ON LINE CLIENT-WISE COHORT SET SELECTION FOR SPEAKER VERIFICATION USING ITERATIVE NORMALIZATION OF CONFUSION MATRICES

Srikanth Nagineni and Rajesh M Hegde

Department of Electrical Engineering
Indian Institute of Technology, Kanpur
{*srikanth,rhegde*}*@iitk.ac.in*

ABSTRACT

T-normalization is a widely used method for normalizing the scores in a speaker verification system in order to reduce undesirable variation arising from acoustically mismatched conditions. In this paper we propose a particular form of T-normalization using iterative normalization of confusion matrix generated from impostor trials for each client speaker. The normalized confusion matrix along with a simple distance metric is then used to select a cohort set based on similarity modeling for each client speaker. The normalization statistics thus computed from this cohort set is used for both impostor and claimant scoring. Experiments on the NIST 2004 SRE data demonstrate reasonable improvements in terms of the equal error rate(EER) computed from the detection error trade(DET) curves, when compared to the baseline GMM-UBM schemes. Encouraging improvements in terms of DCF over the general T-normalization schemes are also illustrated for 8C-1C and 1C-1C conversation conditions.

1. INTRODUCTION

Speaker verification [1], is a task of identifying whether an unknown speech utterance was uttered by a claimed speaker or not. The general approach used in the speaker verification system is to apply a likelihood ratio test to an input utterance to determine if the test claimed speaker is accepted or rejected. The likelihood ratio essentially measures how much better claimant model scores for the test utterances compared to some non claimant model. The decision threshold is then set to adjust the trade off between rejecting true claimant utterances(false rejection errors) and accepting non claimant utterances(false acceptance errors). An important issue in the statistical approaches to speaker verification is that of score normalization which is used to reduce environmental and variability effects on the verification decision. Hence finding an optimal strategy for classification and score normalization is a significant problem in speaker verification systems. Both offline [2], and on line [3], methods have been used earlier for impostor cohort set selection for further use in the T-normalization process. In this paper we propose an approach which uses normalized confusion matrices for each claimant speaker which is herein called the client speaker. For each client speaker a normalized confusion matrix is generated using the iterative proportional fitting (IPF) procedure [4], in multiple passes. A simple distance metric [5], is then applied on the confusion matrix at each pass to find the most similar set of impostors to the client speaker. Once the most similar set of speakers are found, the initial (conventional) cohort model set is pruned based on the selected number of impostor speakers. Only these set of cohort models are then used in computing the statistics for T-normalization of claimant and impostor scores during the final testing phase. We first briefly discuss the issue of score normalization in speaker verification. We then propose the overview of the new approach of selecting the cohort set to be used in the final T-normalization process. This is followed by a discussion on the use of IPF based normalized confusion matrices for each client speaker to come up with the pruned cohort set. Speaker verification experiments conducted on the NIST 2004 SRE data [6], are then discussed. Reasonable improvements over the conventional GMM-UBM modeling and the existing T-normalization methods are also illustrated as DET curves [7]. The results are further substantiated using equal error rate (EER) and decision cost functions (DCF) on the 8C-1C and 1C-1C conversation sides of the NIST 2004 SRE corpus. A Bayesian interpretation of the proposed approach is also described in Appendix I.

2. SCORE NORMALIZATION TECHNIQUES

The GMM-UBM system [1], is a likelihood-ratio detector in which the likelihood ratio is computed for an unknown test utterance Y_{test} between a speaker-independent acoustic distribution (UBM) and a speaker-dependent acoustic distribution i.e client (claimed) speaker S. The general block diagram of speaker verification system illustrating score normalization is shown in figure1. Score normalization technique is used to normalize the log likelihood ratio score for a test utterance Y and target model S as

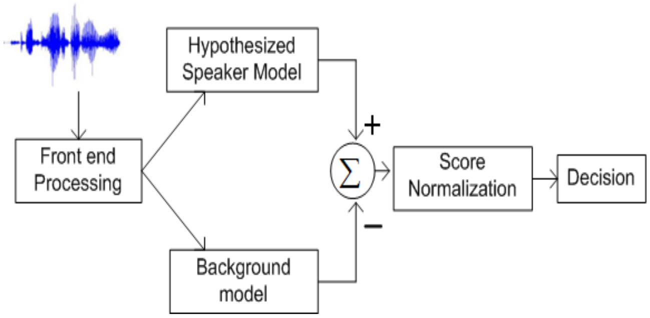

Figure 1: Block diagram of a speaker verification system illustrating score normalization.

$$LLR(Y_{test},S)_{norm} = \frac{LLR(Y_{test},S) - \mu}{\sigma} \tag{1}$$

© EURASIP, 2010 ISSN 2076-1465

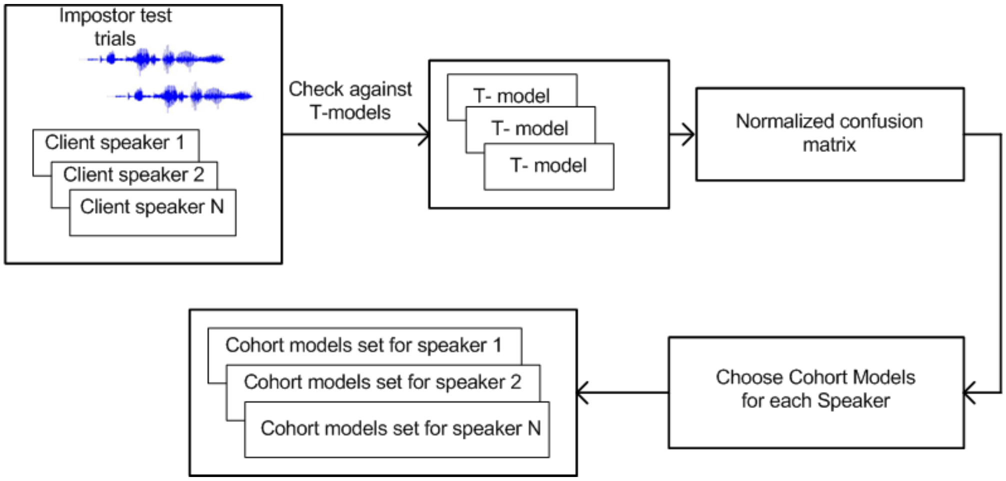
Impostor test trials
Client speaker 1
Client speaker 2
Client speaker N
Check against T-models
T- model
T- model
T- model
Normalized confusion matrix
Choose Cohort Models for each Speaker
Cohort models set for speaker 1
Cohort models set for speaker 2
Cohort models set for speaker N

Figure 2: Block diagram illustrating the proposed client-wise cohort selection and normalization (CWCS-NORM) method.

where (μ) is the mean and (σ) is the standard deviation as computed from the cohort models. Different score normalization techniques can be performed either with respect to the speaker model or with respect to the test data. Techniques like Z-NORM and T-NORM [2] are widely used in this context. While T-norm sets use some broad speaker-specific information, data driven T-norm selection approaches have received little attention. We therefore propose an approach to cohort set selection for score normalization which is essentially data driven albeit by the use of normalized confusion matrix for each speaker. The discussion on the development of this approach for client-wise cohort set selection for T-normalization ensues in the following Section.

3. ON LINE CLIENT-WISE COHORT SET SELECTION USING NORMALIZED CONFUSION MATRICES

In this Section, we present an overview of the proposed technique for the on line selection of the cohort set for each client speaker. In this approach we select different set of cohort models for each client speaker model. Note that we are using the term client model in place of claimed or hypothesized model as is generally used in speaker verification terminology. If the cohort models are chosen during the test phase (on line), the selected speakers will be more meaningfully similar depending on the test utterance. This leads to a more efficient score normalization when compared to conventional offline T-norm score normalization using a fixed set of impostor models for all test utterances. The block diagram illustrating the selection of cohort models using the proposed approach is shown in Figure 2. In order to select the best cohort models for each client speaker we start with the impostor trials of each client speaker and test it against all the cohort models. A confusion matrix is generated and normalized using the iterative proportional fitting procedure (IPF) [4], which is described in detail in the succeeding Section. A simple distance metric is used in multiple passes to select the most closest impostor cohort set until convergence in terms of a adaptive threshold is met. After the selection, each client model will have its own set of cohort models for each target speaker. This on line cohort set selection method for each client speaker leads to better system performance compared to conventional cohort model selection which do not perform any similarity modeling and use the same cohort set for all test utterances.

3.1 Normalizing Confusion Matrices via iterative proportional fitting

In this Section we describe the use of confusion matrices for selecting the most similar cohort set of models each client speaker. A confusion matrix lists the values for known types of the reference data in the columns and for the classified data in the rows. The columns indicate actual data of the reference classes while rows indicate the classifications that result from using a specific classifier. The main diagonal of the matrix lists the correctly classified data. The Table1, represents a $3x3$ confusion matrix for the classes A,B, and C. In case of finding the best cohort set for each target speaker, the confusion matrix is the client models versus the cohort models. From the table1 we observe that the cell entries need to be converted to probability values to make it more convenient to compare each cell value irrespective of the number of test examples used to derive the confusion matrix. The process of normalization will balance each cell value in the matrix by its corresponding row and column. This will ensure that both the cross classification errors and the row-column ambiguity are taken care of unlike conventional approaches. Iterative Proportional Fitting (IPF) [4] algorithm, can be used to normalize the confusion matrices, by estimating cell probabilities in a confusion matrix by forcing each row and column sum to one. Suppose there are $n_{ij} > 0$ observations in a confusion matrix$(r \times c)$, where

	A	B	C
A	6	2	1
B	1	7	1
C	2	2	5

Table 1: An example of a $3x3$ confusion matrix.

$$\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} = n. \qquad (2)$$

The cell probabilities are estimated by minimizing the following criterion

$$\sum_{i=1}^{r}\sum_{j=1}^{c}(n_{ij}-np_{ij})^2/n_{ij} \tag{3}$$

assuming the following fixed marginal totals.

$$p_{i+}=\sum_{j=1}^{c}p_{ij}(i=1,2,....,r) \tag{4}$$

$$p_{+j}=\sum_{i=1}^{r}p_{ij}(j=1,2,....,c) \tag{5}$$

where $p_{i+}=1$ and $p_{+j}=1$. An involved discussion on the IPF procedure can be found in [4]. To illustrate IPF, we consider an example normalized confusion matrix of three classes (A,B, and C) as shown in Table 2. From Table 2 it can be noted that IPF alleviates the differences in test examples and also makes the off diagonal values more indicative of the cross classification errors.

	A	B	C
A	0.6771	0.1314	0.1925
B	0.1685	0.6875	0.1439
C	0.1554	0.1811	0.6635

Table 2: An example normalized confusion matrix computed using IPF.

3.2 Cohort set selection using a distance metric

In order to compute the cohort set from normalized confusion matrices a distance metric that quantifies the similarity in terms of the cell probabilities values is required. The L1 distance measure [5] is used in our work. The L1 distance measure totals the absolute differences of corresponding coordinate values between two vectors $(rows)$, then takes a pair wise similarity between all pairs of classes to prepare a upper triangular matrix, called the incidence matrix.

$$L1_{ij}=\sum_{k=1}^{c}abs(a_{ik}-a_{jk}); i\neq j;$$

$$L1_{i,j}=0; i\geq j$$

where $1\leq i\leq r, 1\leq j\leq c$ and $L1_{ij}$ are the elements in incidence matrix L1. $L1_{ij}=0$: $i\geq j$ to get a upper triangular incidence matrix. The incidence matrix for L1 measure using the normalized confusion matrix in Table 2 is. The confusion matrices in conjunction with the distance metric is used in multiple passes if necessary to find a cohort set for each client speaker. Fixing the small fractional probability threshold is completely data driven and is based on a convergence measure in terms of the cell probability values. It must be noted here that the row wise candidates of the confusion matrix are the claimed identities. Hence the possibility of pruning one closest speaker in the selection of the cohort set is possible. However possiblitiy of such errors are larger in other on line cohort set selection [3], techniques.

0.0000	1.1124	1.0415
0.0000	0.0000	1.0392
0.0000	0.0000	0.0000

Table 3: Incidence matrix corresponding to the normalized confusion matrix in Table 2 using the L1 distance measure.

4. PERFORMANCE EVALUATION

The baseline GMM-UBM system used in the speaker verification experiments has already been illustrated in Figure 1 A Gaussian mixture model (GMM) is trained using pooling data from many different speakers to create a universal background model(UBM). The target speaker models are trained by maximum a posteriori(MAP) adaptation of the background model to the training data. For a given test sample, the accumulated and averaged log likelihood ratio for the target model and the background model is used as a score. The features used for the experiments in this system are the thirteen dimensional Mel frequency cepstral coefficients (MFCCs), without the zeroth order coefficient, and appended with velocity and acceleration coefficient's, resulting in thirty nine dimensional feature vectors. The features are modeled by 512 mixture component GMMs. Only the GMM means are adapted to the observed data. The cohort models are also trained in a similar manner as the target models for a set of impostor models. In following Section, we describe the NIST 2002 SRE one-speaker detection task data corpus and NIST 2004 SRE Mixer data corpus used to carry out the speaker verification experiments.

4.1 Organization of the NIST 2002 and 2004 data sets for performance evaluation

Two data sets are used in this work. The first data set corresponds to the one-speaker limited data detection task of the NIST SRE 2002. This database consists of 191 female and 139 male speakers. The data is recorded using the Switchboard methodology and consists of excerpts from cellular telephone conversations. The second evaluation database used is the NIST 2004 SRE Mixture data corpus. Of the many conditions evaluated there [6], we will focus on the 8C-1C and 1C-1C conversation side conditions. In these conditions the test segments contained a whole speaker conversation side, and the model training material consisted of either one conversation side (of approximately 5 minutes) or 8 conversation sides, respectively. The data is part of the MIXER data corpus. The evaluation contains different languages and includes many trials for which the training and testing material consists of different languages. Also, there is a great variation in handset and channel type within the database. NIST 2002 SRE data was used for training the background models. For evaluation purposes NIST SRE 2004 data has been used. The SRE 2004 data was split as per gender in two separate groups of speakers. The selection of speakers was random. The first split was used for computing T-normalization statistics. The second split was used for client model training and also for testing. The UBM model was trained from the female subset of the NIST 2002 SRE one-speaker limited data set. It is modeled by a GMM with 512 mixtures. Ten iterations were sufficient for parameter convergence. T-normalization models were trained using one set of NIST 2004 female SRE data set. The T-normalization models are adapted using utterances from T-normalization speakers in a similar manner as is done with the client models. In our training and testing phases the NIST 2004 SRE 1C-1C and 8C-1C training/testing female conversation side condition set served as the data set. The client models were built by doing MAP adaptation on the UBM.

4.2 Experimental Results

Experiments on the NIST 2004 SRE data for speaker verification are conducted using four methods

- The GMM-UBM method (GMM-UBM) : In this method the conventional GMM-UBM method of speaker verification is used.
- The client-wise hypothesized set method (CWHS) : In this method we replace the hypothesized speaker model by a client-wise hypothesized (CWHS) model. This CWHS model is trained using the most similar speakers as computed from the normalized confusion matrix method as described earlier in Section 3.
- The CWHS followed by T-normalization method (T-NORM) : This method follows the CWHS approach followed by the T-normalization method.
- The client-wise cohort set selection followed by T-normalization method (CWCS-NORM) : In this method we use the aforementioned client-wise cohort set selection method followed by T-normalization. Note that a common set cohort set is not used for score normalization (T-NORM) as is done in conventional T-normalization schemes.

The experimental results obtained with the four methods are given as detection error trade off (DET) curves and also illustrated in terms of the decision cost function (DCF). The detection error trade-off (DET) plots for the NIST 2004 SRE for the 8-conversation side training with 1 conversation side test (8C-1C) condition is shown in Figure 3. Similarly the DET plots for 1-conversation side training condition with 1-conversation side test(1C-1C) condition is shown in Figure 4. From Figures 3 and 4, we observe from the DET curves that the proposed CWCS-NORM method shows reasonable improvements in both the 8C-1C and 1C-1C conditions. In Figures 5 and figure6, the corresponding equal error rate (EER) and decision cost function (DCF) values for all the four methods under consideration are illustrated. It can be observed that the proposed CWCS-NORM method gives consistently reasonable improvements over other methods in terms of both EER and the DCF. It can also be noted that the conventional T-NORM method displays characteristic improvement in the low false-alarm region and nominal gain at the EER point over CWHS. In contrast, the CWCS-NORM method shows improvement across the entire false alarm region when compared with all the other methods.

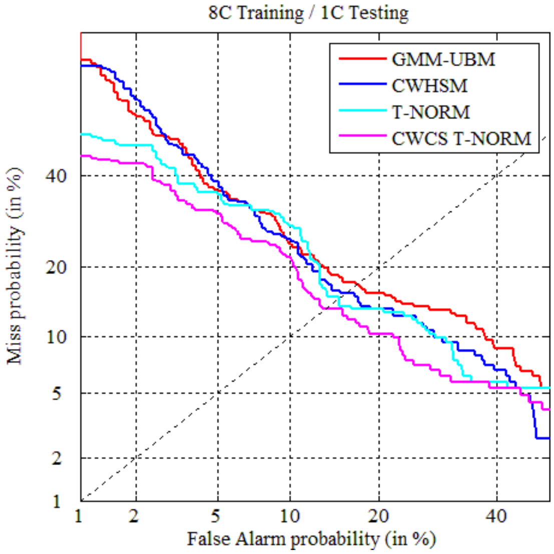

Figure 3: DET plots for the NIST speaker-recognition evaluations for the 8-conversation-side training condition with 1 conversation side test for all the four methods used for comparison.

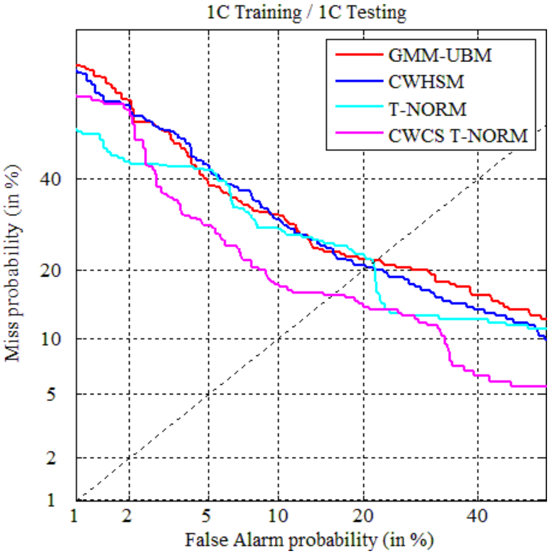

Figure 4: DET plot for the NIST speaker-recognition evaluations for the 1-conversation-side training condition with 1 conversation side test.

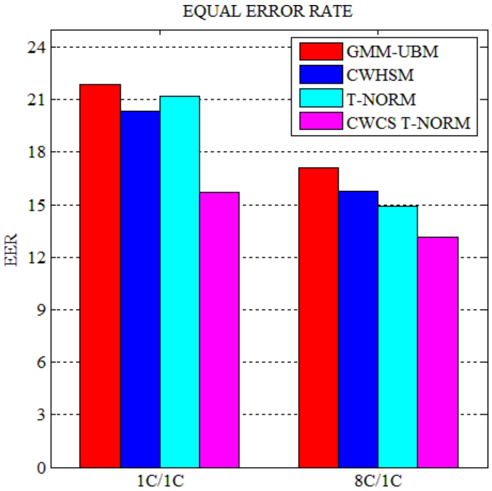

Figure 5: Bar chart of the EER for 1C/1C and 8C/1C conditions on the NIST 2004 SRE data.

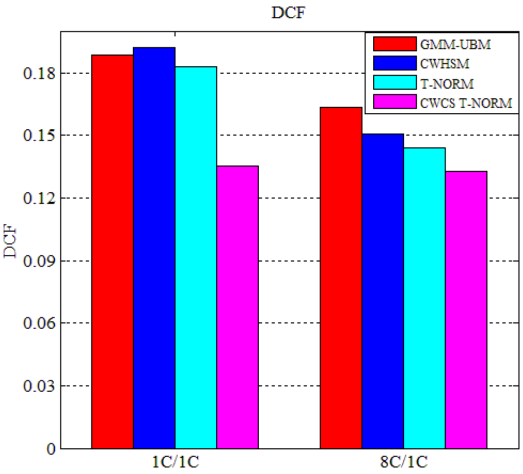

Figure 6: Bar chart of the DCF for 1C/1C and 8C/1C conditions on the NIST 2004 SRE data.

5. CONCLUSION

In this paper a new approach to cohort model selection depending on the claimed or client speaker is presented. The advantage of this approach is that the same set of cohort models are not used for all test utterances. This approach is also on line i.e, the cohort set is selected based on the test utterance. The primary contribution of this work is also a new method of similarity modeling via the iterative use of normalized confusion matrices. The experimental results on the NIST 2004 SRE data are encouraging. A Bayesian interpretation of the approach is also presented in Appendix I, to illustrate the motivation for the use of confusion matrices in a a multi pass fashion. We are currently working on improving the real time performance of the proposed approach.

6. APPENDIX I : BAYESIAN INTERPRETATION OF THE PROPOSED SIMILARITY MODELING FOR COHORT SET SELECTION

Similarity modeling can be used to select the best cohort set for each client speaker involves selecting the the cohort models that are closest to the test utterance in question. This can be viewed as a classification problem. Using a Bayesian approach for this problem one can proceed as follows. Let $\omega_l, l = 1,..,N$ be the N classes in the classification problem. Assuming $C_m, m = 1,..,L$, denote the classes in the initial classification pass where $L < N$, the initial classes C_m form a partition of the original classes ω_l, as $C_m = [\omega_{m1}, \omega_{m2}, .., \omega_{mk}]$ and the classes C_m are mutually disjoint. A diagrammatic illustration of this is shown in Figure 7. When a Bayesian classifier is used, we start by computing class likelihood given the feature vector X as

$$P(\omega_l \mid X) = \sum_{m=1}^{L} P(\omega_l, C_m \mid X) \qquad (6)$$

Applying Bayes rule on Equation 6 reduces it to

$$P(\omega_l \mid X) = \sum_{m=1}^{L} \frac{P(X \mid C_m)P(C_m)}{P(X)}(P(\omega_l \mid C_m, X)) \qquad (7)$$

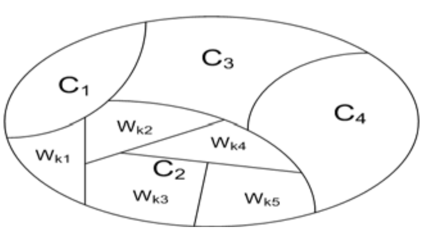

Figure 7: Illustration of the similarity modeling approach for cohort set selection.

Further assuming that $\omega_l \in C_i$,

$$P(\omega_l \mid C_m, X) = \left\{ \begin{array}{cc} 0; & m \neq i \\ p, p > 0; & m = i \end{array} \right\}$$

This assumption also implies that the first pass has to be designed to be robust. With this assumption

$$\begin{aligned} P(\omega_l \mid X) &= \tfrac{P(X|C_i)P(C_i)}{P(X)}(P(\omega_l \mid C_i, X)) && (8) \\ &= \tfrac{P(X|C_i)P(C_i)}{P(X)} \tfrac{P(X|\omega_l,C_i)P(\omega_l|C_i)}{P(X|C_i)} && (9) \end{aligned}$$

Note that we have made successive use of Bayes rule and assumption that once the classification decision on C_i is made, the finer classification decides on w_l. Further once C_i is decided during the first pass, the finer classification on w_l requires computing

$$P(X \mid \omega_l, C_i)P(\omega_l \mid C_i) \qquad (10)$$

This ensures that if the first pass classification is assumed to be robust, the next pass has the advantage of working with smaller and manageable class sizes.

REFERENCES

[1] F. Bimbot, J.F. Bonastre, et al., ”A Tutorial on Text-Independent Speaker Verification”,EURASIP Journal on Applied Signal Processing, Vol. 4 (2004), pp. 430-451.

[2] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, ”Score Normalization for Text-Independent Speaker Verification System”, Digital Signal Processing 10, pp.42-54 ,2000.

[3] T. Kinnunen, E. Karpov, and P. Franti, ”Efficient on line Cohort Selection Methods for Speaker Verification”, in Proc. INTERSPEECH 2004, Jeju Island, Korea, Oct. 4-8, 2004, Vol. III, pp. 2401-2402.

[4] S. E. Fienberg, An Iterative Procedure for Estimation in Contingency Tables. ”Annals of Mathematical Statistics”, Vol. 41, No.3, pp. 907-917, June 1970.

[5] K. Fukunaga, Introduction to Statistical Pattern Recognition. Academic Press, Boston, 1990.

[6] NIST Speech Group, ”The 2004 NIST Speaker Recognition Evaluation Plan”, *www.itl.nist.gov/iad/mig/tests/sre/2004/*

[7] A. Martin, and G.Doddington, ”The DET Curve in Assessment of Detection Task Performance”, in Proc. EUROSPEECH, 1997, Vol.4, pp. 1895-1898.