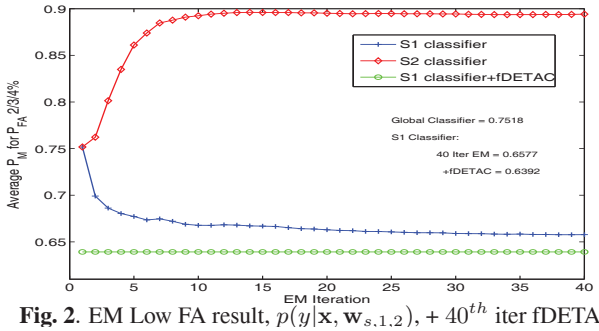

Fig. 2. EM Low FA result, $p(y|\mathbf{x}, \mathbf{w}_{s,1,2})$, + 40^{th} iter fDETAC.

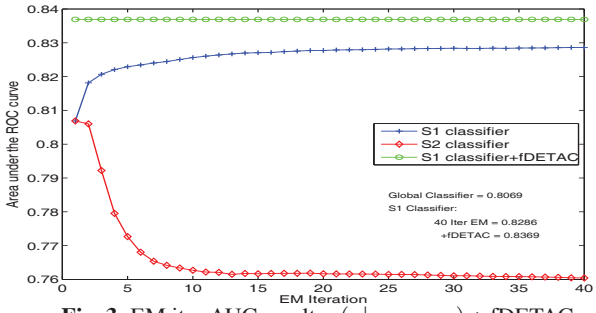

Fig. 3. EM iter AUC result, $p(y|\mathbf{x}, \mathbf{w}_{s,1,2})$ + fDETAC.

AUC=0.8038, so adding parameters doesn't help (small degradation vs. 2 mixture global could be overfitting with MCE).

6. CONCLUSIONS

We have presented a framework for unsupervised retraining of a classifier along with a technique to optimize this classifier in a desired FA region. We feel these are extensions to the mixture regression covariate shift framework [5] and Bayes error reshaping in the unsupervised condition [6]. To our knowledge, this is the first time such an approach is applied to an ASR task.

7. REFERENCES

[1] H. Shimodaira. *Improving predictive inference under covariate shift by weighting the log-likelihood function*. Journal of Statistical Planning and Inference, 90:227-244, 2000.

[2] B. Zadrozny. *Learning and evaluating classifiers under sample selection bias*. Proceedings ICML, 2004.

[3] C. Cortes, M. Mohri, M. Riley, A. Rostamizadeh. *Sample Selection Bias Correction Theory*. ALT 2008.

[4] M. Sugiyama, et al. *A Density-Ratio Framework for Statistical Data Processing*. Trans. on Computer Vision and Applications. Sept. 2009.

[5] A. Storkey, M. Sugiyama. *Mixture regression for covariate shift* Advances in Neural Information Processing Systems, 19, (NIPS-2006).

[6] J. Navratil, G. Ramaswamy. *DETAC: A Discriminative Criterion For Speaker Verification*. ICSLP-2002.

[7] Michael Mozer et. al. *Prodding the ROC Curve: Constrained Optimization of Classifier Performance*. NIPS-2001.

[8] B. Maison, R. Gopinath. *Robust Confidence Annotation and Rejection for Continuous Speech Recognition*. ICASSP 2001.

[9] M. Andorno, P. Laface, R. Gemello. *Experiments in Confidence Scoring for Word and Sentence Verification*. ICSLP-2002.

[10] C. White, J. Droppo, A. Acero, J. Odell. *Maximum Entropy Confidence Estimation for Speech Recognition*. ICASSP 2007.

[11] S. Bickel, et al. *Discriminative Learning under Covariate Shift*. Journal of Machine Learning Research, 10(2009) 2137-2155.