Unsupervised Evaluation of Speaker Verification Systems

Niko Brümmer [1] *and Jason Pelecanos* [2]

DataVoice Systems, Stellenbosch, South Africa [1]
niko.brummer@datavoice.co.za
Speech Research Lab, Queensland University of Technology [2]
GPO Box 2434, Brisbane, Australia
j.pelecanos@qut.edu.au

Abstract

A method for blind estimation of DET curves for speaker verification systems is proposed. Verification error probabilities are estimated on a database where speaker identities are unknown. The database must provide a set of impostor-only tests as well as a set of mixed impostor and target tests. This method is tested on 9 speaker verification systems that were scored on the NIST 2000 database. Good DET estimates are obtained for systems with low error rates, while poorer estimates are obtained for systems with high error rates.

1. Introduction

Databases for speaker verification (SV) system development and testing are expensive because: they have to be large, they have to be *supervised* (the identity of speakers have to be known) and they are not portable (a database collected in one environment may not be a valid test for a system that is to be used under different circumstances).

If reasonable performance estimates of verification systems could be made with *unsupervised* databases (where speaker identities are not known), performance evaluation might become feasible in cases where it would not otherwise be. The authors are unaware of previous work that has addressed this problem.

A method is proposed where an impostor score distribution model is estimated from a set of pure impostor test scores. Starting with this distribution, a model for the distribution of combined impostor and target scores is estimated from a set of mixed scores. The DET curve for the system is analytically determined from the model parameters. The combined model also has a mechanism to compensate for small changes in the impostor distribution between the two score sets.

Below we discuss the prerequisites, describe the procedure and present experimental data of 9 SV systems tested on the NIST 2000 database. Model estimation details are given in the appendix section.

2. Prerequisites

Prerequisites for the proposed method, required of the evaluation database and the verification system under test are listed below:

2.1. Unsupervised database

The required properties of the unsupervised database are:

(a) The database must contain single-speaker utterances from many speakers, where the identities of the speakers of the utterances need not be known.
(b) The database must provide *test pairs* of utterances, where a pair consists of a *training* utterance and a *test* utterance. A significant proportion of these pairs must be *impostor* pairs (where the two utterances are not spoken by the same speaker), and a significant proportion must be *target* pairs (where both utterances are spoken by the same speaker).
(c) The database must also provide a separate set of pure impostor pairs.

Property (b) may be difficult to obtain in an unsupervised manner. Consider as an example the recording of conversations of 100 people who phone with equal frequencies. If random pairs of calls are chosen, the probability of target pairs will be 1/100, which is probably too low for this evaluation procedure. As the number of recorded speakers increases, the problem worsens.

(Recording a test database for same-number tests would not be a problem however. A database could consist of many subsets of calls, where each subset is recorded from a small number of telephones, and test pairs are chosen from within each subset.)

Property (c) may be easier to obtain, by tapping two different environments where speakers are very unlikely to overlap. Since these pairs will be used to estimate the impostor distribution, it is however important that these pairs come from the same conditions as the pairs from (b). SV systems that use impostor normalization methods such as H-norm and T-norm [2,3] must have such normalization data available in any case.

2.2. Verification system

For the purpose of this article we define a speaker verification system as having the following properties:

(a) It creates a speaker model given a training utterance of a speaker.
(b) It outputs a verification score given a speaker model and a test utterance. The score must be typically larger for *target* (same-speaker) tests and smaller for *impostor* (different speaker) tests.

3. Error estimation

The steps used to obtain a DET curve for an SV system, given an unsupervised database are as follows:

3.1. Impostor distribution estimation

Use the pure impostor pairs from the database and the SV system under test to generate a set of pure impostor scores. Model the impostor distribution with an *n*-component, 1-dimensional Gaussian mixture model (GMM), where the likelihood of a score, *x*, is:

$$p(x) = \sum_{i=1}^{n} q_i N(x, \mu_i, \sigma_i) \qquad (1)$$

$$\sum_{i=1}^{n} q_i = 1 \qquad (2)$$

where $N(x,\mu,\sigma)$ is the normal distribution with mean μ and standard deviation σ.

As a first try, this model was initialized with a binary-splitting k-means vector-quantization algorithm. This resulted in a GMM having components with small variances and with means spread over the score range. It proved not to work well for the rest of the procedure. Instead, a *concentric* initialization was adopted where all components are initialized with the sample mean of the score set and the variances are spread over a range. The range begins smaller than the sample variance and ends larger than the sample variance.

Next, the model parameters are adapted with several iterations of the EM algorithm. During the EM iteration, the concentric property is lost – the means spread out somewhat, but the variances are generally larger than in the k-means initialized case.

3.2. Mixed score model estimation

The mixed pairs from the database and the SV system under test are used to generate a set of mixed impostor and target scores. This distribution is modeled with a structured GMM of the form:

$$p(x) = P_{imp} \sum_{i \in \mathbb{I}} q_i N(x, \alpha + \gamma\mu_i, \gamma\sigma_i) + P_{tar} \sum_{i \in \mathbb{T}} r_i N(x, \beta_i, \delta_i) \qquad (3)$$

where

$$P_{imp} + P_{tar} = 1 \qquad (4)$$

$$\sum_{i \in \mathbb{T}} r_i = 1 \quad \text{and} \quad \sum_{i \in \mathbb{I}} q_i = 1 \qquad (5)$$

Here, P_{imp} is the fraction of impostor scores in the mixed score set and P_{tar} the fraction of target scores. This model has two sets of components: $\mathbb{I}$ the impostor components and $\mathbb{T}$ the target components. The impostor component parameters are formed from the previously estimated parameters and are left unchanged throughout re-estimation. A global offset parameter α and global scale parameter γ are added to modify the impostor distribution. These are to allow for possible change in the impostor distribution between the pure impostor set and the mixed score set. Note that $N(x,\alpha+\gamma\mu,\gamma\sigma) = N((x-\alpha)/\gamma,\mu,\sigma)/\gamma$, effecting a transformation of the random variable *x*. The target components are left to adapt freely.

3.2.1. Initialization

The *a priori* parameters P_{imp} and P_{tar} are initialized with guesses. Adaptation parameter α can be set to zero and γ to one. Impostor parameters $\{q_i,\mu_i,\sigma_i\}$ are fixed as previously estimated. An equal number of target components are initialized from the impostor components, with an offset and enlarged variances: $r_i = q_i$, $\beta_i = \mu_i + s$, $\delta_i^2 = 20\,\sigma_i^2$, where the offset *s* can be roughly estimated by inspection of the mixed score histogram.

3.2.2. EM re-estimation

All parameters except the original impostor parameters $\{q_i,\mu_i,\sigma_i\}$, are re-estimated with several iterations of the EM algorithm on the mixed score data. The re-estimation formulae are given in the Appendix.

3.3. DET calculation

The detection error tradeoff (DET) curve for a detection (or verification) system is a non-linear scaling of the receiver operating curve (ROC), where the threshold of a detection system is varied to produce a curve of miss probability against false acceptance probability. The DET transform is designed to display a straight line for systems having both a normal impostor and a normal target score distribution [4,3].

Given the impostor and target parts of the estimated structured GMM, we can analytically calculate the DET curve. For a threshold *t*, we have error probabilities:

$$P_{miss} = \sum_{i \in \mathbb{T}} r_i \int_{x=-\infty}^{t} N(x, \beta_i, \delta_i)\, dx \qquad (6)$$

$$P_{fa} = \sum_{i \in \mathbb{I}} q_i \int_{x=t}^{\infty} N(x, \alpha + \gamma\mu_i, \gamma\sigma_i)\, dx \qquad (7)$$

The integrals are evaluated using the error function [7]. Varying *t* over a range of values and applying the DET transform, produces the estimated DET curve.

4. Experiments

4.1. Database

The proposed blind DET estimation method was tested on 9 different speaker verification systems that produced verification scores on the "1-speaker detection" part of the NIST 2000 Speaker Recognition Evaluation Database [1]. The authors wish to acknowledge NIST and 7 of the other NIST 2000 participants for making this data available. DET curves of these systems are presented anonymously here.

This database provides *different-number* tests (implying different telephones were used between training and testing). We limited our experiments to the subset of these tests where both training and test utterances had been automatically identified as electret [8]. The database is partitioned into male and female portions and we present experiments based on male data only here.

Since we used handset and gender information, our test database was strictly not unsupervised. However, the handset labels were automatically produced [9] and an automatic gender detector would also be feasible.

This database has no pure impostor score set as required by Section 2.1(c). To provide this, the speakers were partitioned into two sets of roughly equal size. This provided two impostor score sets, *A* and *B*. All target tests were put into one set. The initial impostor distribution was estimated from the *B* set. The mixed score set was the union of the target set and the *A* impostor set.

4.2. Determination of GMM size

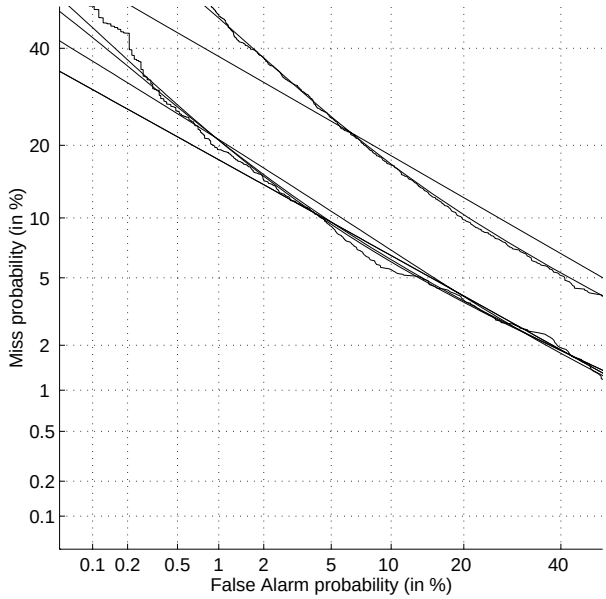

Figure 1: Supervised GMM-based DETs

A *supervised* test (using separate impostor and target score sets) was performed to determine how well GMM-derived DET curves match DET curves obtained directly from the data and what a suitable size for the GMMs would be.

A DET curve can be obtained directly from a set of impostor and a set of target scores, by (a) sorting both sets, (b) using every score in the (usually smaller) target set as a threshold, and (c) counting the fraction of both types of scores on either side of the threshold.

The GMM-based DET curves were produced by training two separate *n*-component simple GMMs on the target and impostor sets respectively and then using equations 6 and 7.

Figure 1 has plots of two different SV systems and a range of values for the number of Gaussian mixtures, *n*. The two irregular curves are the DETs obtained directly from the data. For the lower error-probability system, four cases are plotted with $n \in \{1,2,5,10\}$. Note that $n=1$ produces a straight line, but as *n* increases, the match improves. For the higher error probability system, the DETs for two cases, $n \in \{1,2\}$, are plotted. In the remaining experiments we used $n=10$.

4.3. Impostor distribution adaptation

Another experiment was conducted (using the two separate sets of impostors) to investigate the adaptation mechanism of the impostor distribution via the parameters α and γ. See Figure 2: The solid plot with the lower peak was made from a 10-component GMM trained on impostor set *B* (of one of the SV systems). The solid plot with the higher peak was made from a 10-component GMM trained on impostor set *A*. It is evident that these two distributions differ slightly in mean and variance. Next, a structured GMM with $\mathbb{T}=\varnothing$ (Equation 3), was initialized from the former model (*B*), and α and γ were adapted on the *A* data. Clearly, the adapted model fits the data better. The aim is to allow for small changes in the impostor score distribution, while retaining most of the information gained from the prior impostor distribution estimation.

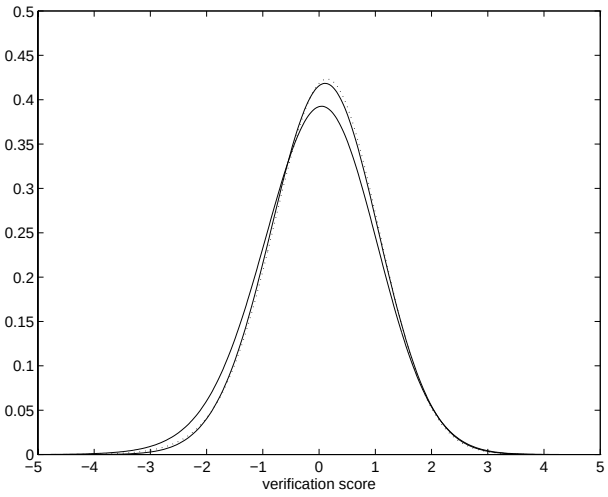

Figure 2 : Impostor distribution adaptation

4.4. Blind DET estimation

The results of the *unsupervised* DET estimation experiments on 9 different SV systems are presented. We used impostor score set *B* of every system to train a 10-component simple GMM. This was used as initialization for a 20-component structured GMM that was trained on the unsupervised mixed score set, from which the estimated DETs were calculated. See Figures 3 and 4: The solid curves are the DET curves obtained directly from the data (in a supervised way), while the dotted curves are the blind DET estimates.

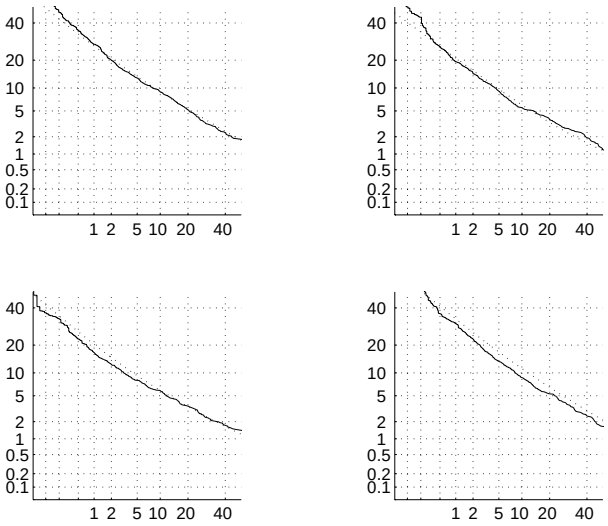

Figure 3: DET estimates

Figure 3 shows the better estimated DET curves. Note that these were obtained for the systems with the better error rates. In these four cases, the estimates for the proportion of target speakers P_{tar} (Equation 3) were close to correct.

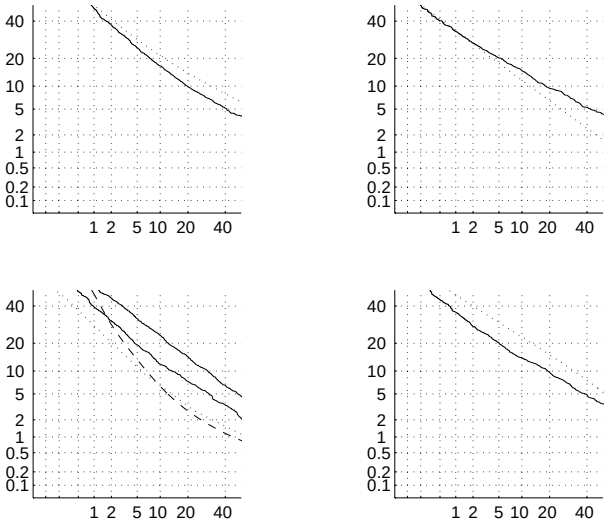

Figure 4: DET Estimates

Figure 4 shows DET estimates with various degrees of failure. In the top two cases, estimates for P_{tar} were close to the true values, but the slopes of the DET curves were inaccurate. The bottom left plot shows two cases where the P_{tar} values were underestimated, giving overoptimistic DET estimates. (The dashed curve is for the high error-probability system). In the bottom right case, P_{tar} was overestimated, giving a pessimistic DET estimate.

5. Discussion

In these limited number of experiments we find that the better systems (equal error rate < 10%) are well evaluated with this method. However, the higher the error rates become, the more likely an inaccurate DET estimate will be obtained. Below some of the factors leading to inaccurate estimates are considered.

5.1. Target ratio

The estimation of the ratio of target to impostor tests is crucial. In a situation where this ratio is known accurately (but speaker identities are not available), P_{imp} and P_{tar} should be fixed during re-estimation. This information was used for the three systems in the bottom two plots of Figure 4. (The two other systems in Figure 4 where the P_{tar} estimate was close to correct can't be improved much in this way.) The improved DET estimates are plotted in Figure 5.

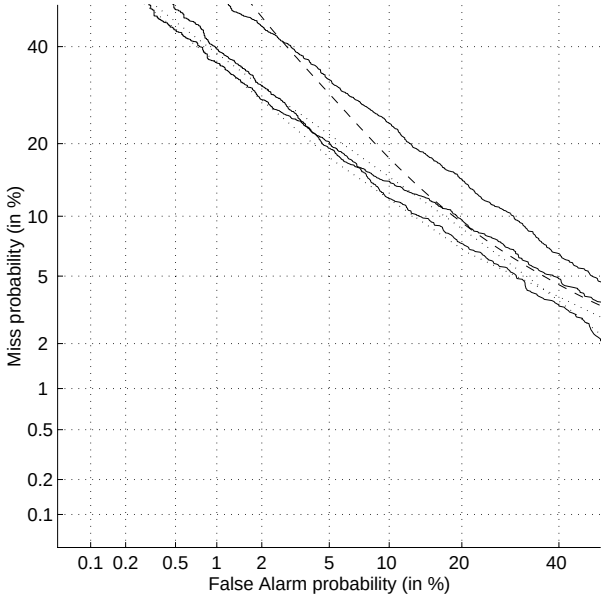

Figure 5:Target ratio fixed

Two of the systems have satisfactory DET approximations under these conditions (dotted curves), but the system with the highest error rate still has a poor estimate (dashed curve).

5.2. Target distribution

In an attempt to analyze the source of the problem with the top right system in Figure 4, we provided (a) accurate P_{imp} and P_{tar} estimates, (b) an accurate impostor distribution and (c) an accurate target distribution. Only the information provided by (c) produced a significantly improved DET estimate. Figure 6 shows solid plots of distributions derived from two simple GMMs trained on separate impostor and target scores, while the dotted plots are the distributions obtained with the blind estimation procedure. The left plots are the impostor distributions and the right plots the target distributions. The impostor distribution is estimated more successfully than the target distribution. The impostor distribution estimation is more accurate because its shape is fixed by the prior estimation step and because there is more impostor data. The target distribution estimation is relatively more inaccurate because it relies on less target data and has no prior shape information. When too much of the target distribution is obscured by the overlapping impostor distribution, inaccurate target distribution estimation can result.

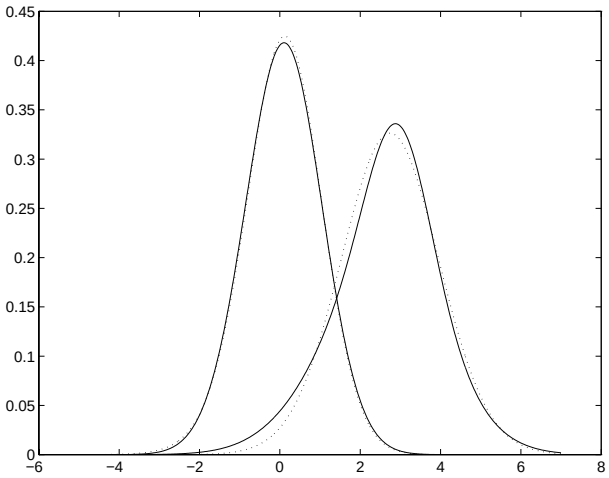

Figure 6: Scaled target and impostor score distribution estimates

5.3. Smaller target-to-impostor ratios

In addition to performing poorer with high error-rate systems under test, this estimation method is also dependent on the relative sizes of the impostor and target sets. If either set is too small compared to the other, the distribution of the smaller set will be obscured by the distribution overlap. A small target set means the DET approximation is also sensitive to the estimate of the impostor distribution. The top left system of Figure 3 was retried with both sets of impostors (instead of only one) used in the mixed set of impostors. This reduced P_{tar} from the previous 0.165 to 0.092. See Figure 7: The solid plot is obtained directly from the data, the dotted plot is blind estimation, while the dashed plot is blind estimation with the true P_{tar} fixed.

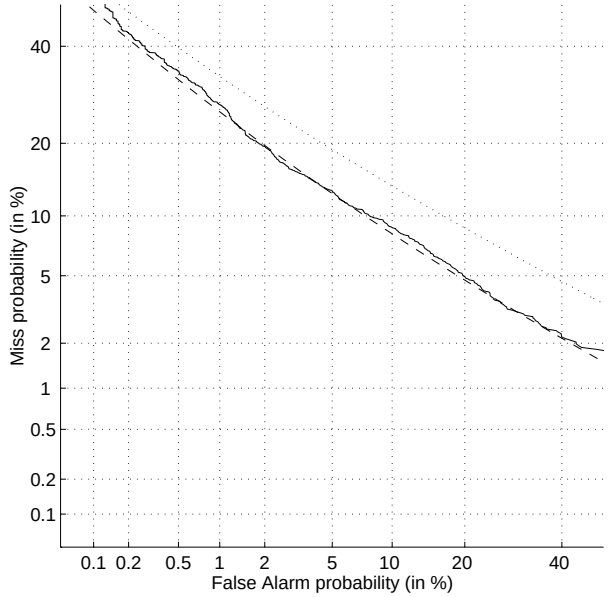

Figure 7: Effect of smaller target-to-impostor ratios

The smaller ratio of target tests resulted in blind estimation failure, but addition of the knowledge of P_{tar} improved the estimate again.

6. Conclusions

It was shown that blind DET estimation of speaker verification systems is possible under certain favourable conditions. These conditions include:

(a) Small overlap of impostor score and target score distributions. This is only achieved when both (i) the error rate of the system under test is low enough (ii), and when the difference between the numbers of impostor and target tests in the evaluation database is not too extreme.

(b) The shape of the impostor distribution stays unchanged between the pure impostor score set and the mixed score set. (Variation in the mean and variance of the impostor distribution can however be compensated for by the proposed method.)

Furthermore it was shown that extra information, like accurate knowledge of the proportion of target tests can aid the estimation in some cases of unfavourable distribution overlap.

In addition to providing DET estimates, this method could also provide estimates of the number of impostor attacks (P_{imp}) and hence also undetected impostor attacks on a system during a given period. This would be useful for risk management.

Useful further work would be to find a way of estimating confidence in the estimated DET curve. Some kind of bootstrapping, where synthetic data is generated according to the estimated distributions, could be considered.

Other possible uses of the concepts used here, applied to speaker verification, include:

(a) More sophisticated impostor score normalization – H- and T-norm are essentially based on normal score distributions.

(b) Using a multidimensional generalization of the α-γ adaptation to adapt GMM speaker models during recognition, to compensate for condition mismatches.

7. Appendix: Re-estimation formulae

For convenience, the derivation for the EM re-estimation formulae for a simple GMM in the form of equations 1 and 2 is outlined. (See [5] for the general EM algorithm and [6] for the GMM-specific algorithm.) The result of a similar derivation on the structured GMM of equations 3 through 5 is then presented.

7.1. EM derivation outline

The EM algorithm is used to obtain a local maximum of the likelihood of the observed data $\{x_t\}$, given a model λ for the data. Specifically, $\prod_t p(x_t|\lambda)$ is maximized with respect to λ, by iteratively re-estimating the parameters of λ. An iteration of the EM algorithm starts with a model $\underline{\lambda}=\{q_i,\underline{\mu}_i,\underline{\sigma}_i\}$ and produces a new model $\lambda=\{q_i,\mu_i,\sigma_i\}$, so that

$$\prod_t p(x_t|\lambda) \geq \prod_t p(x_t|\underline{\lambda}) \qquad (8)$$

In the case of a simple GMM, we have:

$$p(x_t|\lambda) = \sum_i q_i N(x_t, \mu_i, \sigma_i) \quad (9)$$

$$\sum_i q_i = 1 \quad (10)$$

It can be shown [5,6] that the inequality of equation 8 is ensured by maximizing the auxiliary function $Q(\underline{\lambda}, \lambda)$ with respect to λ, where

$$Q(\underline{\lambda}, \lambda) = \sum_t \sum_i P(i|x_t, \underline{\lambda}) \log(q_i N(x_t, \mu_i, \sigma_i)) \quad (11)$$

and where

$$P(i|x_t, \underline{\lambda}) = \underline{q}_i N(x_t, \underline{\mu}_i, \underline{\sigma}_i) / p(x_t|\underline{\lambda}) \quad (12)$$

is the posterior probability of component i, given the data and the old model. $Q(\cdot)$ is augmented by adding a Lagrange-multiplier term to ensure the constraint of equation 10 and is globally maximized by setting its partial derivatives, with respect to each of the parameters in λ, to 0.

7.2. Formulae for structured GMM re-estimation

In the case of the stuctured GMM of equations 3 through 5, the auxiliary function becomes:

$$Q(\underline{\lambda}, \lambda) = \sum_t \sum_{i \in \mathbf{I}} P_{it} \log(P_{imp} q_i N(x, \alpha + \gamma\mu_i, \gamma\sigma_i))$$

$$+ \sum_t \sum_{i \in \mathbf{T}} P_{it} \log(P_{tar} r_i N(x, \beta_i, \delta_i)) \quad (13)$$

where we write the posterior component probabilities, given the old model, for brevity as:

$$P_{it} = P(i|x_t, \underline{\lambda}) \quad (14)$$

Here we need three Lagrange-multipliers for the three constraints (equations 4 and 5). After differentiating and solving we obtain the required formulae:

$$P_{imp} = (\sum_t \sum_{i \in \mathbf{I}} P_{it}) / (\sum_t \sum_{i \in \mathbf{I} \cup \mathbf{T}} P_{it}) \quad (15)$$

$$P_{tar} = (\sum_t \sum_{i \in \mathbf{T}} P_{it}) / (\sum_t \sum_{i \in \mathbf{I} \cup \mathbf{T}} P_{it}) \quad (16)$$

$$r_k = (\sum_t P_{kt}) / (\sum_t \sum_{i \in \mathbf{T}} P_{it}) \quad (17)$$

$$\beta_k = (\sum_t P_{kt} x_t) / (\sum_t P_{kt}) \quad (18)$$

$$\delta_k^2 = ((\sum_t P_{kt} x_t^2) / (\sum_t P_{kt})) - \beta_k^2 \quad (19)$$

$$[-E]\gamma^2 + [AC/B - F]\gamma + [D - A^2/B] = 0 \quad (20)$$

$$\alpha = (A - \gamma C)/B \quad (21)$$

where

$$A = \sum_t \sum_{i \in \mathbf{I}} P_{it} x_t / \sigma_i^2 \quad (22)$$

$$B = \sum_t \sum_{i \in \mathbf{I}} P_{it} / \sigma_i^2 \quad (23)$$

$$C = \sum_t \sum_{i \in \mathbf{I}} P_{it} \mu_i / \sigma_i^2 \quad (24)$$

$$D = \sum_t \sum_{i \in \mathbf{I}} P_{it} x_t^2 / \sigma_i^2 \quad (25)$$

$$E = \sum_t \sum_{i \in \mathbf{I}} P_{it} \quad (26)$$

$$F = \sum_t \sum_{i \in \mathbf{I}} P_{it} x_t \mu_i / \sigma_i^2 \quad (27)$$

Note that equation (20) has two solutions for γ. We found in all cases of solving these equations, that γ had one positive and one negative solution. (No attempt was made to prove this analytically. A look at the second derivatives of $Q(\cdot)$ with respect to γ would be a good idea). The positive solution was always used, and as control it was verified that inequality (8) was always met.

8. References

[1] http://www.nist.gov/speech/tests/spk/index.htm

[2] Reynolds, D.A., et al., "Speaker Verification Using Adapted Gaussian Mixture Models", *Digital Signal Processing*, **10**(1-3):19-41, 2000.

[3] Auckenthaler, R., Carey, M., and Lloyd-Thomas H., "Score Normalization for Text-Independent Speaker Verification Systems", *Digital Signal Processing.***10** (1-3):42-54, 2000.

[4] Martin, A., Doddington, G., et al., "The DET curve in assessment of detection task performance", *Proc. Eurospeech 1997, Rhodes*: 1895-1898, 1997.

[5] Dempster, A., Laird, N., and Rubin, D., "Maximum likelihood from incomplete data via the EM algorithm", *J. Roy Stat. Soc*. **39:**1-38, 1977.

[6] Reynolds, D.A, and Rose R.C., "Robust text-independent speaker identification using Gaussian mixture speaker models", *IEEE Trans. Speech Audio Process*. **3:**72-83, 1995.

[7] Press, W. H., Teukolsky S.A, Vetterling W. T., Flannery B.P., *Numerical Recipes in C*: 220-221, Cambridge University Press, 1992.

[8] Martin, A., and Przybocki, M.,"The NIST 1999 speaker recognition evaluation – an overview", *Digital Signal Processing.***10** (1-3)1:18, 2000.

[9] Quatieri T., Reynolds D. and O'Leary G., "Magnitude-only estimation of handset nonlinearity with application to speaker recognition", *Proceedings ICASSP-98*: 745-748, 1998.