OPTIMIZATION OF THE DET CURVE IN SPEAKER VERIFICATION UNDER NOISY CONDITIONS

Leibny Paola Garcia Perera
Tecnologico de Monterrey
Computer Science Department
Monterrey Nuevo Leon, Mexico

Bhiksha Raj
Carnegie Mellon University
Language Technology Institute
Pittsburgh, PA, USA

Juan Arturo Nolazco Flores
Tecnologico de Monterrey
Computer Science Department
Monterrey Nuevo Leon, Mexico

ABSTRACT

The increasing need for secure authentication systems has motivated recent interest in effective algorithms for Speaker Verification (SV). In particular, there is increasing need for *noise robust* algorithms for SV, which will allow SV systems to operate successfully in real conditions, which are typically noisy.

Speaker verification addresses a pattern classification problem, in which there is a tradeoff between false acceptance and false rejections. Traditional approaches optimize the parameters of a classifier for a single operating point emobidied by the proportions of positive and negative examples in the training data, or by learning the parameters without considering the tradeoff.

In a real situation where noise is present, the operating point is effectively unknown and may not match training conditions. We believe that for such situations the optimization of the parameters should not be limited to a single operating point, and that a more robust strategy is to optimize the parameters for all operating points by minimizing the area under the detection error tradeoff curve. In this paper we investigate the minimization of the area under the detection error curve in noisy conditions. Experiments performed on the database NIST2008 show our method improves the performance with respect to conventional methods.

Index Terms— Speaker verification, robustness, minimum verification error, joint factor analysis, detection error tradeoff.

1. INTRODUCTION

Speaker verification (SV) systems perform a voice-based biometric authentication task. Given a recording of a spoken utterance and a claimed identity for the speaker, they must verify that the utterance was indeed spoken by the claimed speaker.

The state-of-the art in SV systems is quite advanced, and extremely good performance can be obtained on clean speech recordings. However, when the speech recordings have been corrupted by noise, performance degrades remarkably even for state-of-art algorithms, as documented by Ferrer et al. [1], who demonstrate (on a new noisy database that they introduce for the research community) that the performance obtained with an I-vector front-end followed by PLDA falls with decreasing SNR. A variety of approaches have consequently been proposed in the literature to deal with the problem of noise in speech-classification tasks including SV, such as denoising the signal [2], reducing the noise in speech feature vectors [3, 4, 5, 6, 7], or simply projecting out noise factors from the data [8]. Other methods, such as the algorithm presented by [9], attempt to ameliorate the effect of noise by utilizing noise-robust feature representations, obtained for instance by replacing DFT spectral estimation with temporally weighted linear prediction (LP).

But what about the basic pattern classification paradigm used for verification itself? How does *that* affect the robustness of verification systems to noise? We address this problem in this paper. Specifically, we demonstrate that through a modification of the training paradigm to consider all possible operating points, the classifier itself can be made more robust to noise, sometimes in a surprising way. To explain, let us consider errors in SV systems. SV systems can commit two kinds of errors: false acceptance (FA), *i.e.* erronous acceptance of imposters, and false rejection (FR), i.e. erroneous rejection of valid speakers. These can be traded off for one another. Each combination of FA and FR represents an *operating point*. On a plot of FA vs. FR, the continuum of possible operating points can be represented as a curve, commonly known as the *detection error tradeoff* or DET curve. Ideally, the system would be capable of perfect performance, with zero FA and FR. Thus the DET curve would pass through the origin and lie directly *on* the FA/FR axes. In practice, the curve usually lies somewhat away from the axes as illustrated by the examples in Figure 1. Nonetheless, the closer the DET curve is to the axes, the more accurate the system will be.

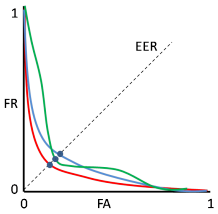

Fig. 1. DET curves. The better classifier has the lower curve that is closer to the axes. The "Equal Error Rate" is the operating point shown by the intersection of the dotted diagonal line and the curve.

Speaker verification is usually formulated as a likelihood-ratio test [10], computed using the estimated distribution of data from the target speaker and that from impostors. Traditional learning algorithms for SV systems such as maximum likelihood (ML) estimation [10, 11] and factor analysis [12, 13, 14, 15, 16] learn the parameters of these distributions to "fit" the training data. Discriminative training paradigms attempt instead to optimize classification performance. Eventually, the performance of SV systems is evaluated in terms of how far the DET curve is from the axes, as characterized by the equal error rate (EER) – the operating point at which the false acceptance and false rejection probabilities are equal. Intuitively, the lower the EER is, the closer the DET curve is to the axes.

However, this introduces a dichotomy between the evaluation and training procedures. Firstly, we note that the EER is not always a good proxy for the distance of the DET curve from the axes; the green curve in Figure 1 is clearly inferior to the blue one, yet has a better EER. A better measure of the overall performance of the system is the area under the curve (AUC) which measures the area between the axes and the DET curve. Secondly, the training procedures themselves do not actually consider the DET curve, nor even the actual operating point at which the SV system is evaluated. Rather, they consider the specific operating point exemplified by the proportion of "positive" (from the target speaker) and "negative" (not from the target speaker) training instances presented to them, which is just a single operating point that may represent neither the EER nor the overall DET curve. In prior work we, and others, have shown [17, 18, 19] that training methods that explicitly optimize the DET curve can result in improved verification performance over conventional training procedures.

We now revert to our original problem of dealing with noise. Noise increases the inherent variability of the signal, potentially shifting it from any operating point a classifier may be optimized for, possibly in an unpredictable manner. In this paper we demonstrate that an operating-point-agnostic training paradigm that optimizes the entire DET curve can result in classifiers that are significantly more robust to noise than conventional classifiers *even when the effect of noise is not explicitly considered*. We present a training formalism that optimizes the entire DET curve by minimizing the AUC. To obtain an analytical solution, we use the well known Wilcoxon Mann Whitney (WMW) statistic [20] as a proxy to the AUC The parameters of the distribution are estimated using a (GPD) generalized gradient descent algorithm. The AUC-minimization training paradigm can be applied to most current learning methods for SV systems including discriminative Minimum Verification Error training, Joint-Factor Analysis (JFA) [12, 13, 14, 15, 16] and its various extensions and modifications.

Models trained in this manner are seen to be more robust to noise than models trained by conventional maximum likelihood or discriminative training. Moreover, we also observe a surprising inversion of behavior in training paradigms. In conventional systems, discriminative training schemes such as minimum-verification error training [21], although more effective than simple maximum-likelihood training, are found to be less effective than factor analyzed models [12, 13, 16], particularly when the latter too are discriminatively trained. However, models are trained to minimize the AUC, an inversion happens on noisy data – the performance of MVE-trained classifiers surpasses that of factor-analyzed classifiers!

The rest of the paper is as follows: in Section 2 we outline conventional training methods for speaker verification systems, in Section 3 we describe our operating-point-agnostic training approach, in Section 4 we describe our experiments and finally in Section 5 we present our conclusions.

2. MODELS FOR SPEAKER VERIFICATION

The main purpose of a SV systems is to verify the identity of a speaker providing a reliable decision (accept or reject a speaker) given a claimed identity and a spoken phrase. Conventionally, this is generally treated as a hypothesis testing problem. We define two hypotheses: a) the null hypothesis H_0 accepts the speaker as legitimate and b) the alternative hypothesis H_1 rejects him as impostor. The actual testing is performed using a likelihood ratio test. A parametric model with parameters Λ_S is defined for the distribution of data from the target speaker S. An *imposter* model with parameters $\lambda_{\bar{S}}$ is specified for the class of impostors the aggregate of all speakers who are not S. The actual likelihood-ratio is stated as follows:

$$\begin{array}{l} \theta_S\left(X\right) = \log\left(P\left(X|\Lambda_S\right)\right) - \log\left(P\left(X|\Lambda_{\bar{S}}\right)\right) \\ \quad \text{accept } H_0 \quad \text{if } \theta_S(X) > \tau \\ \quad \text{accept } H_1 \quad \text{otherwise} \end{array} \tag{1}$$

The problem now reduces to finding an appropriate model for $P(X|\Lambda_S)$ and $P(X|\Lambda_{\bar{S}})$. For this, each recording χ is transformed into a sequence of feature vectors, typically mel-frequency cepstral coefficient vectors, augmented by their delta (velocity) and double delta (acceleration) coefficients. Thus, $\chi = X_1, X_2, \cdots, X_T$, where X_i is the i^{th} feature vector in the sequence. The vectors X_i are assumed to be IID and have a a Gaussian mixture distribution given by,

$$P(\chi; \Lambda_C) = \prod_i \sum_k w_k^C \mathcal{N}(X_i; \mu_k^C, \Sigma^C, k),$$

where C is either S or $\bar{S}$, and w_k^C, μ_k^C and Σ_k^C are the mixture weight, mean and covariance (usually assumed to be a diagonal matrix) of the k^{th} Gaussian in the mixture. Thus $\Lambda_C = \{w_k^C, \mu_k^C, \Sigma_k^C \ \forall k\}$.

The problem of *training* the system is reduced to learning the GMM parameters Λ_S and $\Lambda_{\bar{S}}$ for the true speaker S and impostor $\bar{S}$ respectively. The imposter model $\Lambda_{\bar{S}}$ is typically trained from a large corpus of imposter recordings, using the Expectation Maximization (EM) algorithm [22]. Since this model represents the "universal" speaker, it is frequently called the "Universal Background Model" or "UBM".

Typically, the amount of enrollment training data available from the target speaker is insufficient to train Λ_S directly using EM. In the ideal case, where the recording conditions for the test data from the speaker are identical to those in the enrollment data, the UBM $\Lambda_{\bar{S}}$ can be *adapted* to the enrollment training data using a *maximum a posteriori* (MAP) adaptation procedure [11] to obtain Λ_S.

When recording channel mismatches are expected between test and enrollment data for the speaker, the state of the art uses *joint factor analysis* (JFA) [14, 12] to train Λ_S. The JFA approach *decomposes* the parameters of the distribution into two sets of factors – one representing the speaker and the second representing the channel. Channel factors, that contain no information about the speaker, are marginalized out when performing the likelihood ratio test.

For JFA, the means $\{\mu_i^{S,H}\}$ of all the Gaussians in a GMM $\Lambda_{S,H}$, representing the distribution for speaker S recorded over channel H are concatenated into a single vector, termed a *supervector* : $M_{S,H} = [\mu_1^{S,H} \| \mu_2^{S,H} \| \cdots]$. The supervector $M_{S,H}$ is further assumed to be composed from a collection of factors as:

$$M_{S,H} = m + V y_S + U x_{S,H} + D z_S, \tag{2}$$

where m is a global mean across all speakers (commonly the mean of the UBM), V is the *loading matrix* that represents a speaker-specific subspace, U is the loading matrix that represents a channel-specific subspace, and D is a diagonal matrix. y_S is known as the speaker factor vector belonging to speaker S; it is normally distributed with mean 0 and unit variance. $x_{S,H}$, is a channel factor, specific to speaker S recorded over channel H. Finally, z represents the residual error. The various loadings are learned from a large collection of recordings over many speakers and channels, using EM. Thereafter, to adapt the model to the target speaker, recorded over a given channel, only factors y_S and $x_{S,H}$ need be estimated. The procedures for learning the loadings and factors, as well as for performing classification with the resultant model are well described in [12, 14].

3. LEARNING THE PARAMETERS TO MINIMIZE THE AUC OF THE DET CURVE UNDER NOISY CONDITIONS.

Conventional training methods for learning the distribution parameters Λ_S and $\Lambda_{\bar{S}}$ fall into two categories: generative and discriminative. Generative procedures including MAP and JFA attempt to estimate the distribution of each class S and $\hat{S}$ individually, without regard to the actual classification performance obtained with them. Discriminative approaches do optimize classification performance, however they do so primarily at the operating point embodied in the relative proportions of training data from S and $\hat{S}$,

In contrast, we aim to optimize the performance over the entire DET curve. We will do this is by minimizing the AUC - the area under the DET curve. Consider a binary classifier to distinguish between classes S and $\bar{S}$. Let $\mathcal{H}$, and $\mathcal{W}$ be the two sets of data belonging to S and $\bar{S}$ respectively. The *empirical* AUC for the classifier computes a score $\theta(\chi)$ to determine if χ belongs to S as:

$$G(\Lambda) = 1.0 - \frac{\sum_{\chi\in\mathcal{H}}\sum_{\hat{\chi}\in\mathcal{W}} 1(\theta(\chi) > \theta(\hat{\chi}))}{|\mathcal{H}||\mathcal{W}|} \tag{3}$$

where 1 is the indicator function. The second term on the right hand side comes from the Wilcoxon-Mann-Whitney statistic [20] to compute the area under the curve. $G(\Lambda)$ is a function of $\Lambda = \Lambda_S \cup \Lambda_{\bar{S}}$.

The empirical AUC of equation 3, however, is discontinuous, thanks to the indicator function. We therefore "smoothen" it by replacing the indicator function by the following sigmoid:

$$R(a,b) = \frac{1}{1+\exp\left(-\gamma\ \ \varphi(a,b)\right)}, \tag{4}$$

where γ governs the steepness of the sigmoid and φ is the distance $\varphi(a,b) = a - b$. Introducing this into Equation 3, we obtain the following approximation:

$$G(\Lambda) = 1.0 - \frac{\sum_{\chi\in\mathcal{H}}\sum_{\hat{\chi}\in\mathcal{W}} R(\theta(\chi),\theta(\hat{\chi}))}{|\mathcal{H}||\mathcal{W}|}. \tag{5}$$

Equation 5 now forms our modified objective function which must be minimized to learn Λ_S and $\Lambda_{\bar{S}}$. This can be minimized using the generalized probabilistic descent algorithm (GPD). Let $\mathbf{X} = \mathcal{H} \cup \mathcal{W}$ be the complete set of all training instances. The GPD update rules to learn $\Lambda = \Lambda_S \cup \Lambda_{\bar{S}}$ are given by:

$$\Lambda_{t+1} = \Lambda_t - \epsilon\nabla L(\mathbf{X},\Lambda) \tag{6}$$

$$\nabla L(\mathbf{X},\Lambda) = -\frac{1}{|\mathcal{H}||\mathcal{W}|}\sum_{\chi\in\mathcal{H}}\sum_{\hat{\chi}\in\mathcal{W}}\gamma(1-R)\left[\frac{\partial\theta(\chi)}{\partial\Lambda} - \frac{\partial\theta(\hat{\chi})}{\partial\Lambda}\right] \tag{7}$$

where R denotes $R(\theta(\chi),\theta(\hat{\chi}))$ and ϵ is a learning rate parameter.

We now explain how these rules can be applied to learn model parameters within both, the minimum-verification error and JFA formalisms.

3.1. Minimum Verification Error (MVE)

Minimum verification error training for SV systems adapts a UBM simultaneously to a set of imposter data and data from the target speaker, to obtain a user-specific imposter model $\Lambda^S_{\bar{S}}$ and a user model Λ_S that together provide the the best empirical verification error on the training data. We replace the empirical verification error objective function used in MVE training by the following:

$$G(\Lambda) = 1.0 - \frac{\sum_{\chi\in\mathcal{H}}\sum_{X\in\chi}\sum_{\hat{\chi}\in\mathcal{W}}\sum_{\hat{X}\in\hat{\chi}} R(\theta(X),\theta(\hat{X}))}{\sum_{\chi\in\mathcal{H}} L_\chi \sum_{\hat{\chi}\in\mathcal{W}} L_{\hat{\chi}}} \tag{8}$$

where L_χ is the number of feature vectors in χ. For any GMM parameter ϕ, the GPD update rule is given by, $\phi_{t+1} = \phi_t - \epsilon\nabla_\phi L(\mathbf{X},\Lambda)$. Representing $\sum_{\chi\in\mathcal{H}} L_\chi = |\mathcal{H}|$ and $\sum_{\chi\in\mathcal{W}} L_\chi = |\mathcal{W}|$, the gradient $\nabla_\phi L(\mathbf{X},\Lambda)$ is given by:

$$\nabla_\phi L(\mathbf{X},\Lambda) = -\frac{1}{|\mathcal{H}||\mathcal{W}|}\sum_{\chi\in\mathcal{H}}\sum_{X\in\chi}\sum_{\hat{\chi}\in\mathcal{W}}\sum_{\hat{X}\in\hat{\chi}}\gamma(1-R)\nabla_\phi l(X,\hat{X},\Lambda) \tag{9}$$

where $\nabla_\phi l(X,\hat{X},\Lambda)$ is a local gradient with respect to ϕ at $\chi,\hat{\chi}$ and has the form $\nabla_\phi l(X,\hat{X},\Lambda) = -\frac{\partial\theta(X)}{\partial\phi} + \frac{\partial\theta(\hat{X})}{\partial\phi}$, where $\frac{\partial\theta(X)}{\partial\phi}$ represents the derivative of the log-likelihood-difference given by the Gaussian mixture models for the target speaker and the imposter model for vector X with respect to ϕ.

The update rules for each of the parameters w^S_k, μ^S_k and Σ^S_k can be easily obtained by computing the derivatives, $\frac{\partial\theta(X)}{\partial w^S_k}$, $\frac{\partial\theta(X)}{\partial\mu^S_k}$ and $\frac{\partial\theta(X)}{\partial\Sigma^S_k}$. The update rules for the parameters of the user-specific imposter model $\Lambda_{\bar{S}}$ can be similarly obtained.

3.2. Joint Factor Analysis (JFA)

For the case of joint factor analysis, the learning is divided into two parts. First, the global loading matrices, V (speaker loading), U (channel loading) and D (the diagonal uniqueness) are learned from a large collection of speaker recordings over a variety of channels. The global parameters learn the general characteristics of the speaker and channel subspaces. Second, we estimate the particular parameters: y_S, which represents the target speaker factor; x_S which characterizes the session or channel specific H factor for speaker S, and z is the uniqueness factor. These specific parameters are customized to a specific speaker.

The discriminant objective function, takes the form of Equation 10.

$$\begin{aligned}\theta(\chi) = &\log\ \ p(\chi; V,U,D,y_S(\chi),x_{H(\chi),S(\chi)}) \\ &- \log\ \ p(\chi;\lambda_{\bar{S}},U,x_{H(\chi),S(\chi)})\end{aligned} \tag{10}$$

$S(\chi)$ characterizes the speaker S in the session χ. $H(\chi)$ characterizes the recording channel in χ, m is the global mean computed from the universal background model $\lambda_{\bar{S}}$. Equation 10 computes the log-likelihood for the model of speaker S according to $M = m + Vy_{S(\chi)} + Ux_{S(\chi),H(\chi)} + Dz$. The imposter log-likelihood is computed using $M' = m + Ux_{S(\chi),H(\chi)} + Dz$, which just considers the universal m adjusted to a channel factor given by the information in the recording χ.

3.2.1. Estimating Loading Matrices

To estimate the loadings, let $\mathbf{X}$, be a large collection of recordings from a large number of speakers $\mathcal{S}$. Let $\mathbf{X}_S$ represent the recordings from speaker $S \in \mathcal{S}$ and $\mathbf{X}_{\bar{S}}$ the recordings from all imposters for S, *i.e.* $\bar{S} = \mathcal{S}\backslash S$ and $\mathbf{X} = \mathbf{X}_S \cup \mathbf{X}_{\bar{S}}$. We can define the AUC objective as follows,

$$G(\Lambda) = 1.0 - \sum_{S\in\mathcal{S}} \frac{\sum_{\chi\in\mathbf{X}_S}\sum_{\chi\in\mathbf{X}_{\bar{S}}} R(\theta_S(\chi),\theta_S(\hat{\chi}))}{|\mathbf{X}_S||\mathbf{X}_{\bar{S}}|} \tag{11}$$

The GPD update rule for any global parameter ϕ is given by $\phi_{t+1} = \phi_t - \epsilon\nabla_\phi L(\mathbf{X},\Lambda)$, where

$$\nabla_\phi L(\mathbf{X},\Lambda) = -\sum_{S\in\mathcal{S}} \frac{\sum_{\chi\in\mathbf{X}_S}\sum_{\chi\in\mathbf{X}_{\bar{S}}}\gamma(1-R)\nabla_\phi l(\chi,\hat{\chi},\Lambda)}{|\mathbf{X}_S||\mathbf{X}_{\bar{S}}|} \tag{12}$$

and $\nabla_\phi l(\chi, \hat{\chi}, \Lambda)$ is a local gradient with respect to ϕ at $\chi, \hat{\chi}$, where $\nabla_\phi l(\chi, \hat{\chi}, \Lambda) = -\frac{\partial\theta(\chi)}{\partial\phi} + \frac{\partial\theta(\hat{\chi})}{\partial\phi}$. A solution to JFA was studied in [23].

3.2.2. Estimating Factor Vectors

To estimate the particular parameters or factors for a particular S, the AUC objective function is given by Equation 13.

$$G(\Lambda_S) = 1.0 - \frac{\sum_{\chi\in\mathbf{X}_S}\sum_{\chi\in\mathbf{X}_{\bar{S}}} R(\theta_S(\chi), \theta_S(\hat{\chi}))}{|\mathbf{X}_S||\mathbf{X}_{\bar{S}}|} \quad (13)$$

Note that, Equation 13 considers just one speaker S. The GPD update rule uses the following gradient:

$$\nabla_\phi L(\mathbf{X}, \Lambda) = -\frac{\sum_{\chi\in\mathbf{X}_S}\sum_{\chi\in\mathbf{X}_{\bar{S}}} \gamma(1-R)\nabla_\phi l(\chi, \hat{\chi}, \Lambda)}{|\mathbf{X}_S||\mathbf{X}_{\bar{S}}|}.$$

where $\nabla_\phi l(\chi, \hat{\chi}, \Lambda)$ is defined as $\nabla_\phi l(\chi, \hat{\chi}, \Lambda) = -\frac{\partial\theta(\chi)}{\partial\phi} + \frac{\partial\theta(\hat{\chi})}{\partial\phi}$. To update the parameters y_S, $x_{S,H}$ and z_S, we can compute the derivatives $\frac{\partial\theta(\chi)}{\partial y_S}$, $\frac{\partial\theta(\chi)}{\partial x_{S,H}}$ and $\frac{\partial\theta(\chi)}{\partial z_S}$ and employ the GPD update rule. In practice these factors can be estimated using conventional EM algorithm.

This methodology can also be extended to i-vector feature extraction and PLDA [16], customizing the update rules to employ the partial derivatives for their particular parameters.

4. EXPERIMENTS AND RESULTS

We conducted a set of experiments to evaluate the proposed AUC-minimization approach. In the first, we compared the performance of conventional MAP and JFA based learning with JFA optimized using the AUC criterion on speech recordings, where the noise conditions in the training and test data were matched. In the second, we compared the performance of AUC-minimization base MVE against conventional methods on mismatched conditions, where the test data are noisy.

4.1. Experimental Setup

We employed the NIST Speaker Evaluation 2004, 2005, 2010 and 2008 database [24] to complete this study. We followed the evaluation rules (*e.g.* not using any target speaker in the test set as an imposter for other target speakers). For the feature extraction, a short-time 256-pt Fourier analysis is performed on a 25 ms analysis window with a frame shift of 10 ms between analysis windows. The feature vectors (token) are 39-dimensional, comprising Mel Frequency Cepstral Coefficients (MFCCs), and their delta and double delta coefficients. We included a frame removal criterion that to eliminate low energy frames that do not provide information about the identity of the person.

4.2. Baseline framework

We first obtain a baseline result using both MAP and JFA to model each speaker [25]. We first compute a gender-dependent and target-independent UBM trained from a pool of raw speech (NIST Speaker Evaluation 2004 core database). This model captures the characteristics of all the data vectors of the users not belonging to the target set of speakers to be evaluated. The *expectation maximization* (EM) algorithm is used to estimate the GMM parameters of the UBM. For the MAP-based baseline, the models for target speakers in the evaluation set are obtained by MAP adaptation of the UBM. For the JFA baseline, the speaker and channel factors were learned from the pool of impostors using EM and adapted to individual speakers by estimation of factors. The code for JFA was obtained from the implementation of the Speech Processing Group at the Brno University of Technology [26] and used in part or whole by [27, 28, 29] well known sites. All verification tests were performed under the hypothesis testing framework.

4.3. Experiments on noisy speech

In this experiment we compared baseline techniques to AUC-minimized training on noisy speech. Experiments were performed using speech corrupted to a variety of SNRs (10dB, 15dB, and a cocktail of 0-15dB), all of them using babble noise.

For all experiments, we used 100 male registered users. Following NIST 2008 Evaluation rules, the probability of being a target, P_{target}, is 0.01 and the probability of being a impostor, $P_{impostor}$, is 0.99. We use a 512 component GMM in all cases. No normalization was used after the score computation to observe the full effect of MVE and JFA approach.

The first column of Table 1 shows the results obtained for clean data. The results are consistent with comparisons performed by other researchers: JFA outperforms both baseline MAP learning as well MVE learning significantly. In both cases, the models learned via AUC-minimization somewhat outperform conventionally trained models. All performance numbers are noted to improve as the amount of training data used to learn the base UBM increases.

The remaining columns of Table 1 compares MAP, MVE, JFA, AUC-optimized JFA and AUC-optimized MVE on noisy speech of various SNRs. We observe that AUC-optmized learning consistently outperforms conventional training in all cases. Moreover, the best results are obtained with AUC-optimized MVE. The results are consistent across all noise conditions. This contravenes the observation on clean speech, where the best performance is obtained with JFA.

System	**clean**	10dB	15dB	0-15dB (cocktail)
MAP	15.95	18.01	17.48	35.7
MVE	13.51	17.67	17.15	28.1
JFA	12.07	17.23	16.79	27.3
JFA-AUC	11.93	16.51	16.22	24.0
MVE-AUC	13.21	15.93	15.78	22.8

Table 1. EER of the noisy task (babble noise).

5. CONCLUSION AND DISCUSSION

The results in Section 4 are consistently significant. Clearly, a learning paradigm that optimizes the entire DET curve results in better performance than that obtained with conventional maximum-likelihood of discriminative training methods. What is interesting, however, is that the improvements obtained from AUC minimization are actually significantly greater on noisy speech than on clean speech. A possible reason is that AUC minimization naturally accounts for any shift in the data away from the intended operating point where performance is measured. More curiously, the performance of MVE, which functions over the entire space of data, benefits significantly more than JFA, which factors out subspaces. Thus, the final performance of AUC-optimized MVE on noisy data is significantly superior to that obtained with similarly trained JFA. Moreover the gains increase with increasing noise level.

As part of the future research, we will also investigate other objective functions that assigns weights to the DET curve so that we control not just the area under the curve, but the curve at each operating point.

6. REFERENCES

[1] L. Ferrer, H. Bratt, L. Burget, H. Cernocky, O. Glembek, M. Graciarena, A. Lawson, Y. Lei, P. Matejka, O. Plchot, et al., "Promoting robustness for speaker modeling in the community: the prism evaluation set," in *Proceedings of SRE11 Analysis Workshop*, 2011.

[2] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 27, pp. 113–120, 1979.

[3] P.J. Moreno, B. Raj, and R.M. Stern, "Data-driven environmental compensation for speech recognition: A unified approach," *Speech Communication*, vol. 24, no. 4, pp. 267–285, 1998.

[4] J. Dropo, L. Deng, and A. Acero, "Evaluation of the SPLICE algorithm on the Aurora2 database (web update)," in *Eurospeech*, Aalborg, Denmark, 2001, vol. 2, pp. 217–220.

[5] L. Deng, A. Acero, J. L. Droppo, and J. X. Huang, "High-performance robust speech recognition using stereo training data," in *Proc. ICASSP*, Salt Lake City, UT, May 2001, pp. 301–304.

[6] A. Moreno-Daniel, J. A. Nolazco-Flores, T. Wada, and B.-H. Juang, "Acoustic model enhancement: An adaptation technique for speaker verification under noisy environments," in *Proc. ICASSP*, Honolulu, USA, April 2007.

[7] J.A. Nolazco-Flores and S. Young, "Continuous speech recognition in noise using spectral subtraction and hmm adaptation," in *Proc. ICASSP*, Adelaide, Australia, 1994, pp. 409–412.

[8] T. Hasan and J.H.L. Hansen, "Integrated feature normalization and enhancement for robust speaker recognition using acoustic factor analysis," in *Proc. Interspeech*, 2012.

[9] C. Hanilçi, T. Kinnunen, R. Saeidi, J. Pohjalainen, P. Alku, and F. Ertas, "Regularization of all-pole models for speaker verification under additive noise," in *Odyssey 2012-The Speaker and Language Recognition Workshop*, 2012.

[10] 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.

[11] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of markov chains," *IEEE Trans. on Speech and Audio Processing*, vol. 2, pp. 291–299, Apr. 1994.

[12] P. Kenny, P. Oueleet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of inter-speaker variability in speaker verification," *IEEE Trans. ASLP*, vol. 16, pp. 980–988, 2008.

[13] Najim Dehak, Patrick J. Kenny, Rda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, 2010.

[14] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[15] P Kenny, "Bayesian Speaker Verification with Heavy-Tailed Priors," *keynote presentation, Odyssey Speaker and Language Recognition Workshop Brno, Czech Republic*, 2010.

[16] P. Matejka, O. Glembek, F. Castaldo, MJ Alam, O. Plchot, P. Kenny, L. Burget, and J. Cernocky, "Full-covariance UBM and heavy-tailed PLDA in i-vector speaker verification," in *IEEE Intl. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*, 2011.

[17] L.P. Garcia-Perera, J.A. Nolazco-Flores, R. B. Raj, and R. Stern, "Optimization of the det curve in speaker verification," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*. IEEE, 2012.

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

[19] D. Zhu, H. Li, B. Ma, and C.H. Lee, "Discriminative learning for optimizing detection performance in spoken language recognition," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 4161–4164.

[20] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Annals of Mathematical Statistics*, vol. 18:1, pp. 50–60, 1947.

[21] CH Lee, "A unified statistical hypothesis testing approach to speaker verification and verbal information verification," in *Proc. COST,Workshop on Speech Technology in the Public Telephone Network: Where are we today?*, Greece, September 1997, vol. 250, pp. 63–72.

[22] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Annals of the Royal Statistical Society*, vol. 39, pp. 1–38, Dec. 1977.

[23] L.K. Saul and M.G. Rahim, "Maximum likelihood and minimum classification error factor analysis for automatic speech recognition," *Speech and Audio Processing, IEEE Transactions on*, vol. 8, no. 2, pp. 115–125, 2000.

[24] A.F. Martin and C.S. Greenberg, "NIST 2008 Speaker Recognition Evaluation: Performance Across Telephone and Room Microphone Channels," in *Proc. Interspeech*, 2009.

[25] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, no. 1-2, pp. 91–108, 1995.

[26] L. Burget, M. Fapso, and V. Hubeika, "BUT system for NIST 2008 speaker recognition evaluation," in *Interspeech*, 2009.

[27] N. Scheffer, L. Ferrer, M. Graciarena, S. Kajarekar, E. Shriberg, and A. Stolcke, "The SRI NIST 2010 speaker recognition evaluation system," in *IEEE Intl. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*, 2011, pp. 5292–5295.

[28] O. Glembek, L. Burget, N. Dehak, N. Brummer, and P. Kenny, "Comparison of scoring methods used in speaker recognition with joint factor analysis," in *IEEE Intl. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*, 2009.

[29] L. Burget, O. Plchot, S. Cumani, O. Glembek, P. Matejka, and N. Brummer, "Discriminatively trained probabilistic linear discriminant analysis for speaker verification," in *IEEE Intl. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*, 2011.