KTH Computer Science
and Communication

3. The speaker verification systems

HÅKAN MELIN

KTH/CSC/TMH
Stockholm, Sweden 2007

This is a self-contained re-print of Sections 3.1–3.5 from Chapter 3 in:
Melin, H. (2006). Automatic speaker verification on site and by telephone: methods, applications and assessment. Doctoral Thesis, KTH, Stockholm, Sweden 2006. ISBN 978-91-7178-531-2.

Contents

Chapter 3

The speaker verification systems

3.1 Introduction

This chapter describes the speaker verification research systems used in this thesis: a text-dependent (sub-)system based on word-level hidden Markov models (HMM), a text-independent (sub-)system based on Gaussian mixture models (GMM), and a score level combination of the two. The HMM system, and variants of it, are used stand-alone in Chapter 8 that compares different prompting strategies, and Chapter 9 that looks at variance estimation techniques for HMMs. The combined system is used as a component in the PER system and is used for experiments in Chapter 7 on robust error estimation techniques, and Chapter 10 with various results from the PER system and data collected with it.

All speaker verification research systems used in this thesis are built on GIVES, a generic framework for speaker verification systems developed by the author at KTH Center for Speech Technology (CTT). This framework is shortly described.

In addition to the research systems described in this chapter, a commercial speaker verification system has also been used. Results for this system are included in Chapters 7 and 10 in addition to results from the research systems. For reasons of proprietary interests the design of this commercial system cannot be described here, nor may the identity of the system be disclosed.

3.2 Notation and common features

The HMM and GMM subsystems share several features. These features are described in this section together with some notation used later. In the following, the letter ξ will be used to refer to a subsystem, with $\xi = \mathrm{H}$ for the HMM subsystem and $\xi = \mathrm{G}$ for the GMM subsystem.

3.2.1 Feature extraction

The input signal[1] is pre-emphasized and divided into one 25.6 ms frame every 10 ms. A Hamming window is applied to each frame. 12-element[2] mel-frequency cepstral coefficient (MFCC) vectors are then computed for each frame using a 24-channel, FFT-based, mel-warped, log-amplitude filter bank between 300-3400 Hz followed by a cosine transform and cepstral liftering. Both subsystems use these MFCC vectors, while their use of energy terms, delta features and feature post-processing differ, mainly as a result of the subsystems having been optimized rather independently during separate threads of development (see also Section 10.2.1).

3.2.2 Classifier units

Classifiers in both subsystems share a basic classifier unit structure. Refer to a classifier unit in subsystem ξ as $\psi = \xi_u$, where u is an index that uniquely identifies the classifier unit within the subsystem. This classifier unit has one target model and two gender-dependent background models. Target models represent the voices of particular speakers (legitimate users of the system, or *clients*), while background models represent the voices of universal groups of speakers, in this case male and female speakers. Background models are used for two purposes: as seed models during the training phase, and for score normalization during the verification test phase.

Each classifier unit ψ defines one or more likelihood functions $P^{\psi}(\mathbf{O}|\boldsymbol{\lambda})$ used to evaluate the similarity between an observation sequence $\mathbf{O}$ and the model $\boldsymbol{\lambda}$. In the following, $\boldsymbol{\lambda}_{\psi}$ will denote parameters of the target model for a particular target speaker (the client whose identity is claimed during an enrollment or test session) while $\boldsymbol{\lambda}_{\psi}^{\text{male}}$ and $\boldsymbol{\lambda}_{\psi}^{\text{female}}$ will denote parameters of the two background models in classifier unit ψ.

The data and operation of classifier units within the system are independent of each other during both the training and verification test phases: units share no model parameters and the data processing within one unit takes no input from the processing in other units. Units may operate on the same part of input speech, though.

3.2.2.1 Training phase

Assume that all relevant word repetitions and their boundary locations in the enrollment speech are known from the output of an automatic speech recognizer[3].

[1] at 8 kHz sampling rate (see Section 5.2 for particulars about the on-site PER system)

[2] does not include the 0'th cepstral coefficient

[3] cf. Section 5.6 for the procedure used in the PER system

Denote all valid[4] enrollment data from a given enrollee

$$\overline{\mathbf{O}}^{\text{enroll}} = \bigcup_{w \in \mathbf{W}} \{\mathbf{O}_{w,1}, \ldots, \mathbf{O}_{w,R_w}\}$$

where w is a word in the application vocabulary $\mathbf{W}$[5], $\mathbf{O}_{w,r} = \{\mathbf{o}_1^{(w,r)} \ldots \mathbf{o}_{N_{w,r}}^{(w,r)}\}$ is the observation sequence corresponding to the r'th valid repetition of word w (a word segment), $N_{w,r}$ is the length of that observation sequence, and R_w is the number of valid repetitions of word w.

Since classifier units may be trained on different subsets of the data, introduce $\overline{\mathbf{O}}_{\psi}^{\text{enroll}}$ to denote the subset of $\overline{\mathbf{O}}^{\text{enroll}}$ used to train unit ψ. Rather than training a target model directly from this data, *an adaptation procedure is used.* While the actual adaptation method depends on the implementation of the classifier unit, the first step in the adaptation procedure is the same for all classifier units. Based on the enrollment data, one of the two background models is selected as a seed model $\boldsymbol{\lambda}_{\psi}^{g_{\psi}^{\text{seed}}}$, using an automatic gender detector

$$g_{\psi}^{\text{seed}} = \underset{g \in \{\text{male,female}\}}{\arg\max} \; P^{\psi}(\overline{\mathbf{O}}_{\psi}^{\text{enroll}} | \boldsymbol{\lambda}_{\psi}^{g}). \tag{3.1}$$

That is, if the male model fits better to the data, the male model is chosen, otherwise the female model is chosen. Note that no *a priori* information about the gender of the enrollee is used in this selection, and that gender selection in one classifier unit is independent of other classifier units in the system.

The seed model is then used as a basis for target model adaptation as described for each of the two subsystems below.

3.2.2.2 Verification test phase

To test a claim for a given target identity put forward by a claimant speaker, a test utterance is first collected. Again assuming all relevant word repetitions and their boundary locations are known from the output of an automatic speech recognizer, a test utterance with L words is denoted $\overline{\mathbf{O}}^{\text{test}} = \{\mathbf{O}_1 \ldots \mathbf{O}_L\}$, where $\mathbf{O}_i = \{\mathbf{o}_1^{(i)} \ldots \mathbf{o}_{N_i}^{(i)}\}$ is the vector sequence corresponding to the i'th word segment in the utterance. Denote as $w(i)$ the word spoken in segment i. The exact function used by classifier units to score a test utterance given an identity claim varies between units, but it always has the form

$$z_{\psi} = \mathcal{F}\left(\overline{\mathbf{O}}^{\text{test}} | \boldsymbol{\lambda}_{\psi}, g_{\psi}(\overline{\mathbf{O}}^{\text{test}})\right), \tag{3.2}$$

[4] assuming the application somehow checks collected utterances for validity; see for example Section 5.5 for the procedure used in the PER system

[5] for example $\mathbf{W} = \{0, \ldots, 9, \text{name}\}$ as in the PER system

where $\boldsymbol{\lambda}_\psi$ is the model created for the target identity from the target's enrollment data, and

$$g_\psi(\overline{\mathbf{O}}^{\text{test}}) = \underset{g\in\{\text{male,female}\}}{\arg\max} P^\psi(\overline{\mathbf{O}}^{\text{test}}|\boldsymbol{\lambda}_\psi^g) \tag{3.3}$$

is a gender detector like the one used in the training phase, but it uses test data instead of enrollment data to make the gender selection.

This method of selecting a background model has been referred to as an unconstrained cohort by other authors (Ariyaeeinia and Sivakumaran, 1997). It differs from the traditional cohort method (Higgins et al., 1991; Rosenberg et al., 1992) in that the selection is based on similarity to a test segment rather than to enrollment data. However, our method differs slightly from both the traditional cohort method and the unconstrained cohort method in that the competing models are only two and represent groups of speakers (genders) rather than individual speakers.

3.3 The text-dependent HMM system

The HMM subsystem is text-dependent and operates in a prompted mode with digit string utterances only[6]. Except for how background models are selected during the test phase, the system is the same as the baseline system described and tested in (Melin et al., 1998) and (Melin and Lindberg, 1999). In this section, the design of the HMM subsystem is described relative to the common subsystem features described in Section 3.2. The modified background model selection method is described and evaluated.

3.3.1 Feature extraction

The basic 12-element MFCC vector (Section 3.2.1) is extended with the 0'th cepstral coefficient (frame energy). Cepstral mean subtraction is applied to this 13-element static feature vector, and first and second order deltas are appended. The total vector dimension is 39.

3.3.2 Classifier units

The HMM subsystem contains ten classifier units $\psi = \mathrm{H}_0 \ldots \mathrm{H}_9$, one classifier unit per digit word. Models are continuous word-level left-to-right HMMs with 16 Gaussian terms per phoneme in the represented word distributed on two states per phoneme[7] with an eight-component Gaussian mixture observation probability density function (pdf) per state. Gaussian components have diagonal covariance matrices. The choice of 16 terms per phoneme is based on development experiments on Gandalf data in preparation for previous work (Melin et al., 1998), while their partitioning into two states with eight terms each is somewhat arbitrary as shown

[6] it ignores the name parts of enrollment and test data in the PER case

[7] Swedish digit words have between two and four phonemes per word.

by Bimbot et al. (2000). Denote the target HMM in classifier unit H_w for a given client as $\boldsymbol{\lambda}_{\mathrm{H}_w} = \{\mathbf{c}_w, \mathbf{m}_w, \boldsymbol{\sigma}_w^2, \mathbf{A}_w\}$ where $\mathbf{c}_w$, $\mathbf{m}_w$ and $\boldsymbol{\sigma}_w^2$ are vectors of all mixture weights, mean values and variance values, respectively, and $\mathbf{A}_w$ is the matrix of transition probabilities.

3.3.2.1 Training phase

A target model $\boldsymbol{\lambda}_{\mathrm{H}_w}$ for the word w and a given client is trained on all R_w valid examples of the word spoken during the client's enrollment session[8]. That is, training data $\overline{\mathbf{O}}_{\mathrm{H}_w}^{\mathrm{enroll}}$ for classifier unit H_w is a subset of $\overline{\mathbf{O}}^{\mathrm{enroll}}$ such that $\overline{\mathbf{O}}_{\mathrm{H}_w}^{\mathrm{enroll}} = \{\mathbf{O}_{w,1}, \ldots, \mathbf{O}_{w,R_w}\}$, where observations are 39-dimensional feature vectors as described in the previous section.

Given the training data, one of the gender-dependent background models is first selected as a seed model using a gender detector (Eq. 3.1). The seed model is then used as a basis for target model training: transition probabilities and variance vectors are left as they are, while mean vectors and mixture weights are trained from the data. Training is performed with the Expectation Maximization (EM) algorithm to optimize the Maximum Likelihood (ML) criterion

$$(\hat{\mathbf{c}}_w, \hat{\mathbf{m}}_w) = \arg\max_{(\mathbf{c}_w, \mathbf{m}_w)} P(\overline{\mathbf{O}}_{\mathrm{H}_w}^{\mathrm{enroll}} | \mathbf{c}_w, \mathbf{m}_w, {\boldsymbol{\sigma}_w^{\mathrm{seed}}}^2, \mathbf{A}_w^{\mathrm{seed}}), \tag{3.4}$$

where ${\boldsymbol{\sigma}_w^{\mathrm{seed}}}^2$ and $\mathbf{A}_w^{\mathrm{seed}}$ are the fixed variance and transition probabilities taken verbatim from the seed model $\boldsymbol{\lambda}_{\mathrm{H}_w}^{g_{\mathrm{H}_w}^{\mathrm{seed}}}$. The seed means and mixture weights are used as starting values in the first iteration of the EM algorithm (Rosenberg et al., 1991).

Background models were trained with the EM-algorithm and the ML criterion. After initializing models with a single Gaussian per state, Gaussians were split into $2 \rightarrow 4 \rightarrow 6 \rightarrow 8$ Gaussians per state and re-estimated with up to 20 EM iterations after each splitting operation. A fixed variance floor of 0.01 was used, but only 0.1% of all variance parameters received a value less than twice the floor.

3.3.2.2 Verification test phase

The likelihood function implemented by the classifier unit (during the verification test phase) is the Viterbi approximation of the probability of observation data given a model, i.e. the probability of observations given the model and the most likely path:

$$P^{\mathrm{H}_w}(\mathbf{O}|\boldsymbol{\lambda}) = \max_{\mathbf{S}\in\boldsymbol{\Omega}} P(\mathbf{O}|\boldsymbol{\lambda}, \mathbf{S}) \tag{3.5}$$

where $\mathbf{S}$ is a certain path through the HMM $\boldsymbol{\lambda}$ and $\boldsymbol{\Omega}$ is the set of all possible paths. The notation $P^{\mathrm{H}_w}(\mathbf{O}|\boldsymbol{\lambda})$ is used to indicate this is the likelihood function used in classifier unit $\psi = \mathrm{H}_w$ (cf. Section 3.2.2).

[8] in the PER-system, $R_w = 5$ (for digits)

Given a target model and a test utterance, a classifier unit produces an output score value s_{H_w} for each word segment i for which $w(i) = w$:

$$s_{\mathrm{H}_w}(i) = \frac{1}{N_i}\left(\log P^{\mathrm{H}_w}(\mathbf{O}_i|\boldsymbol{\lambda}_{\mathrm{H}_w}) - \log P^{\mathrm{H}_w}(\mathbf{O}_i|\boldsymbol{\lambda}^g_{\mathrm{H}_w})\right) \tag{3.6}$$

and for the entire test utterance

$$z_{\mathrm{H}_w} = \begin{cases} \frac{1}{L_{\mathrm{H}_w}} \sum_{i:w(i)=w} s_{\mathrm{H}_w}(i), & L_{\mathrm{H}_w} > 0 \\ 0, & L_{\mathrm{H}_w} = 0 \end{cases} \tag{3.7}$$

where

$$g = g_{\mathrm{H}_w}\left(\overline{\mathbf{O}}^{\mathrm{test}}_{\mathrm{H}_w}\right)$$

is the gender detected for the test utterance in the same classifier unit, Eq. (3.3), and N_i is the number of observation vectors in word segment i. $\overline{\mathbf{O}}^{\mathrm{test}}_{\mathrm{H}_w} = \{\mathbf{O}_i : w(i) = w\}$ is the subset of the test utterance where word w is spoken, and L_{H_w} the number of word segments in this subset (i.e. the number of repetitions of word w).

The score output value z_{H} from the entire HMM subsystem for a test utterance $\overline{\mathbf{O}}^{\mathrm{test}}_{\mathrm{H}} = \left\{\mathbf{O}_i : w(i) \in \{0\ldots9\}\right\}$ (the subset of $\overline{\mathbf{O}}^{\mathrm{test}}$ where a digit word is spoken) is

$$z_{\mathrm{H}} = \frac{1}{L_{\mathrm{H}}}\sum_{u=0}^{9} L_{\mathrm{H}_u} z_{\mathrm{H}_u} = \frac{1}{L_{\mathrm{H}}}\sum_{i:w(i)=\{0\ldots9\}} s_{\mathrm{H}_w}(i), \tag{3.8}$$

where L_{H} is the number of word segments in $\overline{\mathbf{O}}^{\mathrm{test}}_{\mathrm{H}}$.

3.3.3 Background model selection

The background model selection method in this system is different from the one used in our previous publications (Melin et al., 1998) and (Melin and Lindberg, 1999), where the background model was chosen based on similarity to enrollment data like in the traditional cohort method. The purpose of selecting a background model based on similarity to the test segment is to circumvent a well-known problem with traditional cohorts and dissimilar impostors. If the background model is trained on data “close” to the target speaker, then both the target model and the background model will be poor models in regions of the sample space “far away” from the target speaker. Hence, the likelihood ratio test will not be a good test for dissimilar speakers, such as cross-sex impostors. By selecting the background model that is closer to the test segment, the likelihood ratio test is more likely to reject a dissimilar impostor. The advantage of the used method is evident from Figure 3.1, where same-sex (on the left) and cross-sex (on the right) DET curves are shown for both methods. These curves are from experiments on the Gandalf corpus with identical enrollment and test sets as were used in (Melin and Lindberg, 1999). Results show that the unconstrained cohort method reduces cross-sex imposture rate considerably, at no loss in same-sex imposture rate.

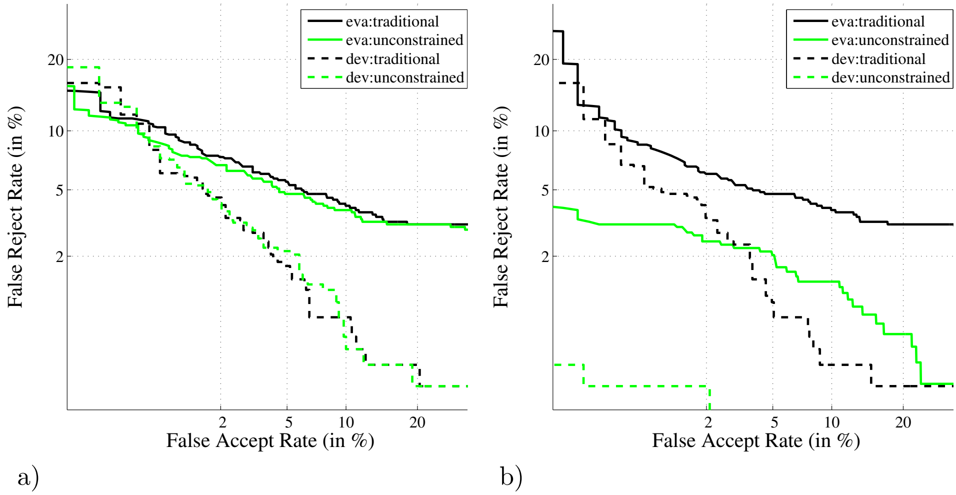

Figure 3.1: DET plots for the HMM subsystem with two different methods for selecting from one of two gender-dependent background models: by similarity to enrollment data (traditional cohort) or by similarity to test data (unconstrained cohort). Test data is from the Gandalf corpus with single-session, one-minute enrollment and two four-digit test utterances. DET curves are shown for both the development (dev) and the evaluation (eva) sets. Curves in a) are based on same-sex impostor attempts, while curves in b) are based on cross-sex impostor attempts. True-speaker tests are the same in a) and b).

3.4 The text-independent GMM system

The GMM subsystem is inherently text-independent, though in this thesis it is used in a prompted, text-dependent way in the sense that enrollment and test utterances are always composed of words from the same vocabulary[9]. Background models are still used text-independently, however. The GMM-specific modules for the GIVES framework were initially developed as part of a student project (Neiberg, 2001), and then extended by Neiberg in conjunction with CTT's participation in the 1-speaker detection cellular task in the NIST 2002 Speaker Recognition Evaluation (NIST, 2002). Experiments on a PER development set of Gandalf data (Section 10.2.1) were then used as the basis for selecting the particular configuration of the GMM subsystem used in this work. This section describes the design and configuration of the subsystem in detail. It is included for completeness since the GMM system is used in the thesis and because not all parts of the description were published elsewhere.

[9] proper name and digits in the PER case

3.4.1 Feature extraction

The basic 12-element MFCC vectors (Section 3.2.1) are RASTA-filtered (Hermansky et al., 1991) and first order deltas are appended. The total vector dimension is 24.

3.4.2 Classifier unit

The GMM subsystem contains a single classifier unit $\psi = \mathrm{G}_0$, where target and background models are 512-component Gaussian mixture pdfs with diagonal covariance matrices, also known as GMMs (Rose and Reynolds, 1990; Reynolds, 1995). Denote the parameters of the target GMM in the classifier unit as $\boldsymbol{\lambda}_{\mathrm{G}_0} = \{c_k, \mathbf{m}_k, \boldsymbol{\sigma}_k^2\}_{k=1}^K$, where c_k is the weight and $\mathbf{m}_k$ and $\boldsymbol{\sigma}_k^2$ the vectors of mean and variance values of mixture term k, and $K = 512$ is the number of terms[10] in the model

$$p(\mathbf{o}|\boldsymbol{\lambda}_{\mathrm{G}_0}) = \sum_{k=1}^{K} c_k \phi\Big(\mathbf{o}|\mathbf{m}_k, \boldsymbol{\sigma}_k^2\Big). \tag{3.9}$$

$\phi()$ denotes the multivariate normal density function.

3.4.2.1 Training phase

A target model $\boldsymbol{\lambda}_{\mathrm{G}_0}$ for a given client is trained on all valid enrollment data from the client, i.e. $\overline{\mathbf{O}}_{\mathrm{G}_0}^{\mathrm{enroll}} = \overline{\mathbf{O}}^{\mathrm{enroll}}$ with observation vectors being 24-dimensional feature vectors as described above. Note that observation vectors from non-speech segments are not included in training data (provided word boundaries are correctly estimated).

Given the training data, one of the gender-dependent background models is first selected as a seed model using a gender detector (Eq. 3.1). The target model is then created from the seed model using the following maximum a posteriori (MAP) like update formulas (Reynolds et al., 2000):

$$\begin{aligned} c_k &= \big(\alpha_k \eta_k / N + (1-\alpha_k) c_k^g\big)\gamma && (3.10)\\ \mathbf{m}_k &= \alpha_k E_k(\mathbf{o}) + (1-\alpha_k)\mathbf{m}_k^g && (3.11)\\ \boldsymbol{\sigma}_k^2 &= \alpha_k E_k(\mathbf{o}^2) + (1-\alpha_k)\big({\boldsymbol{\sigma}_k^g}^2 + {\mathbf{m}_k^g}^2\big) - \mathbf{m}_k^2 && (3.12) \end{aligned}$$

where

$$\alpha_k = \frac{\eta_k}{\eta_k + r} \tag{3.13}$$

is a data-dependent adaptation coefficient with relevance factor $r = 16$,

$$\gamma = \frac{1}{\sum_{k=1}^{K} c_k} \tag{3.14}$$

[10] "term" and "component" are used interchangeably in this thesis when referring to the terms of the sum in (3.9)

assures target model weights sum to unity, and[11]

$$\eta_k = \sum_{n=1}^{N} \eta_{k,n} \tag{3.15}$$

$$E_k(\mathbf{o}) = \frac{1}{\eta_k}\sum_{n=1}^{N} \eta_{k,n}\mathbf{o}_n \tag{3.16}$$

$$E_k(\mathbf{o}^2) = \frac{1}{\eta_k}\sum_{n=1}^{N} \eta_{k,n}\mathbf{o}_n^2, \tag{3.17}$$

where

$$\eta_{k,n} = \frac{c_k\phi\left(\mathbf{o}_n|\mathbf{m}_k^g, \boldsymbol{\sigma}_k^{g\,2}\right)}{\sum_{l=1}^{K} c_l\phi\left(\mathbf{o}_n|\mathbf{m}_l^g, \boldsymbol{\sigma}_l^{g\,2}\right)} \tag{3.18}$$

is the *a posteriori* weight of mixture term k given an observation vector $\mathbf{o}_n$ and the seed model. Training data $\overline{\mathbf{O}}_{\mathrm{G}_0}^{\mathrm{enroll}}$ have here been viewed as a single vector sequence $\mathbf{O} = \{\mathbf{o}_1 \ldots \mathbf{o}_N\}$ with

$$N = \sum_{w\in\mathbf{W}}\sum_{r=1}^{R_w} N_{w,r} \tag{3.19}$$

where $N_{w,r}$ is the length of the observation sequence from the r'th valid repetition of word w, and $\mathbf{W}$ is the vocabulary (cf. Section 3.2.2.1).

Background models were trained with the EM-algorithm and the ML criterion. First a gender-independent "root" GMM was initialized from a VQ codebook and then trained on pooled male and female data with eight EM iterations. Centroids of the VQ codebook were initialized from 512 equidistant (in time) training vectors and then trained with the generalized Lloyd algorithm (e.g. Gersho and Gray, 1992) using the Mahanalobis distance measure. The root GMM was then used as the starting point for training a male GMM on male data and a female GMM on female data with three iterations for each gender model.

3.4.2.2 Verification test phase

The classifier unit is tested on all available speech segments in a test utterance, i.e. $\overline{\mathbf{O}}_{\mathrm{G}_0}^{\mathrm{test}} = \overline{\mathbf{O}}^{\mathrm{test}}$. Given a target model and a test utterance, a classifier unit produces an output score value

$$z_{\mathrm{G}_0} = \frac{1}{N_{\mathrm{G}_0}^{\mathrm{test}}}\left(\log P^{\mathrm{G}_0}\left(\overline{\mathbf{O}}_{\mathrm{G}_0}^{\mathrm{test}}|\boldsymbol{\lambda}_{\mathrm{G}_0}\right) - \log P^{\mathrm{G}_0,\mathrm{gps}}\left(\overline{\mathbf{O}}_{\mathrm{G}_0}^{\mathrm{test}}|\boldsymbol{\lambda}_{\mathrm{G}_0}^{g}\right)\right) \tag{3.20}$$

where $N_{\mathrm{G}_0}^{\mathrm{test}}$ is the number of observation vectors in the test utterance and

$$g = g_{\mathrm{G}_0}\left(\overline{\mathbf{O}}_{\mathrm{G}_0}^{\mathrm{test}}\right)$$

[11] $\mathbf{o}^2$ is a shorthand for $\mathrm{diag}(\mathbf{o}\mathbf{o}^{\mathrm{T}})$ (from Reynolds et al. (2000))

is the gender detected for the test utterance in the same classifier unit, Eq. (3.3). z_{G_0} is also used as the output score value of the GMM subsystem, i.e. $z_{\mathrm{G}} = z_{\mathrm{G}_0}$.

The likelihood function $P^{\mathrm{G}_0}\left(\overline{\mathbf{O}}^{\text{test}}_{\mathrm{G}_0}|\boldsymbol{\lambda}_{\mathrm{G}_0}\right)$ used with the target model is the probability of test data given the model, i.e.

$$\log P^{\mathrm{G}_0}\left(\overline{\mathbf{O}}^{\text{test}}_{\mathrm{G}_0}|\boldsymbol{\lambda}_{\mathrm{G}_0}\right) = \sum_{i=1}^{L}\sum_{n=1}^{N_i}\log\left(\sum_{k=1}^{K} c_k\phi\left(\mathbf{o}_n^{(i)}|\mathbf{m}_k,\boldsymbol{\sigma}_k^2\right)\right) \tag{3.21}$$

where $\mathbf{o}_n^{(i)}$ is an observation vector in the i'th word segment $\mathbf{O}_i$ in the test utterance (cf. Section 3.2.2.2).

A modified likelihood function $P^{\mathrm{G}_0,\mathrm{gps}}\left(\overline{\mathbf{O}}^{\text{test}}_{\mathrm{G}_0}|\boldsymbol{\lambda}_{\mathrm{G}_0}\right)$ is used with background models in (3.20). It uses a Gaussian pre-selection (gps) method to reduce the number of calculations relative to (3.21). Each time (3.21) is evaluated for an observation vector in segment $\mathbf{O}_i$, the index k of the $C = 6$ top contributing mixture terms for that observation vector is stored into an N by C matrix $\boldsymbol{\kappa}^{(i)}$, and the likelihood for a background model is calculated as

$$\log P^{\mathrm{G}_0,\mathrm{gps}}\left(\overline{\mathbf{O}}^{\text{test}}_{\mathrm{G}_0}|\boldsymbol{\lambda}^g_{\mathrm{G}_0}\right) = \sum_{i=1}^{L}\sum_{n=1}^{N_i}\log\left(\sum_{j=1}^{C} c^g_{\kappa^{(i)}(n,j)}\phi\left(\mathbf{o}_n^{(i)}|\mathbf{m}^g_{\kappa^{(i)}(n,j)},{\boldsymbol{\sigma}^g_{\kappa^{(i)}(n,j)}}^2\right)\right) \tag{3.22}$$

where

$$\boldsymbol{\lambda}^g_{\mathrm{G}_0} = \left\{c^g_k, \mathbf{m}^g_k, {\sigma^g_k}^2\right\}_{k=1}^{K}$$

are the parameters of the background model for gender g.

This gps-method is a modified version of a method suggested by Reynolds (1997) based on the assumption that a mixture term of an adapted GMM has a relation to the corresponding term in the GMM it was adapted from (the *parent* model), such that the two terms are "close" compared to other terms. Call this a parent relation. While Reynolds evaluated all mixture terms of a (single) background model and only selected terms in the target model, we used a variant where all terms of the target model are evaluated and only selected terms of the two background models. A similar variant was previously tested with a single background model by Navrátil et al. (2001), who showed that the modification results in a clock-wise rotation of the DET curve relative to the original method, i.e. reduced false accept rates at low false reject rates.

With our use of two background models in gender-detection during the test phase (Eq. 3.3), evaluating all terms in the target model and only a few in background models is a logical choice, since more computations are saved compared to fully evaluating both background models. To allow this, we create both background models through the adaptation of a common (gender-independent) "root" model as described above. Analogous to the mentioned parent relation, such background models have a *sibling relation*. We assume that siblings have a similar kind of closeness relation as parent-child, though weaker in strength. The sibling relation between background models is needed to use indexes of top-scoring mixture terms in a target model to pick mixture terms for Eq. (3.22) with both background models, because the target model has a parent relation to (was adapted from) only one of the background models (Eq. 3.1).

With our use of Gaussian pre-selection, the number of evaluated mixture terms is $K + 2C$ per observation vector compared to $3K$ for a full evaluation of target model and both background models, a reduction of 66%.

3.5 The combined system

3.5.1 Score fusion

The HMM and GMM subsystems are fused at the score level. The system output score value, or decision variable, z for a test utterance is a linear combination of subsystem score values

$$z = \omega_{\mathrm{H}} z_{\mathrm{H}} + \omega_{\mathrm{G}} z_{\mathrm{G}}. \tag{3.23}$$

Combination weights ω_ξ are computed as

$$\omega_\xi = \frac{1}{\sum_{\zeta \in \{\mathrm{H},\mathrm{G}\}} (1 - \epsilon_\zeta)/\sigma_\zeta} \cdot \frac{1 - \epsilon_\xi}{\sigma_\xi} \tag{3.24}$$

where ϵ_ξ and σ_ξ are determined empirically through a development experiment [12] with the individual subsystems, as their respective equal error rate and standard deviation of observed values for z_ξ. The rationale for (3.24) is that scores from each of the subsystems are first scaled to have unit variance (on development data) and are then weighted such that the subsystem with lower EER gets a higher weight.

3.5.2 Classification

The actual classifier decision is taken by comparing the value of the decision variable z to a speaker-independent threshold θ:

$$z \quad \underset{\text{reject}}{\overset{\text{accept}}{\begin{matrix} > \\ \leq \end{matrix}}} \quad \theta. \tag{3.25}$$

The value of the threshold is also determined empirically from a development experiment[12].

[12] cf. Section 10.2 for the PER system

3.6 References

Ariyaeeinia, A. and Sivakumaran, P. (1997). Analysis and comparison of score normalisation methods for text-dependent speaker verification. In *Proc. 5th European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 1379–1382, Rhodes, Greece. [4]

Bimbot, F., Blomberg, M., Boves, L., Genoud, D., Hutter, H.-P., Jaboulet, C., Koolwaaij, J., Lindberg, J., and Pierrot, J.-B. (2000). An overview of the CAVE project research activities in speaker verification. *Speech Communication*, 31(2-3):155–180. [5]

Gersho, A. and Gray, R., editors (1992). *Vector Quantization and Signal Compression*. Springer. [9]

Hermansky, H., Morgan, N., Bayya, A., and Kohn, P. (1991). Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP). In *Proc. 2nd European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 1367–1370, Genova, Italy. [8]

Higgins, A., Bahler, L., and Porter, J. (1991). Speaker verification using randomized phrase prompting. *Digital Signal Processing*, 1:89–106. [4]

Melin, H., Koolwaaij, J., Lindberg, J., and Bimbot, F. (1998). A comparative evaluation of variance flooring techniques in HMM-based speaker verification. In *Proc. 1998 International Conference on Spoken Language Processing (ICSLP)*, pages 1903–1906, Sydney, Australia. [4, 6]

Melin, H. and Lindberg, J. (1999). Variance flooring, scaling and tying for text-dependent speaker verification. In *Proc. 6th European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 1975–1978, Budapest, Hungary. [4, 6]

Navrátil, J., Chaudhari, U., and Ramaswamy, G. (2001). Speaker verification using target and background dependent linear transformations and multi-system fusion. In *Proc. 7th European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 1389–1392, Aalborg, Denmark. [10]

Neiberg, D. (2001). Text independent speaker verification using adapted gaussian mixture models. Master's thesis, KTH/TMH, Stockholm, Sweden. [7]

NIST, editor (2002). *NIST Speaker Recognition Workshop, Informal proceedings*, Vienna VA, USA. NIST. [7]

Reynolds, D. (1995). Speaker identification and verification using gaussian mixture speaker models. *Speech Communication*, 17(1-2):91–108. [8]

Reynolds, D. (1997). Comparison of background normalization methods for text-independent speaker verification. In *Proc. 5th European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 963–966, Rhodes, Greece. [10]

Reynolds, D., Quatieri, T., and Dunn, R. (2000). Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10:19–41. [8, 9]

Rose, R. and Reynolds, D. (1990). Text-independent speaker identification using automatic acoustic segmentation. In *Proc. 1990 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 293–296, Albuquerque NM, USA. [8]

Rosenberg, A., DeLong, J., Lee, C.-H., Juang, B.-H., and Soong, F. (1992). The use of cohort normalized scores for speaker verification. In *Proc. 1992 International Conference on Spoken Language Processing (ICSLP)*, pages 599–602, Banff, Canada. [4]

Rosenberg, A., Lee, C.-H., and Gokcen, S. (1991). Connected word talker verification using whole word hidden markov models. In *Proc. 1991 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 381–384, Toronto, Canada. [5]