

# DETAC: A DISCRIMINATIVE CRITERION FOR SPEAKER VERIFICATION

Jiří Navrátil

Ganesh N. Ramaswamy

IBM Thomas J. Watson Research Center  
Rt. 134, Yorktown Heights, NY 10598, USA

## ABSTRACT

This paper introduces a general criterion applicable to discriminative training of detection systems, and discusses its particular implementation in GMM-based text-independent speaker verification. Based on an analysis of the detection error trade-off curve of a baseline system, we argue that the new criterion extends several conventional methods such as the maximum posterior training by logistic regression and the linear discriminative analysis projection, by a second aspect - “reshaping” the Bayes error area in favor of a relevant operating range. Optimization results with relative error reduction of up to 16% are presented on the cellular task of the NIST-2001 speaker recognition evaluation.

## 1. INTRODUCTION

The enticing idea of discriminative training (DT) is to extract features or to estimate model parameters such that the resulting class overlap is minimized. Over the widely used maximum-likelihood (ML) paradigm, DT has gained some success in the area of speaker identification [1, 2]. A difficulty persistent to DT is the fact that training data requirements grow as the competing classes are involved in the training, and so grows the risk of over-training if data sparseness exists. Another aggravating circumstance occurs specifically in detection (verification) systems, namely a high degree of asymmetry in terms of model coverage. While the target class (speaker) is well defined and represented by data, the non-target (world) class is usually approximated by a finite set of other speakers, who will not likely appear in the test. The DT, focusing by principle on distinguishing better between classes on the provided data, will tend to over-train.

Two DT methods specifically designed for verification were described in [3][4]. Rosenberg et al. [3] defined a smooth loss function with an embedded misverification measure based on likelihoods and carried out a minimum classification error training of the model parameters via gradient descent. Similarly, Heck et al. [4] approximated the error probabilities directly using sets of target and impostor tests and used minimum verification cost training of the model parameters. Both systems aim at minimizing a predefined detection cost function which is a linear combination of the two types of detection errors, i.e. False Alarms and Target Misses.

Based on an analytic description of the Detection Error Trade-Off (DET) curve, we introduce a new DET Analysis Criterion (DETAC) for discriminative training. Rather than optimizing the system for a specific operating point or detection cost, the criterion allows for improvements in pre-selectable ranges of operating points. We show that DETAC aims at minimizing the class overlap as well as at trading-off errors between operating regions, and that it ex-

hibits a good generalization behavior.

## 2. EVALUATING DETECTION SYSTEMS

The quality of a detection system is typically evaluated by measuring the rates of false alarm (type-I error,  $P_{FA}$ ) and target miss (type-II error,  $P_M$ ), given some test trials using a varying detection threshold. A set of operating points can then be displayed on the  $\{P_{FA}, P_M\}$  coordinates, obtaining a Receiver Operating Characteristics (ROC) curve that characterizes the system by its performance. The ROC can be plotted on special, non-linearly scaled coordinates, such that systems producing normally distributed detection scores for the target and the impostor trials will produce straight lines as the ROC. Such a representation, originally termed the “Burdick’s Chart” [5], was introduced into the speaker verification as the Detection Error Tradeoff (DET) plot by Martin et al. [6] and has been used in the annual Speaker Recognition Evaluations organized by the National Institute of Standards and Technology (NIST). The DET plot generally offers a better viewability and assessment of system with close-to-normal score distributions. By associating certain costs with the two types of error, the performance can also be evaluated by searching for the minimum achievable cost. Given a detection cost function (DCF), which is linear combination of the two error rates, specific areas of the DET curve become more relevant, such as the low FA area if the cost of a FA is higher than that of a Miss.

## 3. THE DETECTION ERROR TRADE-OFF ANALYSIS CRITERION (DETAC)

The first step towards our criterion is motivated by an analysis of the DET curve given the normal distribution assumption. Let  $\Phi(t)$  be a normal error function with a threshold variable  $t$  defined as  $\Phi(t) = \int_{-\infty}^t \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$ . Assuming the scores of impostor trials are distributed around mean  $\mu_1$  with standard deviation  $\sigma_1$  and those of targets with  $\mu_2, \sigma_2$ , the false alarm  $P_{FA}$  and miss error  $P_M$  probabilities can be written as

$$\begin{aligned} P_{FA}(t) &= \int_t^{\infty} \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2} dx = \Phi\left(\frac{\mu_1 - t}{\sigma_1}\right) \\ P_M(t) &= \int_{-\infty}^t \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2} dx = \Phi\left(\frac{t - \mu_2}{\sigma_2}\right) \end{aligned}$$

Using inversion and by eliminating the threshold  $t$ , the two errors can be related:

$$\begin{aligned} \frac{\mu_1 - t}{\sigma_1} &= \Phi^{-1}(P_{FA}), & \frac{t - \mu_2}{\sigma_2} &= \Phi^{-1}(P_M), \\ \Phi^{-1}(P_M) &= -\frac{\sigma_1}{\sigma_2} \Phi^{-1}(P_{FA}) + \frac{\mu_1 - \mu_2}{\sigma_2} \end{aligned} \quad (1)$$

Obviously, the functional relationship is linear. Since  $\Phi^{-1}$  leads to linear curves for normal distributions, it is identical to the DET scaling described above, and hence two parameters can be identified as directly related to the DET curve: 1) The Sigma-Ratio  $\frac{\sigma_1}{\sigma_2}$  that corresponds to the DET *slope* and 2) the normalized Delta-Term  $\frac{\mu_1 - \mu_2}{\sigma_2}$  corresponding to the *bias* of the DET line. While a bias change reflects uniform performance changes across all operating points (and is related to the normalized distance between the population means), a Sigma-Ratio change reflects a relative improvement in the low/high  $\{P_{FA}, P_M\}$  region versus the high/low region. For example, keeping the Delta-Term value constant, reducing the Sigma-Ratio can be accomplished solely by reducing the variance of the impostor distribution

If a DCF shifts the area of interest away from the equal-error point of operation, in either direction, then the Sigma-Ratio becomes a promising optimization parameter as its controlled modification can rotate the DET in favor of that DCF. Later in this paper, we argue that targeted rotational optimization exhibits a better generalization behavior than the discriminative one and represents a novel part of the DETAC.

Given a particular system with a DET curve, an operating area of interest, and an optimization parameter set  $\theta$ , we formulate the Detection Error Trade-Off Analysis Criterion (DETAC) as

1. Constrained minimization of either:

- (a) the Sigma-Ratio

$$\theta^* = \arg \min_{\theta} \pm R(\theta), \text{ subj. to } D(\theta) \leq 0 \quad (2)$$

- (b) the Delta-Term

$$\theta^* = \arg \min_{\theta} D(\theta), \text{ subj. to } R(\theta) = 0 \quad (3)$$

$$\text{with } R(\theta) = \frac{\sigma_1(\theta)}{\sigma_2(\theta)} - C_R; \quad D(\theta) = \frac{\mu_1(\theta) - \mu_2(\theta)}{\sigma_2(\theta)} - C_D$$

2. or Unconstrained minimization

$$\theta^* = \arg \min_{\theta} w_R \left[ \frac{\sigma_1(\theta)}{\sigma_2(\theta)} - C_R \right] + w_D \left[ \frac{\mu_1 - \mu_2}{\sigma_2} - C_D \right], \quad (4)$$

where  $C_D, C_R$  are values of the Sigma-Ratio and the Delta-Term at the beginning of optimization, and  $w_D, w_R$  may be used to regulate the two parts. While the unconstrained DETAC aims at both rotating and shifting the DET, the constrained face (2) minimizes primarily the slope (+ for a counterclockwise, and - for a clockwise DET rotation). Similarly (3) reduces the bias while keeping the slope constant. The constrained criteria *guarantee* improvements in both training errors in their respective targeted operating regions for systems with normally distributed scores. The above nonlinear constraints can be adapted in a straightforward way, e.g. to allow for shifts of the center of rotation or to combine both types of movement. In our experiments, the unconstrained face (4) was used because of a better stability of unconstrained optimization routines.

### 3.1. Feature Space optimization for GMM systems (fDETAC)

While the DETAC is a suitable criterion for detection tasks in general, the functional relationship  $\sigma(\theta), \mu(\theta)$  needs to be determined for the individual model type. We use a system for text-independent speaker verification based on Gaussian

Mixture Models (GMM) and a likelihood-ratio detector [7]. Herein, for each trial, a lower bound on the average log likelihood-ratio (LLR) is calculated on the target and the Universal Background Model (UBM) pair with ‘‘coupled’’ components.

$$\frac{1}{N} \sum_t \log \frac{\sum_i p_{i2} p_2(x_t|i)}{\sum_i p_{i1} p_1(x_t|i)} \geq \frac{1}{N} \sum_{t,i} \gamma_{ti} \log \frac{p_{i2} p_2(x_t|i)}{p_{i1} p_1(x_t|i)}$$

with  $\gamma_{ti} = P_1(i|x_t) = p_1(x_t, i) / \sum_j p_1(x_t, j)$ ,  $N$  vector count (index 1 pertains to the UBM, index 2 to the target model). This *average componentwise* ratio,  $\bar{l}$ , simplifies the subsequent analysis while giving performance comparable to that of LLR. (in [7] only the maximum scoring component of the UBM and  $\gamma$  indicator function was used, whereby the bound equality holds). Due to the fact that the target GMM is a mean-only Maximum A-Posteriori adaptation of the UBM, the  $\bar{l}$  for a vector  $x_t$  given a coupled model  $M$  can be written as

$$l(x_t|M) = \sum_i \gamma_{ti} \delta_i^T \Sigma_i^{-1} x_t + \sum_i \gamma_{ti} d_i,$$

$$\delta_i = \mu_{i2} - \mu_{i1}, \quad d_i = \frac{1}{2} \mu_{i1}^T \Sigma_i^{-1} \mu_{i1} - \frac{1}{2} \mu_{i2}^T \Sigma_i^{-1} \mu_{i2}$$

with  $\mu_{i1}, \mu_{i2}$  denoting componentwise mean vectors of the UBM and the target respectively, and  $\Sigma_i^{-1}$  the component precision matrix. Note that  $\bar{l}$  is a linear function of the feature vector  $x$ , since the Gaussian pairs share the same covariance matrix and priors. Then, the average  $\bar{l}$  for a test vector sequence is:

$$\begin{aligned} \bar{l}(X|M) &= \frac{1}{N} \left( \sum_{i,t} \gamma_{ti} \delta_i^T \Sigma_i^{-1} x_t + \sum_{i,t} \gamma_{ti} d_i \right) = \sum_i y_i^T \bar{x}_i + c \\ c &= \sum_i n_i d_i / N, \quad n_i = \sum_t \gamma_{ti} \\ y_i^T &= \delta_i^T \Sigma_i^{-1}, \quad \bar{x}_i = \sum_t \gamma_{ti} x_t / N \end{aligned}$$

We now seek to apply a global full-rank transform  $\mathbf{A} \in \mathbf{R}^{d \times d}$  to the componentwise distributed average feature vectors  $\bar{x}_i$ , as the  $\theta$  set in (4). Such transformed LLR can be brought into a compact form

$$\begin{aligned} \bar{l}(X|M, A) &= \sum_i y_i^T A \bar{x}_i + c = \text{tr} \left\{ A \sum_i y_i \bar{x}_i^T \right\} + c = \\ &= \text{tr} AB + c. \end{aligned} \quad (5)$$

with computation of  $O(d^2)$ , where  $d$  is the vector dimensionality,  $\text{tr}$  the trace operator, and  $B \in \mathbf{R}^{d \times d}$  a precomputed transform-invariant matrix. Using a set of  $N_2$  true target trials  $T = \{\alpha_1, \dots, \alpha_{N_2}\}$  (with  $\alpha_k = \text{tr} AB_k + c_k$ ) and a set of  $N_1$  impostor trials  $\bar{T} = \{\alpha_1, \dots, \alpha_{N_1}\}$ , the minimization of DETAC (4) with respect to a feature space transform  $A$  can now be carried out via a nonlinear optimization using the fDETAC gradient function

$$\begin{aligned} \frac{\partial F}{\partial \mathbf{A}} &= \frac{w_R}{\sigma_1 \sigma_2} \sigma_{\alpha} \mathbf{B}_1 - \left( \frac{w_R \sigma_1 + w_D \delta}{\sigma_2^3} \right) \sigma_{\alpha} \mathbf{B}_2 + \frac{w_D}{\sigma_2} \Delta \\ \sigma_{\alpha} \mathbf{B}_k &= \frac{1}{N_k} \sum_i \alpha_i \mathbf{B}_i^T - \frac{1}{N_k^2} \sum_i \alpha_i \sum_j \mathbf{B}_j^T, \quad k = \{1, 2\} \\ \Delta &= \frac{\partial \delta}{\partial \mathbf{A}} = \frac{1}{N_1} \sum_{i \in \bar{T}} \mathbf{B}_i^T - \frac{1}{N_2} \sum_{j \in T} \mathbf{B}_j^T \end{aligned} \quad (6)$$

where  $F$  is the objective of (4) and symbols in **boldface** are  $d \times d$  matrices.

This feature-space DETAC (fDETAC) optimization starts with  $A = \mathbf{I}$  and aims at finding  $A^*$  so as to further reduce the DET offset and to achieve a favorable rotation towards lower cost on top of the maximum-likelihood pre-trained GMM system.

### 3.2. Linear Projection (pDETAC)

Another implementation of DETAC is to seek a vector  $a \in R^{M \times 1}$  to be applied on score level, i.e. to linearly combine multiple detection systems. Assume we have  $M$  systems supplying scores for a trial, i.e. forming a score vector  $s = [s_1, \dots, s_M]^T$ . Using the target and impostor sets  $T, \bar{T}$  the means  $\mu_{1,2} \in R^{M \times 1}$  and covariances  $S_{1,2} \in R^{M \times M}$  of  $s$  can be estimated, and the pDETAC objective function can be written as

$$F = w_R \left( \sqrt{\frac{a^T S_1 a}{a^T S_2 a}} - C_R \right) + w_D \left[ \frac{a^T (\mu_1 - \mu_2)}{\sqrt{a^T S_2 a}} - C_D \right] \quad (7)$$

Projected scores  $a^T s$  can then be used as basis for the threshold decision.

### 3.3. Discussion

An interesting comparison of the DETAC objective to several known methods can be made. An alternative to fDETAC offers itself in the form of Logistic Regression (Maximum Entropy training). Here, we look for a full-rank  $A$ , such that the log posterior of the target/impostor class for the training sets  $T, \bar{T}$  is maximized, i.e. the following function is minimized

$$F = -\log \left\{ \prod_{i \in T} \frac{e^{\alpha_i}}{1 + e^{\alpha_i}} \prod_{j \in \bar{T}} \frac{1}{1 + e^{\alpha_j}} \right\} + w \|A - I\|^2 \quad (8)$$

$$\frac{\partial F}{\partial A} = \frac{1}{N_2} \sum_{i \in T} \frac{B_i^T e^{\alpha_i}}{1 + e^{\alpha_i}} - B_i^T + \frac{1}{N_1} \sum_{j \in \bar{T}} \frac{B_j^T e^{\alpha_j}}{1 + e^{\alpha_j}} + 2w(A - I)$$

with  $w \|A - I\|^2$  a weighted regulator term to enforce a solution close to the unity. Assuming the LLR values  $\alpha$  are globally distributed in  $(-\epsilon, \epsilon)$  close to zero (which may be achieved to an arbitrary degree by a global shift and scaling), the objective (8) can be approximated as follows

$$F = \frac{1}{N_1} \sum_{j \in \bar{T}} \log(1 + e^{\alpha_j}) - \frac{1}{N_2} \sum_{i \in T} [\alpha_i - \log(1 + e^{\alpha_i})]$$

$$\log(1 + e^\alpha) \approx \frac{1}{2} \alpha + \log 2; \quad \alpha \in (-\epsilon, \epsilon)$$

$$F \approx \frac{1}{2N_1} \sum_{j \in \bar{T}} \alpha_j - \frac{1}{2N_2} \sum_{i \in T} \alpha_i + c = \frac{1}{2} (\mu_1 - \mu_2) + c$$

which turns out to be identical to minimizing the numerator of the DETAC Delta-Term in (4). Hence, the logistic regression aims primarily at increasing the distance between the class distributions to reduce the Bayes error. DETAC, in addition to doing the same, is also looking at ways to “reshape” the Bayes error area of these distributions.

In case of equal covariances  $S = S_1 = S_2$  in (7), the square of the Delta-Term becomes

$$a^T \delta (a^T S a)^{-\frac{1}{2}} (a^T S a)^{-\frac{1}{2}} \delta^T a = \frac{a^T B a}{a^T W a} \quad (9)$$

with  $\delta = \mu_1 - \mu_2$ ,  $\mu_2 > \mu_1$  and  $B, W$  denoting the between- and within-class covariances. For  $a \in R^{M \times 1}$ , the form (9) is the well-known LDA objective with  $B$  of rank 1 for binary classification problems and a single solution  $a^* = c \cdot \delta S^{-1}$ . The equal covariances are an intrinsic assumption of the LDA and therefore the LDA *cannot* change the Sigma-Ratio. Interestingly, for  $S = S_1 = S_2$ , the squared Delta-Term also equals the Bhattacharyya distance, i.e. corresponds to minimizing an upper bound on the Bayes error, and it further equals measures such as the Kullback-Leibler divergence and the Mahalanobis distance.

The above comparison highlights the novel part of the DETAC - the access to the DET slope, in addition to the DET offset that seems to be focused on by many conventional techniques.

An extension of DETAC to non-Gaussian distributions exists [8], however its description exceeds the scope of this paper.

## 4. EXPERIMENTS

### 4.1. Database

The performance of the described steps was experimentally evaluated using data from the cellular part of the Switchboard (SWB) telephone corpus, in particular the set defined by the NIST for the 1-speaker cellular detection task (CT) in the 2001 Speaker Recognition Evaluation (SRE) [9]. This set consists of 60 development, 174 test speakers, and a total of 20380 verification trials. The trial test was split into two non-overlapping subsets (both in speakers and test recordings) to allow for cross-evaluating the discriminative optimization. The SWB-I (1996 SRE) part containing landline-telephone recordings (4 hrs), and an internal cellular database (2 hrs) were used to create the UBM of the baseline systems (see below). Further details about the NIST SRE CT can be found in [9].

### 4.2. Baseline Systems

Three GMM systems with coupled UBM-Target modeling and a single maximum-likelihood linear transform as described in [7] served as the baseline systems, differing solely in the dataset used to create the respective UBM: System 1 (1536 Gaussians) UBM was trained using the 60 development speakers of the 2001 CT plus 2 hrs of data from the internal data collection, System 2 (2048 Gaussians) had the 60 speakers plus the 1996 SRE landline-telephone recordings post-processed by a GSM-encoder, and System 3 (2048 Gaussians) used the 1996 SRE landline data only. The front-end was implemented as described in [10]. The scoring in all three systems was carried out as in (5) with the posteriors  $\gamma_{i,t}$  being approximated by the indicator function of the maximum scoring component [7].

The detection cost function (DCF) defined in the 2001 SRE [9] served as the primary evaluation measure. This DCF weights the two error types according to  $Cost = 0.99 * \epsilon_{FA} + 0.1 * \epsilon_{Miss}$  and thus shifts the operating point towards low  $P_{FA}$  and makes a counterclockwise DETAC rotation desirable.

### 4.3. Results

The fDETAC transformation matrix was optimized for each of the three systems according to Eqs.(4), (6) using a non-linear optimization tool package ( $w_D, w_R$  were set both to 1.0). The performance was cross-evaluated on the two subsets 1 and 2 and the held-out set performance averaged. Figure 1 illustrates the minimum DCF value and the Equal-Error Rate (EER) on the training and test data for System 2 as the optimization process evolves. Irrespective of the absolute performances, a trend regarding the generalization behavior can be observed in this example: While the

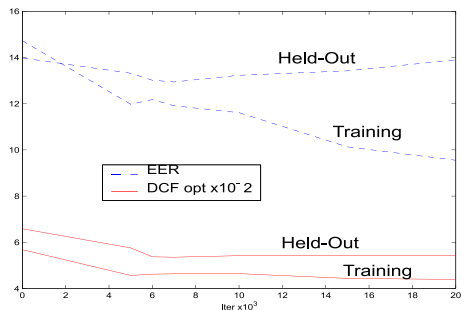


Figure 1. Optimum DCF and EER of the fDETAC iterations

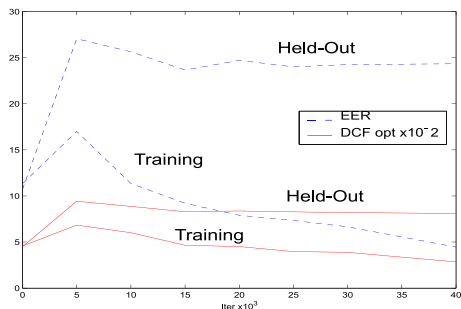


Figure 2. Optimum DCF and EER for iterative Logistic Regression

training apparently reduces both the DCF and the EER on the training set, the main gain on held-out data seems to hold only for the DCF measure (note the dynamic range of the DCF shows up narrower than that of the EER in the plot). The DET curves shown in Fig. 3 (System 3) illustrate why this happens: A distinct rotational gain “survives” the generalization from training to held-out set. Same observations can be made on the remaining two systems as well. The DCF performances of the three systems with and without fDETAC are summarized in Table 1. It has to be noted that the iterative optimization process was terminated with Sigma-Ratio gain reaching a value of 0.15 on the training set. This value, although not critical in the range (0.05, 0.3), was determined as the stopping criterion for all systems and data sets globally. The pDETAC combination could further reduce the DCF to  $41 \cdot 10^{-3}$ , although same gain was also achieved by an LDA projection for these three systems.

The results of a comparative experiment using the Lo-

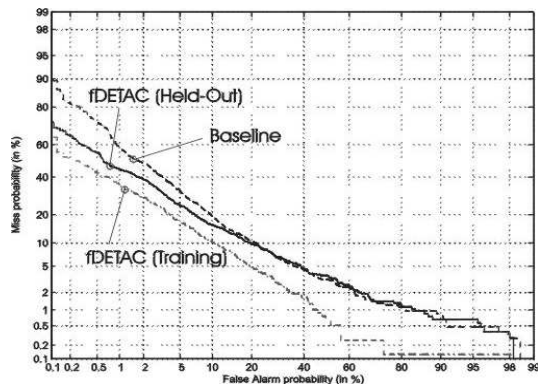


Figure 3. The fDETAC optimization curves (System 3)

	opt. DCF $10^{-3}$		
	Sys. 1	Sys. 2	Sys. 3
Baseline	45.6	49.8	63.4
+fDETAC	42.7	43.8	50.8
pDETAC comb.	41.0		

Table 1. DCF Performance of the three GMM systems with feature-space DETAC and projectional DETAC

gistic Regression optimization according to (8) is shown in Figure 2, in which obviously no generalization could be achieved on the held-out set. Increasing the weight  $w$  on the regulator term of (8) caused the optimization to converge earlier and closer to the unity matrix, but did not alleviate the overtraining problem.

## 5. CONCLUSION

The promise of the new DETAC seems to be in its generalizing nature: This criterion extends the classic discriminative objective (reducing the class overlap) by another aspect (reshaping the class overlap). In our experiments, DETAC exhibited better generalization behavior, compared to logistic regression, the main gain being due to the second aspect mentioned above. Favorable generalization may also be supported by the fact that only two simple trends (linear slope and bias) are contained in the DETAC objective function, as opposed to specific operating points or cost function definitions. We show that DETAC can be implemented as a feature space transform or a system combiner and improvements of 6% to 16% rel. on the NIST 2001 cellular task can be achieved depending on the baseline system.

## REFERENCES

- [1] Q. Jin and A. Waibel. Application of LDA to speaker recognition. In *Proc. of the ICSLP-00*, Beijing, China, October 2000.
- [2] S. Fine, J. Navrátil, and R.A. Gopinath. A hybrid gmm/svm approach to speaker identification. In *Proc. of the ICASSP-01*, Salt Lake City, Utah, May 2001.
- [3] A.E. Rosenberg, O. Siohan, and S. Parthasarathy. Speaker verification using minimum error training. In *Proc. of the ICASSP-98*, Seattle, WA, May 1998.
- [4] L. Heck and Y. Konig. Discriminative training of minimum cost speaker verification systems. In *Proc. RLA2-ESCA*, pages 93–96, Avignon, France, 1998.
- [5] K. Fukunaga. *Statistical Pattern Recognition*. Academic Press, 1990. 2nd Edition.
- [6] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The DET curve in assessment of detection task performance. In *Proc. of the EUROSPEECH-97*, pages 1895–8, Rhodes, Greece, September 1997.
- [7] J. Navrátil, U.V. Chaudhari, and G. Ramaswamy. Speaker verification using target and background dependent linear transforms and multi-system fusion. In *Proc. of EUROSPEECH-01*, Aalborg, Denmark, September 2001.
- [8] J Navrátil. Generalized DET analysis criterion. IBM-internal presentation, January 2002.
- [9] <http://www.nist.gov/speech/tests/spk/index.htm>.
- [10] B. Xiang, U.V. Chaudhari, Navrátil J., G.N. Ramaswamy, and R.A. Gopinath. Short-time gaussianization for robust speaker verification. In *Proc. of the ICASSP-02*, Orlando, FL, May 2002.