UNIVERSITY OF JOENSUU
DEPARTMENT OF COMPUTER SCIENCE

Report Series A

The Mystery of Cohort Selection

Tomi Kinnunen and Ismo Kärkkäinen
and Pasi Fränti

Report A-2005-1

ISBN 952-458-676-2
ISSN 0789-7316
ACM I.2.7, I.5.1, I.5.2, I.5.4

Joensuun yliopistopaino
Joensuu
2005

The Mystery of Cohort Selection

Tomi Kinnunen, Ismo Kärkkäinen, Pasi Fränti
Speech and Image Processing Unit, Department of Computer Science
University of Joensuu, Finland

Abstract

In speaker verification, *cohort* refers to a speaker-depended set of "anti-speakers" that are used in match score normalization. A large number of heuristic methods have been proposed for the selection of cohort models. In this paper, we use genetic algorithm (GA) for minimizing a cost function for a given security-convenience cost balance. The GA jointly optimizes the cohort sets and the global verification threshold. Our motivation is to use GA as an analysis tool. When comparing with heuristic selection methods, GA is used for obtaining a lower bound to error rates reachable by MFCC-GMM verification system. On the other hand, we analyze the models selected by GA, attempting to gain understanding into how cohort models should be selected for an application with given security-convenience tradeoff. Our findings with a subset of the NIST-1999 corpus suggest that in user-convenient application, the cohort models should be selected more close to the target than in secure application. The lower bounds in turn show that that there is a lot of room for further studies in score normalization, especially in the user-convenient end of the detection error tradeoff (DET) curve.

1 Introduction

Speaker verification [1] is the task of deciding whether a given speech utterance was produced by a claimed person (*target*). In biometric verification, two errors are possible: *false acceptance* (FA) and *false rejection* (FR). The former means accepting an impostor, and the latter refers to rejecting a genuine speaker. By adjusting the verification threshold, the system administrator can balance between the error types. By lowering the threshold, the number of false rejections can be reduced ("user-convenient" applications), but with the cost of increased number of false acceptances. By setting a high threshold, the number of false acceptances can be reduced ("secure" application).

In state-of-the-art verification systems, the features extracted from the unknown speaker's utterance are matched against the target and nontarget models. *Normalized score* [2, 3, 4, 5] is a function of the two scores, and it is compared with the verification threshold. The rationale is to make the match score relative to other models so that it is more robust against acoustic mismatches between

training and recognition. Setting of speaker independent verification threshold becomes also easier because the scores are in common range.

The nontarget hypothesis represents the possibility that anyone else expect the target produced the unknown utterance. Thus, in principle the nontarget model should be composed of all possible speakers. Two popular approaches for approximating the nontarget likelihood are *world modeling* [5] and *cohort modeling* [6, 3, 7, 8, 9, 4, 10, 11], see Fig. 1. The world model, or *universal background model* (UBM), represents "the world of all possible speakers", and it is represented by a single model, which is same for all speakers. In the cohort approach, nontarget likelihood is approximated using a small number of speaker-depended "antispeakers", called the *cohort set* of the speaker.

The UBM normalization is straightforward and computationally efficient, but there are two motivations to study cohort selection more closely. Firstly, since the normalization depends on the speaker, it can change speaker rankings and could be also applied in the identification task (1:N matching); the UBM normalization does not help in this because the match scores are scaled by the same number. The second motivation comes from the field of forensic speaker identification [12]. In forensic cases, the acoustic evidence must be contrasted against a relevant background population (e.g. speakers of same gender and dialectal region) to estimate the likelihood of a random match. Cohort selection could be applied to find the background population automatically from a database of several thousands of speakers.

In addition to the verification threshold, the selection of cohort models has influence on the accuracy. Traditionally, the balancing between FA/FR errors has been tackled by adjusting the verification threshold. However, the FA and FR errors are functions of both the score distributions *and* the verification threshold, and therefore, should be optimized jointly when setting up the verification system for a certain application.

Our goal is to gain some insight into the selection of the cohort models for a given secure-convenience balance. We approach the problem from two directions. Firstly, we give experimental comparison of existing cohort selection methods by comparing their performance at three different operating points. Secondly, we consider the cohort selection as a combinatorial optimization problem which we attack by a genetic algorithm. Both the cohort sets and the verification threshold are jointly optimized to minimize detection cost function (DCF). In this way, we can estimate a lower bound reachable by the acoustic features and model if the cohort models would be selected optimally. We also analyze the distances of the selected cohort models to the target speaker.

The rest of the paper is organized as follows. In Section 2 we review the background of GMM-based speaker verification. In Section 3 we define the optimization problem and formula the GA for solving it. Section 4 includes experiments and discussion. Finally, conclusions are drawn in Section 5.

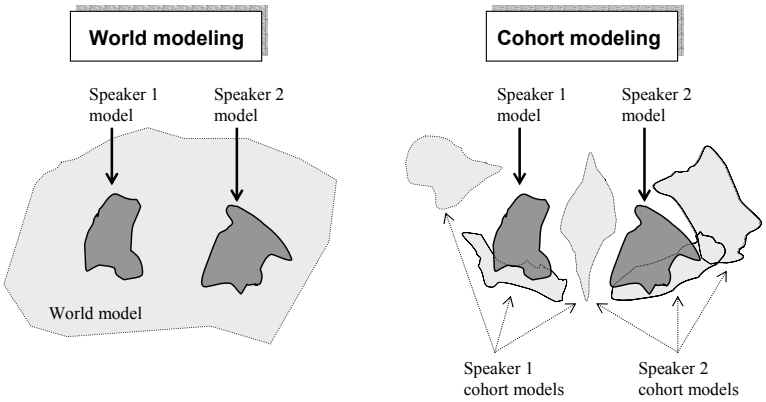


Figure 1: Illustration of the world and cohort modeling approaches.

2 Verification Background

2.1 GMM Speaker Modeling

The state-of-the-practise text-independent speaker model is the *Gaussian mixture model* (GMM) [13, 5]. GMM is well-suited for modeling of short-term spectral features like mel-frequency cepstral coefficients (MFCC) and linear predictive cepstral coefficients (LPCC) (see [14]), possibly appended with the corresponding dynamic features [15, 16].

A GMM of speaker i, denoted as $\mathcal{R}^{(i)}$, consists of a linear mixture of K Gaussian components. Its density function is

$$p(\boldsymbol{x}|\mathcal{R}^{(i)}) = \sum_{k=1}^{K} P_k^{(i)} \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k^{(i)}, \boldsymbol{\Sigma}_k^{(i)}), \tag{1}$$

where $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k^{(i)}, \boldsymbol{\Sigma}_k^{(i)})$ denotes multivariate Gaussian density function with mean vector $\boldsymbol{\mu}_k^{(i)}$ and covariance matrix $\boldsymbol{\Sigma}_k^{(i)}$. $P_k^{(i)}$ are the component prior probabilities (mixing weights) and they are constrained by $P_k^{(i)} \geq 0$, $\sum_{k=1}^{K} P_k^{(i)} = 1$.

Assuming independent and identically distributed (i.i.d.) observations $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$, the likelihood given a GMM $\mathcal{R}^{(i)}$ is

$$p(X|\mathcal{R}^{(i)}) = \prod_{t=1}^{T} p(\boldsymbol{x}_t|\mathcal{R}^{(i)}) = \prod_{t=1}^{T} \sum_{k=1}^{K} P_k^{(i)} \mathcal{N}(\boldsymbol{x}_t|\boldsymbol{\mu}_k^{(i)}, \boldsymbol{\Sigma}_k^{(i)}), \tag{2}$$

and the log-likelihood is

$$\log p(X|\mathcal{R}^{(i)}) = \sum_{t=1}^{T} \log \sum_{k=1}^{K} P_k^{(i)} \mathcal{N}(\boldsymbol{x}_t|\boldsymbol{\mu}_k^{(i)}, \boldsymbol{\Sigma}_k^{(i)}). \tag{3}$$

Usually GMM is trained with *maximum a posteriori adaptation* (MAP) from a *universal background model* (UBM) [5]. The UBM is a GMM trained from a large pool of different speakers and it is supposed to represent the distribution

of speech parameters in general. In this way, the amount of training data can be small since the parameters are not estimated from scratch. A *relevance factor* parameter is used for balancing between the background model and the new data.

2.2 Bayesian Framework

In speaker verification, we are given an input sample $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$, and an identity claim. The verification is defined as a two-class classification problem (or *hypothesis testing*) with the following possible decisions:

$$\begin{cases} \textit{Accept} \text{ identity claim, i.e. classify } X \rightarrow \text{Target} \\ \textit{Reject} \text{ identity claim, i.e. classify } X \rightarrow \text{Nontarget.} \end{cases}$$

We set nonnegative *decision costs* C_{FR} and C_{FA} for the FA and FR error types. As an example, for a high security system, we might set $C_{\text{FR}} = 1$ and $C_{\text{FA}} = 10$, i.e. accepting an impostor is ten times more costly than rejecting a true speaker. According to *Bayes' rule for minimum risk classification* [17], speaker is accepted if

$$\frac{p(X|\text{Target})}{p(X|\text{Nontarget})} \geq \frac{P(\text{Nontarget})}{P(\text{Target})} \cdot \frac{C_{\text{FA}}}{C_{\text{FR}}}, \tag{4}$$

where $p(X|\cdot)$ are the likelihoods and $P(\cdot)$ are the prior probabilities. Notice that the right hand side of (4) does not depend on X, and therefore, decision rule is of the form $l(X) \geq \Theta$, where

$$l(X) = \frac{p(X|\text{Target})}{p(X|\text{Nontarget})} \tag{5}$$

is the *likelihood ratio* and Θ is the verification threshold. Equivalently, for the log likelihood ratio, we accept speaker if

$$\log p(X|\text{Target}) - \log p(X|\text{Nontarget}) \geq \log \Theta. \tag{6}$$

The likelihood ratio concept is intuitively easy to understand: when the evidence in favor of the target hypothesis is large while the evidence for the nontarget hypothesis is small, we are confident that the speaker is the one who he claims to be. On the other hand, when $l(X) \ll 1$, we are confident that the speaker is not the claimed one, and the case $l(X) = 1$ corresponds to the most uncertain case ("no decision").

The likelihood ratio $l(X)$ is called *normalized score* as it is a relative score computed by normalizing the target score by the nontarget score. Score normalization is expected to reduce the acoustic mismatch between training and testing. When the acoustic conditions change, both the target and nontarget scores change but the relative score is expected to remain unchanged [18]. The same idea can be applied to other than likelihood scores. In addition to cohort and world modeling approaches, the scores can be normalized using impostor score distribution mean and variance [2, 4]. Some of the various background normalization methods have been compared experimentally in [19, 20, 10, 21].

2.3 World and Cohort Normalization

In the world modeling (UBM) approach, nontarget likelihood is computed using a single world model $p(X|\mathcal{R}^{\mathrm{UBM}})$. Thus, the log likelihood ratio for speaker i is simply

$$\log l(X) = \log p(X|\mathcal{R}^{(i)}) - \log p(X|\mathcal{R}^{\mathrm{UBM}}). \tag{7}$$

In the cohort approach, each speaker has a set of personal cohort[1] models which we index by $\mathcal{C}_i$. In addition to the target likelihood $p(X|\mathcal{R}^{(i)})$, we have the cohort likelihoods $p(X|\mathcal{R}^{(j)})$, where $j \in \mathcal{C}_i$. The nontarget likelihood can be approximated by applying geometric mean [7], arithmetic mean [3] or maximum [18] to the cohort likelihoods. For cohort size $M = |\mathcal{C}_i|$, the log likelihood ratios for these are given respectively by

$$\log l(X) = \log p(X|\mathcal{R}^{(i)}) - \frac{1}{M}\sum_{j\in\mathcal{C}_i} \log p(X|\mathcal{R}^{(j)}) \tag{8}$$

$$\log l(X) = \log p(X|\mathcal{R}^{(i)}) - \log \frac{1}{M}\sum_{j\in\mathcal{C}_i} p(X|\mathcal{R}^{(j)}) \tag{9}$$

$$\log l(X) = \log p(X|\mathcal{R}^{(i)}) - \max_{j\in\mathcal{C}_i} \log p(X|\mathcal{R}^{(j)}). \tag{10}$$

Different normalization approaches have been proposed e.g. in [22, 23, 24].

The world model approach is more popular because of the following reasons. Firstly, in the MAP adaptation [5], the world model is needed anyway, so it integrates into the GMM framework naturally without extra storage requirements. Secondly, there is no ambiguity in defining the normalized score, whereas the cohort approach requires selection of the cohort speakers and fixing both the normalization formula and the cohort size. However, the cohort approach is intuitively reasonable, and because of the flexibility, it is potentially more accurate.

2.4 Cohort Selection

A large number of cohort selection methods have been proposed [6, 3, 7, 8, 25, 9, 4, 26, 10, 11]. Closest speakers to the target are the most competitive ones, and they are good candidates for the cohort speakers. This approach [6, 25, 8, 27, 4, 21] is the most commonly used one, and will referred here to as the *closest impostors* (CI) method. One problem with this approach is that it prepares for impostor attacks only against "similar" speakers. However, if the impostor is dissimilar (e.g. another gender), the data will be in the tails of both target and nontarget distributions, giving rise to poorly estimated likelihood ratio [28]. Thus, the cohort should include models both from close and far from the target [3].

[1] According to *Oxford English Dictionary*, cohort was a body of infantry in the Roman army, of which there were ten in a legion, each consisting of from 300 to 600 men. In demography, cohort refers to a group of persons having a common statistical characteristic, for instance, being born in the same year.

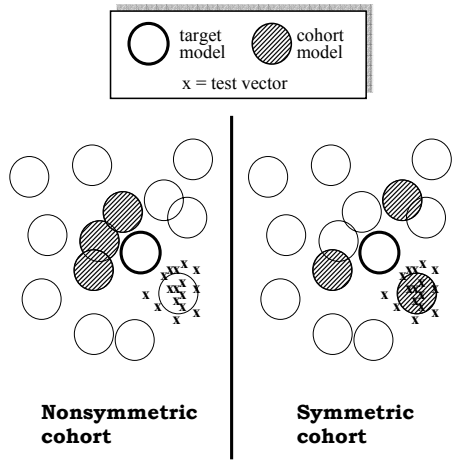


Figure 2: Problem of redundant cohort models.

If the cohort size is small, selection of redundant models should be avoided, see Fig. 2 for an illustration. Approaches presented in [3, 10] prevent adding redundant models into the cohorts. In both studies, initial cohort candidate set is first constructed, and the final cohort set is obtained by pruning out similar models [3] or by clustering them [10].

Cohort speakers are usually selected in the training phase because of computational reasons. *Unconstrained cohort selection* (UCN) that selects the competing models based on the test utterance likelihood is proposed in [8]. This method is computationally expensive, but it can be made more efficient by clustering the test sequence [11]. Usually cohort sets are composed of *full* speaker models; an alternative approach has been proposed in [9, 29], in which the impostor model is built from the individual Gaussian components of different speakers.

In the model selection algorithms, a similarity or distance measure between two GMMs is needed. Rosenberg *et al.* [6] propose the following similarity measure:

$$s(\mathcal{R}^{(i)}, \mathcal{R}^{(j)}) = \frac{1}{2}\Big\{ \log p(X_i|\mathcal{R}^{(j)}) + \log p(X_j|\mathcal{R}^{(i)}) \Big\}, \tag{11}$$

where X_i and X_j are the training data used for constructing the models $\mathcal{R}_i$ and $\mathcal{R}_j$, respectively. Reynolds [3] proposes the following divergence-like dissimilarity measure:

$$d(\mathcal{R}^{(i)}, \mathcal{R}^{(j)}) = \log \frac{p(X_i|\mathcal{R}^{(i)})}{p(X_i|\mathcal{R}^{(j)})} + \log \frac{p(X_j|\mathcal{R}^{(j)})}{p(X_j|\mathcal{R}^{(i)})}. \tag{12}$$

3 Optimization Framework

We assume that the speaker models $\mathcal{R}^{(i)}, i = 1, \ldots, N$ have already been trained. In general, these can be other than GMMs since we operate on the score space. All cohort sets are denoted collectively as $\mathcal{C} = (\mathcal{C}_1, \ldots, \mathcal{C}_N)$. We consider each speaker's model $\mathcal{R}_i$ and the cohort models $\{\mathcal{R}^{(j)}|j \in \mathcal{C}_i\}$ together as a one model, called the *compound model*. The compound model for speaker i is denoted as

$\mathcal{M}^{(i)} = (\mathcal{R}^{(i)}, \{\mathcal{R}^{(j)} | j \in \mathcal{C}_i\})$, and we will denote the normalized match score as $s(X, \mathcal{M}^{(i)})$. The task is to optimize the compound models $\mathcal{M}^{(i)}$ from the existing single models so that a cost function is minimized. In a sense, cohort selection can be seen as *discriminative training* of speaker models.

3.1 False Acceptance and Rejection

The match score $s(X, \mathcal{M}^{(i)}) \in \mathbb{R}$ is a continuous random variable with an unknown probability distribution $p(s)$ which can be divided into genuine and impostor distributions $p(s|\text{genuine}), p(s|\text{impostor})$, see upper panel of Fig. 3. These represent the distributions obtained by matching a random utterance X against genuine speaker model (the speaker who actually produced X) and someone else's model, respectively.

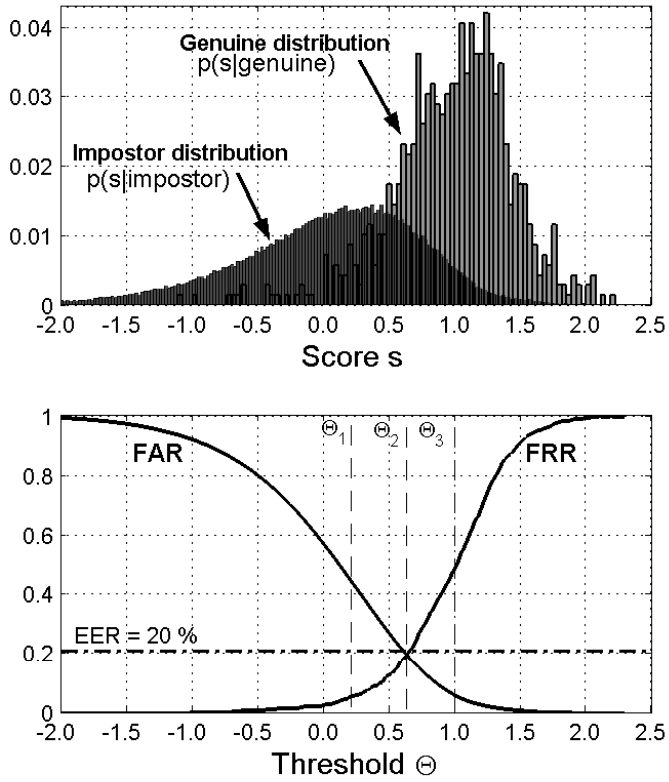


Figure 3: Increasing Θ decreases false acceptances and increases false rejections.

The true distributions $p(s|\text{target}), p(s|\text{nontarget})$ are not available, so we need to estimate them empirically. For this, we use a labeled development set $\mathcal{Z} = \{(X_j, Y_j) | j = 1, 2, \ldots, L\}$, including at least one segment per speaker ($L \geq N$). Here, X_j's are the test segments, and Y_j's are the correct class labels ($Y_j \in \{1, \ldots, N\}$).

First, we define the error counts FR_i and FA_i for each speaker i as follows:

$$\mathrm{FR}_i = \sum_{j=1}^{L} \mathcal{I}\{Y_j = i \wedge s(X_j, \mathcal{M}_i) < \Theta\} \tag{13}$$

$$\mathrm{FA}_i = \sum_{j=1}^{L} \mathcal{I}\{Y_j \neq i \wedge s(X_j, \mathcal{M}_i) \geq \Theta\}, \tag{14}$$

where $\mathcal{I}\{A\} = 1$, if proposition A is true and 0 otherwise. False rejection rate (FRR) and false acceptance rate (FAR) can now be calculated as

$$\mathrm{FRR}(\mathcal{C}, \Theta) = \frac{1}{N \cdot L} \sum_{i=1}^{N} \mathrm{FR}(\mathcal{C}_i, \Theta) \tag{15}$$

$$\mathrm{FAR}(\mathcal{C}, \Theta) = \frac{1}{N \cdot L} \sum_{i=1}^{N} \mathrm{FA}(\mathcal{C}_i, \Theta), \tag{16}$$

where we used the notation to emphasize their dependence on both the cohort sets and the verification threshold Θ. Because the errors depend on both, they should be jointly optimized.

By keeping the cohort sets fixed and sweeping the verification threshold over the real line, we can calculate FRR and FAR at every threshold. By plotting FRR as a function of FAR, we get a curve that shows the trade-off between the two error types. On the other hand, by varying the cohort sets, we get different score distributions. Again, we get a new error trade-off curve by sweeping the threshold over the real line. Each point at each curve corresponds to a certain $(\mathcal{C}, \Theta)$ pair, and the error values $\mathrm{FRR}(\mathcal{C}, \Theta)$, $\mathrm{FAR}(\mathcal{C}, \Theta)$ for this pair are known. The optimization task can be formulated as finding the pair $(\mathcal{C}, \Theta)$ for which an objective function depending on FRR and FAR is minimized.

3.2 Detection Cost Function

Decreased FAR implies increased FRR, and vice versa. In most applications, either one of the error types can be considered more costly than the other one. Following the detection cost function (DCF) defined by NIST [30], we define the optimization problem as finding $(\mathcal{C}, \Theta)$ for which the weighted sum of errors is minimized:

$$\min_{(\mathcal{C},\Theta)} \Big\{ \gamma \cdot \mathrm{FRR}(\mathcal{C}, \Theta) + (1 - \gamma) \cdot \mathrm{FAR}(\mathcal{C}, \Theta) \Big\}, \tag{17}$$

where $0 < \gamma < 1$ is a design parameter controlling the tradeoff between the errors. An illustration of the cost function is shown in Fig. 4.

Since the cohort sets $\mathcal{C}_i$ do not depend on each other, the cost function can be written as a sum of cost functions over different speakers:

$$\min_{(\mathcal{C},\Theta)} \sum_{i=1}^{N} \Big\{ \gamma \cdot \mathrm{FR}(\mathcal{C}_i, \Theta) + (1 - \gamma) \cdot \mathrm{FA}(\mathcal{C}_i, \Theta) \Big\} \tag{18}$$

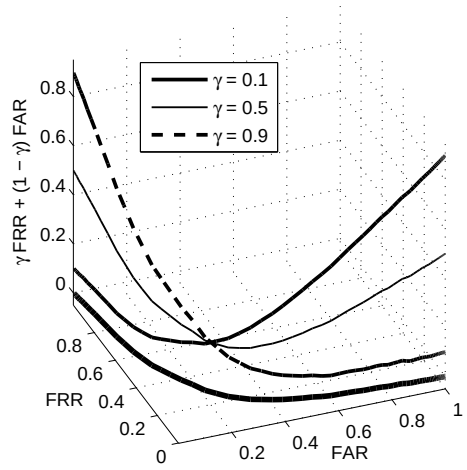


Figure 4: Illustration of the cost function along with the error tradeoff curve in the xy-plane.

We can separate $\mathcal{C}$ and Θ by defining the optimal threshold $\Theta^*(\mathcal{C})$ for a given $\mathcal{C}$ as

$$\Theta^*(\mathcal{C}) = \arg\min_{\Theta} \sum_{i=1}^{N} \Big\{ \gamma \cdot \mathrm{FR}(\mathcal{C}_i, \Theta) + (1-\gamma) \cdot \mathrm{FA}(\mathcal{C}_i, \Theta) \Big\}, \tag{19}$$

which can be found by linear search by sweeping Θ over the genuine and impostor score distributions. The optimization problem becomes

$$\min_{\mathcal{C}} \sum_{i=1}^{N} \Big\{ \gamma \cdot \mathrm{FR}\Big(\mathcal{C}_i, \Theta^*(\mathcal{C})\Big) + (1-\gamma) \cdot \mathrm{FA}\Big(\mathcal{C}_i, \Theta^*(\mathcal{C})\Big) \Big\}. \tag{20}$$

3.3 Genetic Algorithm for Minimizing DCF

Brute force optimization requires evaluating an exponential number of cohort sets and is out of question. We use a *genetic algorithm* (GA) [31] to minimize DCF. We maintain a separate population for each speaker, see Fig. 5 for the data structures. Individuals are integer vectors of dimensionality M (cohort size). The jth individual for speaker i is denoted as $\mathcal{C}_i^j$.

Pseudocode for the GA is given in Algorithm 1. Initialization is done by selecting M disjoint random integers as the individuals. New candidates are generated using crossover and mutation operators, which doubles the sizes of the cohort populations. Next, we compute the normalized match scores using a labeled tuning set $\mathcal{Z}$.

Since computation of the fitness values $\mathrm{DCF}(\mathcal{C}_i^j, \Theta)$ requires the common threshold, we must pool together all genuine and impostor trial scores over all speakers and cohorts. In practise, we use histograms for reducing the number operating points before pooling. As a result, we have the genuine and impostor trial

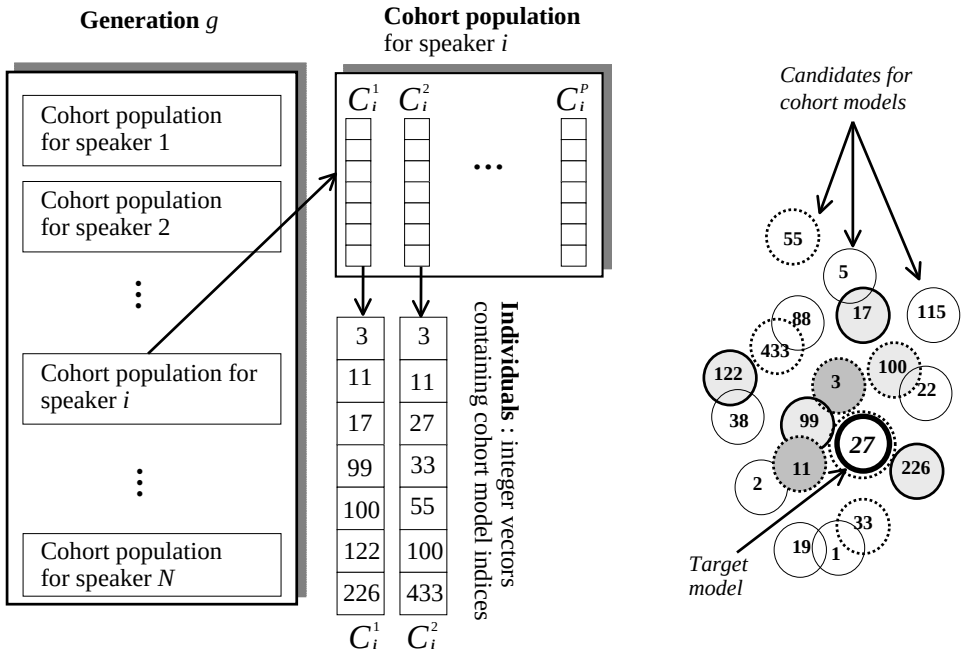


Figure 5: Basic data structures in the GA-based cohort optimization.

Algorithm 1 Outline of the GA-based cohort optimization.

$\mathcal{P} \leftarrow$ InitializePopulations() ;
for $g = 1, 2, \ldots,$NumGenerations **do**
  $\mathcal{P}_{\rm cand} \leftarrow$ GenerateNewCandidates($\mathcal{P}$) ;
  (G,I) $\leftarrow$ ComputeNormalizedScores($\mathcal{R}, \mathcal{P} \cup \mathcal{P}_{\rm cand}, \mathcal{Z}$);
  $\Theta_{\rm opt} \leftarrow$ ComputeOptimalThreshold(G,I,$C_{\rm FA}, C_{\rm FR}$) ;
  $\mathcal{F} \leftarrow$ ComputeDCFValues(G,I,$\Theta_{\rm opt}$) ;
  $(\mathcal{P}, \mathcal{F}) \leftarrow$ SelectSurvivors($\mathcal{P} \cup \mathcal{P}_{\rm cand}, \mathcal{F}$) ;
end for
return $(\mathcal{P}, \Theta_{\rm opt})$;

score distributions (G, I). Using these, we find the optimal threshold as (19). After the threshold $\Theta^*(\mathcal{C})$ is found, the fitness values are calculated as DCF($\mathcal{C}_i^j, \Theta$).

New candidates are generated by pairing the vectors randomly and performing crossover. The parents and the offspring are pooled, and for the pooled population, every vector is mutated with a probability P_m. Crossover is implemented by duplicating the parent vectors into the offspring vectors and swapping their elements with probability P_c. In mutation, we replace a randomly selected index by a random number.

For selection, we sort the vectors according to their fitness (DCF) values. The best individual (smallest DCF) is always selected to the next generation. For the remaining ones, we compare successive pairs, and select the better one. The worst individual dies out.

Table 1: Summary of the corpus.

Language	English
Speakers	207
Speech type	Conversational
Quality	Telephone
Sampling rate	8.0 kHz
Quantization	8-bit μ-law
Training speech (avg.)	119.0 sec.
Evaluation speech (avg.)	30.4 sec.

4 Experiments

4.1 Corpus, Feature Extraction, and Modeling

For the experiments, we use the male subset of *NIST 1999 Speaker Recognition Evaluation* corpus [32]. Both the "a" and "b" files are used for training the 64 component diagonal covariance GMMs, whereas the 1-speaker male test segments are used as the tuning set $\mathcal{Z}$ for the cohorts.

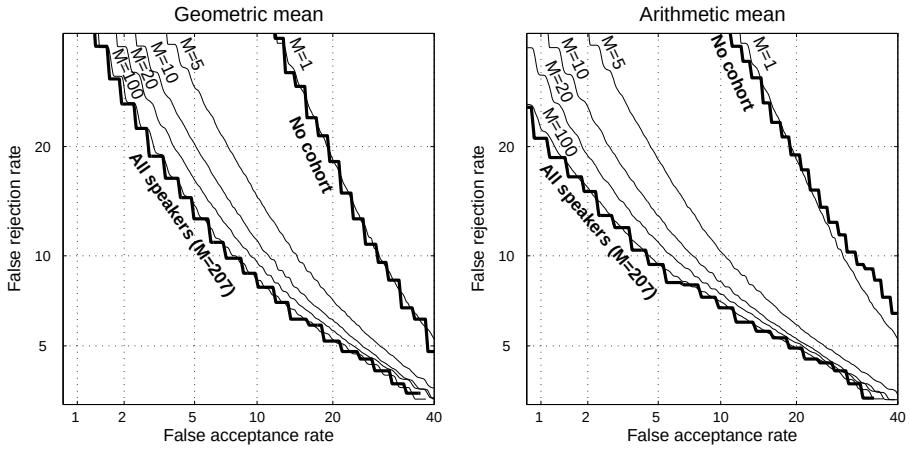
In the current implementation, we use a simple MFCC front end without channel normalization, so we decided to restrict the experiments to matched telephone lines case. There are 230 male speakers in total, and from these 207 fulfill the matched telephone line case.

The UBM is trained by using all the two-speaker detection task files from the same corpus, including both males and females. From this, speaker-depended GMMs are derived by adapting the mean vectors using the MAP procedure [5]. MFCC features are computed from Hamming-windowed and pre-emphasized 30 ms frame with 10 ms overlap. We retain the 12 lowest MFCC coefficinets (excluding c_0) from the log-compressed 27-channel filterbank outputs using DCT.

Throughout the experiments, we consider three operating points corresponding to the following application scenario:

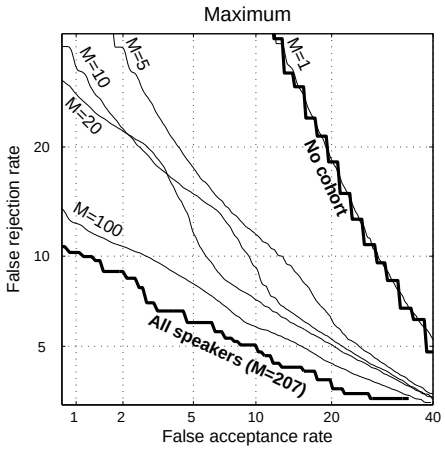
- Secure scenario (low FAR)
- 50-50 scenario (low EER)
- User-convenient (low FRR)

For the secure scenario, we require false acceptance rate to be at most 3 %, and compare the obtained FRRs for different approaches. Similarly, for the user-convenient scenario, we require the FRR to be at most 3 % and compare the obtained FARs.



(a) Geometric mean method

(b) Arithmetic mean method



(c) Maximum method

Figure 6: The effect of the normalization formula and cohort size (randomly selected cohorts, averaged DET curves for 100 repetitions).

Table 2: Standard deviations of errors using random cohort (100 repetitions).

	Secure FRR @ FAR = 3 %			**50-50** EER			**User convenient** FAR @ FRR = 3 %		
Cohort size	5	10	20	5	10	20	5	10	20
Geometric mean	3.9	3.1	2.9	1.0	0.8	0.8	9.1	8.8	8.7
Arithmetic mean	3.8	2.4	1.7	0.9	0.6	0.7	9.6	8.2	8.3
Maximum	3.6	3.0	4.6	1.8	1.0	0.6	10.0	9.9	9.3

4.2 Normalization Formula

First, we study the behavior of the normalization formulae (8)-(10), with the focus on their robustness. For this, we select the cohort models randomly and repeat the procedure 100 times. In this way, we get an idea about the average performance and variance. The average detection error tradeoff (DET) curves [33] for the three normalization methods are shown in Fig. 6 for different cohort sizes. For comparative purposes, we also show the baseline (no score normalization) and the case where all speakers are included in the cohort. Table 2 shows the standard deviations for the three application scenarios and cohort sizes $M = 5, 10, 20$.

We observe that increasing the cohort size improves accuracy for all methods, except for cohort size $M = 1$, for which the baseline gives similar or better results. However, the performance increases rapidly with increasing cohort size in both "secure" and "user-convenient" ends of the curve for all three methods. Increased cohort size reduces also variance, which is due to the fact that larger cohorts include more and more tge same models as the models are selected among the targets.

Regarding the three methods, the ordering is consistent: geometric mean performs the worst and maximum the best on average. However, the variance of the arithmetic mean is smallest, and thus it is expected to be most robust. Because of larger variance, we expect that the geometric mean and maximum methods require more careful selection of the cohort.

Geometric mean and maximum operators are in a sense opposites to each other. Geometric mean gives high nontarget score if the test data yields high likelihood for *all* cohort models ("AND" operator). In contrast, maximum method indicates high nontarget score if there is a single cohort model that has high likelihood ("OR" operator). The arithmetic mean is in between the two extremes, and all the three formulae are special cases of *generalized mean* [34].

Even though performance increases with the cohort size, it must be remembered that large cohort size implies a large number of likelihood calculations and it becomes computationally unfeasible. For this reason, we are interested in smaller cohort sizes.

Table 3: Verification thresholds optimized by GA (log likelihood ratio domain).

	Secure $\gamma = 0.1$			**50-50** $\gamma = 0.5$			**User convenient** $\gamma = 0.9$		
Cohort size	5	10	20	5	10	20	5	10	20
Geometric mean	1.37	1.39	1.4	0.89	0.95	0.97	0.27	0.37	0.42
Arithmetic mean	1.09	1.11	1.11	0.73	0.75	0.77	0.12	0.19	0.21
Maximum	0.56	0.41	0.27	0.25	0.00	0.00	-0.35	-0.50	-0.64

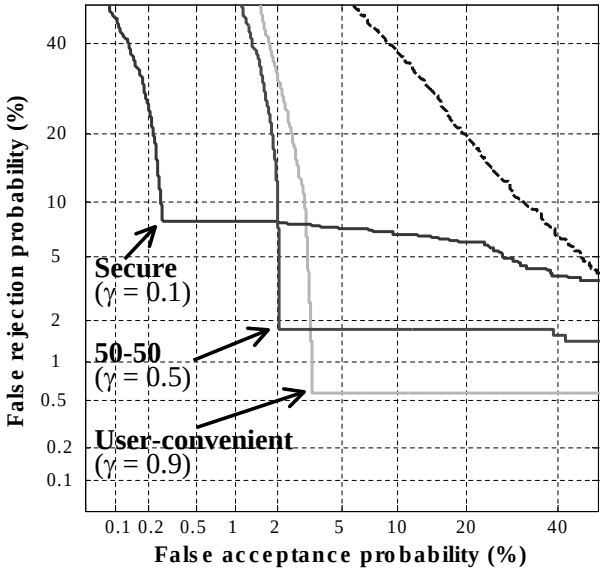


Figure 7: Examples of DET curves obtained by GA (arithmetic mean, cohort size $M = 5$).

4.3 Selection Algorithms

Next, we compare the following heuristic approaches:

Random	Random cohort
CI	Closest impostors selected using (12)
MSC	Maximally spread close [3]
MSCF	Maximally spread close + far [3]
UCN	Unconstrained cohort normalization [8]

Genetic algorithm is optimized for the test data, and its purpose is to provide a lower bound to the error rates reachable by MFCC/GMM combination. It presents an "oracle selection" scheme - the oracle knows exactly what the targets are going to say during verification trial and selects the optimal cohorts for future.

GA finds a single operating point from the error tradeoff curve and is suboptimal in the other regions, see Fig. 7. Examples of thresholds optimized found by GA are listed in Table 3. It can be observed that the threshold increases when moving towards secure applications, which is expected.

The "corner" points in Fig. 7 are the minimum cost function operating points. We set $\gamma = 0.1$, $\gamma = 0.5$, and $\gamma = 0.9$ for the secure, 50-50, and the user-convenient scenarios, respectively. After preliminary experimentation, we fixed the GA parameters as follows: population size 100, the number of generations 500, mutation probability 0.01, and crossover probability 0.5.

The results for the three normalization methods are given in Tables 4-6. The results for baseline (no score normalization) and the UBM [5] are also shown as a reference. Several observations can be made. Firstly, arithmetic mean and

Table 4: Results for geometric mean normalization.

	Secure FRR @ FAR = 3 %			**50-50** EER			**User-convenient** FAR @ FRR = 3 %		
Baseline	69.4			20.2			56.1		
UBM	17.2			8.4			45.8		
Cohort size	5	10	20	5	10	20	5	10	20
Random	38.6	29.6	24.3	12.1	10.5	9.7	45.9	43.2	43.5
CI	20.8	16.7	14.8	9.7	8.1	7.7	41.6	31.6	39.5
MSC	20.7	16.6	14.5	9.2	8.3	7.9	42.6	36.6	35.5
MSCF	34.7	32.1	27.2	12.1	11.0	10.3	49.3	52.7	50.3
UCN	60.9	55.4	47.8	17.6	15.8	14.6	52.7	50.2	44.7
GA reference	3.7	2.6	4.3	3.1	2.8	3.1	13.4	5.0	19.1

Table 5: Results for arithmetic mean normalization.

	Secure FRR @ FAR = 3 %			**50-50** EER			**User-convenient** FAR @ FRR = 3 %		
Baseline	69.4			20.2			56.1		
UBM	17.2			8.4			45.8		
Cohort size	5	10	20	5	10	20	5	10	20
Random	27.3	18.5	14.8	10.1	8.9	8.3	44.2	41.9	40.6
CI	17.5	13.6	11.3	8.8	7.8	7.4	40.8	36.4	40.1
MSC	15.1	11.4	10.2	8.1	7.9	7.2	41.1	35.4	32.8
MSCF	18.4	13.2	11.1	9.2	8.0	7.9	43.2	48.2	49.3
UCN	56.1	48.8	39.5	15.9	14.3	12.7	51.1	49.0	48.7
GA reference	3.9	2.6	4.0	3.1	2.7	4.0	12.0	2.7	30.2

Table 6: Results for maximum normalization.

	Secure FRR @ FAR = 3 %			**50-50** EER			**User-convenient** FAR @ FRR = 3 %		
Baseline	69.4			20.2			56.1		
UBM	17.2			8.4			45.8		
Cohort size	5	10	20	5	10	20	5	10	20
Random	24.7	18.4	19.9	10.9	9.0	7.9	44.9	43.3	44.1
CI	13.9	11.7	10.4	9.2	8.3	7.7	42.8	40.8	49.4
MSC	13.8	11.8	10.8	8.9	8.6	7.9	40.5	51.5	49.5
MSCF	19.4	14.1	11.7	9.9	8.8	8.6	42.2	50.5	58.4
UCN	50.4	39.6	29.0	14.5	14.0	11.3	51.2	46.4	48.0
GA reference	2.8	2.0	3.6	2.9	2.2	3.6	2.9	5.3	24.8

maximum are more accurate than geometric mean. Secondly, comparing the heuristic methods, CI, MSC and MSCF are similar in performance, whereas UCN is worse. Thirdly, comparing the cohort and UBM approaches, UBM outperforms

random cohort, MSCF and UCN in most cases, whereas CI and MSC outperform UBM.

Some interesting observations can be made regarding the application scenario and UBM versus cohort approaches. In the 50-50 case, the differences are small between the methods. However, in the secure and user-convenient scenario, the cohort approach clearly outperforms UBM. In the secure end, UBM reaches an FRR of 17.2 %, whereas the best heuristic cohort selection method reaches 10.2 % (MSC with arithmetic mean, cohort size 20). In the user-convenient end, UBM reaches a FAR of 45.8 %, whereas the best heuristic cohort method reaches 31.6 % (CI with geometric mean, cohort size 10). These observations stress the importance of comparing methods using not only on the EER operating point which is an arbitrary choice.

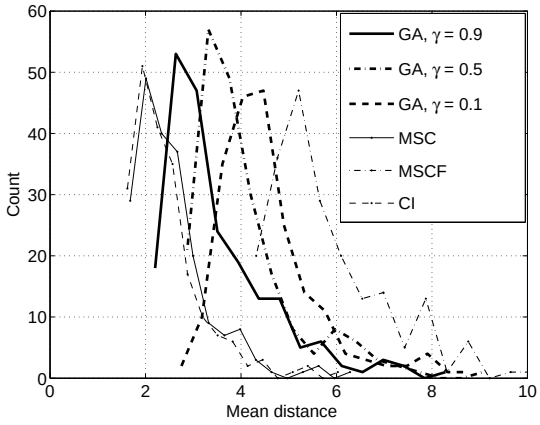
The reference performance given by GA shows that there is much room to improve cohort selection algorithms. In particular, all the studied methods are poor at the user-convenient end. The GA suggests that it would be possible to reach a FAR of 2.7 % at FRR = 3.0 % if the cohorts were selected optimally. The best heuristic reaches as poor as 31.6 % FAR, an order of magnitude worse than GA suggests. Notice however that for GA, increased cohort size reduces the performance, which is contradictory to the results for the heuristic methods. A possible explanation for this is that the parameter space is larger for increased cohort size and GA might not have converged yet. We did not make further attempts in optimizing the number of generations as the simulations take rather long time.

4.4 Analysis of Selected Cohorts

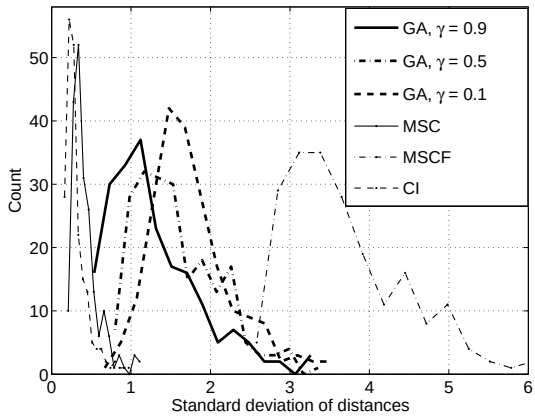
Next, we analyze the cohort sets selected by the genetic algorithm, with the hope to gain understanding on the selection procedure. The GA was optimized for the test data, and now we are interested to see if optimal selection could be predicted from the training conditions only. We use the distance (12) for analyzing the model proximities. We also experimented with the similarity measure (11), and the results were similar.

The distribution of means and standard deviations of the distances from the target to his cohort models are shown in Fig. 8 for the arithmetic mean method and cohort size $M = 20$. The CI, MSC and MSCF are also shown for comparison. We make the following observations. Regarding the distribution of means, the models selected using CI and MSC are closer to target models than for other methods as expected. The models selected using MSCF are further away, and the GA selected models in between. The order of the standard deviations is the same, and holds for all the three application scenarios. These observations suggest that the optimal cohort should contain not "too close" or "too far" models but something in between. Similarly, the optimal cohort should not be too concentrated

or too spread but something in between.



(a) Means



(b) Standard deviations

Figure 8: Distributions of mean and standard deviation of cohort model distances from the target.

According to Fig. 8, in user-convenient scenario, the cohort models should be selected closer to the target than in the secure scenario. Table 7 gives further evidence of this by showing the the number of cases, in which speaker belongs to his own cohort. We observe that in the user-convenient scenario, speaker belongs to his own cohort in 74 % - 97 % of the cases, and the number decreases when moving towards the secure end.

This result might seem counterintuitive at the first glance. In a user-convenient application, it is important that the correct speaker is not rejected; thus, it seems logical to assume that competing models should not be located "too close" to the target. However, by including close models to the cohort, the denominator of the LR will be accurately presented when a genuine speaker is present (likelihood of X for both target and cohorts is accurately computed). In

Table 7: Number of cases (%) where speaker belongs to his own cohort

	Secure $\gamma = 0.1$			**50-50** $\gamma = 0.5$			**User convenient** $\gamma = 0.9$		
Cohort size	5	10	20	5	10	20	5	10	20
Geometric mean	11.0	20.0	24.0	25.0	38.0	46.0	86.0	74.0	74.0
Arithmetic mean	19.0	33.0	48.0	49.0	66.0	76.0	93.0	95.0	95.0
Maximum	0.00	0.00	0.48	0.00	99.0	99.5	95.0	95.0	97.0

the extreme case of cohort size $M = 1$ and speaker in his own cohort, LR for a genuine speaker will be always close to 1 and the threshold is set easily around this value by GA (see Table 3).

By excluding the target from his cohort in the secure scenario, the score for a genuine speaker will be in general larger, which has the effect of shifting the genuine distribution right. On the other hand, (casual) impostor data is far away from the target model in general, and it does not matter if the target is included in the denominator or not - the impostor data will far away from the target model and not be affected by it much. Thus, the impostor distribution will be relatively unchanged regardless of whether target is or is not included in the cohort. Because the genuine distribution shifts up, the distributions will be better separated.

In conclusion, the effect of including target in his own cohort in a user-convenient application makes the genuine distribution centered around LR = 1, and setting of threshold is easier. In the secure application, leaving the speaker out from the cohort has the effect of shifting genuine distribution right while retaining impostor distribution relatively unchanged.

5 Discussion and Conclusions

We have presented a step towards non-heuristic cohort selection based on minimizing a detection cost function. We find the following observations the most interesting ones:

1. UBM and cohort approaches perform similar in 50-50 and user-convenient scenario, whereas cohort is clearly better in the secure scenario.

2. There is lots of room for studying score normalization, especially in the user-convenient end of the DET curve. The results of GA suggest that the MFCC features can reach both low FAR and FRR if the cohorts are well-selected.

The experiments suggest the following design rules for the cohort normalization approach:

1. Randomly selected cohort is better than no cohort. In this case, the cohort size should be as large as possible.

2. In general, larger cohort is better because it reduces the variance of the nontarget scores.

3. Arithmetic mean normalization is most robust and consistent over different selection methods, and we recommend to use it by default.

4. Maximum normalization has the best potential according to the GA reference, but the difference with the arithmetic mean is not large.

5. Of the heuristic methods compared, CI and MSC are both good choices.

6. In a user-convenient and 50-50 applications, it is advantageous to include nearby models into the cohort. In particular, the speaker's own model.

From a practical point of view, we must ask how useful the cohort normalization is in real applications. Sometimes cohort approach is criticized for its computational complexity and memory requirements, which is true if cohort size is large or the cohort models are selected from an external population. However, the results of GA suggest that good cohorts can be selected among the other registrants; in this case, we need to store only the lookup tables for the cohort indices in addition to the models. The results also suggest that small error rates could be reached if we knew how to select the cohorts; the methodology in this study presents an "oracle selection" scheme where the oracle knows exactly what the targets are going to utter during verification trial and selects good cohorts.

We have used GA here merely as an analysis tool. However, it might be used also as a practical cohort selection method. We believe in its potential, because it jointly optimizes the cohort sets and the verification threshold; usually these two are designed independent from each other, although FAR and FRR errors depend on both of them.

To apply GA as a practical cohort selection method, there are two principal issues that need to be studied. Firstly, as seen from Fig. 7, the algorithm optimizes a single point on the tradeoff curve. However, from the system administrator's perspective, it would be good to have the whole tradeoff curve optimized, from which the desired optimal threshold can be selected. For this, the objective function should be modified to minimize the total area under the DET curve for example. The second challenge relates to computational complexity: the simulations made in this study were time- and memory-consuming.

Finally, we wish to emphasize that the optimization was carried out entirely in the score space by having fixed acoustic features and models. The result of the optimization is a set of indices that merely tells against which models the features are to be matched during the verification process. Similar optimization can be

carried out for any biometric authentication problem, in which severe mismatches are expected between training and testing.

References

[1] J. Campbell. Speaker recognition: a tutorial. *Proceedings of the IEEE*, 85(9):1437–1462, 1997.

[2] K.-P. Li and J.E. Porter. Normalizations and selection of speech segments for speaker recognition scoring. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 1988)*, pages 595–598, New York, 1988.

[3] D.A. Reynolds. Speaker identification and verification using gaussian mixture speaker models. *Speech Communication*, 17:91–108, 1995.

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

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

[6] A.E. Rosenberg, J. DeLong, C.-H. Lee, B.-H. Juang, and F.K. Soong. The use of cohort normalized scores for speaker recognition. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP 1992)*, pages 599–602, Banff, Canada, October 1992.

[7] C.-S. Liu, H.-C. Wang, and C.-H. Lee. Speaker verification using normalized log-likelihood score. *IEEE Trans. on Speech and Audio Processing*, 4(1):56–60, 1996.

[8] A.M. Ariyaeeinia and P. Sivakumaran. Analysis and comparison of score normalization methods for text dependent speaker verification. In *Proc. 5th European Conference on Speech Communication and Technology (Eurospeech 1997)*, pages 1379–1382, Rhodes,Greece, 1997.

[9] T. Isobe and J. Takahashi. Text-independent speaker verification using virtual speaker based cohort normalization. In *Proc. 6th European Conference on Speech Communication and Technology (Eurospeech 1999)*, pages 987–990, Budapest, Hungary, 1999.

[10] Y. Zigel and A. Cohen. On cohort selection for speaker verification. In *Proc. 8th European Conference on Speech Communication and Technology (Eurospeech 2003)*, pages 2977–2980, Geneva, Switzerland, 2003.

[11] T. Kinnunen, E. Karpov, and P. Fränti. Efficient online cohort selection method for speaker verification. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP 2004)*, volume 3, pages 2401–2402, Jeju Island, Korea, 2004.

[12] P. Rose. *Forensic Speaker Identification.* Taylor & Francis, London, 2002.

[13] D.A. Reynolds and R.C. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Trans. on Speech and Audio Processing*, 3:72–83, 1995.

[14] X. Huang, A. Acero, and H.-W. Hon. *Spoken Language Processing: a Guide to Theory, Algorithm, and System Development.* Prentice-Hall, New Jersey, 2001.

[15] S. Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29(2):254–272, 1981.

[16] F.K. Soong and A.E. Rosenberg. On the use of instantaneous and transitional spectral information in speaker recognition. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 36(6):871–879, 1988.

[17] R. Duda, P. Hart, and D. Stork. *Pattern Classification.* Wiley Interscience, New York, second edition, 2000.

[18] A. Higgins, L. Bahler, and J. Porter. Speaker verification using randomized phrase prompting. *Digital Signal Processing*, 1:89–106, 1991.

[19] D.A. Reynolds. Comparison of background normalization methods for text-independent speaker verification. In *Proc. 5th European Conference on Speech Communication and Technology (Eurospeech 1997)*, volume 2, pages 963–966, Rhodes,Greece, 1997.

[20] D. Tran and M. Wagner. Fuzzy C-means clustering-based speaker verification. In *Proc. Advances in Soft Computing (AFSS 2002)*, pages 318–324, Calcutta, India, February 2002.

[21] P. Sivakumaran, J. Fortuna, and A.M. Ariyaeeinia. Score normalization applied to open-set, text-independent speaker identification. In *Proc. 8th European Conference on Speech Communication and Technology (Eurospeech 2003)*, pages 2669–2672, Geneva, Switzerland, 2003.

[22] L.F. Lamel and J.L. Gauvain. Speaker verification over the telephone. *Speech Communication*, 31:141–154, 2000.

[23] D. Tran and M. Wagner. Noise clustering-based speaker verification. In *Proc. Advances in Soft Computing (AFSS 2002)*, pages 325–331, Calcutta, India, February 2002.

[24] K.P. Markov and S. Nakagawa. Text-independent speaker recognition using non-linear frame likelihood transformation. *Speech Communication*, 24:193–209, 1998.

[25] R.A. Finan, A.T. Sapeluk, and R.I. Damper. Impostor cohort selection for score normalization in speaker verification. *Pattern Recognition Letters*, 18:881–888, 1997.

[26] N. Mirghafori and L. Heck. An adaptive speaker verification system with speaker dependent a priori decision thresholds. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP 2002)*, pages 589–592, Denver, Colorado, USA, 2002.

[27] T. Pham and M. Wagner. Fuzzy-integration based normalization for speaker verification. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP 1998)*, pages 3273–3276, Sydney, Australia, 1998.

[28] S. Furui. Recent advances in speaker recognition. *Pattern Recognition Letters*, 18(9):859–872, 1997.

[29] T. Isobe and J. Takahashi. A new cohort normalization using local acoustic information for speaker verification. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 1999)*, volume 2, pages 841–844, Phoenix, Arizona, USA, 1999.

[30] M. Przybocki and A. Martin. NIST speaker recognition evaluation chronicles. In *Proc. Speaker Odyssey: the Speaker Recognition Workshop (Odyssey 2004)*, pages 15–22, Toledo, Spain, 2004.

[31] Z. Michalewicz. *Genetic Algorithms + Data Structures = Evolution Programs*. Springer Verlag, Berlin, 3rd revised and extended edition edition, 1996.

[32] A. Martin and M. Przybocki. The NIST 1999 speaker recognition evaluation - an overview. *Digital Signal Processing*, 10:1–18, 2000.

[33] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The DET curve in assessment of detection task performance. In *Proc. 5th European Conference on Speech Communication and Technology (Eurospeech 1997)*, pages 1895–1898, Rhodes,Greece, 1997.

[34] L.I.Kuncheva. *Fuzzy Classifier Design*. Physica Verlag, Heidelberg, 2000.

Julkaisija
Publisher

Joensuun yliopisto
Tietojenkäsittelytieteen laitos
University of Joensuu
Department of Computer Science

Vaihdot

Joensuun yliopiston kirjasto / Vaihdot
PL 107, 80101 Joensuu
Puh. 013-251 2677, fax 013-251 2691
e-mail: vaihdot@joensuu.fi

Exchanges

Joensuu University Library / Exchanges
P.O. Box 107, FI-80101 Joensuu, FINLAND
Tel. +358-13-251 2677, fax +358-13-251 2691
e-mail: vaihdot@joensuu.fi

Myynti

Joensuun yliopiston kirjasto / Julkaisujen myynti
PL 107, 80101 Joensuu
Puh. 013-251 4509, fax 013-251 2691
e-mail: joepub@joensuu.fi

Sales

Joensuu University Library / Sales of Publications
P.O. Box 107, FI-80101 Joensuu, FINLAND
Tel. +358-13-251 4509, fax +358-13-251 2691
e-mail: joepub@joensuu.fi