COMPARISON OF DISCRIMINATIVE TRAINING METHODS FOR SPEAKER VERIFICATION

Chengyuan Ma and Eric Chang

Microsoft Research Asia
{i-chenma, echang}@microsoft.com

ABSTRACT

The *maximum likelihood estimation* (MLE) and Bayesian *maximum a-posteriori* (MAP) adaptation methods for *Gaussian mixture models* (GMM) have proven to be effective and efficient for speaker verification, even though each speaker model is trained using only his own training utterances. Discriminative criteria aim at increasing discriminability by using out-of-class data. In this paper, we consider the speaker verification task using three discriminative training methods to compare performance. Comparisons are discussed for the *maximum mutual information* (MMI), *minimum classification error* (MCE) and *figure of merit* (FOM) criteria. Experiments on the 1996 NIST speaker recognition evaluation data set show that FOM training method outperforms the other two methods for speaker verification in terms of system performance. Meanwhile, logistic regression is investigated and successfully employed as a discriminative score-normalization technique.

1. INTRODUCTION

The speaker verification task is essentially a hypothesis testing problem or a binary classification problem. Both target-speaker model and impostor model are necessary for decision making. The GMM has been widely used as a probabilistic model in most state-of-the-art speaker verification systems [1]. Each speaker is characterized by a GMM. As a generalized *probabilistic density function* (pdf), the parameters of a GMM are estimated using several techniques, which lead to different system performance. Traditionally, a GMM can be estimated using EM algorithm under MLE criterion aiming at maximizing the likelihood for all observations. In order to reduce computation and to improve performance when only a limited number of training utterances are available, some adaptation techniques were proposed, in which MAP adaptation outperforms the other two, *maximum likelihood linear regression* (MLLR) adaptation and the *eigen-voices* method [1] [2]. However, larger likelihood doesn't necessarily mean better discrimination and better system performance. So several discriminative training methods were proposed to increase discriminability, including the MCE [3] [4], MMI [5] and FOM training strategies [6]. Our goal of this paper is to explore which method has the best performance.

For an objective comparison of different training methods, we need performance measures to evaluate each system. There are four common measures including the *receiver operating characteristics* (ROC) curve, *detection error tradeoff* (DET) curve, *equal error rate* (EER) and *detection cost function* (DCF). ROC and DET represent the overall system performance for all possible operating points (threshold for decision making), while EER and DCF indicate the system performance at a specified operating point.

In the next section, MLE training method will be briefly discussed. In Section 3.1 and Section 3.2, MCE and MMI criteria will be presented. The FOM training method will be discussed in Section 3.3. Logistic regression as a discriminative score-normalization technique will be presented in Section 4. We will describe the NIST 1996 SRE data set used in our experiments and front-end processing in Section 5.1. Before drawing conclusions in Section 6, we present experiments results in Section 5.2.

2. MLE TRAINING OF GMM

Given a GMM model Λ with M mixtures and diagonal covariance matrix, and a sequence of feature frames, $\boldsymbol{O} = \{\boldsymbol{o}_1, \cdots, \boldsymbol{o}_T\}$, when we make such an assumption: all feature frames from the same utterance are independent, the probability (likelihood) of the observation $\boldsymbol{O}$ can be obtained as follow:

$$p(\boldsymbol{O}|\Lambda) = \prod_{t=1}^{T} \sum_{m=1}^{M} w_m \cdot \mathcal{N}(\boldsymbol{o}_t; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \quad (1)$$

Here, Λ is also used to represent the model parameters. w_m is the weight of Gaussian mixture $\mathcal{N}(\boldsymbol{o}_t; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$ with mean $\boldsymbol{\mu}_m$ and covariance matrix $\boldsymbol{\Sigma}_m$.

Using the EM algorithm, we can get a local optimum $\hat{\Lambda}$ for Λ under the MLE criterion by maximizing the probabilistic density for all observation frames.

$$\hat{\Lambda} = \arg\max p(\boldsymbol{O}|\Lambda) \quad (2)$$

Both the *universal background model* (UBM) and the target speaker model can be obtained by MLE training.

3. DISCRIMINATIVE TRAINING METHODS

MLE training and MAP adaptation only use training utterances from target speaker himself. Discriminative training methods incorporate other speakers' training utterances into target-speaker model to increase speaker separability.

At first, we will discuss some common issues for all discriminative methods.

Given a certain criterion, the model training is a process of parameters optimization. We use *generalized probability descent* (GPD) to optimize the model parameters. For each training utterance $\boldsymbol{O}$, the gradient for the input data and current model parameters is used to update model parameters.

Sometimes there are only a limited number of training utterances for each target speaker. We need some methods to "re-sample" from the training corpus. We do this by drawing vector groups from each training utterance. A vector group is a series of consecutive feature vectors derived from a training utterance. The starting point of each vector group is randomly selected within the training utterance duration. For each target model, we can get sufficient vector groups from both target training utterances and impostor training utterances by this way.

In the following discussion, $\boldsymbol{\Lambda} = (\Lambda_{\rm tar}, \Lambda_{\rm imp})$ stands for parameters of both target model $\Lambda_{\rm tar}$ and the corresponding impostor model $\Lambda_{\rm imp}$. $\boldsymbol{O}_{\rm pos}$ is the vector group set from the target speaker, and $\boldsymbol{O}_{\rm neg}$ from the impostors.

3.1. MCE Training

The *minimum classification error* (MCE) criterion [3] [4] aims at minimizing the approximation of error rate on the training data ($\boldsymbol{O}_{\rm pos}$, $\boldsymbol{O}_{\rm neg}$). The idea is to define a misclassification measure that is continuous with respect to the model parameters to minimize. There are many possible misclassification measure definitions. The following definition is the most used one and will be used in our experiments.

$$d(\boldsymbol{O}|\boldsymbol{\Lambda}) = \log(p(\boldsymbol{O}|\Lambda_{\rm tar})) - \log(p(\boldsymbol{O}|\Lambda_{\rm imp})) \quad (3)$$

$$\mathcal{E}(\boldsymbol{\Lambda}) = \sum_{\boldsymbol{O}\in\boldsymbol{O}_{\rm pos}} \frac{1}{1+\exp\left(-\gamma\cdot(-d(\boldsymbol{O}|\boldsymbol{\Lambda}))\right)} + \sum_{\boldsymbol{O}\in\boldsymbol{O}_{\rm neg}} \frac{1}{1+\exp\left(-\gamma\cdot d(\boldsymbol{O}|\boldsymbol{\Lambda})\right)} \quad (4)$$

$$\hat{\boldsymbol{\Lambda}} = \arg\min \mathcal{E}(\boldsymbol{\Lambda}) \quad (5)$$

$$\hat{\boldsymbol{\Lambda}}^{k+1} = \hat{\boldsymbol{\Lambda}}^{k} - \eta \frac{\partial \mathcal{E}(\boldsymbol{\Lambda})}{\partial \boldsymbol{\Lambda}}\Big|_{\boldsymbol{\Lambda}=\hat{\boldsymbol{\Lambda}}^k} \quad (6)$$

$d(\boldsymbol{O}|\boldsymbol{\Lambda})$ is the misclassification measure for $\boldsymbol{O} \in \boldsymbol{O}_{\rm neg}$, while $-d(\boldsymbol{O}|\boldsymbol{\Lambda})$ is for $\boldsymbol{O} \in \boldsymbol{O}_{\rm pos}$. γ ($\gamma > 0$) is GPD parameter and η is the step of GPD algorithm.

3.2. MMI Training

The *maximum mutual information* (MMI) training [5] aims at maximizing the mutual information between the observations ($\boldsymbol{O}_{\rm pos}, \boldsymbol{O}_{\rm neg}$) and the corresponding class labels ($\Lambda_{\rm tar}$, $\Lambda_{\rm imp}$). The MMI criterion can be defined by the sum over the logarithms of the posteriori probabilities of each observation $\boldsymbol{O} \in (\boldsymbol{O}_{\rm pos}, \boldsymbol{O}_{\rm neg})$.

$$p(\boldsymbol{O}) = p(\boldsymbol{O}|\Lambda_{\rm tar})\cdot P(\Lambda_{\rm tar}) + p(\boldsymbol{O}|\Lambda_{\rm imp})\cdot P(\Lambda_{\rm imp}) \quad (7)$$

$$\mathcal{F}(\boldsymbol{\Lambda}) = \sum_{\boldsymbol{O}\in\boldsymbol{O}_{\rm pos}} \log\frac{p(\boldsymbol{O}|\Lambda_{\rm tar})\cdot P(\Lambda_{\rm tar})}{p(\boldsymbol{O})} + \sum_{\boldsymbol{O}\in\boldsymbol{O}_{\rm neg}} \log\frac{p(\boldsymbol{O}|\Lambda_{\rm imp})\cdot P(\Lambda_{\rm imp})}{p(\boldsymbol{O})} \quad (8)$$

$$\hat{\boldsymbol{\Lambda}} = \arg\max \mathcal{F}(\boldsymbol{\Lambda}) \quad (9)$$

$$\hat{\boldsymbol{\Lambda}}^{k+1} = \hat{\boldsymbol{\Lambda}}^{k} + \eta \frac{\partial \mathcal{F}}{\partial \boldsymbol{\Lambda}}\Big|_{\boldsymbol{\Lambda}=\hat{\boldsymbol{\Lambda}}^k} \quad (10)$$

$P(\Lambda_{\rm tar})$ and $P(\Lambda_{\rm imp})$ are the prior probabilities. They are assumed to be equal, $P(\Lambda_{\rm tar}) = P(\Lambda_{\rm imp}) = \frac{1}{2}$. The maximization of the MMI criterion tries to simultaneously maximize the class-conditional probabilities of the observation and to minimize a weighted sum over the class-conditional probabilities of all competing classes. Thus, the MMI criterion optimizes the class separability.

3.3. FOM Training

The *figure of merit* (FOM) training for speaker verification is a method proposed recently[6]. Figure 1 shows the definition of FOM. FOM is the area of the shadow region. As previously described, the ROC curve is one of the system-performance indicators. The closer the ROC curve to the axes, the better the system performance. So the larger the FOM, the better the system performance. FOM training directly maximizes the FOM by adjusting model parameters. FOM training is a flexible training algorithm. We can specify a *false acceptance* (FA) and *false rejection* (FR) range to calculate FOM so that we can optimize system performance only in this range. Generally speaking, FOM training is a nice technique for binary classification problems. $\frac{\partial \rm FOM}{\partial \boldsymbol{\Lambda}}$ is the FOM gradient. The model parameters are updated iteratively:

$$\hat{\boldsymbol{\Lambda}}^{k+1} = \hat{\boldsymbol{\Lambda}}^{k} + \eta \frac{\partial \rm FOM}{\partial d(\boldsymbol{O}|\boldsymbol{\Lambda})}\cdot\frac{\partial d(\boldsymbol{O}|\boldsymbol{\Lambda})}{\partial \boldsymbol{\Lambda}}\Big|_{\boldsymbol{\Lambda}=\hat{\boldsymbol{\Lambda}}^k} \quad (11)$$

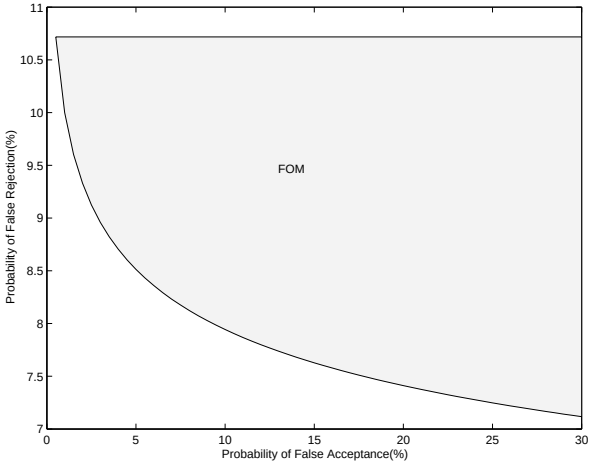

Fig. 1. ROC curve and the definition of FOM

4. SCORE NORMALIZATION

Score-normalization techniques are widely used in speaker verification systems to perform channel and handset compensation. The most frequently used score-normalization techniques are T-Norm and Z-Norm [7]. These two score-normalization methods lead to better system performance but they need additional speech data or external speakers to be computed. In our experiments, logistic regression is used. It doesn't need any extra training data set. The logistic regression model is a discriminative statistic model, in which training observations $\boldsymbol{O}$ from both $\boldsymbol{O}_{\rm pos}$ and $\boldsymbol{O}_{\rm neg}$ are used to get logistic-regression model parameters. As a generalized discriminative normalization technique, logistic regression can be combined with any other speaker-model training methods. The logistic regression model is

$$p(y=\pm 1|\mathbf{x},\mathbf{w}) \quad = \quad \frac{1}{1+\exp\left(-y\mathbf{w}^T\mathbf{x}\right)} \qquad (12)$$

Given the training data set $[(\mathbf{x}_1,y_1),\cdots,(\mathbf{x}_N,y_N)$, we want to find the parameter $\mathbf{w}$ that maximize the log-likelihood:

$$l(\mathbf{w}) \quad = \quad -\sum_{n=1}^{N}\log\left(1+\exp\left(-y_n\mathbf{w}^T\mathbf{x}_n\right)\right) \qquad (13)$$

Here, $\mathbf{x}_n = (p(\boldsymbol{O}_n|\Lambda_{\rm tar}), p(\boldsymbol{O}_n|\Lambda_{\rm imp}))$, and if $\boldsymbol{O}_n \in \boldsymbol{O}_{\rm pos}, y_n = 1$, otherwise, $y_n = -1$. The parameter $\mathbf{w}$ can be estimated under the MLE criterion using training observations $\boldsymbol{O}_{\rm pos}$ and $\boldsymbol{O}_{\rm neg}$.

5. EXPERIMENTS

5.1. Experiment Data Set and Feature Extraction

All experiments are conducted on the 1996 NIST speaker recognition evaluation data set to be consistent with our previous work [6]. Only male speakers are investigated. There are 21 male target speakers and 204 male impostors in the evaluation data set. The training utterances for each target speaker are extracted from two sessions originating from two different handsets, one minute per session. As for the testing, there are 321 target trials and 1060 impostor trials. The duration of each test utterance is about 30 seconds.

MFCCs are used as acoustic features for speaker verification. All utterances are pre-emphasized with a factor of 0.97. A Hamming window with 32ms window length and 16ms window shift is used for each frame. Each feature frame consists of 10 MFCC coefficients and 10 delta MFCC coefficients. Finally, the *relative spectral* (RASTA) filter and *cepstral mean subtraction* (CMS) are used to remove linear channel convolutional effects on the cepstral features.

Table 1. DCF, EER and relative reduction for different training methods

	DCF	rel. red.	EER	rel. red.
MLE(baseline)	0.0457		0.0869	
MMI	0.0452	1.1%	0.0841	3.2%
MCE	0.0428	6.3%	0.0808	7.0%
FOM	0.0384	16.0%	0.0712	18.1%
MLE+LR	0.0435	4.8%	0.0744	14.4%
FOM+LR	0.0412	9.8%	0.0623	28.3%

5.2. Experiment Results

The system performance is evaluated using the DET curve, DCF and EER. DCF can be defined as follow [8],

$$DCF \quad = \quad C_{\rm fr}\cdot p_{\rm fr}\cdot P_{\rm tar} + C_{\rm fa}\cdot p_{\rm fa}\cdot P_{\rm imp} \qquad (14)$$

Here, $p_{\rm fr}$ and $p_{\rm fa}$ are false rejection rate and false acceptance rate respectively at a operating point. $C_{\rm fr}$ and $C_{\rm fa}$ are costs for false rejection and false acceptance. $P_{\rm tar}$ and $P_{\rm imp}$ are the prior probability of target trials and impostor trials. $P_{\rm tar} = 0.01$ and $P_{\rm imp} = 0.99$.

EER and DCF for MLE, MMI, MCE, FOM, MLE with logistic regression (MLE + LR), FOM with logistic regression (FOM + LR) are shown in Table 1. In Figure 2, DET curves for MLE training, MCE training, MMI traing and FOM training are shown. In Figure 3, DET curves for MLE training, MLE training with logistic regression, FOM training, FOM training with logistic regression are shown. (The lines are more visible in color pictures).

From the results, we know that FOM training has the best system performance in terms of EER, DCF and DET curves and logistic regression combined with other training methods can bring benefits as well.

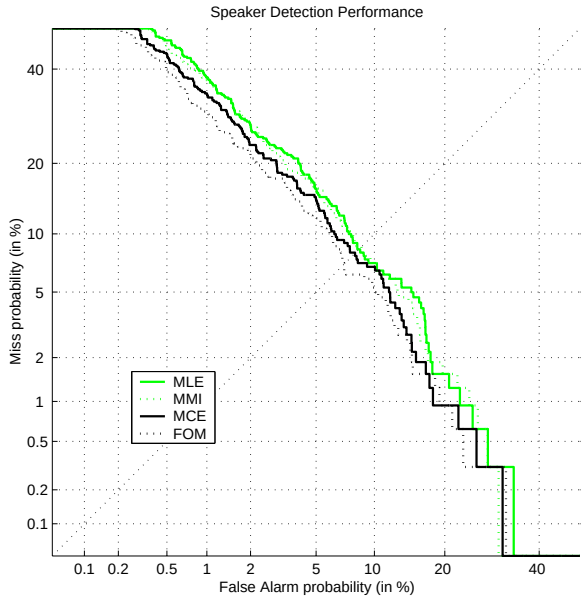

Fig. 2. DET curves for MLE, MMI, MCE and FOM training methods

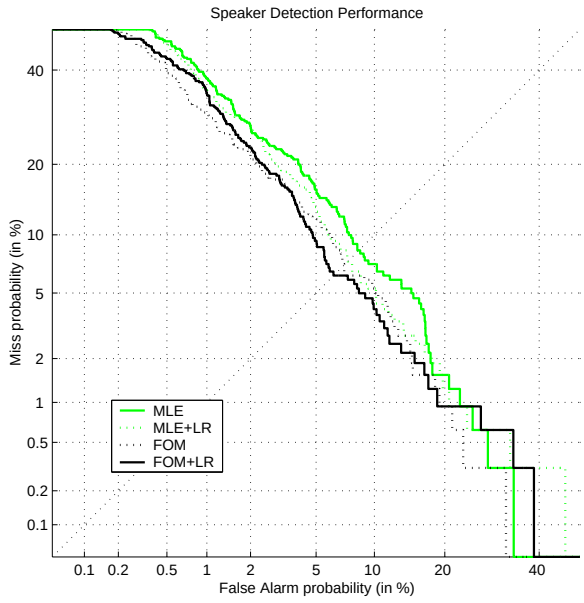

Fig. 3. DET curves for MLE, FOM training methods and logistic regression normalization

6. CONCLUSION

Three discriminative training methods for speaker verification were investigated on the NIST 1996 SRE data set. We derive two main conclusions from our experiments results.

First, FOM training outperforms the other two traditional discriminative training algorithms in terms of system performance.

Second, logistic regression is an effective discriminative score-normalization technique. It can be combined with other model training methods successfully.

Comparing with MLE training, discriminative training methods are computation-consuming. The use of other optimization schemes rather than GPD can be one of the directions for future research.

Efficiently incorporating discriminative training with adaptation methods is also an interesting future research direction.

7. ACKNOWLEDGEMENT

The authors thank Jianlai Zhou, Frank Seide, and other members of the Speech Group at Microsoft Research Asia for many fruitful discussions.

8. REFERENCES

[1] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[2] J. Mariethoz and S. Bengio, "A comparative study of adaptation methods for speaker verification," in *Proc. of ICSLP*, 2002, pp. 581–584.

[3] A. E. Rosenberg, O. Siohan, and S. Parthasarathy, "Speaker verification using minimum verification error training," in *Proc. of ICASSP*, 1998, pp. 105–108.

[4] O. Siohan, A. E. Rosenberg, and S. Parthasarathy, "Speaker identification using minimum classification error training," in *Proc. of ICASSP*, 1998, pp. 109–112.

[5] R. Schlüter and W. Macherey, "Comparison of discriminative training criteria," in *Proc. of ICASSP*, 1998, pp. 493–496.

[6] X.-H. Li, E. Chang, and B.-Q. Dai, "Improving speaker verification with figure of merit training," in *Proc. of ICASSP*, 2002, pp. 693–696.

[7] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification system," *Digital Signal Processing*, vol. 10, pp. 42–54, 2000.

[8] NIST, "the 1996 speaker recognition evaluation plan," http://www.nist.gov/speech/tests/spk/.