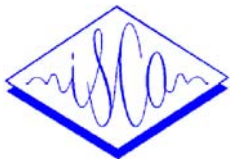

ODYSSEY04 - The Speaker and Language Recognition Workshop
Toledo, Spain
May 31 - June 3, 2004

Evaluation of a Small-Footprint Text and Language Independent Speaker Recognition System on Forensic Data

Suhadi[1], Stephan Grashey[2], Sorel Stan[1], Tim Fingscheidt[1]

[1]Siemens AG, ICM Mobile Phones, 81675 Munich, Germany
[2]Siemens AG, Corporate Technology, 81730 Munich, Germany

{firstname}.{lastname}@siemens.com

Abstract

In this paper we evaluate on a forensic task our text and language independent speaker recognition system, characterized by modest memory requirements and robustness to environment noise. Noise robustness is achieved by employing a Kalman filter-based sequential interacting multiple models (SIMM) algorithm.

The evaluation data was provided by the Netherlands Forensic Institute (NFI) and consisted of telephone conversations in four different languages gathered in real police investigations. The results of NFI evaluation show that our small-footprint system provides competitive equal error rates (EER) for the class of text independent systems operating on telephone speech with strong channel mismatch.

1. Introduction

In autumn 2003, the Netherlands Forensic Institute (NFI) and the Netherlands organisation for applied research (TNO) jointly organized an evaluation of text independent automatic speaker recognition systems for forensic applications [1]. The speech material of this evaluation consisted of data gathered in real police investigations. The intention was on one hand to determine the state of the art of automatic speaker recognition systems, on the other hand to verify the possibility of using the results of such systems for investigative purposes in police enquiries.

We participated with a text-independent speaker recognition system, based on Gaussian mixture models (GMM) and a standard mel-frequency cepstral coefficients (MFCC) front-end [2]. Although this system was designed for personal use within a mobile phone [3, 4] (constraints: low complexity approach with a moderately long enrollment utterance and only a short testing utterance), the NFI/TNO evaluation was a good opportunity to test our system under real world conditions and to compare it with more complex approaches.

Since the operating environments of mobile phones are typically noisy, an algorithm originally developed for noise robust speech recognition – the Kalman filter-based SIMM [5] is employed in order to increase the robustness in noise.

Our paper is organized as follows: in Section 2 we describe briefly the SR system employed, then we present in Section 3 the noise compensation algorithm, followed by the experimental results in Section 4, and the conclusions in Section 5.

2. Speaker Recognition Approach

2.1. Feature Extraction

We employ a standard front-end processing based on MFCC features [6]. The speech signal sampled at 8 kHz is divided into overlapping frames of 32 ms length with a frame shift of 15 ms. Energy-based voice activity detection (VAD) is performed on each frame to discard silence segments.

Each frame is multiplied by a Hamming window prior to transformation by a 256-point FFT. The squared spectral magnitudes are pre-emphasized by a first order FIR filter and then transformed to the Mel-frequency domain using $P = 15$ triangle-shaped filters. After taking the logarithm a vector of 15 *log-spectral* coefficients $\mathbf{x}_t$ is obtained for frame t. Subsequent application of the discrete cosine transformation (DCT) yields a vector of $C = 12$ *cepstral* coefficients $\mathbf{x}_t^{\text{cep}}$.

The convolutional effects of a linear transmission channel appear in the cepstral domain as additive biases, which can be eliminated by estimating the mean and subtracting it from each cepstral feature vector. Cepstral mean subtraction (CMS) is part of the standard front-end processing in SR systems [2], and we also apply it to obtain channel-compensated feature vectors.

2.2. Models and Training

The world and the speaker are represented in the *cepstral* domain using GMMs with $J = 64$ Gaussian probability density functions (PDF) each. A GMM is completely characterized by the parameter set $\lambda =$

$\{w_j, \mu_j, \mathbf{\Sigma}_j\}_{j=\overline{1,J}}$ and the density

$$p(\mathbf{x}_t^{\text{cep}} \mid \lambda) = \sum_{j=1}^{J} w_j \mathcal{N}(\mathbf{x}_t^{\text{cep}}; \mu_j, \mathbf{\Sigma}_j) \ , \qquad (1)$$

where $\mathcal{N}(\cdot)$ is a Gaussian PDF and w_j is the mixing weight satisfying $\sum_{j=1}^{J} w_j = 1$.

One evaluation rule for the NFI/TNO Forensic Speaker Recognition Evaluation was, to use a world model λ_w obtained from

- speech data not originating from this evaluation,
- where Dutch is not the spoken language [1].

Given these constraints, a German and a English world model were trained in order to capture a richer phonetical structure accounting for different languages.

The German world model λ_w^{DE} was based on about 14 hours of speech data pooled together from phonetically rich sentences of 1262 male and female speakers of the German Speechdat II – Mobile Database [7].

The English world model λ_w^{UK} is based on about 12 hours of speech data pooled together from phonetically rich sentences of 1000 male and female speakers of the British English Speechdat II – Mobile Database [7]. Both world models were obtained via Expectation Maximization (EM) training [8] in the cepstral domain.

Similarly to the world model, the speaker model can be obtained via EM training from speaker's enrollment data. Alternatively, the speaker model can be derived from the world model using Bayesian adaptation of means and variances [2].

We use the latter approach as it requires only accumulating statistics of speaker's enrollment data instead of buffering all cepstral vectors, which is certainly not an option for our system designed for mobile phones. We build language-dependent speaker models λ_s^{DE} and λ_s^{UK} based on adaptation sequences of 30, 60 and 120 seconds obtained from NFI evaluation data.

2.3. Acceptance/Rejection Criterion

Given a sequence of cepstral vectors $\mathbf{X}^{\text{cep}} = \{\mathbf{x}_t^{\text{cep}}\}_{t=\overline{1,T}}$ the decision rule for identifying the authorized user or the impostor is based on comparing the log-likelihood ratio (LLR) with a threshold θ

$$\text{LLR} = \log \frac{p(\mathbf{X}^{\text{cep}} \mid \lambda_s)}{p(\mathbf{X}^{\text{cep}} \mid \lambda_w)} \gtrless \theta \ , \qquad (2)$$

where "$>$" means acceptance and "$<$" means rejection. It is common practice to make the approximation

$$\log p(\mathbf{X}^{\text{cep}} \mid \lambda) \approx \sum_{t=1}^{T} \log p(\mathbf{x}_t^{\text{cep}} \mid \lambda) \ , \qquad (3)$$

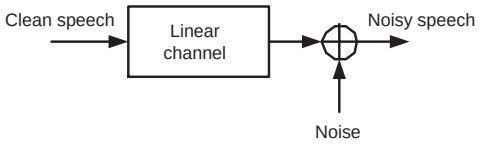

Figure 1: Environment Model.

which holds exactly if and only if the feature vectors are statistically independent.

Since we have two sets of world and speaker models for German and English, we need first to determine which of these languages should be used in Eq. 2. Let us denote the log-likelihoods of the observed vector sequences by

$$\text{LL}(\lambda_w^{DE}) = \log p(\mathbf{X}^{\text{cep}} \mid \lambda_w^{DE}) \qquad (4)$$
$$\text{LL}(\lambda_w^{UK}) = \log p(\mathbf{X}^{\text{cep}} \mid \lambda_w^{UK}). \qquad (5)$$

We make a decision based on

$$\log p(\mathbf{X}^{\text{cep}} \mid \lambda_w) = \max \left(\text{LL}(\lambda_w^{DE}), \text{LL}(\lambda_w^{UK})\right) , \quad (6)$$

and take the respective speaker model. This approach resembles a simple "*language detector*" based on the maximum log-likelihood value of the world models.

3. Noise Compensation Approach

3.1. Environment Model

The environmental effect on the clean speech can be modelled in the time domain by a linear transmission channel and additive noise as shown in Fig. 1. After performing CMS for the channel compensation, only the effect of the additive noise will be considered.

Although linear in the time domain, the noise contamination rule in the log-spectral domain changes to the non-linear function

$$\mathbf{z}_t = \mathbf{f}(\mathbf{x}_t, \mathbf{n}_t) = \mathbf{x}_t + \log[\mathbf{1} + \exp(\mathbf{n}_t - \mathbf{x}_t)], \qquad (7)$$

where $\mathbf{x}_t$, $\mathbf{n}_t$, and $\mathbf{z}_t$ denote the clean speech, noise, and noisy speech *log-spectral* vectors, respectively [9].

3.2. Linear Approximation

Assuming that the PDF of clean speech log-spectra is represented by a GMM with $K = 64$ components

$$p(\mathbf{x}_t) = \sum_{k=1}^{K} w_{\mathbf{x},k} \mathcal{N}(\mathbf{x}_t; \mu_{\mathbf{x},k}, \mathbf{\Sigma}_{\mathbf{x},k}) \qquad (8)$$

and the PDF of noise log-spectra is a single Gaussian

$$p(\mathbf{n}_t) = \mathcal{N}(\mathbf{n}_t; \mu_{\mathbf{n}}, \mathbf{\Sigma}_{\mathbf{n}}) \, , \qquad (9)$$

we would like to find the distribution of the noisy speech log-spectral vectors $\mathbf{z}_t$.

If we approximate the noise contamination rule in each GMM component by a linear model

$$\mathbf{z}_t \approx \mathbf{A}_k \mathbf{x}_t + \mathbf{B}_k \mathbf{n}_t + \mathbf{c}_k, \quad k = 1, ..., K, \tag{10}$$

then each Gaussian of clean speech log-spectra will transform into a corresponding Gaussian of noisy speech log-spectra with mean vector and covariance matrix given by

$$\begin{aligned} \mu_{\mathbf{z},k} &= \mathbf{A}_k \mu_{\mathbf{x},k} + \mathbf{B}_k \mu_{\mathbf{n}} + \mathbf{c}_k \\ \Sigma_{\mathbf{z},k} &= \mathbf{A}_k \mathbf{\Sigma}_{\mathbf{x},k} \mathbf{A}_k' + \mathbf{B}_k \mathbf{\Sigma}_{\mathbf{n}} \mathbf{B}_k' , \end{aligned} \tag{11}$$

where the prime denotes transposition. Note that $\mathbf{A}_k$ and $\mathbf{B}_k$ are $P \times P$ matrices and $\mathbf{c}_k$ is a $P \times 1$ vector. If the matrices are diagonal, then the transformation models the shifting of means and scaling of variances.

We conclude that the distribution of the noisy speech log-spectra can be approximated by the GMM

$$p(\mathbf{z}_t) = \sum_{k=1}^{K} w_{\mathbf{z},k} \mathcal{N}(\mathbf{z}_t; \mu_{\mathbf{z},k}, \mathbf{\Sigma}_{\mathbf{z},k}) \tag{12}$$

with weights $w_{\mathbf{z},k} = w_{\mathbf{x},k}$ and means and covariances given by Eq. 11.

We need to estimate the parameters of the linear model based on some criterion of optimality. Kim [10] proposes statistical linear approximation (SLA) in order to estimate the parameters $\{\mathbf{A}_k, \mathbf{B}_k, \mathbf{c}_k\}_{k=\overline{1,K}}$. SLA computes the parameters which minimize the expected error between a m-th order Taylor polynomial expansion $\mathbf{P}_{\mathbf{f}}^m(\mathbf{x}_t, \mathbf{n}_t)$ of the non-linear noise contamination rule and the linear model

$$\begin{aligned} \{\mathbf{A}_k, \mathbf{B}_k, \mathbf{c}_k\} &= \arg \min_{\{\tilde{\mathbf{A}}_k, \tilde{\mathbf{B}}_k, \tilde{\mathbf{c}}_k\}} E\left[\mid \mathbf{e}_k^{\mathbf{f}(\mathbf{x}_t, \mathbf{n}_t)} \mid^2 \right] \\ \mathbf{e}_k^{\mathbf{f}(\mathbf{x}_t, \mathbf{n}_t)} &= \mathbf{P}_{\mathbf{f}}^m(\mathbf{x}_t, \mathbf{n}_t) - \tilde{\mathbf{A}}_k \mathbf{x}_t - \tilde{\mathbf{B}}_k \mathbf{n}_t - \tilde{\mathbf{c}}_k, \end{aligned} \tag{13}$$

and $\mathbf{P}_{\mathbf{f}}^m(\mathbf{x}_t, \mathbf{n}_t)$ is defined as

$$\begin{aligned} \mathbf{P}_{\mathbf{f}}^m(\mathbf{x}_t, \mathbf{n}_t) = \sum_{k=0}^{m} \frac{1}{k!} \big\{ (\mathbf{n}_t - \mathbf{n}_0) \frac{\delta}{\delta \mathbf{n}_t} + (\mathbf{x}_t - \mathbf{x}_0) \frac{\delta}{\delta \mathbf{x}_t} \big\}^k \\ \cdot f(\mathbf{x}_t, \mathbf{n}_t)|_{\mathbf{n}_t = \mathbf{n}_0, \mathbf{x}_t = \mathbf{x}_0}. \end{aligned} \tag{14}$$

For the k-th mixture component, the center of the Taylor series expansion $\{\mathbf{n}_0, \mathbf{x}_0\}$ is set to the mean vectors $\mu_{\mathbf{n}}$ and $\mu_{\mathbf{x},k}$. We use throughout our experiments a third order Taylor polynomial.

Let us assume that the distribution of clean speech and noise log-spectra are obtained with diagonal covariance matrices. Applying the non-linear noise contamination rule in Eq. 7 and the third order Taylor polynomial, the diagonal matrices $\{\mathbf{A}_k, \mathbf{B}_k\}$ and vector $\mathbf{c}_k$ can be calculated per log-spectral vector dimension i as a scalar operation and it leads to

$$\begin{aligned} \mathbf{A}_{k,ii} &= 1 - \frac{\xi(\mu_{\mathbf{n},i}, \mu_{\mathbf{x},k,i})}{1 + \xi(\mu_{\mathbf{n},i}, \mu_{\mathbf{x},k,i})} \\ &\quad + \frac{\xi(\mu_{\mathbf{n},i}, \mu_{\mathbf{x},k,i}) \cdot [\xi(\mu_{\mathbf{n},i}, \mu_{\mathbf{x},k,i}) - 1] \cdot \Delta\Sigma_{\mathbf{n},k,i}}{2 \cdot [1 + \xi(\mu_{\mathbf{n},i}, \mu_{\mathbf{x},k,i})]^3} \\ \mathbf{B}_{k,ii} &= \frac{\xi(\mu_{\mathbf{n},i}, \mu_{\mathbf{x},k,i})}{1 + \xi(\mu_{\mathbf{n},i}, \mu_{\mathbf{x},k,i})} \\ &\quad + \frac{\xi(\mu_{\mathbf{n},i}, \mu_{\mathbf{x},k,i}) \cdot [1 - \xi(\mu_{\mathbf{n},i}, \mu_{\mathbf{x},k,i})] \cdot \Delta\Sigma_{\mathbf{n},k,i}}{2 \cdot [1 + \xi(\mu_{\mathbf{n},i}, \mu_{\mathbf{x},k,i})]^3} \\ \mathbf{c}_{k,i} &= \mu_{\mathbf{x},k,i} + ln[1 + \xi(\mu_{\mathbf{n},i}, \mu_{\mathbf{x},k,i})] \\ &\quad + \frac{\xi(\mu_{\mathbf{n},i}, \mu_{\mathbf{x},k,i}) \cdot \Delta\Sigma_{\mathbf{n},k,i}}{2 \cdot [1 + \xi(\mu_{\mathbf{n},i}, \mu_{\mathbf{x},k,i})]^2}, \end{aligned} \tag{15}$$

where

$$\begin{aligned} \xi(\mu_{\mathbf{n},i}, \mu_{\mathbf{x},k,i}) &= \exp[\mu_{\mathbf{n},i} - \mu_{\mathbf{x},k,i}] \\ \Delta\Sigma_{\mathbf{n},k,i} &= [\mathbf{\Sigma}_{\mathbf{n},ii} - \mathbf{\Sigma}_{\mathbf{x},k,ii}]. \end{aligned}$$

From Eq. 15, we see that $\mathbf{A}_k + \mathbf{B}_k = I$, where I is the identity matrix.

3.3. SIMM

Given the observed sequence of noisy speech log-spectral vectors $\mathbf{Z} = \{\mathbf{z}_t\}_{t=\overline{1,T}}$ and the GMM parameters of clean speech log-spectra $\lambda_{\mathbf{x}} = \{w_{\mathbf{x},k}, \mu_{\mathbf{x},k}, \mathbf{\Sigma}_{\mathbf{x},k}\}_{k=\overline{1,K}}$, the goal of the SIMM algorithm is to estimate the statistics of noise log spectra $\hat{\lambda}_{\mathbf{n},t} = \{\hat{\mu}_{\mathbf{n},t}, \hat{\mathbf{\Sigma}}_{\mathbf{n},t}\}$ for each frame [5].

The GMM model $\lambda_{\mathbf{x}}$ representing clean speech log-spectra is computed using 17 hours of speech from phonetically rich sentences of 3911 male and female speakers from the German and British English Speechdat II – Databases [7].

SIMM estimates the noise statistics of current frame $\hat{\lambda}_{\mathbf{n},t}$ from that of the previous frame $\hat{\lambda}_{\mathbf{n},t-1}$ by employing a bank of K Kalman filters (one per state), which share the state transition equation but have different observation models

$$\begin{aligned} \mathbf{n}_t &= \mathbf{n}_{t-1} + \mathbf{u}_{t-1} \\ \mathbf{z}_t &= \mathbf{A}_k \mathbf{x}_t + \mathbf{B}_k \mathbf{n}_t + \mathbf{c}_k . \end{aligned} \tag{16}$$

Note that the noise log-spectral vector is treated as the state of interest, and $\mathbf{u}_t$ is a zero-mean Gaussian process with covariance matrix $\mathbf{\Sigma}_{\mathbf{u}}$.

The algorithm is initiated using an initial guess for $\hat{\lambda}_{\mathbf{n},t=0}$ computed from the first few silence frames, then it iteratively estimates the noise statistics $\hat{\lambda}_{\mathbf{n},t}$ in the t-th frame. The estimation for each frame contains four major steps, namely SLA Coefficient Update, Kalman Filtering, Probability Update and Estimate Mixing.

3.3.1. SLA Coefficient Update

For the t-th frame, we first apply the third order SLA to linearize $\mathbf{f}(\mathbf{x}_t, \mathbf{n}_t)$ around the mean log-spectral vectors

$\mu_{\mathbf{x},k}$ and $\hat{\mu}_{\mathbf{n},t-1}$ for each k [10]. This computation yields $\{\mathbf{A}_{k,t-1}, \mathbf{B}_{k,t-1}, \mathbf{c}_{k,t-1}\}_{k=\overline{1,K}}$. Note that the SLA Coefficient Update step is not a part of the original SIMM algorithm [5]. This enhancement is motivated by the observation that the estimate of noise statistics $\lambda_{\mathbf{n},t}$ changes with each frame and this change affects the linear model estimate.

3.3.2. Kalman Filtering

The k-th Kalman filter employs the previous estimate of the noise statistic $\hat{\lambda}_{\mathbf{n},t-1}$ to give an estimate of the k-th statistical parameter $\hat{\lambda}_{\mathbf{n},k,t}$. In each mixture component, the estimation is performed in two steps, namely *Time-Update* and *Measurement-Update*. The first step computes *a priori* estimates $\hat{\mu}^{-}_{\mathbf{n},k,t-1}$ and $\hat{\mathbf{\Sigma}}^{-}_{\mathbf{n},k,t-1}$ as

$$\begin{aligned} \hat{\mu}^{-}_{\mathbf{n},k,t-1} &= \hat{\mu}_{\mathbf{n},t-1} \\ \hat{\mathbf{\Sigma}}^{-}_{\mathbf{n},k,t-1} &= \hat{\mathbf{\Sigma}}_{\mathbf{n},t-1} + \mathbf{\Sigma}_{\mathbf{u}}. \end{aligned} \tag{17}$$

In the second step, *a posteriori* estimate $\hat{\lambda}_{\mathbf{n},k,t} = \{\hat{\mu}_{\mathbf{n},k,t}, \hat{\mathbf{\Sigma}}_{\mathbf{n},k,t}\}$ is obtained as follows

$$\begin{aligned} \hat{\mu}_{\mathbf{n},k,t} &= \hat{\mu}^{-}_{\mathbf{n},k,t-1} + \alpha \mathbf{K}_{k,t} \mathbf{e}_{k,t} \\ \hat{\mathbf{\Sigma}}_{\mathbf{n},k,t} &= \hat{\mathbf{\Sigma}}^{-}_{\mathbf{n},k,t-1} + \alpha \mathbf{K}_{k,t} \mathbf{B}_{k,t-1} \hat{\mathbf{\Sigma}}^{-}_{\mathbf{n},k,t-1}. \end{aligned} \tag{18}$$

In the equation above, the *Kalman Gain* $\mathbf{K}_{k,t}$ is computed from the *innovation* $\mathbf{e}_{k,t}$ and the *innovation covariance* $\mathbf{R}^{\mathbf{e}}_{k,t}$ as

$$\begin{aligned} \mathbf{K}_{k,t} &= \hat{\mathbf{\Sigma}}^{-}_{\mathbf{n},k,t-1} \mathbf{B}'_{k,t-1} (\mathbf{R}^{\mathbf{e}}_{k,t})^{-1} \\ \mathbf{e}_{k,t} &= \mathbf{z}_t - \mathbf{A}_{k,t-1} \mu_{\mathbf{x},k} - \mathbf{B}_{k,t-1} \hat{\mu}^{-}_{\mathbf{n},k,t-1} - \mathbf{c}_{k,t-1} \\ \mathbf{R}^{\mathbf{e}}_{k,t} &= \mathbf{A}_{k,t-1} \mathbf{\Sigma}_{\mathbf{x},k} \mathbf{A}'_{k,t-1} + \mathbf{B}_{k,t-1} \hat{\mathbf{\Sigma}}^{-}_{\mathbf{n},k,t-1} \mathbf{B}'_{k,t-1} \end{aligned} \tag{19}$$

and α referred as *Kalman Gain Shrinking* is lying in (0,1). In our system, an *Adaptive Kalman Filter* scheme is employed to increase the SIMM performance by giving an update of the slow-evolving covariance matrix $\mathbf{\Sigma}_{\mathbf{u}}$ [5].

3.3.3. Probability Update

Having the statistical parameter of all mixture component $\{\hat{\lambda}_{\mathbf{n},k,t}\}_{k=\overline{1,K}}$, probability of k-th Gaussians given the observation data $\mathbf{Z}_t = \{\mathbf{z}_t\}_{t=\overline{1,t}}$ is computed as [5]

$$\begin{aligned} \gamma_{k,t} &= p(k \mid \mathbf{Z}_t) \\ &= \frac{p(\mathbf{z}_t \mid k, \mathbf{Z}_{t-1}) p(k)}{\sum_{\kappa=1}^{K} p(\mathbf{z}_t \mid \kappa, \mathbf{Z}_{t-1}) p(\kappa)} \end{aligned} \tag{20}$$

where

$$\begin{aligned} p(k) &= w_{x,k} \\ p(\mathbf{z}_t \mid k, \mathbf{Z}_{t-1}) &= \frac{\exp\left\{ -\frac{1}{2} \mathbf{e}'_{k,t} \left(\mathbf{R}^{\mathbf{e}}_{k,t} \right)^{-1} \mathbf{e}_{k,t} \right\}}{\left[(2\pi)^P |\mathbf{R}^{\mathbf{e}}_{k,t}| \right]^{\frac{1}{2}}}, \end{aligned} \tag{21}$$

and P is the dimension of the log-spectral vectors.

3.3.4. Estimate Mixing

By using the Kalman filter bank approach we end up with K different estimates for the noise statistics, one per mixture component. Given that our noise model is a single Gaussian, we need to combine the K estimates in the mixing step to obtain a single estimate as follows

$$\begin{aligned} \hat{\mu}_{\mathbf{n},t} &= \sum_{k=1}^{K} \gamma_{k,t} \hat{\mu}_{\mathbf{n},k,t} \\ \hat{\mathbf{\Sigma}}_{\mathbf{n},t} &= \sum_{k=1}^{K} \gamma_{k,t} \left(\hat{\mathbf{\Sigma}}_{\mathbf{n},k,t} + \Delta\hat{\mu}_{\mathbf{n},k,t} \, \Delta\hat{\mu}'_{\mathbf{n},k,t} \right), \end{aligned} \tag{22}$$

where $\Delta\hat{\mu}_{\mathbf{n},k,t} = (\hat{\mu}_{\mathbf{n},k,t} - \hat{\mu}_{\mathbf{n},t})$.

3.3.5. Computational Requirements

Let us analyze the computational requirements of the SIMM algorithm. We consider here the case of diagonal covariance matrices.

First, the third order SLA Coefficient Update requires $15KP$ additions, $8KP$ multiplications, $3KP$ divisions, KP logarithmic operations, and KP exponential operations. For the Adaptive Kalman Filtering, $7(K+1)P$ additions, $(12K+9)P$ multiplications, and $(K+1)P$ divisions are performed for each frame. The Probability Update is computed with $2KP$ additions, $K(2P+1)$ multiplications, $K(P+2)$ divisions, and K exponential operations. Finally, the Estimate Mixing consists of $2(KP+K-1)$ additions and $3KP$ multiplications.

3.4. Clean Speech Estimation

Once the noise statistics $\hat{\lambda}_{\mathbf{n},t}$ for the current frame has been computed, the estimated clean speech log-spectral vector $\hat{\mathbf{x}}_t$ is given by the minimum mean squared error (MMSE) estimator

$$\begin{aligned} \hat{\mathbf{x}}_t &= E\left[\mathbf{x}_t \mid \mathbf{z}_t, \hat{\lambda}_{\mathbf{n},t} \right] \\ &= \int_{\mathbf{X}} \int_{\mathbf{N}} \mathbf{x}_t \, p(\mathbf{x}_t, \mathbf{n}_t \mid \mathbf{z}_t, \hat{\lambda}_{\mathbf{n},t}) \, d\mathbf{x}_t \, d\mathbf{n}_t \,, \end{aligned} \tag{23}$$

which reduces to

$$\hat{\mathbf{x}}_t = \mathbf{z}_t - \sum_{k=1}^{K} p(k \mid \mathbf{z}_t, \hat{\lambda}_{\mathbf{n},t}) \cdot \left\{ (\mathbf{A}_{k,t} - I)\, \mu_{\mathbf{x},k} - \mathbf{B}_{k,t}\, \hat{\mu}_{\mathbf{n},t} - \mathbf{c}_{k,t} \right\} \quad (24)$$

by using Eq. 10. The probability $p(k \mid \mathbf{z}_t, \hat{\lambda}_{\mathbf{n},t})$ is obtained as

$$p(k \mid \mathbf{z}_t, \hat{\lambda}_{\mathbf{n},t}) = \frac{p(\mathbf{z}_t, k \mid \hat{\lambda}_{\mathbf{n},t})}{\sum_{\varkappa=1}^{K} p(\mathbf{z}_t, \varkappa \mid \hat{\lambda}_{\mathbf{n},t})}, \quad (25)$$

where

$$p(\mathbf{z}_t, k \mid \hat{\lambda}_{\mathbf{n},t}) = w_{\mathbf{z},k} \mathcal{N}\left(\mathbf{z}_t; \hat{\mu}_{\mathbf{z},k,t}, \hat{\mathbf{\Sigma}}_{\mathbf{z},k,t}\right). \quad (26)$$

Note that in Eq. 24 above the parameters $\{\mathbf{A}_{k,t}, \mathbf{B}_{k,t}, \mathbf{c}_{k,t}\}$ and the probability $p(k \mid \mathbf{z}_t, \hat{\lambda}_{\mathbf{n},t})$ are computed based on the new noise statistics estimate $\hat{\lambda}_{\mathbf{n},t}$ obtained from Eq. 22.

4. Experiments

4.1. Description of Data

Both, training and testing data was provided by the NFI. It can be described as follows [1]:

"The speech material used in the NFI/TNO forensic speaker recognition evaluation was taken from real police investigations. This was done in order to obtain field data and to get as close as possible to a real forensic application as possible. It consists of wire tapped cellular GSM to GSM telephone conversations recorded over a 23 month period. All speakers are males. The telephone line quality varies between recordings from excellent to moderate (extremes at the lower end were omitted). The telephone handsets used are unknown. The level and nature of background noises of the material varies and includes slight room reverberations, music in the background of the recording and in some cases background speakers (mostly children playing in the background). Although the speaking style was constant (spontaneous speech, laughter, shouting and whispering was omitted) emotions varied between recordings from relaxed (frequent) to stressed (rare). The distribution of these parameters among speakers is not homogeneous. The range and distribution of recording dates between speakers varies. The material was edited by NFI in order to select single speakers and to make the material anonymous. Care was taken in editing so that no acoustic artifacts were introduced. Signaling noises in the telephone recordings were removed but speaking pauses were not edited out. The languages used are Dutch, English, Sranan Tonga (language spoken in Surinam) and Papiamento (language spoken at the Netherlands Antilles)."

The test samples contained in the test set of the database consists of speech segments of approximately 7, 15 and 30 seconds of length, collected from a single conversation. For testing each sample of the test set is compared only to a predefined subset of the reference templates. During the actual evaluation period of the NFI/TNO evaluation the relation between the speaker-id's of the training set and the id's of the test set was unknown. The results presented in Section 4.2 beneath were obtained with information provided by the NFI some while after the evaluation deadline, which allowed a mapping from the test id's to the training id's.

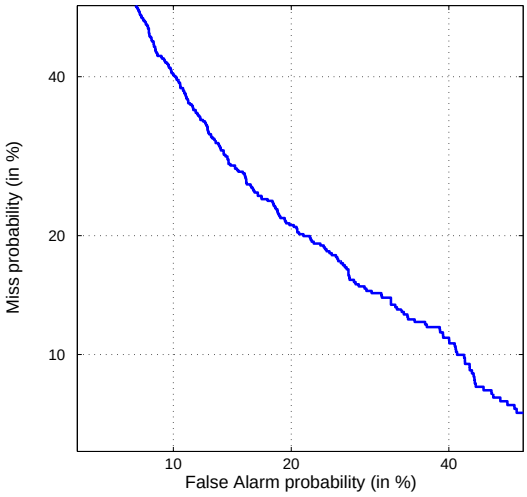

Figure 2: DET Curve.

4.2. Results

Evaluation of the speaker recognition is taken from the database provided by NFI. All the speaker training and testing data is level normalized to -26 dB using the ITU tools [11]. The same data pre-processing is also performed to the training data of the both world models λ_w^{DE} and λ_w^{UK} and the GMM parameters of clean speech log spectra λ_x.

Having the clean speech log spectra model λ_x, all the testing data is noise-compensated by SIMM algorithm in the log-spectral domain. Initialization of the noise statistics $\lambda_{n,t=0}$ is taken from the first five frames, and the Kalman Gain Shrinking is set to 0.1. Subsequent to DCT, all the unidentified testing data is evaluated to numerous speaker models listed on the NFI pre-defined test trials.

In this evaluation, DET curve [12] is employed to evaluate the SR performance. The DET curve is generated based on the true and false LLR score, and it is determined according to the speaker identity given by NFI after the evaluation period. The obtained SR performance is depicted in Fig. 2.

The experimental result shows a moderate performance of our speaker recognition system. Although we provided a low complexity approach, with this result we achieved a middle rank with respect to the EER com-

pared with the preliminary results of 11 other systems who joined the evaluation. This result gives an idea of the current state of the art of text-independent speaker recognition systems with small footprint when real world telephone data is applied.

5. Conclusions

In this contribution we evaluated on forensic data our GMM-based automatic speaker recognition system using a standard MFCC front-end and a feature domain noise compensation technique employing a bank of Kalman-filters.

The evaluation data consisted of short wire-tapped telephone conversation recorded in real police investigations. In this environment, the our system exhibits equal error rates of around $22\,\%$.

6. Acknowledgements

The authors would like to thank Panji Setiawan for his contributions to the development of the speaker recognition system.

7. References

[1] van Leeuwen, D. A. and Bouten J. S., "NFI/TNO Forensic Speaker Recognition Evaluation", <http://speech.tm.tno.nl/aso/>, Rijswijk, Netherlands, 2003.

[2] Reynolds, D.A., Quatieri, T. F., and Dunn R. B., "Speaker Verification Using Adapted Gaussian Mixture Models", Digital Signal Processing, 10:19–41, Oct. 2000.

[3] Setiawan, P., Aalburg, S., Fingscheidt, T., Stan, S., Ruske, G., "A Text-Independent Speaker Verification Approach for Mobile Devices", Elektronische Sprachsignalverarbeitung ESSV 2003, pp. 300–306, Karlsruhe, Germany, Sept. 24–26, 2003.

[4] Suhadi, Stan, S., Fingscheidt, T., Beaugeant, C., "An Evaluation of VTS and IMM for Speaker Verification in Noise", Eurospeech 2003, pp. 1669–1672, Geneva, Switzerland, Sept. 1–4, 2003.

[5] Kim, N. S., "Feature domain compensation of non-stationary noise for robust speech recognition", Speech Communication, 37:231–248, 2002.

[6] Davis, S. B. and Mermelstein, P. "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Trans. on ASSP, 28(4):357–366, Aug. 1980.

[7] URL: <http://www.speechdat.org/>

[8] Dempster, A., Laird, N., and Rubin, D., "Maximum likelihood from incomplete data via the EM algorithm", J. Royal Stat. Soc., vol. 39, 1977.

[9] Moreno, P. J., "Speech Recognition in Noisy Environments", Ph.D. Thesis, Carnegie Mellon University, 1996.

[10] Kim, N. S., "Statistical Linear Approximation for Environment Compensation", IEEE Signal Processing Letters, 5(1):8–10, 1998.

[11] ITU-T, "SVP56: The Speech Voltmeter", in *Software Tool Library 2000 User's Manual*, pp. 151–161, Geneva, Switzerland, December 2000.

[12] Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M., "The DET Curve in Assessment of Detection Task Performance", Eurospeech 1997, 4:1895–1898, 1997.