Discriminative Training of Gaussian Mixture Speaker Models: A New Approach

Srikanth M R, Hema A Murthy
Department of Computer Science and Engineering,
Indian Institute of Technology, Madras
Email : srikanth,hema@lantana.tenet.res.in

***Abstract*—Conventional speaker recognition systems use Gaussian mixture models (GMM) to model a speaker's voice based on the speaker's acoustic characteristics. This method is categorized as a non-discriminative training process, as the model-building process does not take into account the negative examples of the speaker. To increase the discriminative properties of a GMM for each speaker, a new approach that includes both positive and negative examples during the speaker training process is proposed. In this approach, speaker models are trained by moving the mixture model's means in such a way as to maximize the likelihood of speaker data while also minimizing the likelihood of negative examples for the speaker. The effectiveness of this approach on classification accuracies on speaker recognition tasks is tested on the NTIMIT database and NIST SRE 2003 corpora. The results indicate improvements in the performance of the system built using this new approach when compared to the traditional GMM-based speaker recognition systems.**

***Index Terms*—discriminative training, Gaussian mixture models, speaker recognition**

I. Introduction

The goal of a speaker recognition system is to identify a speaker based on his/her voice-characteristics. This task typically involves two phases - training and testing. In the training phase, a speaker registers with the system by providing sample utterances and statistical models based on those examples are then built. During testing, an utterance is identified (a) as belonging to a particular speaker (identification) or (b) the utterance along with a claim is authenticated (verification). A speaker recognition system is said to be text-independent when the text corresponding to an utterance does not influence the decisions made by the system.

A common way to model a speaker's voice in text-independent speaker recognition systems is to build a Gaussian Mixture Model [1] with the distinctive features extracted from the training samples. Mel-scale cepstral co-efficients are ubiquitous features in speaker recognition that are extracted from the training data by performing a filter bank analysis on a mel-scale. When a test utterance is supplied, the features extracted from it are used to compute the probability that the test sample belongs to a particular speaker. This kind of modelling is a non-discriminative way to build speaker models as it does not take into account the negative examples of a class. We refer to negative data as those samples that do not belong to the speaker and thus should not be classified as belonging to him/her. GMMs, however, are good approximations of speaker classes [2] and easier to compute. This encourages the development of a new approach for discriminative training within the GMM framework of speaker recognition.

To offset the lack of discriminative qualities of Gaussian mixture models, several approaches have been proposed. UBM-GMM (Universal Background Model - Gaussian Mixture Model) is a popular one among them. UBM is a base model from which all speaker models are adapted by a form of Bayesian adaptation [3]. A UBM is generally built from a large data set containing all probable speakers. During training, every speaker model is adapted from this UBM by performing MAP adaptation on the UBM with the training samples of the speaker. A model of the speaker is thus obtained and testing is done as in any GMM-based speaker recognition system [4]. UBM-GMMs are also common in speaker verification systems, where the likelihood ratio between the claimed speaker's model and the UBM for a test utterance is computed. This ratio is compared to a common or per-speaker threshold.

In [5], discrimination among GMMs have been introduced using the MCE (minimum classification error) criterion. [6] and [7], discuss a Maximum Model Distance algorithm for HMMs (Hidden Markov Models) that has been extended to GMMs in [8]. This approach tries to maximize the distance between each model and a set of competitive speakers' models. All the aforementioned approaches tend to use GMMs built the traditional way and discrimination is performed only across those GMMs, rather than directly build a GMM that has already been modelled to be discriminative. In another approach [9] , GMMs themselves are used as feature vectors, called supervectors, to train Support Vector Machines (SVM). Our primary motivation was to develop a discriminative training technique that still retains the GMM framework for speaker recognition. In the proposed approach, GMMs are built for each speaker discriminatively based on the available positive and negative examples for each speaker. Clearly, the objective would now be to both maximize the likelihood of positive speaker data, as in the existing modelling techniques, and also minimize the likelihood of negative data samples.

The rest of the paper is organized as follows : Section II contains a general introduction to Gaussian mixture models, their application to text-independent speaker identification, and also introduces our approach of discriminative training for GMMs . Experimental results are discussed in Section III and Section IV discusses our conclusions.

II. Proposed Approach

A. Background : Gaussian Mixture Models

Gaussian Mixture Models (GMM) are popular statistical models due to their ability to form good approximations of data and the ease involved in computation. It is a linear combination of multiple Gaussian distributions. They are expressed as

$$p(\bar{x}) = \sum_{k=1}^{K} \pi_k N(\bar{x}|\bar{\mu_k}, \Sigma_k) \tag{1}$$

where

$\bar{x}$	:	a d-dimensional feature vector
π_k	:	weight of k^{th} mixture,
$N(\bar{x}\|\bar{\mu_k}, \Sigma_k)$	:	Unimodal Gaussian distribution with parameters $\bar{\mu_k}, \Sigma_k$
$\bar{\mu_k}$ and Σ_k	:	mean vector and covariance matrix of k^{th} mixture respectively
K	:	number of mixtures per model

It should be noted that

$$\sum_{k=1}^{K} \pi_k = 1$$

The parameters of a GMM for a data set D=$\{\bar{x_1}, \bar{x_2}, ..., \bar{x_n}, ..., \bar{x_N}\}$ are chosen such that

$$\bar{\theta} = \arg\max_{\bar{\theta^i}} P(D|\bar{\theta^i}) \tag{2}$$

where

$$\bar{\theta} = \{\bar{\theta_1}, \bar{\theta_2}, ..., \bar{\theta_K}\} \text{ and } \bar{\theta_k} = \{\bar{\mu_k}, \Sigma_k, \pi_k\} \; (k = 1, 2, ..., K)$$

Generally, implementations of GMM-based speaker recognition systems assume diagonal covariance matrices.

It is more convenient to maximize the log probability than (2).

$$l(\bar{\theta}) = \ln P(D|\bar{\theta}) \tag{3}$$

The feature vectors are assumed to be independent and identically distributed. Parameter estimation is done using the Expectation-Maximization algorithm [10] where likelihood of the training examples is maximized.

B. Discriminative Training

In this paper, a new objective function to be maximized is defined. This incorporates a placeholder for likelihood of negative training samples for a particular speaker. The objective is to maximize the likelihood of positive training samples and simultaneously minimize the likelihood of the negative examples of that class with respect to the same model. That is, we need to find $\bar{\theta}$, the model, that combines both approaches of

$$\bar{\theta} = \arg\max_{\bar{\theta^i}} \ln P(D|\bar{\theta^i}) \tag{4}$$

and

$$\begin{aligned} \bar{\theta} &= \arg\min_{\bar{\theta^i}} \ln P(D^{'}|\bar{\theta^i}) \\ &= \arg\max_{\bar{\theta^i}} - \ln P(D^{'}|\bar{\theta^i}) \end{aligned} \tag{5}$$

where $D^{'} = \{ \; \bar{x_1^{'}}, \bar{x_2^{'}}, ..., \bar{x^{'}_{N'}} \; \}$ is the set of negative examples and $N^{'}$ is the number of negative examples.

We combine (4) and (5) as

$$\bar{\theta} = \arg\max_{\bar{\theta^i}} \{\ln P(D|\bar{\theta^i}) - \ln P(D^{'}|\bar{\theta^i})\} \tag{6}$$

(6), however, requires that the data be balanced. This is not always the case. To deal better with imbalanced data a regularization parameter is introduced that removes the imbalance in data automatically. Our new objective function to be maximized then becomes

$$l(\theta) = \alpha \; \ln P(D|\bar{\theta}) - (1-\alpha) \; \ln P(D^{'}|\bar{\theta}) \tag{7}$$

where $0 < \alpha \leq 1$ is a regularization parameter (a tighter lower bound for α is discussed later). This eliminates the imbalance between positive and negative data samples of a class and the consequent bias.

C. Re-Estimation of Parameters

We now maximize $l(\bar{\theta})$ in (7)

$$\frac{\partial l}{\partial \bar{\mu_k}} = 0, \frac{\partial l}{\partial \Sigma_k} = 0, \frac{\partial l}{\partial \pi_k} = 0$$

and obtain the following set of equations to re-estimate parameters

$$\bar{\mu_k} = \frac{\alpha \; \sum_{n=1}^{N} \gamma_{nk} \bar{x_n} - (1-\alpha) \; \sum_{n=1}^{N^{'}} \gamma^{'}_{nk} \bar{x^{'}_n}}{\alpha \; N_k - (1-\alpha) \; N^{'}_k} \tag{8}$$

$$\Sigma_k = \{\alpha \; \sum_{n=1}^{N} \gamma_{nk} (\bar{x_n} - \bar{\mu_k})(\bar{x_n} - \bar{\mu_k})^T - $$

$$\frac{(1-\alpha) \sum_{n=1}^{N^{'}} \gamma^{'}_{nk} (\bar{x'}_n - \bar{\mu_k})(\bar{x^{'}_n} - \bar{\mu_k})^T\}}{\alpha \; N_k - (1-\alpha) \; N^{'}_k} \tag{9}$$

$$\pi_k = \frac{\alpha \; N_k - (1-\alpha) N^{'}_k}{\alpha \; N - (1-\alpha) \; N^{'}} \tag{10}$$

where

$$\begin{aligned} \gamma_{nk} &= \frac{P(\bar{x_n}|\bar{\mu_k}, \bar{\Sigma_k})}{\sum_{j=1}^{K} \; P(x_n|\bar{\mu_j}, \Sigma_j)} \\ \gamma^{'}_{nk} &= \frac{P(\bar{x^{'}_n}|\bar{\mu_k}, \Sigma_k)}{\sum_{j=1}^{K} \; P(\bar{x^{'}_n}|\bar{\mu_j}, \Sigma_j)} \\ N_k &= \sum_{n=1}^{N} \gamma_{nk} \\ N^{'}_k &= \sum_{n=1}^{N^{'}} \gamma^{'}_{nk} \end{aligned}$$

The parameters in (8), (9) and (10) are estimated in an iterative fashion. The initial values for these parameters can be assigned in several ways. K-Means [11] clustering and the LBG algorithm [12] are common choices for initializing the model parameters. Once the model is initialized, the parameters are re-estimated using the above equations until the algorithm reaches a convergence criterion that there is no further increase in the likelihood of the data.

D. Choosing α

The choice of the regularization parameter, α, is critical as it determines the extent to which the negative samples influence the model-building process. It can be noticed that when α is 1, the technique reverts back to conventional GMM approach. α has to be chosen dynamically based on the amount of data available.

A tighter lower bound for α can be obtained using the condition that all π_k values are non-negative and normalized. That is, by applying $\sum_{k=1}^{K} \pi_k \geq 0$, we get

$$\alpha > \frac{N^{'}}{N + N^{'}} \tag{11}$$

Even with (11), there is a possibility that some cluster weights can become negative. This is because, the inequality does not depend on the effective number of samples (N_k and $N_k^{'}$) belonging to each cluster. We define an even tighter bound by constraining α such that

$$\alpha > (\frac{N_k^{'}}{N_k + N_k^{'}}) \tag{12}$$

$k = 1, 2, ..., K$

From (12), we derive

$$\alpha > \max_k(\frac{N_k^{'}}{N_k + N_k^{'}}) \tag{13}$$

Equation (13) ensures that the cluster weights and variance values are non-negative.

E. Speaker Identification

Given a test utterance $D_T = \{\bar{x}_1^T, \bar{x}_2^T, ..., \bar{x}_t^T\}$, a class label, the identity of the speaker, is assigned according to the Bayes' decision rule

$$i = \arg\max_j P(j|D_T) \tag{14}$$

where i and j are class labels. The Baye's decision rule ensures that the risk involved in classifying is minimal [11].

F. Speaker Verification

The speaker verification task involves validating a claim attached to a test utterance. The utterance is tested against the claimed speaker's model and the score obtained is compared to a threshold for the claim to be validated. Practically, this task is not simple because of the natural variability in the scores. This encourages the need to normalize the scores. Several score normalization techniques have been proposed [13]. T-Norm is a commonly used technique. Given a score, T-Norm normalization is given as

$$S = \frac{\log(P(\theta|D_T)) - \mu_I}{\sigma_I} \tag{15}$$

where

S : normalized score
θ : model corresponding to the claim
μ_I : mean of scores obtained from the cohort set
σ_I : standard deviation of scores obtained from the cohort set

III. Experimental Results

The new approach discussed in the previous section was tested for it's accuracy in speaker identification on a subset of 200 speakers from the NTIMIT database [14] and on a subset (all 149 male speakers) of NIST SRE 2003 corpora [15].

In the NTIMIT subset chosen for experimentation, each speaker had 10 utterances in the database. Six utterances were used for training; all utterances in sa#, si# and one utterance from sx# were used training. The rest of the utterances in sx# were tested on the models built. For the NIST SRE 2003 corpora chosen, as mentioned earlier, the entire male data set was chosen for experimentation. Further, speaker verification experiments were also conducted on the NIST SRE 2003 data set.

A. Baseline system

Traditional GMM-based speaker identification system was used as the baseline system. MFCC features with 23 cepstral co-efficients per feature vector were extracted. The utterances were bandlimited in the range (100 - 4000 Hz). 32-mixture GMMs were built and the models were tested against the test utterances for each of the speakers. The system had a classification accuracy of 41.375% on NTIMIT database, while a classification accuracy of 45.04% was obtained on NIST SRE 2003 database. The baseline system's performance on NTIMIT does not compare to that mentioned in [2] due to the differences in composition of training and test data.

B. Proposed System

Models built using the proposed approach had the same training data as the baseline system. A critical step while building the models was setting a value for α. Clearly, the effective number of positive examples per cluster (N_k) and it's negative counterpart ($N_k^{'}$) need to be balanced. Thus, α was evaluated for each cluster, and their maximum was chosen for the entire training process for each speaker (refer section II.D for details). This is essential in order for the variances and cluster weights to remain positive.

The choice of negative examples for a speaker play an important role in the training process. The entire of the rest of training data was chosen as anti-speakers (199 speakers for each speaker in NTIMIT database and 148 speakers for each speaker in NIST SRE 2003 data set) for the intial set of negative examples. This was dynamically filtered by applying a threshold on the probability that a negative sample belongs to a Gaussian mixture of the class. This was necessary as a larger amount of negative samples could induce unintended movement of cluster means. Also, farther the negative samples are from the class, lesser their likelihood; obviating a necessity to include them in the negative data set. Negative data were

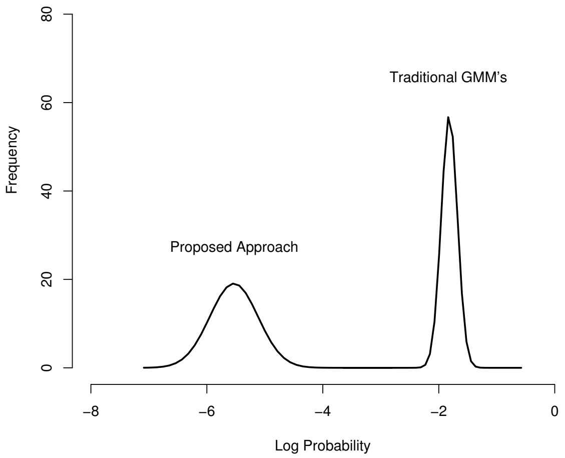

Fig. 1. *Comparison of histograms of (averaged) log probabilities for test utterances from anti-speakers for a speaker i (in the NTIMIT database) between the traditional and the proposed approach. The distribution from proposed approach is in the left and to it's right is that from the traditional GMM-based approach.*

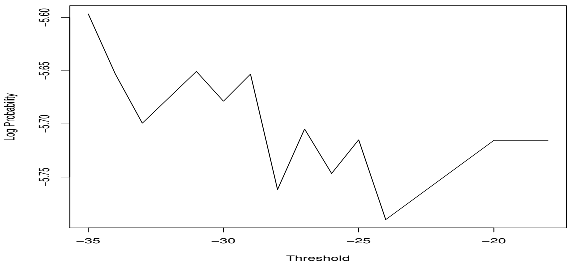

Fig. 2. *Illustration of changes in distortion values (log probabilities) with threshold for a speaker for a single utterance (from an anti-speaker)*

chosen such that their likelihood corresponding to the class was high. Again, the value of this threshold is critical as it influences the regularization parameter, α. A larger threshold would induce a bias on cluster means towards the negative samples, while a smaller distance threshold will include very little negative data.

C. Analysis

Figure (1) shows a probability density histogram of average log probabilities of speaker j ($j = 1, 2, ..., 200$ and $j \neq i$) with speaker models of i built with the NTIMIT data set. It shows decrease in probability values of test utterances from anti-speaker , $j (\neq i)$, compared to that in the traditional GMM approach. It was observed that the distortion values of test utterances of speakers closer to speaker i in the feature space decreased more with respect to i than for others. Such a decrease in probability value implies that the possibility of a test utterance being classified to speaker i that does not belong to him/her is lesser.

1) Effect of Threshold: As mentioned earlier, a large negative data set might influence unnecessary movement of cluster means. Thus a need for subset selection from the negative data set arises. This was done by constraining a threshold on the probability that a negative sample belongs to the class. Figure 2 shows the effect of the threshold on distortion values of test utterances. The threshold and it's effect on the classification accuracy are, of course, dependent on the data set at hand as the range of applicable threshold values could differ based on the data set and the type of distortion measure applied.

The value of the threshold was obtained empirically by performing a line search in the range of -39 to -15 (where each threshold value denotes log probabilities) in steps of 1. Figure 2 shows how the distortion value changes with respect to threshold for one particular speaker (in the NTIMIT database) and a fixed test utterance (not belonging to the speaker). The Y-axis corresponds to the log probability that a data point belongs to a cluster. Clearly, from the figure, one can observe that the best threshold corresponds to a local minimum. The choice of the threshold consequently affects the number of negative examples chosen to discriminate while training. A large number of examples can introduce an undue bias and can hurt the system's performance. Therefore, a judicious choice of the set of negative examples must be made.

We also tested a dynamic threshold selection strategy where the thresholds were set at cluster levels. For each cluster, we selected negative data whose responsibility values ($\gamma_{nk}^{'}$) for that cluster were at most the maximum of responsibility values of the positive data (γ_{nk}) in that cluster. This selection strategy showed reasonable performance improvements, too.

2) Results: The baseline system's accuracy is given in Table I. The results of the system built using the proposed approach at some of thresholds are given in Table II (NTIMIT) and Table III (NIST SRE 2003).

TABLE I
Baseline performances of the traditional GMM system

Database	Classification Accuracy (%)
NTIMIT (200 speakers)	41.375
NIST SRE 2003 male speakers	45.04

TABLE II
Summary of results of the proposed approach on NTIMIT database for some threshold values (distortion values are log probabilties)

Distortion value	Classification Accuracy (%)
-24.0	42.00
-25.0	42.875
-26.0	42.75
-27.0	42.25
-28.0	41.75
Dynamic threshold	42.625

A significant performance gain was observed after training speaker models discriminatively although the results are highly influenced by the subset selection strategy on the negative data set.

Further expermentation was conducted to observe the influence of the proposed discriminative training approach on

TABLE III

Summary of results of the proposed approach on NIST SRE 2003 database for some threshold values (distortion values are log probabilties)

Distortion value	Classification Accuracy (%)
-27	45.5696
-28	46.3887
-29	46.2398

speaker verification tasks. The NIST SRE 2003 database's subset used in speaker identification task previously was again used for this purpose. The baseline system was once again the traditional GMM system that was used in the speaker identification task previously. Additionally, a UBM-GMM system (as mentioned in [3]) was developed on the NIST data set. 1024-mixture UBM was built from the entire training data set and speaker models were adapted from it. All means, covariances and mixture weights were adapted. This system's performance was also compared with that of the proposed sytem.

Evaluation key in the NIST SRE 2003 database was used to generate claims for each utterance in the test set. T-Norm based score normalization was performed on the likelihood scores obtained for each test utterance. A cohort set of 50 speakers was chosen for this purpose. The performances of the baseline, proposed and the UBM-GMM system were compared using DET curves [16]. Figure 3 compares the DET curves of these systems. Clearly, the proposed approach shows improvements over the baseline system and matches the performance of UBM-GMM system. It can further be inferred that at higher false alarm probabilities the proposed approach outperforms the UBM-GMM system. It should be emphasized that the proposed approach used only 32-mixutre models while models in the UBM-GMM framework had 1024 mixtures. Consequently, the proposed system required much lesser computation time than the UBM-GMM approach.

IV. Conclusion and Future Work

The discriminative training approach proposed for modelling speaker data using GMMs performs better than conventional GMM-based speaker recognition systems for appropriate choices of regularization parameters. The choice and the amount of negative examples used while training the speaker is critical in determining the model's performance in the system. In this approach, negative data samples that were close to the speaker model were chosen using two different strategies (a) static threshold and (b) dynamic threshold selection. Both selection strategies displayed performance improvements.

The proposed approach can also be brought under UBM-GMM framework of discriminative training. Further analysis of such a methodology will be the focus of future work.

References

[1] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. on Speech and Audio Processing*, vol. 3, pp. 72–83, January 1995.

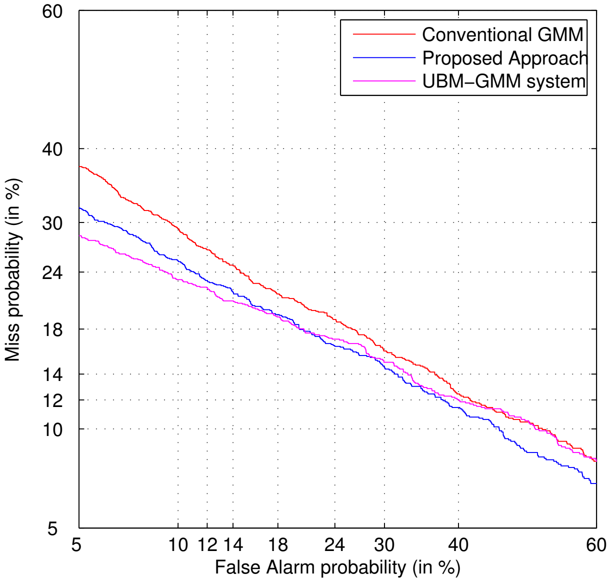

Fig. 3. *DET curves obtained from speaker recognition task on NIST SRE 2003 database*

[2] D. A. Reynolds, "Large population speaker identification using clean and telephone speech," *IEEE Signal Processing Letters*, vol. 2, pp. 46–48, March 1995.

[3] 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.

[4] R. Zheng, S. Zhang, and B. Xu, "Text-independent speaker identification using GMM-UBM and frame level likelihood normalization," *Proc. ICSLP*, pp. 289–292, December 2004.

[5] C. S. Liu, C.-H. Lee, W. Chou, B.-H. Juang, and A. E. Rosenberg, "A study on minimum error discriminative training for speaker recognition," *J.Acoust.Soc.Am*, vol. 97, pp. 637–648, January 1995.

[6] S. Kwong, Q. H. He, K. F. Man, and K. S. Tang, "Improved maximum model distance for HMM training," *Pattern Recognition*, vol. 33, pp. 1749–1758, 2000.

[7] S. Kwong, Q. He, K. Man, and K. Tang, "A Maximum Model Distance approach for HMM based speech recognition," *Pattern Recognition*, pp. 219–229, 1998.

[8] Q. Hong and S. Kwong, "Discriminative training for speaker identification based on maximum model distance algorithm," *Proc. ICASSP*, pp. 25–28, 2004.

[9] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support Vector Machines using GMM Supervectors for Speaker Verification," *IEEE Signal Processing Letters*, vol. 13, pp. 308–311, 2006.

[10] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood for Incomplete Data via the EM algorithm," *Journal of Royal Statistical Society*, vol. 39, pp. 1–38, 1977.

[11] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. Wiley India, 2007.

[12] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Trans. on Communications*, pp. 84–94, 1980.

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

[14] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "NTIMIT: A phonetically balanced, continuous speech telephone bandwidth speech database," *Proc. ICASSP*, pp. 109–112, 1990.

[15] "The NIST year 2003 speaker recognition evaluation plan," 2002. [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/sre/2003/index.html

[16] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," *Proc. EUROSPEECH*, pp. 1895–1898, 1997.